Fine-tuning motif detection among ChIP on Chip DNA fragments

Fabrice Touzain, Théo Mozzanino, Sophie Schbath, Marie Agnes Petit

Journées Ouvertes Biologie Informatique Mathématiques

Institut Pasteur, Paris, 28 juin – 1er juillet 2011

Éditeurs
Emmanuel BARILLOT
Christine FROIDEVAUX
Eduardo PC ROCHA

Journées Ouvertes Biologie Informatique Mathématiques

Institut Pasteur, Paris, 28 juin – 1er juillet 2011

Éditeurs
Emmanuel BARILLOT
Christine FROIDEVAUX
Eduardo PC ROCHA

Journées Ouvertes de Biologie, Informatique et Mathématiques

IV+xiv+422 pages.

Réalisation et mise en page : Pascal Bochet, Frank Rügheimer et Ivan Moszer
Logo JOBIM 2011 : Christophe Louis
Couverture : Valérie Zeitoun

Ce document a été préparé avec la classe LaTeX 2$_\varepsilon$ « Proceedings ».

Impression: 28 juillet 2011
Éditeur : Institut Pasteur, 25-28 rue du Docteur Roux, 75015 Paris, France
Impression : Service Image et Reprographie,
Institut Pasteur, 25-28 rue du Docteur Roux, 75015 Paris, France

Préface

JOBIM célèbre sa douzième année et s'est imposé depuis 2000 comme le lieu privilégié de rencontres et d'échanges de la communauté francophone bio-informatique et biostatistique. Au cours de ces onze années, l'étendue sémantique du terme bio-informatique et le champ de la conférence JOBIM se sont accrus substantiellement. JOBIM couvre actuellement des thèmes se rapportant à la génomique – avec une accumulation et une variété de données sans précédent liées au développement rapide des technologies dites à « haut débit » –, la bio-informatique structurale, l'évolution et la phylogénie, l'analyse fonctionnelle – en particulier l'étude de la régulation (génétique et épigénétique) des processus cellulaires –, mais aussi la génétique des populations et l'écologie. JOBIM est ainsi une occasion unique de présenter des approches méthodologiques originales dans ces domaines d'application, en particulier concernant l'intégration des données et la représentation des connaissances, l'algorithmique (séquences, arbres, réseaux), les statistiques et la modélisation mathématique, l'analyse d'images, ou encore les modèles dynamiques de réseaux d'interactions, utilisés en particulier en biologie des systèmes.

Nous avons reçu cette année 161 soumissions (dont 42 résumés d'affiches lors de l'appel de seconde vague). Suite au travail d'expertise du comité de programme et de quelques relecteurs additionnels, chaleureusement remerciés ici, 19 articles ont été retenus pour des présentations orales en séance plénière, et 24 pour des présentations orales lors de sessions parallèles ; nous avons choisi d'organiser de telles sessions, pour la première fois cette année à JOBIM, dans le but de répondre à la croissance de la discipline tout en offrant davantage de temps pour les exposés et leur discussion. Les présentations orales ont été sélectionnées tant sur articles courts que sur articles longs. Quelque 110 affiches seront également exposées et discutées lors de ces journées.

À ces présentations s'ajoutent les conférences invitées de Matthieu Blanchette, Michael Brudno, Patrick Forterre, Edda Klipp, Marie-France Sagot, Peter Stadler et Sarah Teichmann. Nous leur adressons ici nos plus vifs remerciements pour l'honneur qu'ils nous font en acceptant de venir exposer leurs travaux sur les sujets les plus actuels.

Afin d'inciter les doctorants et post-doctorants à soumettre leurs résultats, nous avons introduit pour cette édition 2011 deux prix à leur intention, décernés l'un pour la meilleure présentation sur affiche, et l'autre pour la meilleure présentation orale. Nous espérons que ce soutien aux jeunes chercheurs de notre communauté sera apprécié, en tant que reconnaissance de travaux de grande qualité scientifique, mais aussi en tant qu'encouragement à poursuivre la valorisation des résultats obtenus.

Nous adressons nos remerciements à l'ensemble des membres du comité d'organisation pour le travail réalisé, tant en amont que pendant ces journées. Nous remercions également nos partenaires institutionnels et industriels, et en particulier l'Institut Pasteur, pour l'accueil et le soutien offerts à l'organisation de ces rencontres, ainsi que la Société Française de Bio-Informatique (SFBI) sous l'égide de laquelle ces journées sont placées.

Nous vous souhaitons la bienvenue à JOBIM 2011 et espérons que ces journées répondront à vos attentes bio-informatiques.

Pour le comité de programme :
Emmanuel Barillot, Institut Curie – Inserm/Mines ParisTech, Paris
Christine Froidevaux, LRI CNRS – U. Paris Sud – INRIA, Orsay
Eduardo PC Rocha, Institut Pasteur – CNRS, Paris

Pour le comité d'organisation :
Hélène Chiapello, INRA, Jouy-en-Josas
Daniel Gautheret, IGM – U. Paris Sud, Orsay
Ivan Moszer, Institut Pasteur, Paris

Comité d'Organisation

Hélène Chiapello, Daniel Gautheret et Ivan Moszer

Pascal Bochet
Caroline Boursaux-Eude
Stéphane Descorps-Declère
Marie-Agnès Dillies
Ghislaine Guigon

Véronique Hourdel
Pierre Lechat
Olivier Lespinet
Alexandra Louis
Christophe Malabat

Thérèse Malliavin
Frank Rügheimer
Erika Souche

Avec la précieuse contribution de Chrystèle Blin et Annie Étienne

Comité de Programme

Emmanuel Barillot, Christine Froidevaux and Eduardo PC Rocha

Guillaume Achaz
Florence D'Alché-Buc
Benjamin Audit
Anne Bergeron
Philippe Besse
Céline Brochier
Anne-Claude Camproux
Hélène Chiapello
Sarah Cohen-Boulakia
Gilbert Deléage
Jean-Daniel Fekete
Olivier Gascuel
Christine Gaspin

Daniel Gautheret
Simonetta Gribaldo
Raphaël Guérois
Hidde De Jong
Fabrice Jossinet
Frédérique Lisacek
Claudine Médigue
Karyn Mégy
Ivan Moszer
Gregory Nuel
Guy Perrière
Yann Ponty
Anne Poupon

Stéphane Robin
Hugues Roest Crollius
Manuel Ruiz
Sophie Schbath
David Sherman
Anne Siegel
Denis Thieffry
Julie Thompson
Jacques van Helden
Yves Vandenbrouck
Jean-Philippe Vert
Stéphane Vialette
Thierry Wirth

Relecteurs additionnels

Justin Bedo
Julie Bernauer
Jeremie Bourdon
Lionel Frangeul
Javier Herrero
Annick Jacq
Andre Khalil

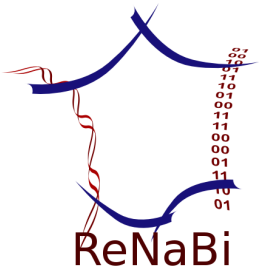
Antonin Marchais
Mahendra Mariadassou
Anthony Mathelier
Stefano Mona
Alban Ott
Hugues Richard
Philippe Roumagnac

Franck Samson
Yves-Henri Sanejouand
Paul Sorba
Claire Toffano-Nioche
Jerome Waldispuhl
Farida Zehraoui
Matthias Zytnicki

Partenaires

Les partenaires institutionnels suivants ont apporté leur soutien à JOBIM 2011 :

MINISTÈRE
DE L'ENSEIGNEMENT SUPÉRIEUR
ET DE LA RECHERCHE



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

Les partenaires industriels suivants ont apporté leur soutien à JOBIM 2011 :

Sommaire

– Avant-Propos –

– Contributions –

–Listes et Index–

Session 1 : Protein Structure

Conférence invitée

Sarah Teichmann

MRC Laboratory of Molecular Biology, Cambridge, UK.

Evolution and Assembly of Protein Complexes

The formation of specific protein complexes is a basic requirement of all biological processes. Proteins interact through surfaces that are complementary at the level of both sequence and structure, and so proteins often undergo changes in conformation and flexibility when they bind to each other. At the same time, "sticky" patches on protein surfaces may lead to spurious interactions between proteins in the cell. We can gain insight into these issues by analyzing the abundance of data on protein interactions and protein complexes, both from conventional small-scale experiments collected over the decades, including three-dimensional structures, and more recently by large-scale functional genomics experiments.

We have analyzed the relationships between the structures of proteins and the conformational changes that they undergo upon binding by comparing crystal structures of free proteins and proteins in complexes. We find that the relative solvent accessible surface area of both free and bound subunits can be used to predict the magnitude of binding-induced conformational changes. We demonstrate that the relative solvent accessible surface area of monomeric proteins is useful as a simple proxy for intrinsic flexibility and for predicting conformational changes upon binding. In addition to the predictive power of this correlation, it reveals a strong connection between the flexibility of unbound proteins and their binding-induced conformational changes, consistent with the conformational selection model of molecular recognition.

Inside the cell, specific interactions compete with non-specific interactions at some level. To what extent are proteins under selection to avoid non-functional and deleterious interactions? To answer this question, we project evolutionary and systems information onto 397, 196, and 701 proteins of known structure from *E. coli*, *S. cerevisiae* and *H. sapiens* respectively. We find that the propensity of proteins to interact in a non-specific manner with other proteins is inversely correlated with their abundance in *E. coli* and *S. cerevisiae*. This tendency is evident at surface residues: high abundance proteins have evolved to have a less sticky surface. In *E. coli* and *S. cerevisiae*, we also find that the evolutionary conservation of an amino acid is positively correlated with the stickiness of the surface environment around it. Thus, residues in sticky surface patches are evolutionarily more constrained, possibly because they are more likely to trigger non-functional interactions if they mutate. Although significant, the impact of protein stickiness is comparatively small in shaping the physico-chemical properties and evolution of *H. sapiens* proteins. This suggests that promiscuous protein-protein interactions are freer to accumulate in species with a small effective population size; a phenomenon akin to junk DNA accumulation.

PEP-FOLD: Biased Approach for the *De Novo* Prediction of Peptide and Miniprotein Structure

Yimin SHEN[1], Julien MAUPETIT[1], Philippe DERREUMAUX[2] and Pierre TUFFÉRY[1]

[1] MTi, UMR-S973 INSERM, Unniversité Paris Diderot, 35 rue H. Brion, 75205 Paris, Cedex 13, France
{yimin.shen,julien.maupetit,pierre.tuffery}@univ-paris-diderot.fr
[2] Laboratoire de Biochimie Théorique, UPR9080 CNRS, Institut de Biologie Physico-Chimique, 13 rue P. et M. Curie, 75005 Paris, France
philippe.derreumaux@ibpc.fr

Keywords Structural alphabet, peptides, miniproteins, de novo modeling.

PEP-FOLD: Prédiction *De Novo* de la Structure de Peptides et Mini-Protéines par une Approche Biaisée d'Echantillonage Conformationnel.

Mots-clés Alphabet Structural, peptides, miniprotéines, modélisation de novo.

1 Introduction

While worldwide effort over the past 20 years (see CASP experiments [1]) has resulted in convincing *ab initio* or *de novo* approaches such a Rosetta [2] for protein structure determination, accurate and fast peptide / mini-protein 3D conformation prediction is still an open challenge. Yet, peptides play many biological functions ranging from hormones, neurotransmitters to antibiotics, among others.

One of the reasons of this limitation is the low rate of peptide structure identification using either NMR spectroscopy or X-Ray crystallography. It prevents the learning of specific sequence structure relationships: in contrast to proteins, short peptides do not systematically adopt stable well-defined tertiary structures [3]. Concomitantly, until recently, the fast and reliable prediction of short peptides conformation has aroused limited effort. Pioneering this domain in the late 90's, Ishikawa and Dill proposed Geocore algorithm to generate peptide conformations. It was followed by PepStr β-turn prediction, Peplook algorithm, the Generalized Pattern Search algorithm (GPS) using secondary structure prediction, and most recently by PEP-FOLD (see [4,5] and references included). We introduce here some early concepts of PEP-FOLD together with its most recent evolutions.

2 PEP-FOLD

Concepts: PEP-FOLD is based on the concept of structural alphabet (SA) - *i.e.* a description of a polypeptidic conformation as a series of local canonical conformations, and uses a HMM (Hidden Markov Chain)-derived SA of 27 letters to describe proteins as series of overlapping fragments of four amino acids [6]. PEP-FOLD is based on a two-step procedure: (i) prediction of a limited set of SA letters at each position from peptide amino acid sequence and (ii) assembly of the prototype fragments associated with each SA letter using a greedy algorithm and a generic protein coarse-grained force field.

Results: Using a benchmark of 25 peptides with 9-23 amino acids, and considering the reproducibility of the runs, the first version of PEP-FOLD[5] identified, on average, lowest-energy conformations differing by 2.6 Å Cα RMS deviation (cRMSd) from the NMR structures. For 13 mini-proteins with 27-49 amino acids, *PEP-FOLD 1* reached an accuracy of 3.6 and 4.6 Å cRMSd for the most-native and lowest-energy conformations, using the non-flexible regions identified by NMR (rigid core). However for several of these mini-proteins, PEP-FOLD was not able to identify the native fold.

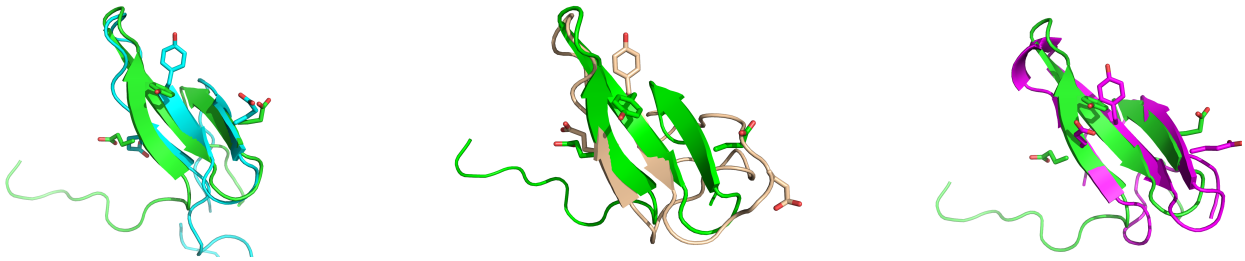


Figure 1. WW domain from the mouse transcription elongation regulator 1 (40 amino acids, PDB: 2ysi). De novo conformations compared to experimental structures. Green: experimental. Blue: PEP-FOLD2, 2.6Å cRMSd. Wheat: PEP-FOLD1, 4.3Å cRMSd. Magenta: Rosetta, 5.2Å cRMSd. For E14, Y23 and E31 side chains orientations clearly show that only PEP-FOLD2 has identified the native fold.

We have recently reconsidered the prediction of the SA letters from the amino acid sequence. We find that the posterior analysis of the probabilities of the SA letters at each position of experimental structures makes possible to bias the set of SA letters used for the 3D assembly. This results in a dramatic decrease of the number of SA letters to consider at each position (close to 40%). As a consequence, the new biased version of PEP-FOLD can be safely extended for mini-proteins up to 50 amino acids. Considering a set of 56 peptides (*i.e.* the whole collection of soluble (not in membrane) mini-protein structures available in the Protein Data Bank of size 25-53 amino acids, not complexed with macromolecules or ions, at a pH more than 5.5), we find that a protocol using 200 simulations of the new PEP-FOLD version, *PEP-FOLD 2*, is able to identify conformations approximating the rigid core by 2.6Å cRMSd on average. Only 2 mini-proteins are misfolded. This is an important improvement compared to the first version of PEP-FOLD that identified best conformations having an average rigid core cRMSd of 3.0Å for 12 misfolded mini-proteins. Using a comparable *de novo* modeling protocol of 200 simulations, Rosetta identifies conformations approximating the rigid core by 3.0Å cRMSd on average.

3 Perspectives

Using PEP-FOLD, the *de novo* prediction of mini-proteins up to 50 amino acids reaches an unprecedented level of reliability. At the same time, the PEP-FOLD approach is very fast (few minutes only) and opens new perspectives for the *in silico* design of peptides.

Acknowledgements

The authors thank INSERM and CNRS for financial support and Pierre Thévenet for his help in producing the latest results.

References

[1] J. Moult, K. Fidelis, A. Kryshtafovych, B. Rost, T. Hubbard, and A. Tramontano, Critical assessment of methods of protein structure prediction-round vii. *Proteins.*, 69(Suppl 8):3–9, 2007.
[2] C.A. Rohl, C.E. Strauss, K.M. Misura, and D. Baker, Protein structure prediction using Rosetta. *Methods Enzymol.,* 383:66–93, 2004.
[3] S.H. Gellman, D.N. Woolfson, Mini-proteins Trp the light fantastic. *Nat. Struct. Biol.* 9(6), 408–10, 2002.
[4] J. Maupetit, P. Derreumaux and P. Tufféry, A fast method for large-scale de novo peptide and miniprotein structure prediction. *J. Comput. Chem.* 31(4):726-38, 2010.
[5] J. Maupetit, P. Derreumaux and P. Tufféry, PEP-FOLD: an online resource for de novo peptide structure prediction. *Nucleic Acids Res.*, 37:W498-503. 2009.
[6] A.C. Camproux, R. Gautier, and P. Tufféry, A hidden markov model derived structural alphabet for proteins. *J. Mol. Biol.*, 339(3):591–605, 2004.

Conformément au souhait des auteurs, cette contribution n'est pas reproduite dans la version en ligne des actes de JOBIM 2011.

Following the wishes of the authors, this paper is not included in the online version of the JOBIM 2011 proceedings.

Conformément au souhait des auteurs, cette contribution n'est pas reproduite dans la version en ligne des actes de JOBIM 2011.

Following the wishes of the authors, this paper is not included in the online version of the JOBIM 2011 proceedings.

Protein-protein Docking based on Shape Complementarity and Voronoi Fingerprints

Thomas BOURQUARD[1], Jerôme AZÉ[2], Anne POUPON[3] and David W. RITCHIE[1]

[1] INRIA Nancy-Grand Est, LORIA, 615 Rue du Jardin Botanique, 54600 Villers-lès-Nancy, France
{Thomas.Bourquard, Dave.Ritchie}@inria.fr

[2] INRIA AMIB group, Équipe Bioinformatique, CNRS UMR8623 Laboratoire de Recherche en Informatique, Université Paris-Sud, 91405 Orsay Cedex, France
Jerome.Aze@lri.fr

[3] Bios group, INRA, UMR85, Unité de physiologie de la Reproduction et des Comportements, F-37380 Nouzilly, France ; CNRS, UMR6175, F-37380 Nouzilly, France ; Université François Rabelais, 37041 Tours, France
Anne.Poupon@inra.fr

Abstract *Predicting the three-dimensional structures of protein-protein complexes is a major challenge for computational biology. Using a Voronoi tessellation model of protein structure, we showed previously that it was possible to use an evolutionary algorithm to train a scoring function to distinguish reliably between native and non-native docking conformations. Here, we show that this approach can be further improved by combining it with rigid body docking predictions generated by the Hex docking algorithm. This new approach is able to rank an acceptable or better conformation within the top 10 predictions for 7 out of the 9 targets available from rounds 8 to 18 of the CAPRI docking experiment.*

Keywords Protein-protein Docking, Evolutionary Algorithms, Hex, CAPRI.

Amarrage Protéine-Protéine par couplage de la Complémentarité de Forme et des Empreintes Voronoï

Résumé *La prédiction de la structure tri-dimensionnelle des complexes protéine-protéine est un enjeu majeur pour la bioinformatique. Nous avions montré dans des travaux précédents que grâce à la modélisation par un diagramme de Voronoï de la structure des protéines, et à l'utilisation d'algorithmes évolutionnaires, il était possible d'optimiser des fonctions de score permettant de distinguer avec une bonne fiabilité les conformations natives des conformations non-natives. Nous montrons dans cet article que cette approche peut être sensiblement améliorée en combinant celle-ci avec des modèles en corps rigide générés par l'algorithme de docking Hex. Cette nouvelle approche, testée sur les cibles CAPRI des rounds 8 à 18, permet de classer dans les 10 meilleures, une conformation quasi-native pour 7 cibles sur les 9 disponibles.*

Mots-clés Amarrage protéine-protéine, Algorithmes Évolutionnaires, Hex, CAPRI.

1 Introduction

L'intégration des signaux extra-cellulaires en une réponse biologique adaptée repose en grande partie sur l'association de complexes protéine-protéine. La détection et la détermination de l'organisation structurale de ces assemblages moléculaires représente donc une étape essentielle pour la compréhension de ces mécanismes et de leur régulation. Si les techniques qui permettent la détermination expérimentale des structures protéiques ont connu des avancées fondamentales, notamment grâce aux projets de génomique structurale, cette détermination reste délicate voire impossible, surtout lorsque l'objet étudié est un complexe. De plus, il a été démontré expérimentalement que le nombre de complexes existant *in vivo* était bien supérieur au nombre de protéines, rendant inenvisageable le recours systématique à l'expérimentation. L'amarrage protéine-protéine, qui consiste à prédire la structure tridimensionnelle de ces assemblages macromoléculaires à partir des structures des partenaires isolés, serait donc un outil crucial dans l'étude du fonctionnement de la cellule [1].

Les différentes procédures existantes traitent généralement le problème en deux étapes : (i) une première phase au cours de laquelle un grand nombre de conformations sont générées (étape limitante en temps de calcul), (ii) puis une seconde phase au cours de laquelle ces différentes conformations sont évaluées afin d'en extraire un sous-ensemble de conformations proches de la conformation native, que nous appellerons conformations quasi-natives.

L'implémentation de l'algorithme de complémentarité de formes Hex sur cartes graphiques (GPU) a permis de réduire considérablement le temps nécessaire pour l'échantillonnage statistique des quelques 10^9 modes d'associations possibles pour deux protéines de taille moyenne [2]. Cet algorithme est capable de générer et évaluer en quelques secondes plusieurs millions de conformations candidates afin d'en extraire un ensemble réduit de conformations d'intérêt [3]. Cependant la fonction d'évaluation intégrée dans Hex ne permet pas d'identifier de manière fiable une solution quasi-native dans cet ensemble.

Dans des travaux précédents, nous avons pu montrer que la représentation des structures protéiques par un modèle "gros-grain" basé sur la tessellation de Voronoï décrivait particulièrement bien les propriétés physico-chimiques aux interfaces protéine-protéine [4]. Ce modèle, couplé à un algorithme évolutionnaire, permet d'optimiser des fonctions de score pour l'amarrage protéine-protéine [5,4,6]. Néanmoins, ces fonctions de score ne sont pas suffisamment sensibles pour envisager l'exploration de l'interactome à grande échelle.

Dans ce travail, nous montrons que la génération de conformations candidates et l'évaluation de la complémentarité de forme par Hex, couplées à l'évaluation des caractéristiques physico-chimiques par les empreintes Voronoï, permettent une prédiction particulièrement efficace de la conformation de complexes protéine-protéine. Afin d'évaluer cette approche, nous nous sommes placés dans le cadre de l'expérience CAPRI [1] [7]. L'objectif de CAPRI est l'évaluation des méthodes d'amarrage protéine-protéine. Des complexes dont la structure tridimensionnelle a été résolue, mais pas encore rendue publique, sont proposés aux prédicteurs. Le processus se déroule en deux étapes : les prédicteurs proposent 10 candidats. Puis ils déposent une centaine de candidats, qui sont alors proposés aux "scoreurs", ce qui permet de tester les fonctions d'évaluation indépendamment de la génération des conformations candidates [8]. Nous présentons dans cet article les résultats obtenus par notre méthode pour les rounds 8 à 18 de cette expérience de "scoring".

2 Méthodes

2.1 Base d'Apprentissage des Complexes Protéine-protéine

Les complexes utilisés pour les procédures d'apprentissage correspondent à ceux utilisés précédemment [6], auxquels nous avons ajouté les complexes des benchmarks 3.0 et 4.0 proposés par *Hwang et al.*[9,10] qui n'étaient pas déjà présents. Ce jeu d'apprentissage comprend 231 complexes liés-non liés ou non liés-non liés (complexes pour lesquels la structure d'au moins un des partenaires isolé est connue). Tous les complexes retenus ont été comparés deux-à-deux suivant la classification SCOP [11] afin d'éviter toute redondance.

Le jeu d'apprentissage est composé de structures natives, correspondant aux structures expérimentales, et de structures non-natives associées. Les conformations non-natives ont été générées avec le logiciel Hex. Pour un complexe donné, Hex recherche la conformation dans laquelle la complémentarité géométrique est la meilleure. Cela permet de définir un axe de référence reliant les centres de gravité des deux partenaires. Les solutions explorées sont alors celles pour lesquelles l'axe reliant les centres de gravité se trouve dans deux cônes dont les sommets sont les centres de gravité, et dont l'axe central est cet axe de référence. Les angles définissant ces cônes peuvent être choisis par l'utilisateur entre 0 et 180 $^\circ$. Dans cette étude, ces deux angles ont été fixés à 45 $^\circ$ car nous avons pu constater que des valeurs supérieures n'augmentaient pas la probabilité de générer des conformations quasi-natives, mais augmentaient très fortement le nombre de conformations non-natives. Afin d'éliminer les modèles trop proches les uns des autres, le seuil de clustering de Hex a été fixé à $9.0\mathring{A}$ Root Mean Square Deviation (*RMSD*). Les examples négatifs du jeu d'apprentissage ont été choisis dans cet ensemble de conformations, et correspondent aux conformations non-natives (ayant un RMSD avec la conformation native supérieur à $10\mathring{A}$) de plus basse énergie trouvés par Hex, et ayant une surface d'interface supérieure à $400\mathring{A}^2$.

1. **C**ritical **A**ssessment of **PR**edictions of **I**nteractions

10 structures non natives pour chaque structure native ont été incluses dans le jeu d'apprentissage (19 dans la comparaison cœur-couronne).

2.2 Empreintes Voronoï et paramètres d'apprentissage

Le modèle "gros-grain" défini dans [6], basé sur la tesselation de Voronoï, a été utilisé pour représenter les structures des complexes. Pour chaque conformation candidate, la triangulation de Delaunay (duale de la tesselation de Voronoï) est construite par utilisation de la CGAL [12]. L'interface est définie comme l'ensemble des acides aminés d'un partenaire en contact avec l'autre partenaire. Cette interface est, soit restreinte aux acides aminés qui ne sont pas en contact avec le solvant : interface cœur, soit non restreinte : interface cœur plus couronne.

Pour chaque conformation, un vecteur de 96 paramètres est calculé et utilisé dans les procédures d'apprentissage ou de test. Ce vecteur comprend le nombre total de résidus à l'interface, l'aire de l'interface, les fréquences et volumes moyens des cellules de Voronoï de chaque type de résidu, les distances et fréquences de paires de résidus regroupés en six catégories physico-chimiques (hydrophobe (IFMLV), aromatique (FYW), polaire (NQ), chargé positivement (HKR), chargé négativement (DE) et petits (AGSTCP)), les fréquences et les volumes moyens de chaque catégorie de résidus (voir [6]).

2.3 Algorithme Évolutionnaire et Procédure d'Apprentissage

À l'aide des attributs d'apprentissage décrits plus haut, des algorithmes évolutionnaires ont été utilisés afin de trouver un ensemble de fonctions permettant de discriminer les conformations quasi-natives et non-natives. La fonction d'adaptation utilisée est l'aire sous la courbe de ROC (Receiver Operating Characteristic). Les fonctions de score apprises dans cette étude sont de la forme :

$$S_j(conf) = \sum_{i=1}^{96} w_i \, |x_i(conf) - c_i|$$

où pour chaque attribut d'apprentissage X_i, x_i, w_i et c_i représentent respectivement les valeurs, poids et valeurs de centrage associés, w_i et c_i étant optimisés au cours de l'apprentissage. L'algorithme évolutionnaire est de type $\lambda + \mu$, avec $\lambda = 20$ parents $\mu = 120$ enfants. Le maximum de générations a été fixé à 500. Les performances ont été évaluées en validation croisée. Un apprentissage correspond à l'optimisation de 30 fonctions de score, et le rang final d'une conformation correspond à la somme des rangs obtenus après application de chacune des 30 fonctions apprises.

Les fonctions de score sont évaluées par la précision et le rappel :

$$\text{Précision} = \frac{VP}{VP+FP} \qquad \text{Rappel} = \frac{VP}{VP+FN}$$

Où VP : vrais positifs, FP : faux positifs et FN : faux négatifs.

2.4 Gestion des Valeurs Manquantes et Normalisation

Dans des travaux précédents, nous avions constaté que les valeurs manquantes ont un impact négatif très important sur les performances des fonctions apprises. Afin de limiter cet impact, nous avons testé plusieurs méthodes de gestion des valeurs manquantes. Nous avons retenu les méthodes les plus fréquemment utilisées pour gérer des valeurs manquantes [13] : remplacement par une valeur constante (0), par une valeur dépendant des données manipulées (valeur minimale, maximale, médiane ou moyenne de l'attribut considéré) ou par des valeurs obtenues sur un sous-ensemble des données manipulées (calcul des exemples les plus proches et remplacement par la valeur moyenne : $k$nn ou $k$means).

Le remplacement des valeurs manquantes par l'approche $k$nn est réalisée de la manière suivante : pour chaque exemple ayant au moins une valeur manquante, ses k plus proches voisins sont recherchés en utilisant une distance euclidienne calculée uniquement entre les valeurs renseignées de l'exemple considéré et le reste

des données disponibles. Puis, pour chaque attribut non renseigné, la valeur manquante est remplacée par la valeur moyenne de cet attribut dans ses k plus proches voisins. Si l'ensemble des plus proches voisins est vide (trop de valeurs manquantes par exemple), ou que pour un attribut les plus proches voisins sont tous non renseignés, alors les valeurs manquantes sont remplacées par les valeurs moyennes calculées sur l'intégralité des données.

Pour l'approche $k$means, les données sont préalablement réparties en k clusters les plus homogènes possibles. La distance intra-cluster est calculée de la même manière que pour l'approche $k$nn. Les clusters sont initialisés avec les exemples contenant le moins de valeurs manquantes (moins de 10%). Puis, dans chaque cluster, les valeurs manquantes des exemples sont remplacées par les valeurs moyennes calculées sur les exemples du clusters. De manière similaire à l'approche $k$nn, si un attribut n'est jamais renseigné dans le cluster, alors la valeur moyenne globale est utilisée pour remplacer les valeurs manquantes de cet attribut.

Enfin, les intervalles de valeurs admissibles pour chaque paramètre sont par définition très hétérogènes. Bien que ces différences d'échelles soient en partie capturées par l'algorithme évolutionnaire via les valeurs de centrage c_i, l'ensemble des attributs dont les valeurs admissibles sont élevées peuvent atténuer voire complètement masquer les attributs ayant des valeurs plus faibles.

Deux procédures de normalisation des données ont été mises en œuvre afin de réduire ce biais :
- la procédure **minMax**, qui normalise les attributs en fonction du minimum et du maximum observés pour le paramètre :

$$x_i(conf) = \frac{x_i(conf) - min(X_i)}{max(X_i) - min(X_i)}$$

- la procédure **meanStd**, qui normalise les attributs en fonction de la moyenne et l'écart-type :

$$x_i(conf) = \frac{x_i(conf) - \bar{X_i}}{\sigma_i}$$

À l'issue de ces deux étapes de pré-traitement, une dernière étape de sélection d'attributs aurait pu être mise en place et ainsi réduire ce problème de valeurs manquantes. Notre choix de représentation des complexes, et notamment le grand nombre de paramètres utilisés, implique nécessairement qu'une partie de ces paramètres soient non renseignés pour un example donné. Cependant, mis à part les paramètres concernant les acides aminés les plus représentés dans les protéines, les paramètres non renseignés varient d'un exemple à l'autre, reflètant la diversité des modes d'interaction, elle-même liée à la diversité des protéines. Considérons un complexe dont l'interface comporte un tryptophane. L'attribut "volume moyen du tryptophane" est essentiel pour la prédiction de cette interface. Or, le tryptophane est un acide aminé très peu représenté dans les protéines, et les attributs correspondant seraient très certainement éliminés par une sélection de paramètres, rendant difficile la prédiction correcte de la structure de ce complexe.

Ainsi, une phase de sélection d'attributs risquerait de nous faire perdre la capacité de représenter efficacement des complexes faisant intervenir, dans leur interface, des résidus peu fréquents dans l'ensemble des complexes étudiés.

2.5 Classification des Conformations

Pour classer les conformations nous avons utilisé les critères définis dans l'expérience CAPRI :
- Haute qualité : $[fnat \geq 0.5$ et $(I_{RMSD} \leq 1$ ou $L_{RMSD} \leq 1)]$
- Moyenne qualité : $[(fnat \geq 0.3$ et $fnnat < 0.5)$ et $(I_{RMSD} \leq 2.0$ ou $L_{RMSD} \leq 5.0)]$ ou $[fnat > 0.5$ et $(I_{RMSD} > 1.0$ ou $I_{RMSD} > 1.0)]$
- Acceptable : $[(fnat \geq 0.1$ et $fnnat < 0.1)$ et $(I_{RMSD} \leq 4.0$ ou $L_{RMSD} \leq 10.0)]$ ou $[fnat > 0.3$ et $(L_{RMSD} > 5.0$ ou $I_{RMSD} > 2.0)]$

Où $fnat$ est la fraction de contacts natifs présents dans la prédiction, $fnnat$ est la fraction de contacts de la prédiction qui sont natifs, I_{RMSD} est le RMSD entre l'interface prédite et l'interface native, L_{RMSD} est le RMSD entre le ligand prédit et le ligand natif, les récepteurs étant superposés.

3 Résultats

3.1 Interface cœur vs Interface couronne

La première question que nous adressons ici est de savoir si les résidus de la couronne, à savoir les résidus de l'interface qui sont en contact avec le solvant, doivent ou non être pris en compte dans l'apprentissage et l'évaluation. Ne pas inclure ces résidus revient à éliminer environ $2/3$ de l'aire de l'interface, et surtout augmente considérablement le nombre de valeurs manquantes. En effet, le pourcentage de valeurs non renseignées pour les structures natives passe de $12,63\%$ en ne considérant que le cœur, à seulement $3,63\%$ en ajoutant la couronne, et de $16,9\%$ à $5,1\%$ pour les conformations non-natives.

Cependant, de nombreuses études ont montré que le cœur et la couronne présentent des caractéristiques physico-chimiques nettement distinctes, ce qui n'est pas favorable dans notre cas. La prise en compte de ces résidus conduit à définir des interfaces contenant plus de résidus polaires et chargés, des volumes moyens associés aux cellules de Voronoï plus importants ou encore des distances entre paires de résidus en interaction également plus grandes. De même, toutes les déviations standards sont plus élevées.

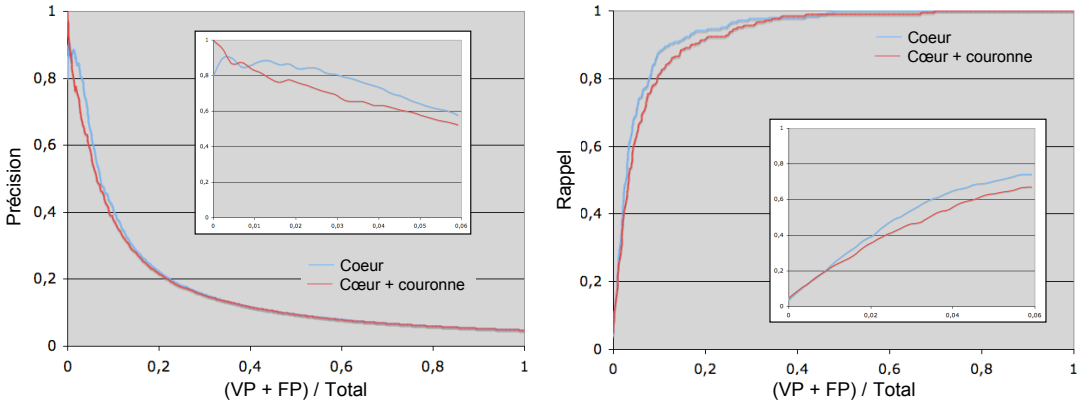


Figure 1. Précision et rappel en fonction de la fraction évaluée positive (VP + FP)/Total, pour un apprentissage en 10 validation croisée, en interface cœur (bleu) ou cœur+ couronne (rouge). La région correspondant aux fractions allant de 0 à 0,06 a été aggrandie (encadrés).

Nous avons réalisé, sur le jeu d'apprentissage en 10-validation croisée, une série d'apprentissages avec les résidus du cœur, et une série d'apprentissages avec les résidus du cœur et de la couronne. Les mesures de précision et rappel montrent que la prise en considération des résidus de cœur uniquement donne de meilleurs résultats (voir Fig. 1). Idéalement, étant donné que le jeu d'apprentissage contient 19 négatifs pour 1 positif, lorsque $(VP + FP)/Total = 0,05$, c'est-à-dire lorsque la fraction de conformations évaluées positives est égale à la fraction de conformations réellement positives, on devrait avoir une précision de 1 (toutes les conformations évaluées positives sont positives), et un rappel de 1 (toutes les conformations positives sont évaluées positives). À cette abscisse, nous obtenons une précision de $0,66$ en interface cœur, contre $0,6$ en interface couronne et des rappels de respectivement $0,69$ contre $0,62$. Ainsi, le "bruit" résultant de la prise en compte des résidus de la couronne à un impact négatif qui est plus important que l'impact positif résultant de la diminution du taux de valeurs manquantes. Par la suite, seuls les résidus du cœur de l'interface seront utilisés.

3.2 Gestion des Valeurs Manquantes et Variants Normalisés

Le fait que les résidus de la couronne ne puissent pas être utilisés rend la gestion des valeurs manquantes d'autant plus importante. Par ailleurs, les différents paramètres ayant des valeurs dans des ordres de grandeurs très différents, il est nécessaire de déterminer si les valeurs doivent être normalisées, et si oui par quelle méthode. Afin de répondre à ces deux questions, nous avons réalisé des apprentissages en 3-validation croisée en faisant varier la normalisation des données et le remplacement des valeurs manquantes.

Les résultats obtenus (Table 1) montrent que la normalisation améliore de manière très sensible les performances des fonctions de score. En effet, quelle que soit la méthode de gestion des valeurs manquante utilisée,

	zéro	min	max	med	moy	$k$means	$k$nn		
Normalisation							$k=3$	$k=5$	$k=10$
aucune	0,62	0,63	0,78	0,72	0,68	0,72	0,68	0,69	0,70
minMax	0,80	0,80	**0,81**	**0,81**	0,80	0,80	0,80	0,80	0,80
meanStd	0,78	**0,81**	0,74	0,79	0,79	0,78	0,78	0,79	0,80

Table 1. Valeurs des critères de ROC obtenus pour différentes méthodes de gestion des valeurs manquantes et avec ou sans normalisation des données.

le critère de ROC est plus élevé avec normalisation que sans. Dans la majorité des cas, la méthode minMax semble plus performante, excepté lorsque les valeurs manquantes sont remplacées par le minimum observé. En ce qui concerne les méthodes de gestion des valeurs manquantes, trois des méthodes donnent des résultats équivalents : remplacement par le minimum, le maximum et la médiane. Dans la suite de cette étude nous avons appliqué le remplacement des valeurs manquantes par le maximum observé et la normalisation via l'approche minMax.

On peut noter ici que la précision obtenue sans normalisation, et en remplaçant les valeurs manquantes par la moyenne, correspondant à la configuration que nous utilisions précédemment, est supérieure à celle que nous avions obtenue sur notre précédent jeu d'apprentissage "(0.62)". Ceci est uniquement dû à l'utilisation des structures non-natives générées par Hex. L'amélioration de la précision est lié au fait que ces conformations non-natives sont plus "vraisemblables" que celles précédemment utilisées : bonne complémentarité géométrique et bonne énergie d'interaction en particulier.

3.3 Résultats sur les cibles CAPRI

Afin de vérifier la validité de notre approche, nous avons repris l'expérience de "scoring" CAPRI des rounds 8 à 18, et comparé les résultats avec ceux obtenus par les autres participants. Certaines cibles ont été éliminées de l'étude :

- Les cibles 23, 24, 26, 27 et 28 car les classements selon les critères CAPRI ne sont pas disponibles.
- La cible 30 car il s'agit d'un homodimère, le fait que ce dimère soit biologique est par ailleurs encore en discussion, et les auteurs n'ont pu le démontrer expérimentalement.
- La cible 31 car la structure native n'est pas disponible, il n'est donc pas possible d'évaluer les résultats.
- Les cibles 36 et 38 car aucune conformation au moins acceptable n'est présente dans les ensembles de conformations.
- Les cibles 33 et 34 car il s'agit de complexes protéine-ARN.

Les résultats obtenus pour les cibles restantes sont présentés dans la Table 2. Notre méthode est capable de classer une solution de qualité moyenne ou haute pour 7 des 9 cibles (voir Fig. 2), ce qui en fait la méthode la plus performante. Le cas de la cible 40 est particulièrement intéressant. En effet, ce complexe est un trimère constitué de l'inhibiteur de protéinase à sérine A, et de deux trypsines cationiques [14]. Il y a ainsi deux interfaces, notées T40A et T40B, et un seul ensemble de conformations candidates. Notre méthode classe une conformation de haute qualité pour l'interface A en première position et une conformation de haute qualité pour l'interface B en seconde position. Il y a par ailleurs dans le top 10 une autre conformation quasi-native pour chacune des deux interfaces.

Le cas de la cible 35[15], pour laquelle aucun groupe n'est parvenu à isoler une solution au moins acceptable dans le top10, est un peu particulier. En effet, il ne s'agit pas réellement de deux protéines, mais de deux domaines de la même protéine qui ont été artificiellement séparés, puis co-cristallisés. Or, nous avons déjà montré dans des travaux précédents que les valeurs moyennes de nos paramètres sont significativement différentes à l'intérieur des protéines et à l'interface entre deux protéines.

Pour la cible 39, aucun participant n'a été capable d'extraire une bonne solution parmi les conformations proposées par l'ensemble des prédicteurs. Ceci s'explique en grande partie par le fait que la structure de l'un des deux partenaire n'était pas connue, et a été modélisée avec un succès relativement mitigé. De ce fait, il y a seulement 4 conformations au moins acceptables dans l'ensemble proposé (3 de qualité moyenne et 1 acceptable) pour 1 296 conformations incorrectes.

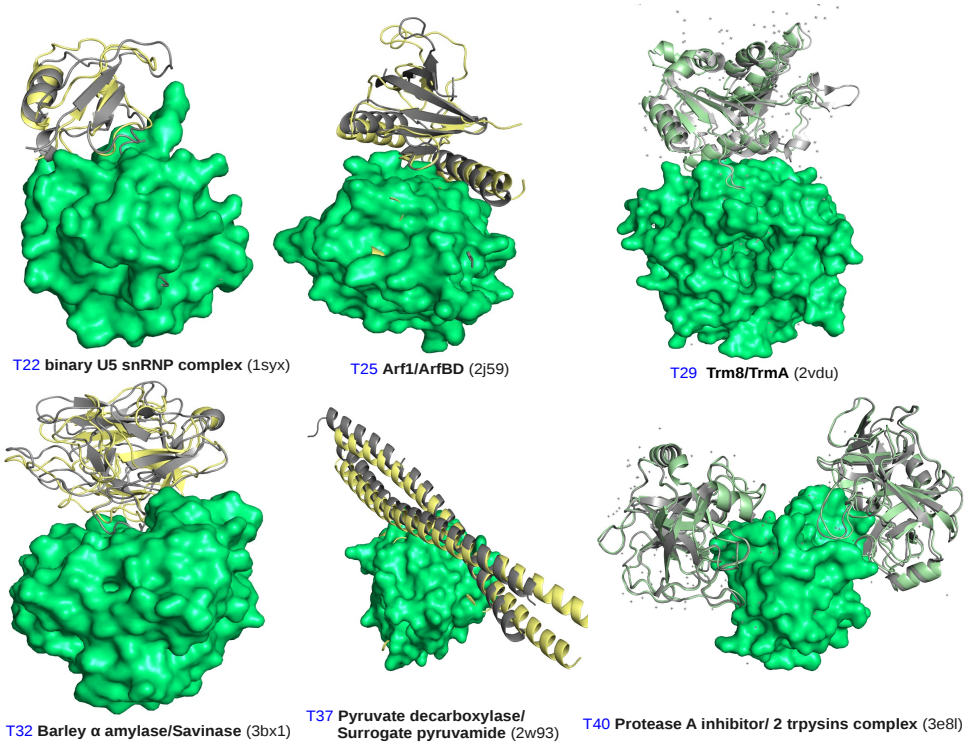


Figure 2. Superpositions des structures des complexes obtenues par cristallographie avec les structures au moins acceptables obtenues par évaluation lors de la seconde phase CAPRI présentes dans les 10 meilleures conformations ou "Top10". Les structures cristallines sont en gris, le recepteur en vert, le ligand représenté en ruban apparaît en jaune (Moyenne qualité) ou vert (Haute qualité).

Groups	T22	T25	T29	T32	T35	T37	T39	T40A	T40B
C Wang	0	**	0	0	*	**	0	***	**
A.M.J.J Bonvin	*	*	**	0	0	*	0	***	**
H. Wolfson	-	**	0	0	0	*	0	*	0
P. A. Bates	-	-	**	0	0	***	0	***	0
Z. Weng	-	-	**	0	0	***	0	***	0
J. F.-Recio	-	**	***	0	0	0	0	0	0
X. Zou	-	-	-	0	0	***	0	***	***
T. Haliloglu	-	-	-	-	-	**	0	***	**
C. J. Camacho	-	-	**	-	-	-	-	***	***
M. Takeda-Shitaka	-	-	0	0	0	-	-	***	**
I. Vakser	-	-	-	**	0	0	0	-	-
VDOCK	0	*	**	***	0	0	0	*	0
VDOCK-Hex Models	**	**	***	**	0	**	0	***	***
	(6)	(1)	(10)	(1)	(145)	(1)	(80)	(1)	(2)

Table 2. Meilleures conformations détectées dans les top 10 des différents scorers. **0** : Aucune solution au moins acceptable n'a été trouvée ; **-** : scorer n'ayant pas participé. Pour notre méthode (VDOCK-Hex Models) lorsqu'aucune conformation au moins acceptable n'est présente dans le top 10 le rang de la première conformation quasi-native est indiqué entre parenthèses.

4 Conclusion

La génération d'exemples négatifs de très basse énergie par Hex, la restriction de l'étude aux résidus appartenant au cœur de l'interface, la normalisation des valeurs des paramètres, et enfin la bonne gestion des valeurs manquantes, nous ont permis d'améliorer considérablement les performances de notre méthode. L'impact, au niveau de l'apprentissage, des exemples négatif générés par Hex est très important.

Cependant, si les choix faits à la suite de cette étude permettent une meilleure performance globale, dans certains cas particuliers ils ont un impact négatif. Par exemple, le remplacement des valeurs manquantes par 0 permet d'améliorer le classement de la première solution de haute qualité pour la cible 29. D'autre part, nous avons pu également montrer que dans certains cas l'utilisation des interfaces cœur plus couronne était plus performante. Il serait donc intéressant de mieux définir les cas dans lesquels ces méthodes alternatives sont plus performantes afin de permettre un choix de la méthode la plus adaptée en fonction du complexe à prédire.

Remerciements

Ce projet a été supporté par le programme ANR-08-CEXC-017-01.

Références

[1] D.W. Ritchie, Recent progress and future directions in protein-protein docking. *Curr Protein Pept Sci*, 9(1) :1–15, 2008.

[2] J.C. Mitchell, R. Kerr and LF. Ten Eyck, Rapid atomic density methods for molecular shape characterization. *J Mol Graph Model*, 19 :325–330, 2001.

[3] D.W. Ritchie and V. Venkatraman, Ultra-fast FFT protein docking on graphics processors. *Bioinformatics*, 26(19) :2398–2405, 2010.

[4] J. Bernauer J., Azé, J. Janin and A. Poupon, A new protein-protein docking scoring function based on interface residue properties. *Bioinformatics*, 23(5) :555–562, 2007.

[5] J. Bernauer, R. P. Bahadur, F. Rodier, J. Janin and A. Poupon, DiMoVo : a Voronoi tessellation-based method for discriminating crystallographic and biological protein-protein interactions. *Bioinformatics*, 24(5) :652–658, 2008.

[6] T. Bourquard, J. Bernauer, J. Azé and A. Poupon, Comparing Voronoi and Laguerre tessellations in the protein-protein docking context. In *Sixth International Symposium on Voronoi Diagrams (ISVD)*, pages 225–232, 2009.

[7] J. Janin, K. Henrick, J. Moult, LT Eyck, MJ Sternberg, S Vajda, I Vakser and SJ. Wodak, CAPRI : a Critical Assessment of PRedicted Interactions. *Proteins*, 52 :2–9, 2003.

[8] J. Janin and S.J. Wodak, The Third CAPRI Assessment meeting. *Structure*, 15 :755–759, 2007.

[9] H. Hwang, B. Pierce, J. Mintseris, and Z. Janin J.and Weng, Protein-protein docking benchmark version 3.0. *Proteins*, 73(3) :705–709, 2008.

[10] H. Hwang, T. Vreven, J. Janin and Z. Weng, Protein-protein docking benchmark version 4.0. *Proteins*, 78(15) :3111–3114, 2010.

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

[12] J.-D. Boissonnat, O. Devillers, S. Pion, M. Teillaud and M. Yvinec, Triangulations in CGAL. *Comput. Geom. Theory Appl.*, 22 :5–19, 2002.

[13] E. Acuna and C. Rodriguez, The treatment of missing values and its effect in the classifier accuracy. *Classification, Clustering and Data Mining Applications*, pages 639–648, 2004.

[14] R. Bao, J.C. Zhou, C. Jiang, S.X. Lin, C.W. Chi and Y. Chen, The ternary structure of double-headed arrowhead protease inhibitor api-a complexed with two trypsins reveals a novel reactive site conformation. *J Biol Chem*, 284 :26676–26684, 2009.

[15] S. Najmudin, BA Pinheiro, JAM Prates, MJ Romao and CMGA Fontes, Putting an n-terminal end to the clostridium thermocellum xylanase xyn10b story : Crystallographic structure of the cbm22-1-gh10 modules complexed with xylohexaose. *Journal of Structural Biology*, 172 :353–362, 2010.

Session 2 : Evolution

Conférence invitée

Patrick Forterre

Institut Pasteur, Paris and Institut de Génétique et Microbiologie, Université Paris-Sud, France.

The Role of Viruses (Virocell) in the Origin and Diversification of Biological Information

Viruses are traditionally considered as vehicles of information from cells to cells (horizontal gene transfer) but not as inventors of new information. Viruses are often considered as byproducts of cellular evolution and non living biological "entities". I will discuss recent concepts on the nature of viruses (the virocell concept) that makes justice of their importance in the biosphere and helps to understand the major role that viruses and derived elements have played in the origin and evolution of biological information, systems and organisms.

Session 3.A : Protein Structure

Conformément au souhait des auteurs, cette contribution n'est pas reproduite dans la version en ligne des actes de JOBIM 2011.

Following the wishes of the authors, this paper is not included in the online version of the JOBIM 2011 proceedings.

Conformément au souhait des auteurs, cette contribution n'est pas reproduite dans la version en ligne des actes de JOBIM 2011.

Following the wishes of the authors, this paper is not included in the online version of the JOBIM 2011 proceedings.

New Developments and Applications for Protein Peeling Algorithm

Jean-Christophe GELLY[1,2,3] and Alexandre G. de BREVERN[1,2,3]

[1] INSERM U665, 6 rue Alexandre Cabanel, 75739, Paris, Cedex 15, France

[2] Univ Paris Diderot, Sorbonne Paris Cité, UMR-S665, Paris, F-75013, France

[3] Institut National de la Transfusion Sanguine, 75739, Paris, Cedex 15, France

{jean-christophe.gelly, alexandre.debrevern}@univ-paris-diderot.fr

Abstract *Analysis of protein structures is of great interest to explain molecular mechanisms in biology. We developed Protein Peeling approach to sub divide a protein structure in smaller sub-units. These elements called Protein Units (PUs) allow a simple and detailed analysis of protein structure organization. We propose here new methods based on Protein Peeling to assist and conduct elaborate studies: unstructured terminal segments recognition, novel scoring function for PUs characterization and lastly structural domains identification.*

Keywords Protein structure, structural domain, structure analysis.

1 Introduction

Studying protein structure anatomy and architecture is fundamental to understand protein folding, stability, function or evolution. The classical approach to describe protein structure is to consider them, at a low level, as series of alpha-helices and beta-sheets, or at a higher level as an arrangement of protein domains. These descriptors are sometimes inadequate for explaining the complexity of protein structure arrangement.

In order to better explain and describe the structure of proteins, we have proposed a new and intermediate view, the Protein Units (PUs) [1]. This is a novel level of description between secondary structures and domains. PUs are linear and compact sub-region of protein structure defined by high number of intra-PU contacts and low number of inter-PU contacts. The Protein Peeling (PP) method has been developed to identify PUs and has been implemented in a web server [2].

We propose here a new version of Protein Peeling [3] incorporating new functionalities:

- Structural domains identification
- Novel scoring function for PUs characterization
- Unstructured terminal segments recognition

2 Material and Methods

We have developed a new bottom up algorithm called Domain Reconstruction (DR) that provides domains identification. The principle is to combine PUs to identify structural domains accordingly to Contact Ratio and Contact Probability Density criterions. One of the main interests of this method is to suggest alternatives domains delineation.

To characterize stability of PU, a pseudo-energetic criterion based on statistical potentials computed on carbon alpha has been also proposed.

Finally we have defined a new assignment method [4] to identify unstructured N or C termini segments, based on a new refined non-redundant protein structure databank: PUs are identified as unstructured if they are isolated at the first cutting event and no more thereafter.

3 Results

Protein Peeling (PP) and Domain Reconstruction (DR) methods have been both implemented in a web server [3] (http://www.dsimb.inserm.fr/dsimb_tools/peeling3). A protein structure is submitted to the server in PDB format then is analyzed by PP and DR algorithms. Results page give splitting events of structure into PUs using various visual outputs: dendrogram representing cutting events with the different generated PUs, secondary structure contents, visual representation of PUs and domains. Different measures to characterize PUs and domains are also proposed.

Domains proposed by DR have been compared to classical protein domains benchmark datasets. Figure 1 shows interest to provide alternative domain delineations on Actin structure (PDB code 1ATNA). Our method is able to identify different alternatives both on number and on boundaries of domains.

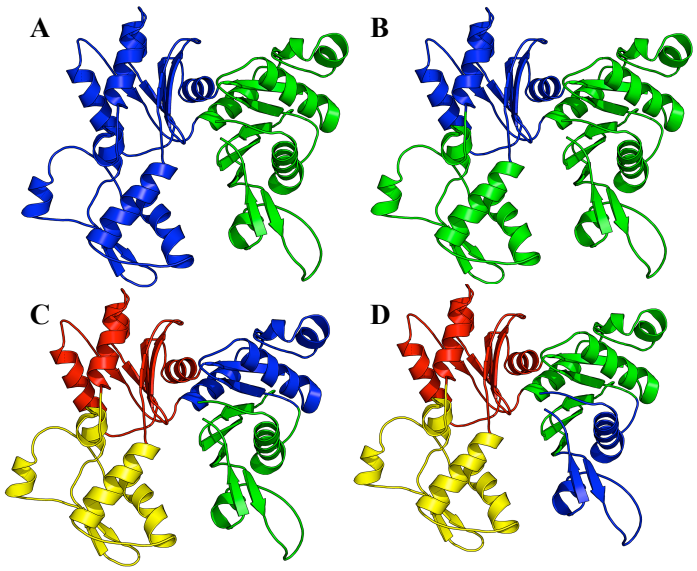


Figure 1. Comparisons of 4 different domain delineation proposed by DR algorithm for Actin (pdb code 1ATNA). (A) is identical to SCOP [5] delineation, (C) to DALI [6] and (D) to Crystallographers [7].

Acknowledgements

This work was supported by the French National Institute for Blood Transfusion (INTS), the French National Institute for Health and Medical Research (INSERM), the University of Paris Diderot - Paris 7.

References

[1] J.C. Gelly, A.G. de Brevern and S. Hazout, 'Protein 'Protein Peeling': an approach for splitting a 3D protein structure into compact fragments. *Bioinformatics*, 22:129-33, 2006.

[2] J.C. Gelly, C. Etchebest, S. Hazout, and A.G. de Brevern, Protein Peeling 2: a web server to convert protein structures into series of protein units. *Nucleic Acids Res*, 34:W75-8, 2006.

[3] J.C. Gelly and A.G. de Brevern, Protein Peeling 3D: New tools for analyzing protein structures. *Bioinformatics*, 27:132-133, 2011.

[4] G. Faure, A. Bornot, and A.G. de Brevern, Analysis of protein contacts into Protein Units. *Biochimie*, 91:876-887, 2009.

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

[6] L. Holm, and C. Sander, C. Parser for protein folding units. *Proteins*, 19:256-68, 1994.

[7] S. Jones, M. Stewart, A. Michie, M.B. Swindells, C. Orengo, and J.M. Thornton, Domain assignment for protein structures using a consensus approach: characterization and analysis. *Protein Sci*, 7:233-242, 1998.

Conformément au souhait des auteurs, cette contribution n'est pas reproduite dans la version en ligne des actes de JOBIM 2011.

Following the wishes of the authors, this paper is not included in the online version of the JOBIM 2011 proceedings.

Conformément au souhait des auteurs, cette contribution n'est pas reproduite dans la version en ligne des actes de JOBIM 2011.

Following the wishes of the authors, this paper is not included in the online version of the JOBIM 2011 proceedings.

Conformément au souhait des auteurs, cette contribution n'est pas reproduite dans la version en ligne des actes de JOBIM 2011.

Following the wishes of the authors, this paper is not included in the online version of the JOBIM 2011 proceedings.

Conformément au souhait des auteurs, cette contribution n'est pas reproduite dans la version en ligne des actes de JOBIM 2011.

Following the wishes of the authors, this paper is not included in the online version of the JOBIM 2011 proceedings.

Conformément au souhait des auteurs, cette contribution n'est pas reproduite dans la version en ligne des actes de JOBIM 2011.

Following the wishes of the authors, this paper is not included in the online version of the JOBIM 2011 proceedings.

Conformément au souhait des auteurs, cette contribution n'est pas reproduite dans la version en ligne des actes de JOBIM 2011.

Following the wishes of the authors, this paper is not included in the online version of the JOBIM 2011 proceedings.

Conformément au souhait des auteurs, cette contribution n'est pas reproduite dans la version en ligne des actes de JOBIM 2011.

Following the wishes of the authors, this paper is not included in the online version of the JOBIM 2011 proceedings.

Conformément au souhait des auteurs, cette contribution n'est pas reproduite dans la version en ligne des actes de JOBIM 2011.

Following the wishes of the authors, this paper is not included in the online version of the JOBIM 2011 proceedings.

Session 3.B : Regulation and Pathways

Identification of Cis-Regulatory Elements Involved in Zygotic Genome Activation During Early *Drosophila melanogaster* Embryogenesis

Elodie DARBO[1], Thomas LECUIT[2], Denis THIEFFRY[3] and Jacques VAN HELDEN[4]

[1] TAGC, UMR628 INSERM, 163 avenue de Luminy, Parc Scientifique de Luminy, 13288, Marseille, Cedex 09, France
darbo@tagc.univ-mrs.fr

[2] IBDML, UMR 6216, 63 avenue de Luminy, Parc Scientifique de Luminy, 13288 Marseille Cedex 09, France
lecuit@ibdml.univ-mrs.fr

[3] IBENS, INSERM 1024 – CNRS 8187, 46, rue d'Ulm, F-75230 Paris cedex 05, France
thieffry@ens.fr

[4] BiGRe, Boulevard du Triomphe, Campus Plaine, Université Libre de Bruxelles, CP263, B-1050 Bruxelles, Belgium
jvhelden@ulb.ac.be

Keywords cis-regulatory elements, Drosophila, transcriptome, motif discovery, pattern matching, RSAT.

1 Introduction

In all metazoa, transcription is inactive during the early stages of embryonic development. Zygotic Genome Activation (ZGA) is triggered at a specific stage in each species. Some of the molecular actors are likely to be conserved from Drosophila to Human, but most of the involved regulatory mechanisms remain to be discovered. In Drosophila embryos, no transcription occurs during the first seven cell cycles after fertilization. The control of the early steps of development is ensured by maternal mRNAs loaded in the egg during oogenesis. ZGA occurs in two waves: the first wave involves about 60 genes, whose transcription is activated during mitotic cycle 8; a second wave involving more than 300 genes occurs at the 14th mitotic cycle. Using Drosophila as model organism, we attempt to unravel the regulatory mechanisms involved in Zygotic Genome Activation.

2 Results

From transcriptome data [1], we selected 169 genes activated during ZGA. In order to find over represented motifs shared by the selected genes, we analyzed their non coding regions (5kb upstream TSS, first intron, UTRs) with discovery and research approaches (*RSATools* suite [2]; *CisTargetX* [3]). Two known motifs were found by both methods. First, the TAGteam motif known to be involved in the minor wave of ZGA [4]. This motif is bound by Zelda, an activator of pre-cellular blastoderm genes [5], as well as by Grh, a repressor for some zygotic genes [6]. The second motif corresponds to Trithorax-like (Trl), which is involved in chromatin modulation, genes activation and repression. We discovered three novel motifs with no correspondence in motifs databases. Next, functional enhancers are regularly formed by multiple transcription factor binding sites (TFBS). Thus, we searched for cis-regulatory elements enriched regions (CRERs) combining significant clusters of putative TFBS. The analysis of the five previous motifs resulted in 421 identified CRERs. To reinforce previous results, we used available ChIP-seq data. Using ChIP-seq data for Trl in the 0-8h embryo, we observed a significant enrichment of Trl peaks in the non coding regions of the genes activated during ZGA. Moreover, high Trl binding-signals significantly discriminated CRERs from random regions. Since CBP is known to interact with Trl, we conducted the same analysis with CBP ChIP-seq data from 0-4h embryo. In fact, CBP is known to interact with some members of the trithorax group, particularly with Trl in the hsp70 gene promoter, where Trl/HSP facilitate CBP recruitment for transcriptional activation [7]. As in the case of Trl, CBP peaks are significantly enriched in non coding regions of selected genes and high CBP binding-signals also significantly discriminate CRERs from random regions. Finally, using the novel workflow *peak-motifs* [8], which combines several motifs discovery algorithms, to analyze the collections of CBP and Trl peaks overlapping non coding sequences of ZGA-induced gene, we found the Trl binding motif and the CAGGTAG (which belongs to TAGteam motifs), respectively.

3 Discussion and perspective

Altogether, these results lead us to formulate several hypotheses. First, the presence of novel significantly over-represented motifs could imply the participation of the corresponding factors in the activation of genes during ZGA. These factors could cooperate with Zelda or act independently (some genes do not contain any TAGteam sites). Next, Trl stands out as a good candidate as a general activator during ZGA. Indeed, Trl is produced during oogenesis and it is uniformly distributed in the early embryo. An homolog of Trl is involved in the early zygotic activation of hsp70.1 gene in the mouse [9]. Moreover, Trl activity is repressed by TTK [10], which is a well known repressor of zygotic activation titrated by increased amount of DNA during nuclear clivages in the syncytial blastoderm [11]. Finally, already known to be involved in the dorso-ventral patterning through its recruitment by Dorsal, CBP could have a more general role in ZGA. Indeed, the very significant enrichment of CBP sites in non coding regions of ZGA-induced genes (122 genes over 169) and the presence of a TAGteam motif at the center of the peaks could be explained by the recruitment of CBP by Zelda. However, some CRERs do not contain any TAGteam sites and overlap with strong signals of CBP binding. CBP could thus also be recruited by unidentified factors binding the novel discovered sites.

To address these pending questions and test our hypotheses, we have selected a thirty of CRERs, which will be inserted in reporter constructs to assess their potential regulatory roles during ZGA.

References

[1] F. Pilot, J.M. Philippe, C. Lemmers, J.P. Chauvin and T. Lecuit, Developmental control of nuclear morphogenesis and anchoring by charleston, identified in a functional genomic screen of Drosophila cellularisation. *Development*, 133(4):711-23, 2006.

[2] M. Defrance, R. Janky, O. Sand and J. van Helden, Using RSAT oligo-analysis and dyad-analysis tools to discover regulatory signals in nucleic sequences. *Nat. Protoc.*, 3, 1589-1603, 2008.

[3] S. Aerts, X.J. Quan, A. Claeys, M. Naval Sanchez, P. Tate, J. Yan and B. Hassan. Robust target gene discovery through transcriptome perturbations and genome-wide enhancer predictions in Drosophila uncovers a regulatory basis for sensory specification, *PLoS Biology,* 8(7):e1000435, 2010.

[4] J.R. ten Bosch, J.A. Benavides and T.W. Cline, The TAGteam DNA motif controls the timing of Drosophila pre-blastoderm transcription. *Development,* 133(10):1967-77, 2006.

[5] H.L. Liang, C.Y. Nien, H.Y. Liu, M.M. Metzstein, N. Kirov and C. Rushlow, The zinc-finger protein Zelda is a key activator of the early zygotic genome in Drosophila, *Nature*, 456(7220):400-3 ,2008.

[6] M.M. Harrison, M.R. Botchan, T.W. Cline and Grainyhead and Zelda compete for binding to the promoters of the earliest-expressed Drosophila genes. *Dev. Biol*., 345(2):248-55, 2010.

[7] S.T. Smith, S. Petruk, Y. Sedkov, E. Cho, S. Tillib, E. Canaani and A. Mazo, Modulation of heat shock gene expression by the TAC1 chromatin-modifying complex, *Nat. Cell Biol.*, 6 (2):162-7, 2004.

[8] M. Thomas-Chollier , M. Defrance, A. Medina-Rivera, O. Sand, C. Herrmann, D. Thieffry and J. van Helden, RSAT 2011: Regulatory Sequence Analysis Tools, *Nucleic Acid Res.*, submitted.

[9] A. Bevilacqua, M.T. Fiorenza and F. Mangia, A developmentally regulated GAGA box-binding factor and Sp1 are required for transcription of the hsp70.1 gene at the onset of mouse zygotic genome activation. *Development,* 127 (7):1541-51, 2000.

[10] S. Pagans, M. Ortiz-Lombardía, M.L. Espinás, J. Bernués and F. Azorín, The Drosophila transcription factor tramtrack (TTK) interacts with Trithorax-like (GAGA) and represses GAGA-mediated activation. *Nucleic Acids Res.*, 30 (20):4406-13, 2002.

[11] W. Tadros and H.D. Lipshitz, The maternal-to-zygotic transition: a play in two acts, *Development,* 136, 3033-3042, 2009.

An Integrative Signaling Pathway Analysis for Determining Master Regulators and Dysregulated Pathways in Her2 Over-Expressed Human Breast Cancer

Paola VERA-LICONA[1,2,3], Andrei ZINOVYEV[1,2,3], Inna KUPERSTEIN[1,2,3], Olga KEL[4], Alexander KEL[4], Thierry DUBOIS[5], Gordon TUCKER[6] and Emmanuel BARILLOT[1,2,3]

[1] Institut Curie, 26 rue d'Ulm, 75248 Paris, France

Paola.Vera-Licona@curie.fr

[2] INSERM U900, 75248 Paris, France

[3] Mines Paris Tech, 77300 Fontainebleau, France

[4] geneXplain GmbH, Am Exer 10b, 38302 Wolfenbuettel, Germany

[5] Institut Curie, Département de Transfert, 75248 Paris, France

[6] Institut de Recherches Servier, 11 rue des Moulineaux, 92150 Suresnes, France

With the increase in high-throughput data, signal transduction pathways' reconstruction and analysis approaches are used to gain understanding of the molecular mechanisms involved in cancer diseases.

In the specific scenario of gene expression data, integrative approaches need to be further developed to study molecular processes simultaneously at both transcriptional and post-transcriptional level.

At the transcriptional level, protein-protein interaction (PPI) networks provide a global picture of cellular function and biological processes. However relevant functional changes of transcription factors (TFs), which commonly require post-transcriptional modifications or are regulated by protein-protein interactions, will not be captured by typical microarray experiments and thus will be missed by network construction methods that rely on such observational data.

We present a combined method based on the integration of protein-protein interaction (PPI) network analysis in BiNoM[1] and TFs analysis in Explain[2]. First we construct the signal transduction pathways controlling the activities of the corresponding TFs and identify master regulators for the observed transcriptomic microarray patterns.

We apply this pipeline to study human epidermal growth factor receptor over-expressing (Her2+) breast cancer. We use transcriptomic microarrays to compare Her2+ data with that obtained from normal breast tissue samples. We identify master regulators of the ERBB family pathways together with less expected molecular mechanisms potentially involved in the molecular pathology of Her2+ breast cancer. These components provide new insights and potentially reveal new therapeutic approaches, including those based on synthetic lethality paradigm.

[1] **BiNoM** http://bioinfo-out.curie.fr/projects/binom/

[2] **Explain** http://www.biobase-international.com/index.php?id=572

Identification of Shortened 3'Untranslated Regions and Impact on MicroRNA Regulation

Loredana MARTIGNETTI[1,2,3], Karine LAUD-DUVAL[1,4] , Franck TIRODE[1,4] , Emmanuel BARILLOT[1,2,3], Olivier DELATTRE[1,4] and Andrei ZINOVYEV[1,2,3]

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

[2] INSERM U900, Paris, F-75248

[3] Mines Paris Tech, Fontainebleau, F-77300

[4] INSERM U830, Paris, F-75248

{loredana.martignetti, karine.laud, franck.tirode, emmanuel.barillot, olivier.delattre, andrei.zinovyev}@curie.fr

Keywords 3'UTR, alternative polyadenylation, microRNAs.

1 Introduction

In eukaryotes, genes are regulated at many different levels to produce the appropriate set of proteins for specific cell types. The discovery of RNA silencing pathways focused the attention on post-transcriptional control as a key layer of regulation in several biological processes. Untranslated regions of eukaryotic mRNAs contain motifs that are essential to regulate post-transcriptional processes (e.g. mRNA processing, export, surveillance, silencing by microRNAs and turnover). At the end of every mRNA there is a signal indicating that the end of the mRNA is reached (the polyadenylation signal). In many genes, two or more polyadenylation signals are found in the 3′ UTR, so that different isoforms with different 3'UTR length can be expressed. This mechanism, called alternative polyadenylation (APA) is quite common in human mRNAs and it is subject to tissue or condition specificity [1]. Recently it has been shown that cancer cells often expressed substantial amounts of mRNA isoforms with shorter 3' UTRs [2]. This is relevant from the point of view of post-transcriptional regulation because if the 3'UTR of a mRNA is shorter or missing, miRNAs and other regulatory proteins are not longer able to bind.

2 Methods and Results

We present here a computational procedure for systematically identifying APA events by Affymetrix GeneChip microarrays. The advantage of this technology compared with more recent and promising ones such as exon arrays and RNA-Seq is that, given the relatively small cost, a typical study includes a considerably higher number of experiments. Moreover, the design of Affymetrix Gene Chips is well-suited for 3'UTR analysis of a large number of genes. The proposed approach requires as input the expression profile from Affymetrix GeneChip array for the samples of interest. As final result of our analysis we obtain a set of genes expressing short 3'UTR isoform in a minimum number of the analyzed samples.

Initially, Affymetrix GeneChip single probes are assigned to CDS or 3'UTR of the transcript, according to NCBI RefSeq database annotation (Release 45). Then we define for each RefSeq two distinct meta-probesets, the first one including probes covering specifically the CDS and a second one including probes covering specifically the 3'UTR. The expression ratio between these two meta-probesets is expected to be equal to one in case the 3'UTR is not subject to shortening. A high value of CDS:3'UTR expression ratio is indicative of variation in the expression between the CDS and the 3'UTR and it can be interpreted as an event of short 3'UTR isoform expression.

The procedure has been applied to expression data from 75 samples of Ewing's sarcoma patients generated by Affymetrix U133A microarray. We used all sequences supported by RefSeq and required at least four probes in both CDS and 3'UTR meta-probesets for each gene. Among the 5500 genes selected in this way, we extracted a list of 266 genes showing short 3'UTR expression in at least 10% of Ewing's sarcoma patient samples. We checked whether the extracted genes have multiple annotated 3'UTR isoforms. The extracted gene list has been crossed with gene entries with multiple polyadenylation signals confirmed

by both AltTrans and AltPas polyA site databases (2856 entries) [3]. The overlapping list contains 74 genes ($Pv \sim 10^{-9}$), confirming that our procedure enables us to identify candidate 3'UTR shortening events.

The impact of 3'UTR shortening on microRNA regulation was evaluated by crossing results of meta-probeset expression analysis with predictions of microRNA binding sites by TargetScan 5.1 algorithm. Predicted microRNA binding sites have been distinguished by their position with respect to the polyadenylation cleavage signals annotated for the corresponding gene. We defined as alternative microRNA binding sites that sites located between two alternative polyadenylation cleavage sites. A list of alternative microRNA binding sites has been compiled for all genes showing shortened 3'UTR expression in Ewing's sarcoma data. As an example, the gene DCN (Bone proteoglycan II, a negative modulator of TGF-beta) is associated with very high CDS:3'UTR expression ratio in Ewing' sarcoma patients (Fig. 1a). Figure 1b shows that microRNA binding sites for hsa-miR-496 and hsa-miR-376c are located between two alternative polyadenylation sites in DCN 3'UTR and they are lost when the short 3'UTR isoform is expressed.

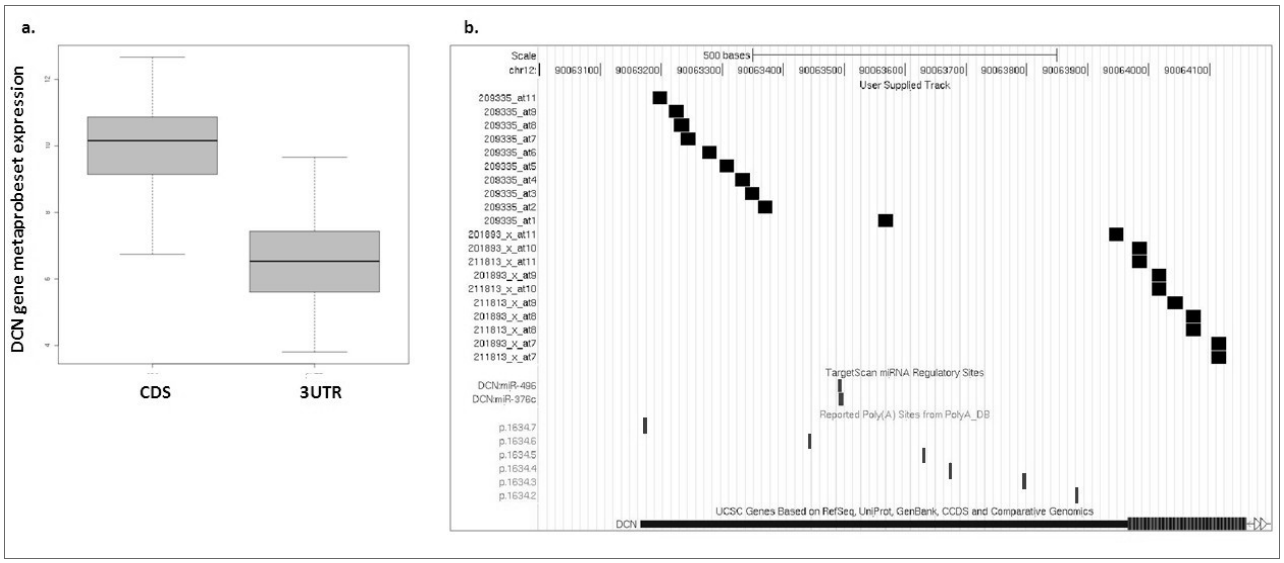


Figure 1. DCN gene results for meta-probeset expression analysis in Ewing's sarcoma (a) and for alternative microRNA binding sites identification (b).

3 Conclusion

These results show that the proposed approach, based on appropriate meta-probeset definition to target specifically CDS or 3'UTR, is able to identify valid candidate events of shortened 3'UTR expression. The results obtained from the analysis of Ewing's sarcoma data overlap significantly with available annotations of APA events, confirming the interest of the proposed approach.

References

[1] F. Ozsolak, P Kapranov, S. Foissac, S.W. Kim, E. Fishilevich, A.P. Monaghan, B. John and P.M. Milos, Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell,* 143(6):1018-29, 2010.

[2] C. Mayr and D.P. Bartel, Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell,* 138(4):673-84, 2009.

[3] V. Le Texier, J.J. Riethoven, V. Kumanduri, C. Gopalakrishnan, F. Lopez, D. Gautheret and T.A. Thanaraj, AltTrans: transcript pattern variants annotated for both alternative splicing and alternative polyadenylation. *BMC Bioinformatics.* 7:169, 2006.

Session 3.C : Gene and Genome Function

Conformément au souhait des auteurs, cette contribution n'est pas reproduite dans la version en ligne des actes de JOBIM 2011.

Following the wishes of the authors, this paper is not included in the online version of the JOBIM 2011 proceedings.

Conformément au souhait des auteurs, cette contribution n'est pas reproduite dans la version en ligne des actes de JOBIM 2011.

Following the wishes of the authors, this paper is not included in the online version of the JOBIM 2011 proceedings.

Conformément au souhait des auteurs, cette contribution n'est pas reproduite dans la version en ligne des actes de JOBIM 2011.

Following the wishes of the authors, this paper is not included in the online version of the JOBIM 2011 proceedings.

Conformément au souhait des auteurs, cette contribution n'est pas reproduite dans la version en ligne des actes de JOBIM 2011.

Following the wishes of the authors, this paper is not included in the online version of the JOBIM 2011 proceedings.

Skew N-domains and Replication U-domains: Gradients of Replication Fork Polarity in the Human Genome

Chun-Long CHEN[1], Lauranne DUQUENNE[1,2], Yves d'AUBENTON-CARAFA[1], Claude THERMES[1], Arach GOLDAR[3], Guillaume GUILBAUD[4], Aurélien RAPPAILLES[4], Olivier HYRIEN[4], Antoine BAKER[5], Benjamin AUDIT[5] and Alain ARNEODO[5]

[1]CENTRE DE GENETIQUE MOLECULAIRE, UPR CNRS 3404, 91198 Gif-sur-Yvette, France
{chen, daubenton, thermes}@cgm.cnrs-gif.fr

[2]UMR CNRS 5558, LBBE, UCB Lyon1, 43 bd du 11 Novembre 1918, 69622 Villeurbanne, France
duquenne@biomserv.univ-lyon1.fr

[3]COMMISSARIAT A L' ENERGIE ATOMIQUE, iBiTecS, 91191 Gif-sur-Yvette, France, France
arach.goldar@cea.fr

[4]ECOLE NORMALE SUPERIEURE DE PARIS, UMR CNRS 8541, 46 rue d'Ulm 75005 Paris, France
{guilbaud, rappaill, hyrien}@biologie.ens.fr

[5]LABORATOIRE JOLIOT CURIE, ECOLE NORMALE SUPERIEURE DE LYON, CNRS, 69364 Lyon, France
{antoine.baker, benjamin.audit, alain.arneodo}@ens-lyon.fr

Keywords human genome, DNA replication, mutational strand asymmetry, DNA combing.

1 Introduction

Previous analyses of the nucleotide compositional skew profile (S=(G-C)/(G+C)+(T-A)/(T+A)) along the human genome allowed us to reveal large domains of ~1Mb exhibiting a characteristic N-shaped pattern and covering more than 1/3 of the genome, called N-domains [1-3]. We have further shown that upward jumps of the skew profile (S-jumps) at N-domain borders are significantly associated with peaks of earlier replication initiation zones and result from mutational strand asymmetries associated with replication in germline cells [4]. The striking linear decrease of the skew inside the N-domains raises the following questions: does it result from mutational asymmetries associated with replication? To what extent does this reflect a specific spatio-temporal replication organization? Using both genome-scale experimental approaches and sequence analyses, we investigated here the spatio-temporal replication program in the human genome and set up a model of replication that reproduces the N-shaped pattern of the skew profile.

2 Results and Discussion

We examined the replication timing profile of embryonic stem cells along N-domains: the mean profile presents a "U-shape" pattern (Fig. 1A) with time values decreasing from early to late from border to centre. We developed a wavelet-based method to detect U-shaped patterns and applied it to define replication U-domains in timing profiles of seven cell types [5-7]. These replication U-domains cover about half of the genome and are significantly ($P<10^{-3}$) co-localized with N-domains, showing that *N-domains likely correspond to U-domains of the germline cell timing profile*. Early replication initiation zones at U-domain and N-domain borders are significantly enriched in open chromatin marks suggesting that the replication program within replication domains is mediated by a gradient of open chromatin conformation.

We showed previously that S-jumps result from different substitution rates on the leading and lagging DNA strands of the replication fork [4]. Here, we computed neutral substitution rates along N-domains: replication-associated asymmetries decrease from maximum values at the left end to zero at the centre and to negative values at the right end. The skew at equilibrium that would be produced by the observed substitution rates acting over long evolutionary times is strongly correlated with the observed skew S and displays an N-shaped pattern along N-domains (Fig. 1B). This strongly suggests that the S decrease in N-domains reflects a progressive change in replication fork polarity. To test this hypothesis, we performed a genome-wide quantitative analysis of replicating DNA molecules stretched by DNA combing at different stages of the S phase in Hela cells [8], which revealed that the velocities of single forks remain constant during S phase. We then demonstrated that the mean fork polarity can be extracted from the derivative of the timing profile: the derivative of the U-shaped timing profile of N-domains is an "N", strongly supporting that *the N-shaped skew profile results from a replication fork polarity gradient*.

2.1 Model of the Replication Program of N-domains

We used the continuous wavelet transform to obtain a well defined and numerically stable measurement of the local slope of the Hela cell timing profile which corresponds to the inverse of the apparent replication speed, leading to a space-scale map of apparent replication speed. Multi-scale analysis showed a broad distribution of apparent replication speeds with practically no regions larger than 100 kb replicating at less than 2 kb/min and a higher proportion of fast-replicating regions in the late S phase, which cannot be explained by the range of single fork velocities (mean 0.7 kb/min obtained by DNA combing). DNA combing data also revealed that replication origins are spaced at mean 40 kb intervals and the global fork density increases during S phase because more replicon clusters and more origins within clusters become active as S phase progresses. We propose a domino model in which replication forks progressing from early origins stimulate initiation in nearby DNA: replication first initiates at U-domain extremities and secondary origins fire coordinately from the borders to the centers (Fig. 1C). This generates a linear gradient of replication fork polarity and a N-shaped skew. Given that U-domains are observed in human and mouse and N-domains are present in all studied mammals, it is likely that *this replication program has been conserved at least since the mammalian radiation*.

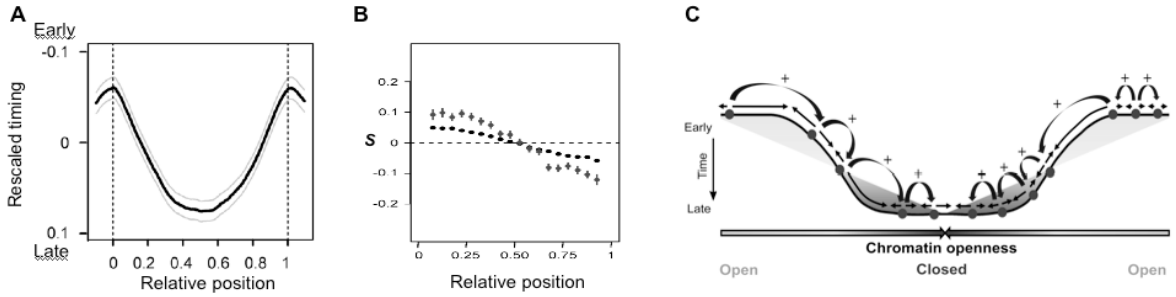


Figure 1. N-domain replication. (A) Replication timing (mean±SEM) along the N-domains (length is rescaled to 1) (B) Skew *S* (black) and *S* at equilibrium (grey) along the N-domains (C) Domino model of the replication program.

3 Materials and Methods

The replication timing values of human cell types [5, 6] were computed as in [5]. Initiation zones [4] and U-domains [7] were detected using a continuous wavelet transform. N-domains were retrieved from [3]. Nucleotide substitution rates and skew at equilibrium were computed as in [4]. DNA combing were performed on Hela cells sorted into four temporal compartments of S phase as in [8].

Acknowledgements

This work was supported by the CNRS, ANR (NT05-3_41825) and grants from ARC, LCC and FRM.

References

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

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

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

[4] C.L. Chen, L. Duquenne, B. Audit, G. Guilbaud, A. Rappailles, A. Baker, M. Huvet*, et al.*, Replication-associated mutational asymmetry in the human genome. *Mol Biol Evol*, in press.

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

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

[7] A. Baker, B. Audit, C.L. Chen, B. Moindrot, A. Leleu, A. Rappailles, G. Guilbaud*, et al.*, Replication domains are self-interacting chromatin structural units. *Genome Res*, under review.

[8] A. Rappailles, G. Guilbaud, A. Baker, C.L. Chen, A. Arneodo, G. Arach, Y. d'Aubenton-Carafa*, et al.*, Sequential and increasing activation of replication origins along replication timing gradients in the human genome. *PLoS Genet*, in revision.

Session 3.D : Algorithmic Development

Conformément au souhait des auteurs, cette contribution n'est pas reproduite dans la version en ligne des actes de JOBIM 2011.

Following the wishes of the authors, this paper is not included in the online version of the JOBIM 2011 proceedings.

Conformément au souhait des auteurs, cette contribution n'est pas reproduite dans la version en ligne des actes de JOBIM 2011.

Following the wishes of the authors, this paper is not included in the online version of the JOBIM 2011 proceedings.

Conformément au souhait des auteurs, cette contribution n'est pas reproduite dans la version en ligne des actes de JOBIM 2011.

Following the wishes of the authors, this paper is not included in the online version of the JOBIM 2011 proceedings.

Conformément au souhait des auteurs, cette contribution n'est pas reproduite dans la version en ligne des actes de JOBIM 2011.

Following the wishes of the authors, this paper is not included in the online version of the JOBIM 2011 proceedings.

Conformément au souhait des auteurs, cette contribution n'est pas reproduite dans la version en ligne des actes de JOBIM 2011.

Following the wishes of the authors, this paper is not included in the online version of the JOBIM 2011 proceedings.

Conformément au souhait des auteurs, cette contribution n'est pas reproduite dans la version en ligne des actes de JOBIM 2011.

Following the wishes of the authors, this paper is not included in the online version of the JOBIM 2011 proceedings.

Conformément au souhait des auteurs, cette contribution n'est pas reproduite dans la version en ligne des actes de JOBIM 2011.

Following the wishes of the authors, this paper is not included in the online version of the JOBIM 2011 proceedings.

Conformément au souhait des auteurs, cette contribution n'est pas reproduite dans la version en ligne des actes de JOBIM 2011.

Following the wishes of the authors, this paper is not included in the online version of the JOBIM 2011 proceedings.

Waffect: a Method to Simulate Case-Control Samples in Genome-Wide Association Studies

Vittorio PERDUCA[1], Raphaël MOURAD[2], Christine SINOQUET[3] and Gregory NUEL[1]

[1] MAP5 - UMR CNRS 8145, Université Paris Descartes, 45 Rue des Saints Pères, 75006 Paris Cedex 06, France
vittorio.perduca@gmail.com; gregory.nuel@parisdescartes.fr

[2] LINA - UMR CNRS 6241, Ecole Polytechnique de l'Université de Nantes, Rue Christian Pauc, BP 50609, 44306 Nantes Cedex 3, France
raphael.mourad@univ-nantes.fr

[3] LINA - UMR CNRS 6241, Université de Nantes, Rue de la Houssinière, BP 92208, 44322 Nantes Cedex, France
christine.sinoquet@univ-nantes.fr

Abstract *In a Genome-Wide Association Study (GWAS) the genomes of a large group of individuals are examined to establish the presence of a significant association between a disease and particular genes. The group of individuals is divided in cases (people with the disease) and controls (people without). The association is assessed through statistical hypothesis testing. The distribution under the null hypothesis H0 is empirically studied shuffling uniformly affectation status (case and control memberships), while complex genetic models are usually used to simulate under the alternative hypothesis H1. We have developed an alternative approach to this problem. The idea is to mimic the H0 simulations by affecting status to individuals under the constraint that the probability to be a case is consistent with the chosen disease model and that the total number of cases is fixed. We suggest a simple but efficient algorithm to perform this constrained affectation. We apply our algorithm to a real data set from the* 1000 Genomes Project *to compare the accuracy of different methods for identifying causal genetic markers and show that accuracy quickly decreases as the candidate regions get wider.*

Keywords Forward-backward sampling, phenotype, power, ROC curve.

Waffect: une Méthode pour Simuler des Données Cas-Témoins dans les Etudes d'Associations à Grande Echelle

Résumé *Dans une étude d'association à l'échelle du génome, les génomes d'un grand groupe de personnes sont examinés afin d'établir si une association significative existe entre une maladie et les gènes. Le groupe de personnes est divisé en cas (les personnes atteintes) et témoins (ceux qui ne sont pas atteints). L'association est évaluée à travers des tests d'hypothèses. La distribution sous l'hypothèse nulle H0 est étudiée empiriquement en permutant le statut cas/témoins, tandis que pour simuler sous H1 des modèles génétiques complexes sont généralement utilisés. Nous avons développé une approche alternative à ce problème. L'idée est d'imiter les simulations sous H0 en affectant le statut des individus sous la contrainte que la probabilité d'être un cas est compatible avec le modèle de maladie choisi et que le nombre total de cas est fixé. Nous proposons un algorithme simple mais efficace pour effectuer cette affectation. Nous appliquons notre algorithme à des données réelles issues du* 1000 genomes project *afin de comparer la précision de différentes méthodes pour identifier des marqueurs génétiques causaux. Les résultats montrent que la précision diminue rapidement quand la longueur des régions candidates augmente.*

Mots-clés Échantillonnage, phénotype, puissance, forward-backward, courbe ROC.

1 Introduction

Genome-wide association studies (GWAs) are a widely-used approach to address the localization of causal mutations responsible for common complex genetic diseases, [1]. Such studies involve the investigation of

hundred of thousands to millions of genetic markers (such as single nucleotide polymorphisms – SNPs), for a cohort of cases and controls whose sizes range in the thousands to tens of thousands individuals. GWASs have met many successes, most notably for type 1 and type 2 diabetes, inflammatory bowel disease, prostate cancer and breast cancer, [2].

In GWASs, very high false positive rates are expected unless a correction for multiple testing is performed. Symmetrically, control for true negative rate - or power - is necessary. Power estimation is the key to evaluate the efficiency of GWAS methods, [3]. The correct estimation of both rates must take into account the existence of high-dependency patterns between SNPs, or linkage disequilibrium (LD). The accurate estimation of the family wise type I error risk in presence of LD consists in sampling the H0 distribution through permutations of phenotypes, [4]. Thus, any association between loci and the phenotype is broken. This permutation strategy is implemented as a gold standard in numerous dedicated packages, together with software suites designed for GWASs [5,6,7,8,9].

Power is a still more complicated function of several factors: study design, correlation patterns in the genotypic data, sizes of cohorts, frequency of the causal allele, relative risk conferred by the causal factor, genetic model (additive, dominant, recessive, multiplicative), [10]. As a consequence of this complexity, the analytical computation of power necessarily relies on simplifying assumptions, including the approximation of the H1 distribution of the statistic test through a probability law, [11]. Most power calculators based on analytical approaches are used for two-stage GWAS design, e.g. [12,13]. Recently, an analytical approach has been proposed to account for LD, under either H0 or H1 approximation [14]: a fixed-size sliding window locally accounts for the inter-marker correlation. Unifying H0 and H1 processing in the same framework, this approach brings an improvement over block-wise strategies [15,16,17]. However, regarding both accuracy and computational burden, the optimal choice for the window size depends on the structure of the data. Moreover, LD blocks are often ambiguous. Thus, the previous sliding window approach cannot account for high order dependences between LD blocks. In particular, this method cannot be used to evaluate the power of any novel approach designed to cope with such high order dependences, in GWASs. In the latter case, the only solution remains using intensive simulations.

Symmetrically to sampling under H0, simulation of the H1 distribution is an appealing means to keep the LD-structured genotypic data. These simulations consist in generating case and control samples which mimic the LD structure in human genome, i.e. in the creation of, say, k datasets under the H1 assumption (at least one SNP is causal). Nonetheless, breaking any association between a locus (or several loci) and the phenotype is far more easy to implement than introducing such an association in a dataset.

Two main strategies have been developed to simulate H1: (i) the prospective one, [18], which first generates a large sample of haplotypes conditional on reference haplotypes such as HapMap haplotypes [19], then pairs haplotypes to build diplotypes and assigns the disease status depending on the penetrance model involving a randomly selected causal SNP, and (ii) the retrospective strategy, [3], which first randomly selects a causal SNP and generates a fixed number of cases and controls, then assigns diplotypes at the causal SNP for cases and controls depending on the penetrance model and finally builds haplotypes (two for each diplotype) for all remaining SNPs of the chromosome, conditionally on reference haplotypes. Nevertheless, both strategies entail implementation problems when applied to power estimation in GWASs. The first strategy presents the drawback to not allow the control of the numbers of cases and controls. To tackle this issue, rejection sampling of case-control samples is used, but leads to a waste of data and time. An illustration of the first strategy is Fregene [18]. The second strategy controls these numbers by first fixing them, but requires to build haplotypes for each simulation. The widely-used simulator Hapgen [20] implements the second strategy. When assessing the performance of gene mapping methods based on LD modeling, such as haplotype block-based methods [21,22] or latent variable-based methods [23], it would be more interesting to use a benchmark dataset of $n + m$ genotypes (unphased data) and then to assign n cases and m controls to the set of $n + m$ genotypes, k times (for k simulations). In this setting, power would be fast to calculate because LD pattern identification, which is computationally expensive, would have to be performed only one time and not k times.

In this scope, we propose a new method able to exactly assign n cases and m controls conditional on $n+m$ simulated or real genotypes. Our idea is to perform the affectation of individual i according to a weight ω_i (for instance the relative risk for individual i) with a constraint on the number of cases. This method provides several advantages. First, it generalizes permutations, the gold standard to generate H0 distribution in order to control type I error in multiple testing. Indeed, permutations represent a particular case of our general approach by using uniform weights. Second, our method is faster than previous ones [3,24] because it does not require a rejection sampling step or the generation of new genotypes for each simulation. Moreover, in the case of LD modeling-based mapping methods, LD pattern identification needs to be performed only one time. Third, no assumption is made on the genotype distribution, because the latter is the same for each simulation. The last advantage, maybe the most important when evaluating the performance of new mapping methods, is that power can be directly assessed using real GWAS datasets, such as those provided by the *1000 Genome Project* [25], because disease status can be generated according to genotypes without loss due to rejection sampling.

We first describe the method itself, which we called WAFFECT, then illustrate its interest on real GWAS data by comparing the power of several approaches. WAFFECT will be soon available as an R package. It can affect the phenotypes of 10,000 individuals with 5,000 cases in 2.2 seconds (on a common - even a bit outdated - workstation).

2 Methods

Let $I = \{1, \ldots, q\}$ be the ordered set of all individuals. We denote P_i, $i \in I$, the random variable accounting for the status (phenotype) of individual i, $P_i \in \{0, 1\}$ where 0 stands for *control* and 1 for *case*. The probability π_i that individual i is a case is proportional to a given weight ω_i: $\mathbb{P}(P_i = 1) = \pi_i \propto \omega_i$. For instance, we can take ω_i to be the relative risk for individual i. We denote N_i, $i \in I$, the random variable counting the total number of cases among individuals indexed by $\{1, \ldots, i\}$: $N_i = \sum_{j=1}^{i} P_j$, with the convention that $N_0 = 0$. Observe that $N_i = N_{i-1} + P_i$ and therefore $N_i \in \{0, \ldots, i\}$ for each i. When all the weights ω_i are given, we are interested in sampling the values of the P_is, given the condition that the total number of cases N_q must be equal to r, i.e. in sampling the distribution $\mathbb{P}(P_1, \ldots, P_q | N_q = r)$. To achieve this goal we find recursive formulas for the probabilities $\mathbb{P}(P_i = 1 | N_{i-1} = m, N_q = r)$, $i \in I$:

THEOREM 2.1. *For each individual $i = 1, \ldots, q$:*

$$\mathbb{P}(P_i = 1 | N_{i-1} = m, N_q = r) = \frac{\omega_i B_i(m+1)}{B_{i-1}(m)}, \tag{1}$$

where the backward quantities B_i *can be computed using the recursive formulas*

$$B_i(m) = \omega_{i+1} B_{i+1}(m+1) + (1 - \omega_{i+1}) B_{i+1}(m), \tag{2}$$

with the following edge conditions: $B_0(0) = \omega_1 B_1(1) + (1 - \omega_1) B_1(0)$ *and* $B_q(m) = \delta(m, r)$, δ *being the Kronecker's symbol.*

The theorem gives a recursive algorithm, which we called WAFFECT, to sample in the space of all possible configurations of the P_is under the condition that the number of cases must be r and knowing for each individual i his weight ω_i (e.g. his relative risk). Starting from $i = q$, simply compute all the backward quantities with Eq. (2). Then starting from the first individual $i = 1$, affect a status for the individual i accordingly to the binomial distribution which depends on the previous affectations $P_i \sim \mathcal{B}\left(\frac{\omega_i B_i(N_{i-1}+1)}{B_{i-1}(N_{i-1})}\right)$. The pseudocode is given below. Observe that if $\omega_i = \omega_0$ for each i, then WAFFECT outputs a permutation of the phenotypes; this is equivalent to simulating under H0.

It is possible to simulate affectations in the case of more than two classes by calling recursively WAFFECT. For instance, in the case of three classes $\{0, 1, 2\}$, start by affecting status 0 versus status $\{1, 2\}$ and then iterate WAFFECT for the individuals with status $\{1, 2\}$ to affect status 1 versus status 2.

Algorithm 1 WAFFECT(ω, r)

Input: vector of weights (e.g. relative risks) $\omega = (\omega_i)_{i=1,\dots,q}$ **, number** r **of cases**
Output: A vector of phenotypes for the individuals

B is $q \times (r+2)$ matrix, $B = 0$ {/* Initialization of B */}
$B_q(r) = 1$
for $i = q-1$ to 0 **do** {/* Iterative computation of B_i */}
  for $m = 0$ to r **do**
    $B_i(m) = \omega_{i+1} B_{i+1}(m+1) + (1-\omega_{i+1}) B_{i+1}(m)$
  end for
end for
Sample P_1 with $P_1 \sim \mathcal{B}\left(\frac{\omega_1 B_1(1)}{\omega_1 B_1(1) + (1-\omega_1) B_1(0)}\right)$ {/* Sampling initialization */}
$N_1 = P_1$
for $i = 2$ to q **do** {/*Sampling of $P_2, \dots, P_q$*/}
  Sample P_i with $P_i \sim \mathcal{B}\left(\frac{\omega_i B_i(N_{i-1}+1)}{B_{i-1}(N_{i-1})}\right)$
  $N_i = N_{i-1} + P_i$
end for
return $P_1, \dots, P_q$

3 Application

We applied WAFFECT on real data to assess the accuracy of different association methods based on the Cochran-Armitage trend test.

The original data set consists on the real genotypes of 629 individuals from the *1000 Genomes Project* [25]. We focused on the first 100,000 SNPs on the X chromosome. In the pretreatment stage, we filtered out all the SNPs with Minor Allele Frequency (MAF) less than or equal to 5%, ending up with 8,048 SNPs.

We arbitrarily defined an additive disease model with two interacting causal SNPs. In particular, we arbitrarily chose two SNPs S_1, S_2 (the ones with base-pair positions 627,641 and 1,986,325) with low MAF and showing no linkage disequilibrium (i.e. correlation). For each individual, we defined the following relative risk: $RR = 1 + 0.1 \times G_{S_1}$ if $G_{S_2} = 0$, $RR = 1 + 0.1 \times G_{S_2}$ if $G_{S_1} = 0$, $RR = 1 + 0.3 + 0.1 \times (G_{S_1} + G_{S_2})$ if $G_{S_1} \times G_{S_2} > 0$, where $G_{S_i} \in \{0, 1, 2\}$ is the number of the less frequent allele in the genotype of S_i. The last expression defines an *epistasis* (interaction) between the two genes. Then, for each individual we computed his relative risk ω_i accordingly to his genotype and the disease model. Running 1,000 times WAFFECT on the vector $(\omega_i)_{i=1,\dots,629}$ and under the constraint that the total number of cases must be 314, we obtained 1,000 simulations of the phenotype of each individual under H1. Similarly, we ran WAFFECT on a vector whose elements have all the same value (e.g. $\omega_i = 1$ for each i) and obtained 1,000 affectations of the phenotypes under H0.

The association analysis was performed running the toolset PLINK v1.07, [9]. In particular, for each SNP we obtained the p-value for the trend statistics under H0 and H1. The association methods we studied consider two intervals centered in the two causal SNPs and having the same length. The statistics is defined as the $\max$ of $-\log$ of the p-values in the overall region. We considered intervals centered in S_1, S_2 with radii $\rho = 0, 1, 5, 10, 50, 100$ and ∞ kb. For each radius, we assessed the accuracy of the method analyzing the trade-offs between false positive rate and true positive rate (power). More precisely, for each ρ we computed the corresponding Receiver Operating Characteristic (ROC) curve, Fig. 1, the empirical Area Under the Curve (AUC) value and an upper bounds σ_m for the standard deviation of the estimate, Table 1, see for instance [26].

The results show that the accuracy is good when the radius is 0 kb, and fair up to a radius of 1 kb. For greater radii, AUC ≤ 0.70 and the accuracy quickly decreases. We conclude that the association method that we have considered is typically suitable for testing the association between a given disease and a pair of candidate genes but not for testing wider regions.

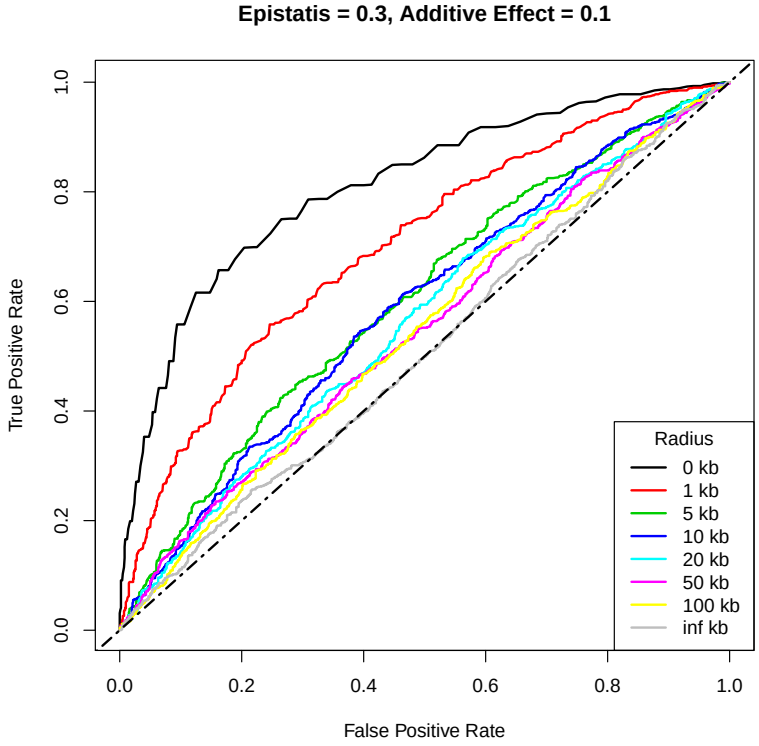


Figure 1. ROC curves at different radii ρ of the intervals centered in the two causal SNPs. The statistics is equal to the $\max$ of $-\log$ of the p-values in the overall region given by the two intervals.

ρ	AUC	σ_m	Accuracy
0 kb	0.80	0.012	good
1 kb	0.70	0.014	fair
5 kb	0.60	0.015	poor
10 kb	0.59	0.015	fail
20 kb	0.56	0.016	
50 kb	0.55	0.016	
100 kb	0.54	0.016	
∞ kb	0.51	0.016	

Table 1. AUCs and s.d. upper bounds σ_m at different radii ρ of the intervals centered in the causal SNPs.

4 Conclusions

We introduced a new iterative algorithm, which we called WAFFECT, to sample case-control affectations under the constraint that the probability to be a case is consistent with the chosen disease model and that the total number of cases is fixed. This method can be used to simulate under H1 for assessing the power of GWAS methods. WAFFECT generalizes the method of permutations, the gold standard for sampling under H0. New values of the disease status are generated by permuting them accordingly to weights proportional to the probabilities to be a case. Similarly to the method of permutations, WAFFECT presents the advantage that genotypes are fixed and only phenotypes are sampled. Two other possible ways to sample case-control affectations along these lines involve rejection and MCMC algorithms. An R package including all these algorithms will be released shortly.

Unlike its competitor HapGen, [20], WAFFECT allows to simulate the phenotypes under any disease model, also in presence of two or more causal genetic markers and when epistasis is taken into account. Not surprisingly, when the disease model consists in only one causal SNP, WAFFECT's results are coherent with the ones found by HapGen. However, WAFFECT is faster than HapGen because it does not generate new genotypes for each simulation.

Our algorithm allows to evaluate the performance of GWAS methods using real GWAS data sets. We applied it to data sets from the *1000 Genomes Project* to assess the accuracy of a region candidate approach and showed that, given the chosen design (629 individual) and the modest effect used for the simulation (relative risk of 1.7 in the most favorable case), we have only power to detect the association signal at the gene candidate level. For wider candidate regions, the power drops quickly.

References

[1] A. P. Morris and L. R. Cardon, Whole genome association, in D. J. Balding, M. Bishop, C. Cannings (eds.), *Handbook of statistical genetics*, volume 2, 3rd edition, pp. 1238-1263, Wiley Interscience, 2007.

[2] L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins and T. A. Manolio, Potential etiologic and functional implications of genome-wide association loci for human diseases and traits, *PNAS*, 106(23):9362-9367, 2009.

[3] C. C. Spencer, Z. Su, P. Donnelly and J. Marchini, Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip, *PLoS Genetics*, 5(5):e1000477+, 2009.

[4] Q. Zhang and J. Ott, Multiple comparison/testing issues, in S. Lin and H. Zhao (eds), *Handbook on Analyzing Human Genetic Data*, pp. 277-287, Springer Berlin Heidelberg, 2010.

[5] Y. S. Aulchenko, S. Ripke, A. Isaacs and C. M. van Duijn, Genabel: an R library for genome-wide association analysis, *Bioinformatics*, 23(10):1294-1296, 2007.

[6] B. L. Browning, PRESTO: rapid calculation of order statistic distributions and multiple-testing adjusted P-values via permutation for one and two-stage genetic association studies, *BMC Bioinformatics*, 13(9):309, 2008.

[7] J. R Gonzalez, L. Armengol, X. Sole, R. Guino, J. M. Mercader, X. Estivill and V. Moreno, SNPassoc: an R package to perform whole genome association studies, *Bioinformatics*, 23:644-5, 2007. Package available at http://cran.r-project.org/web/packages/SNPassoc/

[8] K. S. Pollard, S. Dudoit and M. J. van der Laan, Multiple testing procedures: the multtest package and applications to genomics, in R. C. Gentleman, V. J. Carey, W. Huber, R. Irizarry, and S. Dudoit (eds.), *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, pp. 249-271 Springer, New York, Chapter 15, 2005.

[9] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly and P. C. Sham, PLINK: a toolset for whole-genome association and population-based linkage analysis, *American Journal of Human Genetics*, 81, 2007. Package PLINK available at http://pngu.mgh.harvard.edu/purcell/plink/

[10] G. Lettre, C. Lange and J. N. Hirschhorn, Genetic model testing and statistical power in population-based association studies of quantitative traits, *Genetic Epidemiology*, 31(4):358-362, 2007.

[11] R. J. Klein, Power analysis for genome-wide association studies, *BMC Genetics*, 8(58), 2007.

[12] I. Menashe, P. S. Rosenberg and B. E. Chen, PGA: power calculator for case-control genetic association analyses, *BMC Genetics*, 9(1):36, 2008.

[13] J. P. Steibel and G. R. Abecasis, QpowR: Interactive power calculator for two-stage genetic association studies of quantitative traits, https://www.msu.edu/ steibelj/JP_files/QpowR.html , 2008.

[14] B. Han, H. M. Kang and E. Eskin, Rapid and accurate multiple testing correction and power estimation for millions of correlated markers, *PLoS Genetics*, 5(4):e1000456, 2009.

[15] K. N. Conneely and M. Boehnke, So many correlated tests, so little time! Rapid adjustment of p-values for multiple correlated tests, *American Journal of Human Genetics*, 81:1158-1168, 2007.

[16] D. Lin, An efficient Monte Carlo approach to assessing statistical significance in genomic studies, *Bioinformatics*, 6:781-787, 2005.

[17] S. R. Seaman and B. Müller-Myhsok, Rapid simulation of P values for product methods and multiple-testing adjustments in association studies, *American Journal of Human Genetics*, 76:399408, 2005.

[18] M. C. Hyam, C. Hoggart, P. OReilly, J. Whittaker, M. De Iorio and D. Balding, Fregene: Simulation of realistic sequence-level data in populations and ascertained samples, *BMC Bioinformatics*, 9(1):364+, 2008.

[19] The International HapMap Consortium, A second generation human haplotype map of over 3.1 million snps, *Nature*, 449(7164):851-861, 2007.

[20] Z. Su, J. Marchini and P. Donnelly HapGen v2, http://www.stats.ox.ac.uk/ marchini/software/gwas/hapgen.html , 2010.

[21] B. L. Browning and S. R. Browning, A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals, *American Journal of Human Genetics*, 84(2):210-223, 2009.

[22] C. Pattaro, I. Ruczinski, D. M. Fallin and G. Parmigiani, Haplotype block partitioning as a tool for dimensionality reduction in SNP association studies, *BMC Genomics*, 9:405, 2008.

[23] R. Mourad, C. Sinoquet and P. Leray, Learning hierarchical Bayesian networks for genome-wide association studies, in Y. Lechevallier and G. Saporta (eds.), *Proc. nineteenth International Conference on Computational Statistics, COMPSTAT, France, Paris*, pp 549-556, 2010.

[24] B. Peng and C. I. Amos, Forward-time simulation of realistic samples for genome-wide association studies, *BMC bioinformatics*, 11(1):442+, 2010.

[25] The 1000 Genomes Project Consortium, A map of human genome variation from population-scale sequencing, *Nature*, 467, 1061-1073, 2010.

[26] C. E. Metz, Basic principles of ROC analysis, *Sem Nuc Med*, 8:283-298, 1978.

A Sequencial Monte Carlo Method for estimating Transcriptional Landscape at Basepair level from RNA-Seq data

Bogdan MIRAUTA[1], Pierre NICOLAS[2] and Hugues RICHARD[1]

[1] Génomique des Microorganismes, UPMC UMR7238, 75005 Paris, France
{bogdan.mirauta,hugues.richard}@upmc.fr
[2] Mathématique, Informatique et Génome, INRA UR1077, 78350 Jouy-en-Josas, France
pierre.nicolas@jouy.inra.fr

Keywords RNA-Seq, Sequential Monte Carlo, Hidden Markov model.

1 Introduction

Sequencing technologies applied to transcriptome interrogation (RNA-Seq) permit a high precision in the inference of transcript localization and relative expression level. Namely, RNA-Seq produces millions of reads that, once aligned to the reference genome, provide counts that reflect the transcriptional landscape. This landscape is, even in the presence of post transcription processes, directly correlated to the expression activity. Read counts can serve to estimate, with an existing annotation, gene expression level up to the isoform level [1] assuming homogeneous distribution of the reads inside predefined genome segments. Transcript boundaries can also be identified from abrupt variations in the local abundance of reads using sliding windows. However, realistic probabilistic models that simultaneously account for transcript boundaries and expression levels are still not available to describe RNA-Seq data. This precludes the use of a parametric inference framework to obtain estimate expression at the basepair level.

In this work, we design a strand specific model that includes the changes in expression level along the genome and randomness due to the read sampling. Through Hidden Markov Model formalism this problem reduces to estimating the hidden path for the unobserved variable u_t - the expression level at basepair t. u_t is defined as the product between the relative abundance of the position t and the total number of reads. We then use a Sequential Monte Carlo (SMC) approach [2] to infer hidden expression level $(u_t)_{1:T}$ along a genome sequence of length T from observed read counts $(y_t)_{1:T}$. Model parameters are estimated in Bayesian framework. The SMC approach allows us to specify a reasonable model that fully exploits the complexity of RNA-Seq data.

2 Model and SMC Algorithm

The Hidden Markov model includes two main dependencies: the hidden chain transition kernel that describes changes in expression level and the emission function that relates counts to expression levels. Our choice for the transition between the hidden states (u_t) aims at including abrupt changes characteristic of transcript boundaries, and smooth variations, arising from technological biases or biological processes. Biological effects could be partial termination or degradation, which generate progressive changes in transcription levels. In keeping with [3] we refer these two types of changes as the shifts and the drifts. The transition kernel writes:

$$\begin{aligned}\pi(u_{t+1},u_t) \quad &= \quad \mathbf{1}_{\{u_t=0\}}\cdot\left[(1-\eta)\cdot\delta_0(u_{t+1})+\eta\cdot\frac{1}{c}e^{-\frac{1}{c}\cdot u_{t+1}}\right]\\ &+ \quad \mathbf{1}_{\{u_t>0\}}\cdot\Big[\alpha\cdot\delta_{u_t}(u_{t+1})+\beta\cdot\frac{1}{c}e^{-\frac{1}{c}\cdot u_{t+1}}+\beta_0\cdot\delta_0(u_{t+1})\\ &\quad +\gamma_u\cdot\mathbf{1}_{\{u_{t+1}>u_t\}}\cdot\frac{\lambda_u}{u_t}\cdot e^{-\frac{\lambda_u}{u_t}\cdot(u_{t+1}-u_t)}+\gamma_d\cdot\mathbf{1}_{\{u_{t+1}<u_t\}}\cdot\frac{u_t}{u_{t+1}}\cdot\frac{\lambda_d}{u_{t+1}}\cdot e^{-\frac{\lambda_d}{u_{t+1}}\cdot(u_t-u_{t+1})}\Big]\end{aligned}$$

and is best understood as a mixture of several move types. In expressed regions, expression remains at the same level with probability α; jumps to a non expressed region with probability β_0; changes to a different level exponentially distributed with probability β; and can finally drift upward or downward with probabilities γ_u and γ_d. Small increases or decreases caused by drifts have percentual amplitudes exponentially distributed

with parameters λ_u and λ_d. A jump from a not expressed position into an expressed region has probability η. Read counts (y_t) are considered independent given (u_t). This assumption holds when counting only the first position of the reads. We model the read count y_t as a mixture between a Poisson with expectation u_t [4] and a distribution accounting for technological outliers.

The reconstruction of expression levels is based on a Monte Carlo approach where estimation of $u_{1:T}$ given the observation $y_{1:T}$ relies on the sampling of particles (trajectories) from the target distribution $\pi(u_{1:T} \mid y_{1:T})$. In our context this distribution is complex and T is large. An Importance Sampling scheme is used where sampling is done from a proposal whose outcomes are reweighted. For long sequences, a good importance proposal $q(u_{1:T})$ to approximate $\pi(u_{1:T} \mid y_{1:T})$ is impossible to define directly but the problem remains tractable using a sequential approach [2]. Briefly, at each position t we obtain a sample from $\pi(u_{1:t} \mid y_{1:t})$ using the sample from $\pi(u_{1:t-1} \mid y_{1:t-1})$ obtained at position $t-1$ by drawing u_t and updating importance weights according to the formula $w_t^i = \{\pi(u_t^i \mid u_{t-1}^i)\pi(y_t \mid u_t^i)/q_t(u_t^i; u_{t-1}^i)\}w_{t-1}^i$ where i is the particle index. As t increases, weights w_t^i tend to degenerate leading to a poor estimation. A resampling step is performed avoid this behaviour. The choice of the importance proposal can improve the number of used particles, the resampling frequency and thus the performance of the algorithm. Our proposal $q_t(u_t; u_{1:t-1})$ was designed to approximate $\pi(u_t \mid u_{1:t-1}, y_{1:T})$. This algorithm is a *filtering algorithm* that provides sample approximating $\pi(u_{1:t} \mid y_{1:t})$ for each t and therefore from our target distribution $\pi(u_{1:T} \mid y_{1:T})$ at time T. For large T, when looking backward, the trajectories of the particles coalesce due to resampling and thus make impossible a good estimation of u_t for $t \ll T$. To tackle with this problem a *backward smoothing* is implemented. From the backward sample we can compute both point estimates and credibility intervals of the expression level u_t for each t.

3 Results and Discussion

Relevance of this method is illustrated in Fig. 1 by the application on simulated data for low expressed regions. We used a Gibbs algorithm for the estimation of the parameters.

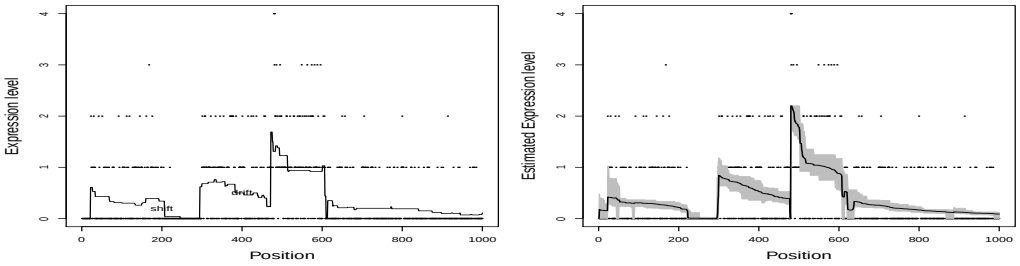


Figure 1. Expression level inference on simulated data. Left panel: simulated expression level (thin line) and read counts (dots). Right panel: estimated expression level (black line) and 95% credibility interval (gray area).

This methodology extends previous work on tiling array data [3]: it introduces a model adapted to RNA-Seq data and it presents an SMC algorithm for estimation of underlying expression levels that overcomes the need for discretization. Better description of the RNA-Seq data is an important step towards disentangling technological artifacts from subtle biological signals. A software is in preparation and will be made available.

References

[1] H. Richard, M. Schulz, M. Sultan et al., Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments, *Nucleic Acid Res.*, 38:e112, 2010.

[2] A. Doucet and A. Johansen, *A tutorial on Particle Filtering and Smoothing: Fifteen years later*, Tehnical report, University of British Columbia, 2008.

[3] P. Nicolas et al., Transcriptional landscape estimation from tiling array data using a model of signal shift and drift, *Bioinformatics*, 25:2341-2347, 2009.

[4] J. Marioni et al., RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays, *Genome research*, 18:1509-1517, 2008.

Session 4 : The Challenges of NGS

Conférence invitée

Michael Brudno

Dept of Computer Science, University of Toronto, Toronto, Canada

Discovering and Visualizing Structural Variation from High Throughput Sequencing Data

High throughput sequencing (HTS) technologies have enabled the inexpensive sequencing of human genomes, and the discovery of some genomic variants from the resulting short read datasets is well underway. In this talk I will present two algorithms for the discovery of structural variants from HTS paired-end data: First, I will describe a method to predict CNVs from paired short reads. Our method combines information from paired short reads to identify variable regions and depth-of-coverage to predict the true copy count in the donor genome. Together, these datasets help overcome both sequencing biases of HTS platforms and spurious read mappings. Our method allows for the detection of CNVs within segmental duplications. We use our method to detect CNVs within the same dataset, and make a total of 5000 calls that show high concordance with previously known CNVs in this individual. I will then describe PRISM, a method to infer the precise breakpoints of structural variants by finding reads that map in two different locations on the genome. We use pairing information to identify putative location to map the split reads. Our preliminary result shows that PRISM outperforms other methods for split read mapping for insertions and deletions, and can detect borders of structural variants, including inversions and duplications, to the base pair. Finally I will give a brief overview of Savant, a method for visualizing HTS data that we have developed, and showcase some of its unique features, including a powerful plug-in framework that allows developers to extend Savant.

An Exact Algorithm for the Segmentation of NGS Profiles using Compression

Exact Segmentation of NGS Profiles

Guillem RIGAILL[1]

Bioinformatics and Statistics, NKI-AVL, Plesmanlaan 121, 1066 CX Amsterdam
g.rigaill@nki.nl

Abstract *Next Generation Sequencing (NGS) is an efficient approach to scan the entire genome for DNA copy number variations or alterations (DNA-seq) or transcribed regions of the genome (RNA-seq). Segmenting or splitting the genome into regions of homogeneous read counts is a natural approach to analyze these profiles. The computational burden is one of the foremost issues. Indeed, due to the very large size of NGS profiles, standard segmentation algorithms cannot be applied to NGS data. We present a fast and exact algorithm to recover the best segmentation of a NGS profile w.r.t. the log-likelihood loss. The algorithm proceeds in two steps. It first compresses the NGS data. Then it uses a fast and exact pruned dynamic programming algorithm that we recently developed. We theoretically demonstrate that the compression does not hamper our ability to recover the best segmentation of the raw NGS profile. We applied our algorithm to real DNA-seq profiles of human tumors and cell-lines and real RNA-seq profiles of bacteria to demonstrate the efficiency of the compression and the competitive runtime of our exact algorithm.*

Keywords Next Generation Sequencing, segmentation, exact algorithm, compression, pruned dynamic programming.

1 Introduction

Next generation sequencing (NGS) has enabled the generation of large-scale genome sequence data. From NGS profiles it is possible to detect DNA copy number alterations or transcribed regions of the genome. A common strategy is to first map reads to a reference genome and then detect regions with unexpectedly high or low number of reads. In the case of DNA-seq data these regions hopefully correspond to DNA copy number amplifications or deletions. In the case of RNA-seq data they hopefully correspond to transcribed regions of the genome. The aim is thus to split the genome into regions of homogeneous read counts. At least in the case of DNA copy number analysis using microarrays, segmentation models [1,2] have been shown to be the best performing methods [3].

For NGS profiles, with many million nucleotides per chromosome (for example 10^8 in the case of human chromosomes), recovering the most likely segmentation or split with respect to maximum likelihood is a challenging optimization problem. Until recently, the fastest exact algorithm to solve this problem had a space complexity of $\Theta(n^2)$ and a time complexity of $\Theta(Kn^2)$, where n is the number of points and K the number of segments or change-points [4]. These space and time complexities are prohibitive for profiles with more than 10^5 points. Thus, even for SNP arrays (with 10^5 up 10^6 points), most segmentation methods rely on heuristic computational schemes to reduce the runtime [5]. However, this is done at the price of some errors as heuristics may not recover the best segmentation but rather a good candidate segmentation. This is clearly a problem for biological interpretation as we cannot guarantee that there is not a better way to segment the data. We recently proposed a new exact algorithm that recovers the best segmentation with a worst space complexity of $\Theta(Kn)$, a time complexity of $O(Kn^2)$ at worst and of order $O(Kn\log(n))$ in practice and that takes a few minutes only to analyze an Affymetrix SNP 6.0 profile with 2.10^6 probes across the human genome ([6] available on arXiv).

Yet, even this new algorithm is too slow in practice to analyze NGS profiles with 10^8 nucleotides or more per chromosome. A common approach to reduce the computational burden is to use non-overlapping windows

[7]. The total number of reads per window is averaged or summarized in some way to recover in the end a much smaller profile. Choosing the size of these windows is obviously not a trivial task and it undoubtedly affects the resolution of the analysis and the final result. More importantly, by averaging over more or less predetermined windows one takes the risk to average the signal in regions where it should not be averaged. Indeed, except for very small window sizes, transitions in the signal are much more likely to occur inside windows rather than at their borders. In other words, it is very likely that these non-overlapping windows will smooth signal transitions and thus hamper our ability to detect them.

Here we propose an exact algorithm to segment a NGS profile into 1 up to K homogeneous regions w.r.t. some loss function (for example the quadratic loss or the negative log-likelihood Poisson loss). The algorithm can be used either on DNA-seq data to detect DNA copy number amplifications and deletions and on RNA-seq data to detect transcribed regions of the genomes in RNA-seq data. It can also be used as a preprocessing step to define homogeneous regions or windows of the genome where the information can indeed be averaged or summarized.

The algorithm proceeds in two steps. It first compresses the data. Then it uses the exact pruned dynamic programming algorithm (DPA) for segmentation that we developed recently [6]. The compression step relies on Theorem 2.2 proven in Section 2. Intuitively, this theorem says that when searching for the best possible segmentation of a profile you should not consider change-points in plateaux of the observed signal. In Section 4, we apply our methodology to some real NGS data (DNA-seq profiles of human cells and RNA-seq profiles of bacteria). In the case of the DNA-seq experiment compressed profiles are around 10^5 to 10^6 nucleotides long and are a hundred times smaller than raw NGS profiles. Given these compressed profiles, the pruned DPA is able to recover the best segmentation in a matter of 5 to 30 minutes on a $2.3Ghz$ processor.

In the following, we briefly describe the statistical framework of our approach, which is similar to the one proposed by [1] for CGH profiles. We then highlight the challenging optimization problem that comes with this framework in the case of NGS data.

1.1 Statistical Framework

Let us assume that we have an ordered sequence of n observations $\{Y_i\}_{i\in[\![1,n]\!]}$, where n is the number of nucleotides on the chromosome and Y_i is either the number of reads starting at nucleotide i or the number of reads covering nucleotide i. A segmentation m of the sequence is defined by a set of K segments and $K+1$ change-points $\{\tau_k\}_{k\in[\![0,K]\!]}$ with the convention that $\tau_0 = 0$ and $\tau_K = n$. The k-th segment r_k of m is delimited by τ_{k-1} and τ_k: $r_k = [\![\tau_{k-1}+1, \tau_k]\!]$. We define $\mathcal{M}_{k,t}$ the set of all possible segmentations in $k > 0$ segments up to point t.

The segmentation model can be written as follows:

$$\{Y_i\} \text{ independant and } Y_i \sim G(\mu_r) \text{ if } i \in r,$$

where i is in segment r and G is some probability distribution depending on parameter μ_r. Typical examples for G are the Poisson or the normal distribution. We denote $p(y_i, \mu_r)$ the likelihood of data-point i in segment r. In a maximum likelihood framework, the goal is to identify the segmentation m in $\mathcal{M}_{k,t}$ of maximum likelihood defined as follows:

$$max_{m\in\mathcal{M}_{k,t}}\{ \sum_{r\,\in\, m} \min_{\mu_r}\{\sum_{i\in r} \log p(y_i,\mu_r)\} \}.$$

More generally one way to identify $k+1$ change-points is to find the segmentation m in $\mathcal{M}_{k,t}$ of minimal cost ([8,9]): $\sum_{r\,\in\, m} c(r)$, with $c(r)$ the cost of segment r of m defined as:

$$\begin{aligned} c(r,\mu) &= \sum_{i\,\in\, r} \gamma(Y_i,\mu) \\ c(r) &= min_{\{\mu\,\in\,\mathbb{R}\}}\{c(r,\mu)\}, \end{aligned}$$

where γ is a loss function (it is used as a measure of fit to the data). A typical example is the quadratic loss, namely: $\gamma(Y_i,\mu) = (Y_i-\mu)^2$ or the likelihood loss, namely: $\gamma(Y_i,\mu) = -\log p(y_i,\mu)$.

1.2 Optimization Problem

From a computational perspective, the goal is then to recover for every k in between 1 and K the best segmentation $M_{k,t}$ with respect to the chosen loss function and its cost $C_{k,t}$:

$$M_{k,t} = \arg\min_{\{m \in \mathcal{M}_{k,t}\}} \left\{ \sum_{r \in m} c(r) \right\} \quad \text{and}$$
$$C_{k,t} = \min_{\{m \in \mathcal{M}_{k,t}\}} \left\{ \sum_{r \in m} c(r) \right\}.$$

Recovering $M_{k,n}$ and $C_{k,n}$ is a difficult problem due to the very large number of possible segmentations in k segments: $\binom{n-1}{k-1}$. A DPA recovers exactly $M_{k,n}$ and $C_{k,n}$ in $O(K.n^2)$ [4], which is prohibitive for NGS profiles. Recently, we proposed a faster pruned and exact DPA to recover $M_{k,n}$ and $C_{k,n}$ [6]. This new algorithm allows the analysis of profiles with a million points in a matter of minutes. Yet, for large NGS data, scanning chromosomes with more than 10^8 base pairs, the empirical runtime of the pruned DPA is still prohibitive, even though its asymptotic complexity is of order $O(K.n.\log(n))$. Thus, there is a need for an even more efficient exact algorithm.

One important feature of NGS data is that they are discrete and that the probability of featuring the same count in two consecutive rows is high. In other words there are plateaux in the observed signal. As we will see in the next section these plateaux allow for an efficient compression of NGS profiles. Importantly, Theorem 2.2 shows that given a compressed profile it is still possible to recover the best segmentation of the raw NGS profiles w.r.t. the considered loss function.

After compression, the signal is small enough to be processed by our exact pruned DPA. However, the compressed signal is still too long to be processed by the original DPA [4]. In other words, our approach truly relies on the complementarity of the compression and the pruned DPA to obtain a reasonable runtime.

2 Signals with Plateaux

In the following section we will consider a sequence of n observations with a plateau, meaning that between two arbitrary t_1 and t_2 ($> t_1$) the signal is constant:

$$\forall t,\ t_1 \leq t \leq t_2,\ y_t = y_{t_1} = y_{t_2}.$$

Using this sequence we will demonstrate Theorem 2.2. This Theorem proves, the arguably intuitive idea that having a change-point between t_1 and t_2 is never beneficial in terms of cost. We first demonstrate this property for $k = 2$ in Lemma 2.1.

2.1 Searching for one Change-point

Let us consider a segmentation in 2 segments with a breakpoint at t. We define $P_t(\mu_1, \mu_2)$, the cost of this segmentation given some parameter μ_1 for the first segment and μ_2 for the second segment:

$$P_t(\mu_1, \mu_2) = \sum_{i=1}^{t} \gamma(y_i, \mu_1) \ + \ \sum_{i=t+1}^{n} \gamma(y_i, \mu_2)$$

The optimal cost P_t is:

$$P_t = min_{\mu_1}\{\sum_{i=1}^{t} \gamma(y_i, \mu_1)\} \ + \ min_{\mu_2}\{\sum_{i=t+1}^{n} \gamma(y_i, \mu_2)\}$$

Having these notations, let us prove the following lemma:

LEMMA 2.1.

- *If* $t_1 = 1$ *and* $t_2 = n$ *then* $\forall\ t\ P_t \geq C_{1,n}$
- *If* $t_1 = 1$ *and* $t_2 < n$ *then* $\forall\ t_1 - 1 \leq t \leq t_2$ *we have* $P_t \geq P_{t_2}$
- *If* $t_1 > 1$ *and* $t_2 = n$ *then* $\forall\ t_1 - 1 \leq t \leq t_2$ *we have* $P_t \geq P_{t_1-1}$
- *If* $t_1 > 1$ *and* $t_2 < n$ *then* $\forall\ t_1 - 1 \leq t \leq t_2$ *we have* $P_t \geq min\{P_{t_1-1}, P_{t_2}\}$

Proof

First scenario [$t_1 = 1$ *and* $t_2 = n$*]* We have:

$$P_t = t.min_{\mu_1}\{\gamma(y_1, \mu_1)\} + (n-t).min_{\mu_2}\{\gamma(y_1, \mu_2)\} = C_{1,n}.$$

Thus we get: $P_t \geq C_{1,n}$.

Second scenario [$t_1 = 1$ *and* $t_2 < n$*]* For any t such that $t \leq t_2$ we have:

$$P_t = t.min_{\mu}\{\gamma(y_1, \mu)\} + min_{\mu}\{(t_2 - t)\gamma(y_1, \mu) + \sum_{i=t_2+1}^{n} \gamma(y_i, \mu)\}$$

Thus we have:

$$P_t \geq t.min_{\mu}\{\gamma(y_1, \mu_1)\} + (t_2 - t).min_{\mu}\{\gamma(y_1, \mu)\} + min_{\mu}\{\sum_{i=t_2+1}^{n} \gamma(y_i, \mu)\}$$

And we get $\forall\ t \leq t_2 \quad P_t \geq P_{t_2}$.

Third scenario [$t_1 > 1$ *and* $t_2 = n$*]* We get $\forall\ t_1 - 1 \leq t \quad P_t \geq P_{t_1-1}$ by reversing the index and using scenario 2.

Fourth scenario [$t_1 > 1$ *and* $t_2 < n$*]* For any t such that $t_1 - 1 \leq t \leq t_2$ we get:

$$\begin{aligned} P_t(\mu_1, \mu_2) \quad = \quad & \sum_{i=1}^{t_1-1} \gamma(y_i, \mu_1) \ + \sum_{i=t_2+1}^{n} \gamma(y_i, \mu_2) \\ & + (t - t_1 + 1)\gamma(y_{t_1}, \mu_1) + \ (t_2 - t)\gamma(y_{t_1}, \mu_2) \end{aligned}$$

Thus, for fixed μ_1 and μ_2 and for $t \in [t_1 - 1, t_2]$, $P_t(\mu_1, \mu_2)$ is a linear function of t. Thus we get that for any μ_1 and μ_2:

$$P_t(\mu_1, \mu_2) \geq min\{P_{t_1-1}(\mu_1, \mu_2), P_{t_2}(\mu_1, \mu_2)\} \geq min\{P_{t_1-1}, P_{t_2}\}.$$

As this is true for any μ_1 and μ_2 we get $P_t \geq min\{P_{t_1-1}, P_{t_2}\}$ ■

2.2 Searching for several Change-points

THEOREM 2.2. *There exists a segmentation* m *in* K *or less segments without any change-point in the plateaux such that the optimal cost of* m *is equal to* $C_{K,n}$.

Proof Assume that we have a segmentation m in $\mathcal{M}_{K,n}$ with a breakpoint τ_k in a plateau. Then applying lemma 2.1 on the sequence $\{y_i\}_{i \in [\![\tau_{k-1}, \tau_{k+1}]\!]}$ we see that τ_k can either be discarded or moved to $t_1 - 1$ or t_2 without increasing the cost. Thus there exists a segmentation in K or less segments without any change-point in the plateau such that its optimal cost is $C_{K,n}$ ■

This theorem is more subtle than we might have thought based on our intuition. It does not mean that a change-point in a plateau is never optimal but only that it is not necessary to have change-points in plateaux to achieve optimality.

2.3 Compression of the Signal

According to Theorem 2.2 when searching for the best segmentation of the data we don't need to look for change-points in plateaux. In other words a plateau starting at position t_1 and ending at position t_2 can be considered as a unique data point with value y_{t_1} and weight $t_2 - t_1 + 1$. At worst the signal cannot be compressed. Thus, the two-step algorithm (compression and pruned DPA) does not change the worst case complexity. Yet, as we will see, very often the compression is efficient and allow for a significant reduction of the overall runtime.

Let us now consider a chromosome with n nucleotides. We have mapped R reads on this chromosome. Let us assume that R is smaller than n and that Y_i is the number of reads starting at nucleotide i. In that case the worst compression is achieved if all reads map different starting nucleotides. If R is smaller than $n - R$ then we can define at most $2R + 2$ plateaux in the signal. If R is bigger than $n - R$ then we can define at most $2(n - R) + 2$ plateaux in the signal. In the end we get at most $2min(R, n - R) + 2$ plateaux allowing for a $2\frac{min(R,n-R)+1}{n}$ compression. So for reads of length 55 and a $5\times$ coverage we overall have $\frac{n}{11}$ reads and we will get at worst a $\frac{2}{11} + \frac{2}{n} \geq 18.2\%$ compression. In the case of RNA-seq we do not expect reads to be evenly scattered, on the contrary they are concentrated in transcribed regions and between those regions we expect large plateaux of 0 allowing for an efficient compression.

3 Segmentation of the Compressed Profile

The compressed profile is still quite large, with 10^5 to 10^6 points. The original DPA [4] cannot be run on such a large profile because of its $\Theta(n^2)$ space and $\Theta(Kn^2)$ time complexities. Thus, we used the exact pruned DPA for segmentation that we developed recently [6]. Here we give a brief overview of this new algorithm and outline its main differences compared to the original DPA.

The segmentation problem involves two different types of parameters: change-points, which are discrete, and the parameters within each region (μ), which are often continuous. First, the original DPA recovers for all $\frac{n(n+1)}{2}$ possible segments the best possible μ and its associated cost or minimal loss. The result is stored in $\Theta(n^2)$ space. Then using a segment additivity property it recovers the best possible segmentations in 1 to K segments. The main idea behind our pruned DPA is to proceed the other way around. First, the pruned DPA optimizes the position of change-points for unspecified μ and then optimizes the value of μ. This way we get a point additivity property and a pruning property that allows for an efficient update and pruning of possible solutions.

4 Application to Real Data

In this section we applied the algorithm to two real datasets. The algorithm was coded in C++ and ran on a laptop with a 2.53 *Ghz* processor and 4 GB RAM.

The first dataset is a nine-sample DNA-seq dataset of human tumors and cell-lines. Human chromosomes are made of 50 million up to 250 million nucleotides. There was around 10 million mapped reads of length 55 per sample. In this DNA-seq experiment the reads are hopefully relatively well scattered across the whole genome. We thus expect the compression of each chromosome to be around $\frac{2\times10^6}{3\times10^9} \approx 0.66\%$. It is indeed what we observed in Fig. 1 (Up). After the compression step we applied the pruned DPA algorithm to recover the segmentation in 1 up to 1000 segments of each chromosome. The runtime of the pruned DPA to analyze one chromosome never exceeded 35 minutes and the relationship between the runtime and the length of the profile after compression seems linear (see Fig. 1 (Bottom)). Note that in this dataset it was not possible to run the pruned DPA on the raw NGS profiles (without compression) due to the very large size of the chromosomes.

We also analyzed a five-sample RNA-seq dataset of bacteria. We analyzed separately the forward and reverse strands making for a total of 10 NGS profiles. The size of the chromosome was $1.3\ 10^6$ base pairs. There was around 7.10^6 mapped reads of length 54 for the 10 profiles. If the mapped reads were scattered across the whole chromosome we would expect a low compression level of around $\frac{2.(1.3-0.7)}{1.3} \approx 92.3\%$. Yet, as expected, reads are concentrated on transcribed regions of the chromosome allowing for a much more efficient

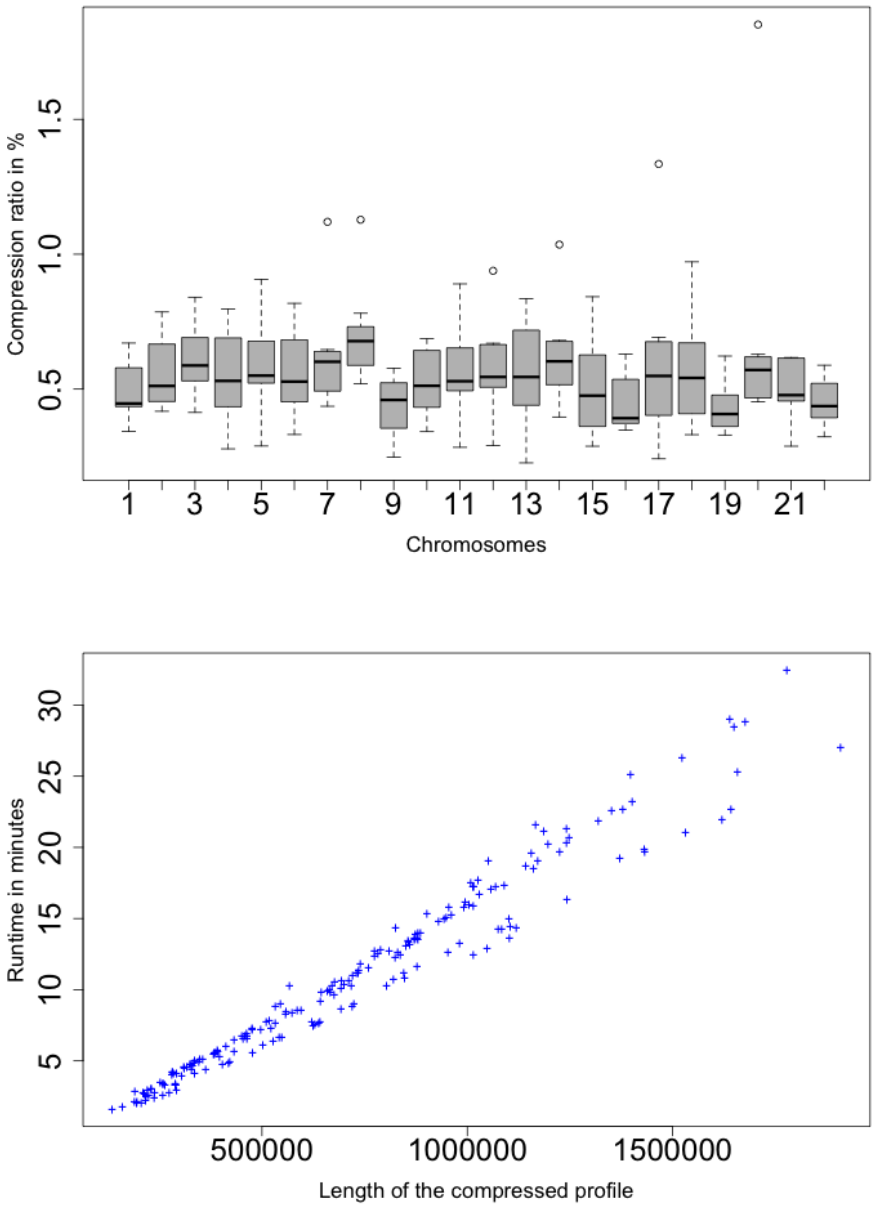


Figure 1. (Up) Compression efficiency in % across the nine samples and for the different chromosomes. The compression ratio is measured as the length of the compressed NGS profile divided by the size of the raw NGS profile. (Bottom) Runtime in minutes of the pruned DPA as a function of the length of the chromosome after compression.

compression. Compressed profiles are 19.6% to 1.8% the length of the raw NGS profiles. On this dataset it was possible to run the pruned DPA with or without compression (due to the smaller size of the chromosome). The compression ratio is smaller than the runtime ratio (runtime with compression over the runtime without compression in %, see Fig. 2 (Left)), still the gain in time is important. The relationship between the runtime and the length of the sequence after compression seems linear similarly as for the DNA-seq dataset (see Fig. 2 (Right)).

Importantly, in both datasets even the compressed profiles are most of the time too large to be analyzed by orginal DPA.

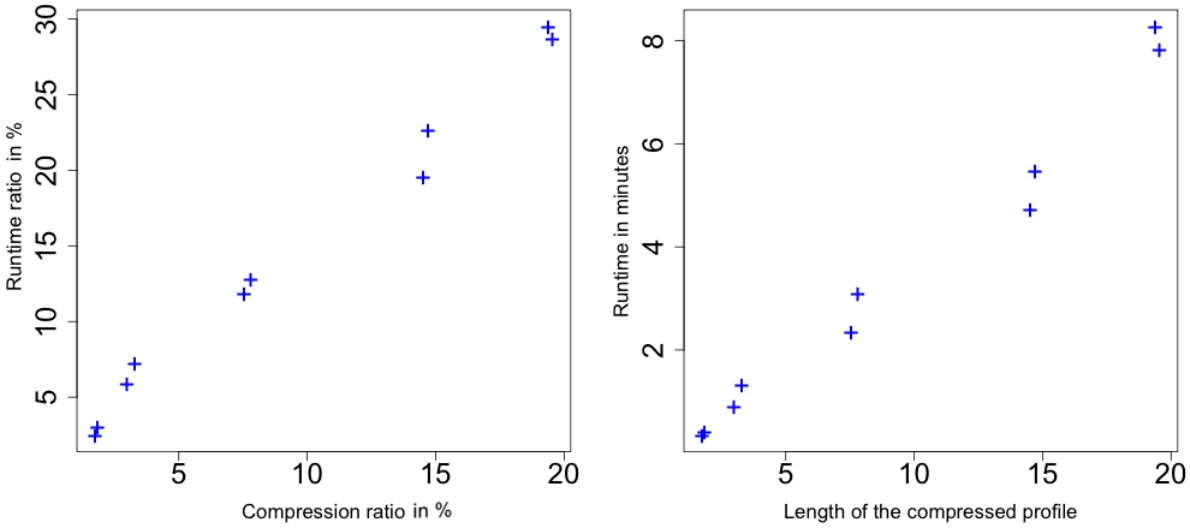


Figure 2. (Left) Runtime ratio in % of the pruned DPA as a function of the compression ratio. The runtime ratio is measured as the runtime of the pruned DPA with compression divided by the runtime without compression. The compression ratio is measured as the length of the compressed NGS profile divided by the size of the raw NGS profile. (Right) Runtime in minutes of the pruned DPA as a function of the length of the chromosome after compression.

5 Conclusion

Segmentation of NGS profiles to discover DNA copy number alterations or transcribed regions of the genome is a computationally difficult problem due to their very large size. Thus many methods rely on heuristic computational schemes. Here we presented an exact algorithm to segment these profiles. Our algorithm recovers the best segmentation in 1 up to K segments with respect to some loss function (either the Poisson or quadratic loss). Our algorithm first compresses the data. After compression, the signal is processed by an exact pruned DPA that we recently developed. The combination of the compression step and the efficient pruned DPA enables the analysis of large NGS profiles. In the case of human DNA-seq experiments the compressed signals are less than 2% the size of the original data. The compressed profiles are still large ($n \approx 10^6$) but small enough for the pruned DPA. Overall our approach allows to recover the best segmentation of an NGS profiles with more than 50 million nucleotides in a matter of minutes.

6 Acknowledgements

This work is funded by NWO/ZonMW through a grant for the Cancer Systems Biology Center at the Netherlands Cancer Institute, Antoni van Leeuwenhoek hospital. I thank S. Robin, M Koskas, E. Lebarbier and L. Wessels for fruitfull discussions.

References

[1] F. Picard, S. Robin, M. Lavielle, C. Vaisse, and J-J. Daudin, A statistical approach for array CGH data analysis. *BMC Bioinformatics*, 6:27, 2005.

[2] A. Olshen, E. Venkatraman, R. Lucito, and M. Wigler, Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics (Oxford, England)*, 5(4):557–572, October 2004.

[3] W. Lai, M. Johnson, R. Kucherlapati, and P. Park, Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics (Oxford, England)*, 21(19):3763–3770, 2005.

[4] R. Bellman, On the approximation of curves by line segments using dynamic programming. *Commun. ACM*, 4(6):284, 1961.

[5] E. Venkatraman and A. Olshen, A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, 23(6):657 –663, 2007.

[6] G. Rigaill, Pruned dynamic programming for optimal multiple change-point detection. *Arxiv:1004.0887*, 2010.

[7] D. Chiang, G. Getz, D. Jaffe, M. O'Kelly, X. Zhao, S. Carter, C. Russ, C. Nusbaum, M. Meyerson, and E. Lander, High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat Meth*, 6(1):99–103, 2009.

[8] M. Lavielle, Using penalized contrasts for the change-point problem. *Signal Processing*, 85(8):1501–1510, 2005.

[9] J. Bai and P. Perron, Computation and analysis of multiple structural change models. *J. Appl. Econ.*, 18:1–22, 2003.

Distributed High Throughput Sequencing Data Analysis on Cloud Computing

Laurent JOURDREN[1], Maria BERNARD[1], Marie-Agnès DILLIES[2] and Stéphane LE CROM[1]

[1] Ecole normale supérieure, Institut de Biologie de l'ENS, UMR8197 CNRS, U1024 INSERM, 46 rue d'Ulm, 75005, Paris, France
{laurent.jourdren, maria.bernard, stephane.lecrom}@ens.fr

[2] Plate-forme Transcriptome et Epigénome, Institut Pasteur, 28 rue du Dr. Roux, 75724 Paris cedex 15, France
{marie-agnes.dillies}@pasteur.fr

Keywords High throughput sequencing, RNA-Seq, Amazon Web Services, cloud computing, MapReduce.

Analyse Distribuée des Données de Séquençage à Haut Débit grâce au Calcul dans le « Nuage »

Mots-clés Séquençage à haut débit, RNA-Seq, Amazon Web Services, cloud computing, MapReduce

1 Introduction

During the last five years, high throughput sequencing techniques evolved to achieve better accuracy, ease of use and yield [1]. As the amount of data generated growths exponentially, the request for computer resources needs to follow these evolutions [2]. Data analysis requires large computer infrastructures that are only profitable for large genomic or informatics centers. In these conditions, how will platforms equipped with one sequencing device or teams working on smaller projects be able to access the computer resources they need occasionally?

Answers to this problem can be found through cloud computing solutions [3]. Here cloud computing refers to on demand access to computer resources using distant networks. Thanks to such data center there is no need to invest in a big computer infrastructure to perform large calculation processes. Through parallelized algorithms, such as MapReduce, data analysis from high throughput sequencing, distributed over several processors at the same time, can be performed in some hours [4]. Meanwhile, programming software using MapReduce algorithm is not widespread among bioinformatics developers, and only few solutions have been made available to deal with sequencing data over distributed networks.

We present Eoulsan, an open source workflow to work on high throughput sequencing data analysis using distributed calculation. This framework has been developed with the aims to automate the analysis of a large number of samples at once, simplify the configuration of cloud computing infrastructure and work with various already available analysis solutions. We first implemented Eoulsan to work on the differential analysis of transcript expression and tried it using Amazon Web Services (AWS) cloud computing facilities. We assess the performances of AWS in order to select the best parameter combination among the type and number of computer servers (instances) that can be used, and we analyze how Eoulsan can deal with the throughput increase that will come from future sequencing devices.

2 Results

Eoulsan works in 5 steps: quality control filtering, mapping, expression calculation, normalization and differential analysis. All information available on the experimental design is gathered in one text file inspired from the one of the limma R package for microarray analysis [5]. All the options needed to run the workflow are gathered in one XML file that allows for the usage of plugin programmed by external developers.

Eoulsan has been adapted for distributed calculation using the Hadoop system, the main open source implementation of the MapReduce algorithm. The workflow runs on AWS as described on Figure 1. It connects to AWS Simple Storage Service (S3) and transfers all the files needed for analysis. AWS books the requested number of instances on AWS Elastic Compute Cloud (EC2). The data are uploaded from S3 to EC2. Eoulsan performs filtering, mapping and expression calculation. The data are then downloaded back to the S3 storage location and AWS shuts down the cluster created on EC2.

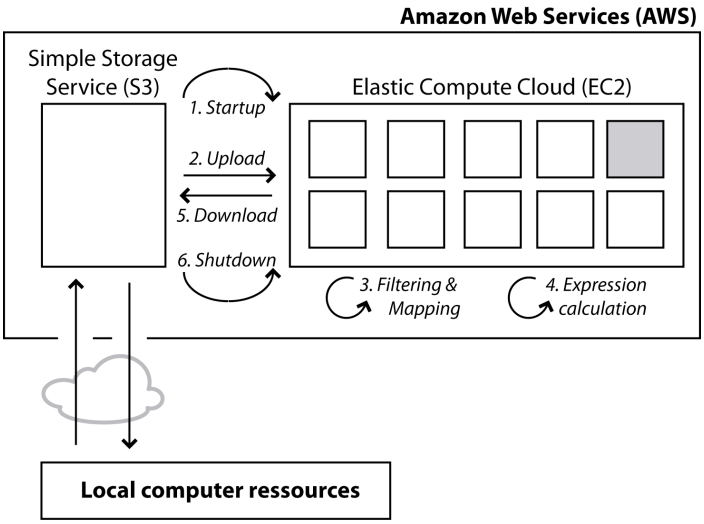


Figure 1. Description of the Eoulsan analysis workflow on Amazon Web Services (AWS).

We ran Eoulsan with 8 mouse samples from RNA-Seq sequencing for a total of 188 million reads using three different read mappers: BWA, Bowtie and SOAP2. We surveyed the time needed to perform the calculation process and the cost charged by AWS on three different EC2 instance types: m1.large, m1.xlarge and c1.xlarge (Table 1). Eoulsan succeeds in calculating the expression of mouse transcripts except for the SOAP2 mapper on c1.xlarge and m1.large instances due to the high memory requirement of SOAP2 algorithm (Table 2). Comparing the time spent using m1.xlarge instance on the three mappers, most of the duration differences come from the mapping step, varying from 39 minutes with Bowtie to 195 minutes with SOAP2. For BWA and Bowtie the fastest result is always obtained from the c1.xlarge instance, followed by m1.xlarge and m1.large. In term of costs, whatever the mapper used, the price of the calculations is always the most expensive using m1.xlarge instances followed by c1.xlarge, with m1.large being the most economic choice.

Instance	Memory (Go)	Virtual cores	I/O performance	Price USD/hour
m1.large	7.5	2	high	$0.44
m1.xlarge	15.0	4	high	$0.88
c1.xlarge	7.0	8	high	$0.88

Table 1. Instance selection for our study was made from the available EC2 servers from AWS accessible in all world regions with high input/output access performances. Prices are given for Ireland based instance location.

We tested how the number of instances booked influences the calculation process by varying the number of m1.large and c1.xlarge instances using the Bowtie mapper. The whole time spent for the calculation follows an exponential decrease curve for both instance types. The main contribution to these curve shapes comes from the mapping and expression steps. The best economic solution is achieved with 6 and 5 instances for m1.large and c1.xlarge respectively. In addition the cost per hour is linear over the number of instances used. This means that the number of instances can be increased with no risk to fall in a suboptimal configuration in order to speed up the data analysis process.

Mapper	Instance	Total (min)	Upload (min)	Mapping (min)	Expression (min)	Download (min)	Startup/Shutdown (min)	Cost
BWA	m1.xlarge	250	16	151	28	43	11	$44.00
BWA	c1.xlarge	187	13	106	23	38	7	$35.20
BWA	m1.large	412	20	300	36	47	8	$30.80
Bowtie	m1.xlarge	121	14	40	27	31	8	$26.40
Bowtie	c1.xlarge	109	14	34	20	32	8	$17.60
Bowtie	m1.large	176	21	70	33	43	11	$13.20
SOAP2	m1.xlarge	270	17	195	25	25	8	$44.00
SOAP2	c1.xlarge	126	18	-	-	-	7	$26.40
SOAP2	m1.large	822	16	-	-	-	11	$61.60

Table 2. Comparison of execution time duration with various read mappers on several AWS EC2 instances. Prices are given for Ireland based instance location.

Finally, to follow throughput evolution we assessed the impact of raw data increase on computational time by running Eoulsan with 16 and 32 samples, respectively 376 and 752 million total reads. The plot of run time against the number of samples for various instance number shows linear relations (data not shown). This clearly demonstrates that Eoulsan is able to handle the increase of raw data coming from future evolutions of sequencing devices.

Eoulsan is distributed under the GNU Lesser General Public License (LGPL) and CeCill-C and is available for download with a complete documentation at http://transcriptome.ens.fr/eoulsan.

3 Discussion

Our framework provides from standalone workstation to cloud computing clusters an integrated and flexible solution for high throughput sequencing data analysis from reads alignment to the list of significant differentially expressed transcripts. With its modular structure and parallel data processing, Eoulsan is ready to fulfill the challenges coming from the massive increase of the amount of data and the new applications of sequencing technologies.

The major limitation of the usage of cloud computing comes from data transfer as the network used to send data to Amazon S3 server storage is shared among all Internet users. However, the distributed calculation process used in Eoulsan is based on Hadoop and it can be installed on numerous cluster server configurations. It would be of interest to create regional genomic computer infrastructures to be shared among several local high throughput sequencing users. With dedicated high-speed networks, this can speed up the time transfer process. In addition, this could also favor the standardization of analysis pipelines developed from the bioinformatics community, making high throughput sequencing technologies really accessible for a wide audience.

Acknowledgements

This work was supported by AWS research grant and IBiSA network.

References

[1] ER. Mardis, A decade's perspective on DNA sequencing technology. *Nature*, 10:198-203, 2011.

[2] E. Pennisi, Human genome 10th anniversary. Will computers crash genomics? *Science*, 331:666-668, 2011.

[3] EE. Schadt, MD. Linderman, J. Sorenson, L. Lee and GP. Nolan, Computational solutions to large-scale data management and analysis. *Nat. Rev. Genet.*, 11:647-657, 2010.

[4] MC. Schatz, B. Langmead and SL. Salzberg, Cloud computing and the DNA data race. *Nat. Biotech.*, 28:691-693.

[5] GK. Smyth, Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol.*, 3, 2004

Conformément au souhait des auteurs, cette contribution n'est pas reproduite dans la version en ligne des actes de JOBIM 2011.

Following the wishes of the authors, this paper is not included in the online version of the JOBIM 2011 proceedings.

Conformément au souhait des auteurs, cette contribution n'est pas reproduite dans la version en ligne des actes de JOBIM 2011.

Following the wishes of the authors, this paper is not included in the online version of the JOBIM 2011 proceedings.

Identification of Regulatory Elements from Gene Expression Data without Clustering

Mathieu LAJOIE[1], Olivier GASCUEL[1] and Laurent BRÉHÉLIN[1]

Méthodes et Algorithmes pour la Bioinformatique, LIRMM, CNRS, Université Montpellier 2, France
{lajoie,gascuel,brehelin}@lirmm.fr

Keywords Gene regulation, motif discovery, transcription factor, k-nearest neighbors.

1 Background

In silico discovery of regulatory motifs (RM) can be seen as a feature extraction problem. Given a set of nucleic acid sequences that are mapped to some expression data, the goal is to find a concise set of motifs that is the most informative with regards to that mapping. Apart from few exceptions (e.g. [1]), one of the most common approaches is to use a clustering algorithm to partition the expression dataset, and to apply on each cluster one of the numerous algorithms that have been designed to find over-represented motifs in a predefined set of sequences, such as MEME [2] or AlignACE [3]. However, the partition induced by this clustering rarely corresponds to a biological reality. Firstly, expression data is inherently noisy, and determining the "real" number of clusters is considered to be one of the most difficult problems in classification. Secondly, different RM usually have overlapping gene sets, which cannot be appropriately modeled by a single partition. In addition, these algorithms rely on statistical models of sequence background, which have been reported to produce many false positives [4,5], especially with repeat-rich and atypical genomes. This is the case, for example, with *Plasmodium falciparum*, whose A+T content almost reaches 90% in the intergenic regions.

The FIRE [4] and GEMS [5] algorithms have been designed for finding RM from whole genomes and high dimensional datasets without models of sequence background. However, they both rely on a clustering of the expression data and are subject to the aforementioned criticisms. The two methods differ in the way the dependency between the presence of a motif and the expression profile of the corresponding gene is measured. GEMS uses the hypergeometric distribution to measure motif enrichment in each co-expression cluster, while FIRE computes the mutual information between the presence/absence of a motif and the cluster membership of the corresponding genes. These two approaches can be seen as two extremes of a simple model, which only assumes that RM must show some kind of statistical dependency with the expression data. The hypergeometric approach is a local criterion, as it considers motif enrichment in a single cluster at a time, while the mutual information approach is a global one including the contributions of all the clusters.

2 Method and Results

In this work, we show that the hypergeometric distribution and the mutual information criteria can be used without requiring any clustering, using the notion of *motif density* in expression space. Namely, rather than considering the number of genes that contain a motif in each cluster, we compute motif densities locally around each gene with a k-nearest neighbors approach. For the hypergeometric criterion, the score of a motif is then defined as the negative logarithm of the lowest p-value observed among all these neigborhoods. For the mutual information, the score is obtained by summing over the density estimate of each gene, instead of each cluster.

We compared the original and new version of both criteria on two *S. cerevisiae* and three *P. falciparum* gene expression datasets. All possible 8-mers were enumerated and scored with the four objective functions, and a false discovery rate (FDR) was estimated for different score thresholds using a shuffling procedure. For the original criteria, we used the k-means algorithm with different number of clusters ($3 - 10, 20, 30, 40$), and kept the clustering that yields the best sensitivity at 0.1 FDR. Fig. 1 shows the number of 8-mers identified by the new

and original criteria under various FDR for two datasets. We see that avoiding the clustering step results in a significant increase of the sensitivity over the original methods. Overall, we observed a significant improvement on the five datasets for the hypergeometric criteria, and on four datasets for the mutual information.

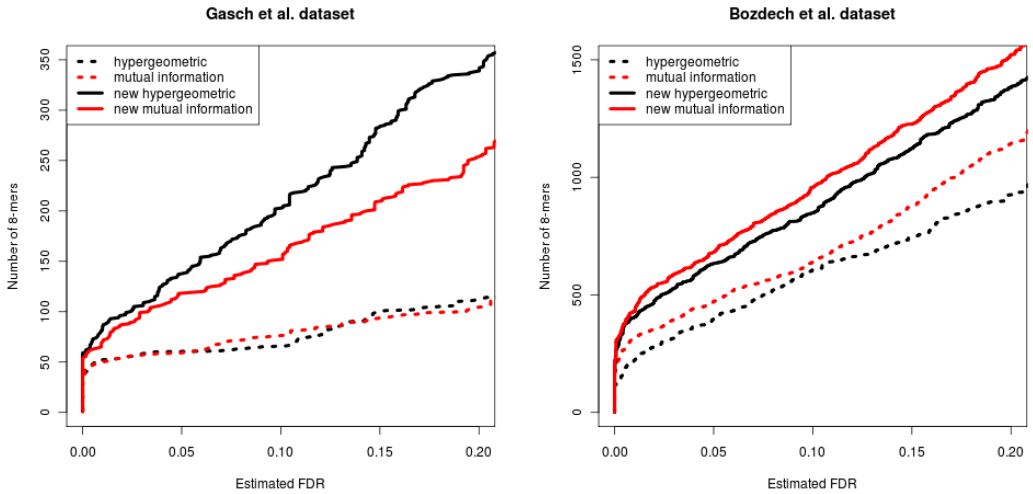


Figure 1. Number of predicted 8-mers, according to different estimated FDR thresholds, for (left) the yeast dataset [6] and (right) the *P. falciparum* dataset [7]. The best results of the original scoring functions (hypergeometric and mutual information) are achieved with 9 and 7 clusters (respectively) in yeast, and 3 and 7 clusters (respectively) in *P. falciparum*.

Using yeast Protein Binding Microarray (PBM) datasets [8], we further show that our continuous approaches also improve prediction. A True Positive Rate (TPR) that measures the fraction of predicted 8-mers bound by a transcription factor in the PBM experiments was computed for the four approaches. As we observed with the estimated FDRs, the new criteria outperform the original ones in this experiment. For the Gasch dataset presented in Fig. 1, the TPR is 55% for the 200 highest scoring 8-mers returned by the original criteria, whereas it reaches 65% for the new versions. Finally, we showed that using motif densities presents several advantages compared to the clustering approach. In addition to the increased sensitivity, it provides a simple way of comparing different motifs and predicting the functionality of individual motifs occurrences. All these methods have been implemented in a software called RED^2, for *Regulatory Elements Discovery from Raw Expression Data*. Motifs are represented as IUPAC strings of arbitrary lenght, allowing an easy and comprehensive analysis of a wide range of expression data.

References

[1] H.J. Bussemaker, H. Li and E.D. Siggia, Regulatory element detection using correlation with expression. *Nature genetics*, 27(2):167–174, 2001.

[2] T.L. Bailey and C. Elkan, Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *ISMB. International Conference on Intelligent Systems for Molecular Biology*, volume 2, page 28, 1994.

[3] J.D. Hughes, P.W. Estep, S. Tavazoie, and G.M. Church, Computational identification of Cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *Journal of molecular biology*, 296(5):1205–1214, 2000.

[4] O. Elemento, N. Slonim, and S. Tavazoie, A universal framework for regulatory element discovery across all genomes and data types. *Molecular cell*, 28(2):337–350, 2007.

[5] J.A. Young, J.R. Johnson, C. Benner, S.F. Yan, K. Chen, K.G. Le Roch, Y. Zhou, and E.A. Winzeler, In silico discovery of transcription regulatory elements in Plasmodium falciparum. *BMC Genomics*, 9(1):70, 2008.

[6] A.P. Gasch, P.T. Spellman, C.M. Kao, O. Carmel-Harel, M.B. Eisen, G. Storz, D. Botstein, and P.O. Brown, Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, 11(12):4241, 2000.

[7] Z. Bozdech, M. Llinás, B.L. Pulliam, E.D. Wong, J. Zhu, and J.L. DeRisi, The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum. *PLoS Biol*, 1(1):E5, 2003.

[8] G. Badis, E.T. Chan, H. van Bakel, L. Pena-Castillo, D. Tillo, K. Tsui, C.D. Carlson, A.J. Gossett, M.J. Hasinoff, C.L. Warren, et al, A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. *Molecular cell*, 32(6):878–887, 2008.

Mapping Genetic Interactions in Various Contexts Provides Complementary Information

Magali MICHAUT and Gary D. BADER

The Donnelly Centre, University of Toronto, 160 College Street, M5S 3E1, Toronto, Ontario, Canada
{magali.michaut, gary.bader}@utoronto.ca

Keywords Genetic interactions, budding yeast, network comparison, phenotype.

1 Introduction

Genetic interactions are useful to study biological processes and their functional relationships, and very powerful to predict gene function [1]. A genetic interaction is a deviation from the expectation for a double mutation in terms of a certain phenotypic readout in a given environment. How different are genetic interaction networks in different conditions or using different phenotypic readouts?

In *Saccharomyces cerevisiae*, most genetic interaction studies assess yeast cell growth in standard laboratory conditions. Recently some studies were performed in DNA damage conditions [2] or were based on a different phenotypic readout [3]. A comparison of genetic interaction networks in standard laboratory conditions and in DNA damage conditions revealed large differences and many interactions specific to each condition [2]. Nevertheless genetic interaction networks are subject to systematic and stochastic errors. Are the differences between the networks only the result of lack of coverage of both networks and errors in the experimental identification of genetic interactions? What do we expect by chance? In addition, we don't know if those results hold more generally for any pairs of conditions or for different phenotypic readouts.

To answer those questions, we compare genetic interaction networks mapped with two different phenotypic readouts in standard laboratory conditions: cell growth (*SGA*) [1] and endocytosis defect (*Burston*) [3]. Using the error rates estimated by Costanzo et al., we estimate the overlap expected by chance between the two data sets in terms of interactions. In addition, we use an alternative genetic interaction network based on cell growth as a control for the comparison (*Collins*) [4].

2 Results

2.1 The Quantitative Scores Are not Correlated

Both *SGA* and *Burston* data sets provide biological insights and are independently biologically informative as shown in the original analysis [1, 3]. To perform a meaningful comparison, all analyses are limited to the set of gene pairs tested in both data sets. Considering all measured scores, we first note that their quantitative scores are not correlated (Pearson r=0.06). This shows that both networks provide different information. As a control, we consider *SGA* and *Collins* networks, which are both based on growth phenotypic readout and in the same conditions. The scores from those data sets are much more correlated (r=0.45), which suggests that genetic interaction networks using same conditions and phenotypic readouts are expected to be correlated.

2.2 The Filtered Binary Interactions Show Low Overlap

We then consider data sets as binary (filtered interactions). We find that the overlap (Jaccard coefficient) between *Burston* and *SGA* is very low (8%) and much lower than the overlap between *Collins* and *SGA* (20%). In addition, we note that both data sets provide unique information. We define the unique ratio as the number of interactions observed in exactly one data set among all observed interactions. *Burston* and *SGA* provide mostly (92%) unique information (Table 1). Surprisingly we find that *Collins* and *SGA* also provide much (80%) unique information. Finally we ask whether both data sets agree on the type of genetic interactions (positive/negative). Among the gene pairs with an edge in both data sets, we compute the percentage of pairs of the same sign. We find that *SGA* and *Burston* agree on 64% of the interactions as opposed to 93% for *SGA* and *Collins* (Table 1). All together, these results show that the *Burston* and *SGA*

data sets overlap less than data sets obtained with the same phenotypic readout and give complementary and unique information.

	Tested	No	One	Two	Agree	Positive	Negative	Agree%	Overlap%	Unique%
Burston	10100	7013	2835	252	162	58	104	64	8.1	92
Collins	101908	84124	14168	3616	3359	741	2618	93	20.3	80

Table 1. Comparison between *SGA* and two other data sets. *Burston* uses a different phenotypic readout whereas *Collins* uses the same and is thus used as a control. No/One/Two indicate how many of the tested gene pairs are identified as interacting in 0/1/2 data sets. Agree is when the observed interactions have the same sign.

2.3 Positive and Negative Networks Overlap Less than Expected by Chance

Since positive and negative interaction networks have different properties and different error rates, we also analyze positive and negative networks separately. Again we find that the overlap between *Burston* and *SGA* is lower than between Collins and *SGA* both for positive (4% vs 9%) and negative interactions (7% vs 24%). Moreover *Burston* and *SGA* provide mostly unique information for both positive (96%) and negative (93%) networks. Recently the comparison of DNA damage to standard conditions revealed a lot of unique information for positive (79%) and negative (61%) genetic interactions [2]. Comparing *SGA* and *Collins* as a control, we find that they also provide mostly unique information for positive (91%) and negative (76%) networks even though the negative networks have a lower rate of unique information.

To assess the overlap expected by chance between both data sets, we model the genetic interaction networks by ordinary graphs, build a stochastic model and use an estimation of the error rates. We find that *Burston* and *SGA* overlap less than expected by chance (Table 2). This suggests that both data sets provide different information. As a control, we compare *SGA* and *Collins* which both map genetic interaction using growth as a phenotypic readout and find that they overlap more than expected by chance.

	No	One	Two	Overlap	Overlap ratio	p-value
Burston-Pos	12262	1645	75	0.044	0.99	8.4e-09
Burston-Neg	12219	1568	195	0.110	0.61	1.8e-17
Collins-Pos	114900	8819	501	0.054	1.68	1.2e-30
Collins-Neg	109623	12650	1947	0.133	1.82	1.8e-146

Table 2. Expected values based on the stochastic model and comparison with the observed overlap.

3 Conclusions

In this work we explore the similarities and differences of genetic interaction networks obtained with different phenotypic readouts. We show that the networks mapped based on cell growth and endocytosis defect provide complementary information, controlling both with a stochastic model of the expected overlap and with an additional network based on the same phenotypic readout. It appears that different networks in the same conditions and using the same readout can be surprisingly complementary as well. Nevertheless, networks based on different readouts show lower overlap and correlation and overlap less than expected by chance while the control overlap more than expected by chance. We are currently investigating more networks and some detailed examples of various biological processes showing significant differences.

References

[1] M. Costanzo, A. Baryshnikova, J. Bellay, Y. Kim, E.D. Spear, C.S. Sevier, H. Ding, J.L. Koh *et al*, The genetic landscape of a cell. *Science*, 327(5964):425-431, 2010

[2] S. Bandyopadhyay, M. Mehta, D. Kuo, M.K. Sung, R. Chuang, E.J. Jaehnig, B. Bodenmiller, K. Licon *et al*, Rewiring of genetic networks in response to DNA damage. *Science*, 330(6009):1385-1389, 2010

[3] H.E. Burston, L. Maldonado-Baez, M. Davey, B. Montpetit, C. Schluter, B. Wendland, E. Conibear, Regulators of yeast endocytosis identified by systematic quantitative analysis. *J Cell Biol*, 185(6):1097-1110, 2009

[4] S.R. Collins, M. Schuldiner, N.J. Krogan, J.S. Weissman, A strategy for extracting and analyzing large-scale quantitative epistatic interaction data. *Genome Biol*, 7(7):R63, 2006

Session 5 : Systems Biology

Conférence invitée

Edda Klipp

Humboldt-Universität zu Berlin, Theoretical Biophysics, Germany.

Mathematical Modeling of Yeast Stress Response

Microorganisms live in changing environments. They have to face nutrient alteration, chemical and physical stresses, and they change themselves over cell cycle and aging. The ability to perceive and respond to information in their environment is one of the most ubiquitous properties of cellular organisms. It is crucial for a cell to react appropriately to changes or signals in its environment. This becomes apparent in many situations such as the search for nutrients, the detection of potentially harmful external conditions and in cell-cell communication as it is required for any multicellular organism. Even though there is a huge selection of perceivable signals the underlying mechanisms are surprisingly alike, which suggests that they are highly conserved in the course of evolution.

Over the last years, we have studied various signal transduction and regulatory pathways in a model organism, the yeast *Saccharomyces cerevisiae*, and investigated the response of cells to external perturbations on various levels. To this end, we have established mathematical models, mainly in form of ordinary differential equation systems, but also as Boolean models, stochastic models or spatial models, which are supported by experimental data. We focus on results with respect to interaction of different signaling and regulatory pathways. Specifically, new aspects in cell cycle regulation and the interaction of stress-activated signaling pathways with cell cycle progression will be discussed. The results indicate that yeast cells have developed different mechanisms for coping with external stress during different periods of their life time.

Session 6.A : Phylogeny and Evolution

PhyleasProg: a User-oriented Web Server for Wide Evolutionary Analyses

Joël BUSSET[1,2,3,4], Cédric CABAU[5], Camille MESLIN[1,2,3,4] and Géraldine PASCAL[1,2,3,4]
[1] INRA, UMR85, Physiologie de la Reproduction et des Comportements, F-37380 Nouzilly, France
[2] CNRS, UMR6175, F-37380 Nouzilly, France
[3] Université François Rabelais de Tours, F-37041 Tours, France
[4] IFCE, F-37380 Nouzilly, France
joel.busset@polytechnique.edu
{camille.meslin, geraldine.pascal}@tours.inra.fr
[5] INRA, SIGENAE, UR83 Recherches Avicoles, F-37380 Nouzilly, France
cedric.cabau@tours.inra.fr

Abstract *Evolutionary analyses of biological data are becoming a prerequisite in many fields of biology. At a time of high-throughput data analysis, phylogenetics is often a necessary complementary tool for biologists to understand, compare and identify the functions of sequences. But available bioinformatics tools are frequently not easy for non-specialists to use. We developed PhyleasProg (http://phyleasprog.inra.fr), a user-friendly web server as a turnkey tool dedicated to evolutionary analyses. PhyleasProg can help biologists with little experience in evolutionary methodologies by analyzing their data in a simple and robust way, using methods corresponding to robust standards. Via a very intuitive web interface, users only need to enter a list of Ensembl protein IDs and a list of species as inputs. After dynamic computations, users have access to phylogenetic trees, positive/purifying selection data (on site and branch-site models), with a display of these results on the protein sequence and on a 3D structure model, and the synteny environment of related genes. This connection between different domains of phylogenetics opens the way to new biological analyses for the discovery of the function and structure of proteins. In future, it will be possible to submit to PhyleasProg, a private sequence as input.*

Keywords Phylogenetic tree, Positive selection, protein three-dimensional structures, synteny.

1 Introduction

Today, more and more eukaryotic genomes have been sequenced thanks to second generation sequencing technologies thereby providing an extraordinary wealth of information for evolutionary analyses. Currently, the GOLD website [1] lists more than 3 000 eukaryotic genomes whose sequencing is complete or ongoing. Under these circumstances, bioinformatics tools could help understand the evolutionary histories of proteins especially by connecting phylogenetics analysis and positive selection calculations. These approaches constitute the core of many biological research areas, and as stated by Theodosius Dobzhansky “Nothing in biology makes sense except in the light of evolution”. Indeed, present protein sequences are the result of a long, complex and extensive evolutionary process. Proteins have different levels of conservation. Active sites or protein–protein interaction domains are often well conserved, while highly variable regions may carry sites under positive selection. Such positively selected sites may be interpreted as being a consequence of molecular adaptation, which may confer an evolutionary advantage to the organism [2-4].

Accordingly, the association of (i) the establishment of orthology and paralogy relationship, (ii) the functional inference by reconstruction of the phylogenetic tree, and (iii) the identification of sites/genes under positive selection is an important step, not only in studies of evolutionary biology, but also in functional studies. By projecting the results of positive selection onto the three-dimensional structure of proteins, this becomes a powerful and very useful tool for biologists. The combined data could help biologists plan site-directed mutagenesis experiments. However, obtaining a phylogenetic tree requires successive computations including identification of homologous sequences, multiple alignment, phylogenetic

reconstructions and graphic representation of the inferred tree. Obtaining positive selection data requires the use of mathematical methods, such as PAML [5], which are designed for specialists.

Several web sites offer phylogenetic tree reconstruction. Some are turnkey systems such as PhyloBuilder [6] and POWER[7]. Some offer a single tool, while others bring together many of the most popular programs for phylogenetic reconstruction such as Mobyle [8]. The web server Phylogeny.fr [9] is designed for non-specialists and has up-to-date programs that are often designed for experts. In parallel, two phylogenetic tree databases, PhylomeDB [10] and TreeFam [11], offer a large number of pre-computed trees based on all genes of all genomes. A number of web sites are also available for analyzing evolutionary forces. The web server Selecton [12] offers a user-friendly tool to compute positive selection and displays results on a three-dimensional structure of proteins. However, it only allows calculation of one set of orthologues. The DataMonkey server [13] enables detection of signatures of positive and negative selection from coding sequence alignments using a wide range of statistical models. The Selectome [14] database provides the results of a branch-site specific likelihood test for positive selection based on whole gene families from the TreeFam database. Phylemom [15] enables experts to build a complete pipeline dedicated to phylogenetics and evolution.

Many tools are already available to reply to phylogenetics and evolutionary questions. However, they are complex to use and do not allow all the necessary computations to be carried out on a single server. Phylogenetic tree reconstruction, positive selection detection and protein three-dimensional structure modelling require (i) installation/use of numerous tools, (ii) knowledge of up-to-date tools and (iii) substantial computational resources. In particular, when biologists analyze several proteins of interest, they want to repeat bioinformatics methods on their data in the same conditions and they want to obtain results in a reasonable amount of time. This is why we built PhyleasProg web server in such a way that it could be used by the largest possible number of biologists. Our aim was to combine usefulness and usability. Such a server is a helpful guide for biologists with little experience in evolutionary methodologies as it can analyse their data in a simple and robust way, using methods corresponding to well-accepted standards.

Via a very simple interface, users enter one or a list of Ensembl protein IDs [16] and choose a set of species about which they wish to obtain evolutionary information among the sequenced vertebrates in Ensembl. There is currently no limit to the number of IDs or the number of selected species. However, we recommend submitting jobs containing less than 20 IDs on all species in order to facilitate analysis of results. Once submitted, each ID is treated independently and the computations are performed on both orthologues and paralogues of the related genes. As output, PhyleasProg provides (i) phylogenetic trees, (ii) positive/purifying selection data (on site and branch-site models) with visualization of these outcomes on the protein sequence and whenever possible, on a 3D structure, and (iii) the genomic environment of related genes. To our knowledge, no other web server performs all these tasks on several input sequences simultaneously. In addition, PhyleasProg computes the degree of purifying selection and positive Darwinian selection for each site in the protein sequence and displays these data on the modelled molecular structure of the protein. To guide users through these different evolutionary methods, which are not always very easy for non-experts, the pipeline only returns results if they are statistically significant.

This unique connection between phylogenetic trees, synteny studies, positive/purifying selection data and 3D structures opens the way to new biological analyses to improve our understanding of function and structure of proteins.

2 Overview

The PhyleasProg pipeline is a combination of Perl modules and external software (Figure 1). As input data, it requires one or a list of Ensembl protein IDs and a list of species selected among completely or partially sequenced vertebrates in Ensembl [16]. Once the process is complete, users can obtain evolutionary results on each ID submitted, treated independently but simultaneously, on orthologues and paralogues of the related genes.

We intentionally chose to not embed an exhaustive number of similar methodologies in our platform. We chose rapid, up-to-date, accurate and proven tools. Multiple sequence alignments are performed by MUSCLE [17] and are refined by GBLOCKS [18], itself improved by a home-made Perl program. TREEBEST (http://treesoft.sourceforge.net/treebest.shtml) reconstructs phylogenetic trees. CODEML, a

PAML program [5], performs positive selection computation. MODELLER [19] builds homology models of the three-dimensional structure of proteins.

Data visualization was an important goal for the development of this platform. JALVIEW [20] is used to display multiple sequence alignments, ARCHAEOPTERYX [21] for interactive manipulation of phylogenetic trees and JMOL [22] to display the 3D structure of proteins. We were careful to present processes and results very simply to enable biologists to navigate through a user-friendly environment. To guide users, the pipeline only returns significant results. Moreover, all input and output data can be downloaded as flat files.

A cluster computer manages the execution of the whole pipeline. This choice allows a very reasonable execution speed and authorizes PhyleasProg to work on several proteins simultaneously. The user interface was optimized for Firefox browser developed in Perl CGI.

3 PhyleasProg Pipeline

3.1 Data Acquisition

3.1.1 Input

For a very simple use of PhyleasProg, only Ensembl IDs of the proteins to be studied and a list of the species with which they should be compared are required as inputs. Protein IDs can be separated by a comma, a space or a new line character. Ensembl protein IDs are unique, they start with 'ENS' and their last letter must be a 'P' (e.g. ENSMUSP00000099398). To choose species for which they want evolutionary results, users simply tick the name of the species in the lists of completely and partially sequenced genomes. The Job summary page summarizes the list of IDs submitted, the selected species, and displays the status of process for each ID.

3.1.2 Interrogation of Ensembl Database

We chose to work with Ensembl protein IDs because Ensembl provides high-quality genome annotation across vertebrate species and allows computer scientists to retrieve a lot of data very quickly, thanks to a Perl application programming interface (API) [23].

Using this API, for each protein ID submitted, we retrieved protein and related transcript sequences, related gene ID, orthologous and paralogous protein IDs, orthologue and paralogue protein sequences and related transcript sequences (Figures 1A, 1A'). Among the numerous orthologues identified in Ensembl, we chose to keep either the one-to-one orthologues or the related gene with the shortest evolutionary distance among the one-to-many or the many-to-many orthologues [24].

3.2 Reconstruction of Phylogenetic Trees

3.2.1 Multiple Sequence Alignment and Refinement

For each protein ID submitted, PhyleasProg reconstructs phylogenetic trees of both orthologues and paralogues. And for each orthologue related to one of the protein IDs submitted, a phylogenetic tree of paralogues is also reconstructed.

As shown in Figure 1B, multiple sequence alignment (MSA) of proteins is generated by MUSCLE. This alignment is then converted into multiple codon alignment by PAL2NAL [25]. As our pipeline offers a turnkey process, we had to pay particular attention to the quality of MSA because this is essential for the quality of the related phylogenetic tree. Thus, GBLOCKS is used to edit MSA. This software removes all sites containing at least one gap and sites that are too divergent because these positions might not be homologous or might be saturated by multiple substitutions. First of all, GBLOCKS is performed with strict parameters (type=codons; maximum number of contiguous non-conserved positions= 8; minimum length of a block= 10; no gaps allowed). After this first step, the generated MSA can be very short, which would seriously damage the rest of the computations in the PhyleasProg pipeline. Consequently, refinement step are performed recursively: if after GBLOCKS, the MSA length is less than 30% of the median length of sequences in the raw MSA, the sequence that induces most of the gaps is removed from the dataset, and a

new MSA is computed. If the length of the clean MSA is between 30% and 50%, a new editing with GBLOCKS is performed on the raw MSA with relaxed parameters (type=codons; maximum number of contiguous non-conserved positions= 10; minimum length of a block= 5; no gaps allowed). If after this last step, the length of the MSA is still too short, computation is aborted. Thus it is important to estimate the quality of the MSA (downloadable through the flat files menu) before analyzing the other results of the pipeline (Figure 2).

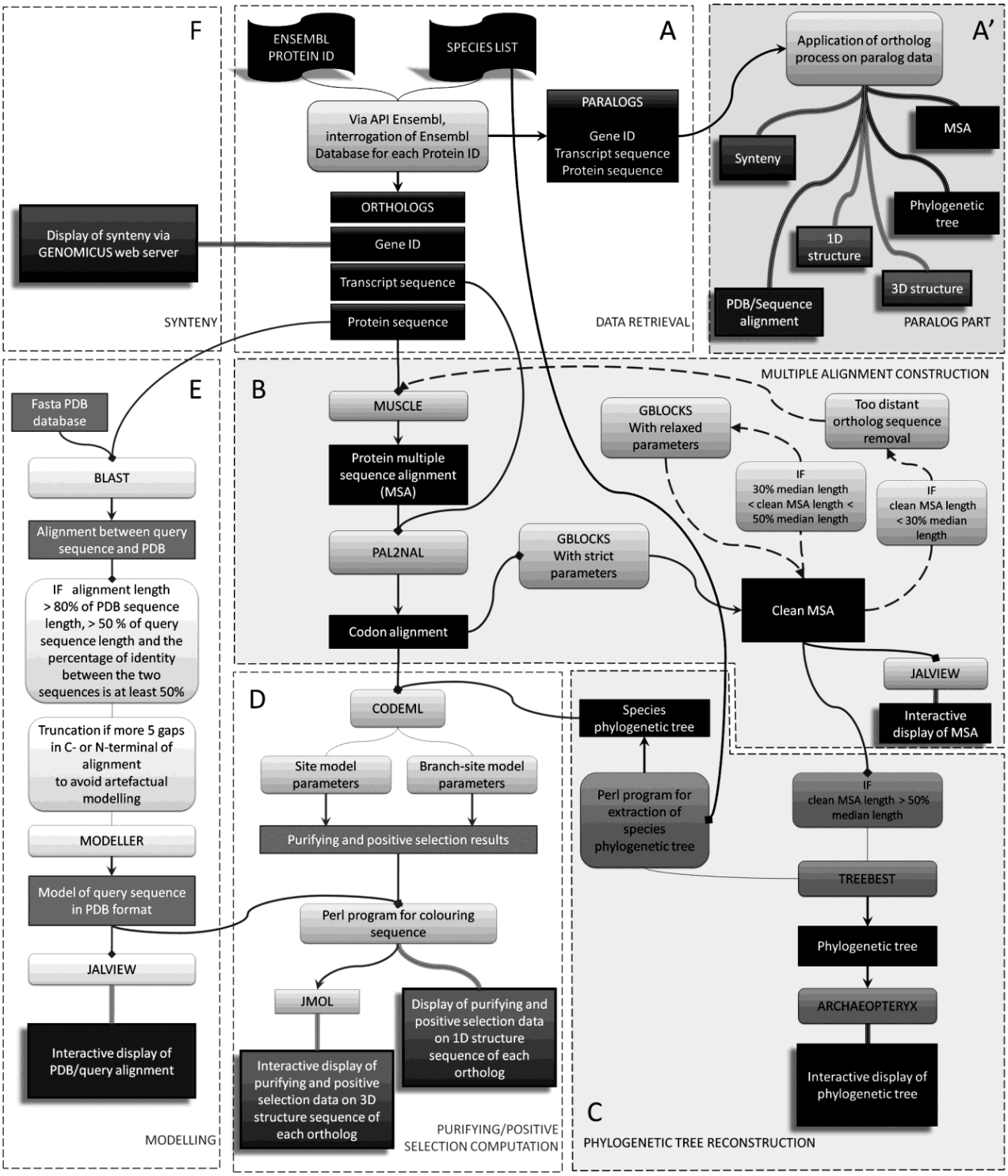


Figure 1. The workflow of PhyleasProg web server.

3.2.2 Phylogenetic Reconstruction

The clean MSA from the previous step is used to reconstruct the phylogenetic tree by TreeBeST (Figure 1C). TreeBeST integrates multiple tree topologies, in particular both DNA level and protein level models and combines them with a species-tree aware penalization of topologies, which is inconsistent with known species relationships. TreeBeST is run with the option best. This enables the combination of (i) a maximum likelihood (ML) tree built using PhyML [26] based on the protein alignment with the WAG model; (ii) a ML tree built using PhyML based on the codon alignment with the Hasegawa-Kishino-Yano (HKY) model; (iii) a neighbour-joining (NJ) tree using p-distance based on the codon alignment; (iv) a NJ tree using dN distance (rate of non-synonymous substitutions) based on the codon alignment; and (v) a NJ tree using dS distance (rate of synonymous substitutions) based on the codon alignment. As TreeBeST runs with a species tree, the final phylogenetic tree is rooted by minimizing gene duplications and then losses, the best rooting strategy for this type of input. TreeBeST produces trees with a bootstrap analysis.

3.2.3 Visualization

Archaeopteryx, the successor of ATV [27], is a Java application used as applet for the display and manipulation of annotated phylogenetic trees.

3.3 Positive/Purifying Selection Calculations

3.3.1 Overview

PhyleasProg gives positive and purifying selection data using maximum likelihood calculations which underlie the stochastic process of evolution. CODEML, from the package PAML (Figure 1D) [5], evaluates the ratio of non-synonymous/synonymous substitution rates (dN/dS), denoted ω, which is a measure of selective pressure. Values of $\omega < 1$, $= 1$, and > 1 are indicators of purifying selection, neutral evolution and positive selection, respectively. Two distinct categories of codon substitution models are used: site models (M1a vs. M2a, M7 vs. M8 and M8a vs. M8) and branch-site models. For the two types of analyses, two models are compared: one model which allows positive selection and one model which does not allow positive selection. For each model, the lnL (log likelihood) value is retrieved (lnL1 for the model allowing positive selection, lnL0 for the other) and a LRT (Likelihood Ratio Test) is calculated (LRT= 2 x (lnL1-lnL0)) to assess the significance of the results. The LRT value follows a χ^2 law which allows the p-value of the LRT to be obtained. If the LRT is significant for the comparison, PhyleasProg lists sites under positive selection detected by Bayes Empirical Bayes (BEB) with posterior probabilities greater than 95% and sites under purifying selection.

As shown in Figure 2, selection pressure data appear in two separate menus. One is dedicated to results of site models and the other one to results of branch-site models. In the second case, these models allow the ω ratio to vary both among sites in the protein and across branches on the tree and aim to detect positive selection affecting a few sites along particular lineages (foreground branches). In the pipeline, all branches of the tree are tested as foreground branches for positive selection. Two models are used, one called Alternative and one called Null. In the alternative model, three classes of sites are admitted for the foreground branch, ω0: $dN/dS < 1$, ω1: $dN/dS = 1$ and ω2: $dN/dS \geq 1$. In the Null model, ω2 is fixed to 1. Significant results with branch-site models are accessible on a clickable tree. Branches under positive selection are represented by a purple star and are highlighted in green. Raw result files (rst) of CODEML are also available.

3.3.2 Visualization

Results of selection pressure calculation with site and branch-site models share the same presentation (Figure 2). Data are visualized on 1D and 3D structures on the same results page. A dropdown menu embedded in the positive selection results web page enables users to visualize data on each protein in the orthologue or paralogue dataset. For the two types of representations, a discrete color scale is used to distinguish the different values of ω for each site. The scale from green to yellow represents purifying selection, i.e. $\omega < 0.3$, while red and orange represent positive selection with posterior probabilities greater than 99% or 95%, respectively. White means that no information is available for this site because no

calculation was performed by CODEML due to at least one gap in the multiple sequence alignment at this position. Grey means results are not significant enough to infer either purifying or positive selection. To locate different amino acids in different organisms, the multiple sequence alignment used for PAML computation is displayed using the JalView applet.

These data can greatly help biologists to plan site-directed mutagenesis experiments to target essential functional residues. This was the main reason to have PhyleasProg display results on a 3D structure, if one can be modelled (Figure 1E). To model the 3D structure, a BLAST [28] search is performed to find a similar structure in the PDB database [29] in order to use it as a template to calculate a model with Modeller. Three-dimensional structure is sometimes difficult to predict, mostly when the template is too distant from the sequence to be modelled. To avoid models of insufficient quality, a model is built only if: (i) the alignment between the sequence to be modelled and the length of the PDB template covers at least 80% of PDB sequence and at least 50% of the query sequence and, (ii) the percentage of identity between the two sequences is at least 50%. If the query sequence is shorter than the template, amino acids in the C- or N-terminal are removed. In order to enable users to locate differences between a raw query sequence and the model, the alignment between the PDB sequence and raw query sequence is displayed using JalView. Hence, when a homology model can be built, evolutionary results are directly visualized on the modelled structure, while if homology modelling is not possible, results are only presented on the 1D sequence.

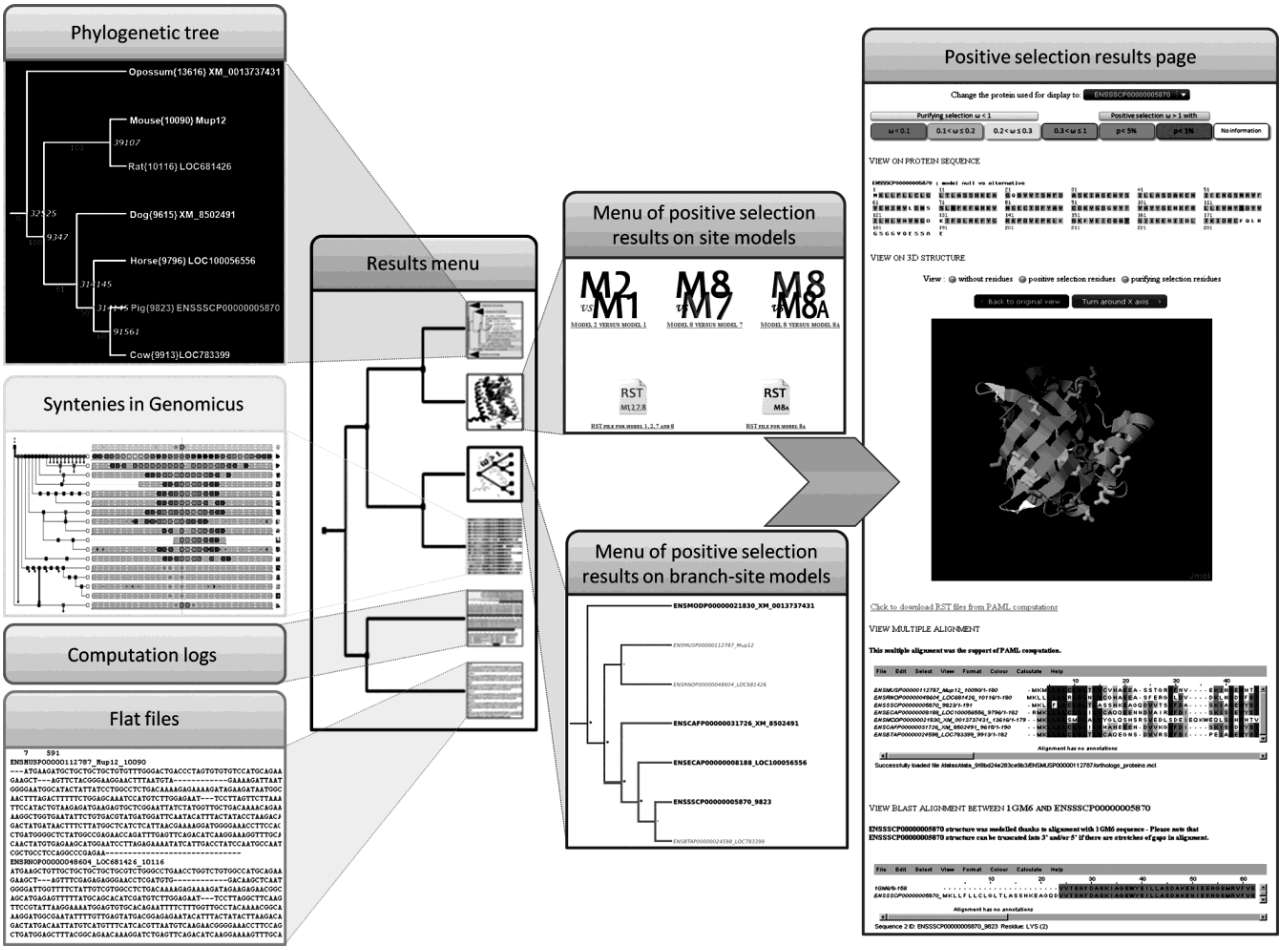


Figure 2. Overview of the results menu of PhyleasProg and its results pages.

3.4 Synteny Exploration

In order to achieve complete evolutionary analysis of the protein submitted, PhyleasProg offers the possibility to explore the genetic environment of related genes. Indeed, in the results menu (Figure 2) the user has a link to Genomicus [30]. This database is a synteny browser that can represent and compare

numerous genomes in a broad phylogenetic view. In addition, Genomicus includes the reconstructed organization of ancestral gene, thus greatly facilitating interpretation of the data. We chose not to develop our own genome browser because this web tool is really accurate, complete, up-to-date, user-oriented and also based on Ensembl data.

4 Conclusion and Future Developments

With PhyleasProg, we offer biologists a tool specially developed for non-specialists of phylogenetics, which is user-oriented, fast, complete, up-to-date, ready-to-use and accessible via a web interface, and allows the user to submit several jobs at the same time. All computations are dynamically produced and displayed as soon as the results are available, so the user can begin to analyze results without waiting for the whole process to end.

Thanks to the modular architecture of our pipeline, it is relatively easy to update and to incorporate new tools. In the short term, our main plan is to extend the range of possible inputs. With the present system, only proteins from organisms available in Ensembl can be treated in PhyleasProg. A FASTA sequence as input, for example, could be useful. We also want to let users upload their own PDB files. In the very near future, we will offer a 3D structure model based on a multiple alignment including several proteins from the PDB database, which would improve the quality of the models. Finally, to provide more accurate pressure selection data, we are already thinking about a way to minimize the GC bias in positive selection results.

Acknowledgements

The project was hosted by the Toulouse Midi-Pyrénées bioinformatics platform. Sincere thanks to Didier Laborie and Sylvain Thomas for their technical support. We thank Anne Poupon for sharing expertise on Modeller, for critical reading and suggestions on the manuscript. Particular thanks go to Delphine Capela for critical reading and suggestions on the manuscript. We would like to thank Philippe Monget for his supportive discussions on project. We are grateful to Alexis Dereeper, Ziheng Yang and Li Heng for helpful discussions on clickable tree, Codeml and Treebest respectively. We are grateful to Daphne Goodfellow for attention to the English-language version. MENRT PhD fellowship (to C.M.).

References

[1] K. Liolios, I.M. Chen, K. Mavromatis, N. Tavernarakis, P. Hugenholtz, V.M. Markowitz and N.C. Kyrpides, The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic acids research*, 38:D346-354, 2010.

[2] A. Eyre-Walker, The genomic rate of adaptive evolution. *Trends in ecology & evolution*, 21:569-575, 2006.

[3] D. Graur and W.H. Li *Fundamentals of molecular evolution*. Sinauer Associates, Sunderland, MA, USA, 2000.

[4] R.A. Studer, S. Penel, L. Duret and M. Robinson-Rechavi, Pervasive positive selection on duplicated and nonduplicated vertebrate protein coding genes. *Genome research*, 18:1393-1402, 2008.

[5] Z. Yang, PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24:1586-1591, 2007.

[6] J.G. Glanville, D. Kirshner, N. Krishnamurthy and K. Sjolander, Berkeley Phylogenomics Group web servers: resources for structural phylogenomic analysis. *Nucleic acids research*, 35:W27-32, 2007.

[7] C.Y. Lin, F.K. Lin, C.H. Lin, L.W. Lai, H.J. Hsu, S.H. Chen and C.A. Hsiung, POWER: PhylOgenetic WEb Repeater--an integrated and user-optimized framework for biomolecular phylogenetic analysis. *Nucleic acids research*, 33:W553-556, 2005.

[8] B. Neron, H. Menager, C. Maufrais, N. Joly, J. Maupetit, S. Letort, S. Carrere, P. Tuffery and C. Letondal, Mobyle: a new full web bioinformatics framework. *Bioinformatics*, 25:3005-3011, 2009.

[9] A. Dereeper, V. Guignon, G. Blanc, S. Audic, S. Buffet, F. Chevenet, J.F. Dufayard, S. Guindon, V. Lefort, M. Lescot, J.M. Claverie and O. Gascuel, Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic acids research*, 36:W465-469, 2008.

[10] J. Huerta-Cepas, S. Capella-Gutierrez, L.P. Pryszcz, I. Denisov, D. Kormes, M. Marcet-Houben and T. Gabaldon, PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. *Nucleic acids research*, 39:D556-560, 2010.

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

[12] A. Stern, A. Doron-Faigenboim, E. Erez, E. Martz, E. Bacharach and T. Pupko, Selecton 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach. *Nucleic acids research*, 35:W506-511, 2007.
[13] W. Delport, A.F. Poon, S.D. Frost and S.L. Kosakovsky Pond, Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics*, 26:2455-2457, 2010.
[14] E. Proux, R.A. Studer, S. Moretti and M. Robinson-Rechavi, Selectome: a database of positive selection. *Nucleic acids research*, 37:D404-407, 2009.
[15] J. Tarraga, I. Medina, L. Arbiza, J. Huerta-Cepas, T. Gabaldon, J. Dopazo and H. Dopazo, Phylemon: a suite of web tools for molecular evolution, phylogenetics and phylogenomics. *Nucleic acids research*, 35:W38-42, 2007.
[16] P. Flicek, M.R. Amode, D. Barrell, K. Beal, S. Brent, Y. Chen, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gordon, M. Hendrix, T. Hourlier, N. Johnson, A. Kahari, D. Keefe, S. Keenan, R. Kinsella, F. Kokocinski, E. Kulesha, P. Larsson, I. Longden, W. McLaren, B. Overduin, B. Pritchard, H.S. Riat, D. Rios, G.R. Ritchie, M. Ruffier, M. Schuster, D. Sobral, G. Spudich, Y.A. Tang, S. Trevanion, J. Vandrovcova, A.J. Vilella, S. White, S.P. Wilder, A. Zadissa, J. Zamora, B.L. Aken, E. Birney, F. Cunningham, I. Dunham, R. Durbin, X.M. Fernandez-Suarez, J. Herrero, T.J. Hubbard, A. Parker, G. Proctor, J. Vogel and S.M. Searle, Ensembl 2011. *Nucleic acids research*, 39:D800-806, 2010.
[17] R.C. Edgar, MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5:113, 2004.
[18] G. Talavera and J. Castresana, Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology*, 56:564-577, 2007.
[19] N. Eswar, D. Eramian, B. Webb, M.Y. Shen and A. Sali *Protein structure modeling with MODELLER*. Humana Press, City, 2008.
[20] A.M. Waterhouse, J.B. Procter, D.M.A. Martin, M. Clamp and G.J. Barton, Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, 25:1189-1191, 2009.
[21] M.V. Han and C.M. Zmasek, phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics*, 10:356, 2009.
[22] A. Herráez, Biomolecules in the computer: Jmol to the rescue. *Biochemistry and Molecular Biology Education*, 34:255-261, 2006.
[23] A. Stabenau, G. McVicker, C. Melsopp, G. Proctor, M. Clamp and E. Birney, The Ensembl core software libraries. *Genome research*, 14:929-933, 2004.
[24] A.J. Vilella, J. Severin, A. Ureta-Vidal, L. Heng, R. Durbin and E. Birney, EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome research*, 19:327-335, 2009.
[25] M. Suyama, D. Torrents and P. Bork, PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic acids research*, 34:W609-612, 2006.
[26] S. Guindon, J.F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk and O. Gascuel, New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology*, 59:307-321, 2010.
[27] C.M. Zmasek and S.R. Eddy, ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics*, 17:383-384, 2001.
[28] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, 25:3389-3402, 1997.
[29] P.W. Rose, B. Beran, C. Bi, W.F. Bluhm, D. Dimitropoulos, D.S. Goodsell, A. Prlic, M. Quesada, G.B. Quinn, J.D. Westbrook, J. Young, B. Yukich, C. Zardecki, H.M. Berman and P.E. Bourne, The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic acids research*, 39:D392-401, 2011.
[30] M. Muffato, A. Louis, C.E. Poisnel and H. Roest Crollius, Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes. *Bioinformatics*, 26:1119-1121, 2010.

Investment in Growth Determines the Evolutionary Rates of Proteins

Sara VIEIRA-SILVA[1] and Eduardo ROCHA[1]

[1] MICROBIAL EVOLUTIONARY GENOMICS, Institut Pasteur, CNRS, URA2171, F-75724, Paris, France
{ssilva, erocha}@pasteur.fr

Keywords Growth, expressivity, evolutionary rate heterogeneity, phylogenetic congruence.

1 Introduction

The molecular clock hypothesis, which states that the rate of molecular evolution is constant in time and across lineages, is rarely observed in molecular data. Heterogeneity in protein evolutionary rates along lineages may result from many factors affecting the balance of mutation, genetic drift and selection. For example, both the increase in mutation rates or the decrease in effective population sizes, (e.g. in maternally inherited endosymbiotic bacteria [1]) result in higher evolutionary rates in lineages. Factors affecting the mutation/selection balance will often result in a genome-wide shift in evolutionary rate. On the other hand, events such as mutational hotspots or diversifying selection acting on specific genes (e.g antigenic proteins [2]) create local rate heterogeneity across lineages.

Besides these lineage-specific rate variations, different proteins systematically evolve at drastically different rates in every lineage. Many determinants of protein evolutionary rates have been proposed, including protein dispensability [3], the number of protein interactions [4] and the level of expression. The latter is the only ubiquitous undisputed determinant of protein evolutionary rates, where highly expressed proteins evolve more slowly than lowly expressed ones. This trend has been found in many diverse species, including bacteria (e.g. *E. coli* and *B. subtilis*), yeast and mammals [5,6].

Ultimately, one of the major goals of evolutionary biology is to link evolutionary rate variations to species physiology, ecology or life-history traits [7]. We hypothesized that minimum generation time, a key life-history trait, may cause protein evolutionary rate heterogeneity. Minimum generation times reflect a gradient of ecological strategies: while fast-growers (copiotrophs) quickly proliferate in high nutrient conditions, slow-growers (oligotrophs) are better competitors in low-nutrient environments [8]. During periods of fast growth, bacterial cells adopt a particular physiology that is highly dedicated to protein translation [9]. The associated selective pressure consistently imprints genome of fast-growers [10]. As a result, the relative weight of highly expressed proteins in the cellular fitness depends on growth rate. Therefore, we propose that the selective pressure associated to gene expression levels may be stronger in organisms subjected to periods of fast growth. That is to say that we expect changes in minimum generation time to impact the evolutionary rate of highly expressed proteins but not, or much less so, that of lowly expressed proteins across lineages.

2 Results

We identified by best bidirectional hit the 61 families of orthologs shared by 74 different proteobacterial species with diverging minimum generation times. We found that indeed the difference between the rates of evolution of highly and lowly expressed essential proteins (i.e. among-protein rate heterogeneity) is greater for rapidly dividing organisms. That is to say that the evolutionary rates of highly expressed proteins decrease relative to that of lower expressed proteins when minimal generation times decrease. We observe that for all pairs of proteins with at least a 5-fold difference in expressivity, the distribution of the correlations between the among-protein rate heterogeneity and the minimum generation times is significantly different from the random expectation. These results were also confirmed by comparing the concatenation of the protein constituents of the ribosome and DNA polymerase. These are well-described essential protein complexes with strikingly different expression levels. These results support our general hypothesis that fast-

growers exhibit stronger purifying selection on highly expressed genes in comparison to lowly expressed genes.

Under the premise that among fast growers the most highly expressed proteins do evolve slower relative to other essential lowly expressed proteins, we also tested that the topology of deep-phylogenetic inferences could be influenced by the choice in phylogenetic marker. To this purpose, we performed phylogenetic reconstructions for 316 bacteria and archaea based on markers with different expression levels. The robustness of the results was evaluated by performing 500 jacknives on each marker set. The results show that clades with a majority of slow-growing representatives, branch deeper in a phylogenetic reconstruction based on highly expressed proteins than in one based on lowly expressed proteins. This is consistent with a effect of minimum generation time creating systematically longer branch lengths in slow-growers and therefore affecting the final topology of trees based on highly expressed proteins. Therefore, the results suggest that minimum generation times and the particular physiology associated to periods of fast growth can severely influence the evolutionary patterns of essential proteins along time and lineages.

3 Conclusions

Cells in exponential growth are mostly devoted to transcription and translation, which are themselves dedicated to the few percent most highly expressed genes. Therefore, the relative cost of translation of highly expressed proteins is increased in cells experiencing high growth rates. As a result, highly expressed proteins evolve slower in fast-growers relative to the average protein in the proteome of fast-growers, and relative to their orthologous counterpart in slow growers. The latter results in a systematic heterogeneity in the evolutionary rates of highly expressed genes across lineages according to their growth capacity. This evolutionary rate heterogeneity leads to topological differences in deep-phylogenetic reconstruction based on highly expressed markers compared to those based on lowly expressed markers. Incidentally, the former correspond to the core proteins involved in informational processes most typically used to reconstruct the evolutionary history of distantly related taxa.

References

[1] N.A. Moran, C.D. Dohlen and P. Baumann, Faster evolutionary rates in endosymbiotic bacteria than in cospeciating insect hosts. *J. Mol. Evol.*, 41: 727--731, 1995.

[2] F.M. Jiggins, G.D. Hurst and Z. Yang, Host-symbiont conflicts: positive selection on an outer membrane protein of parasitic but not mutualistic Rickettsiaceae. *Mol. Biol. Evol.*, 19: 1341-1349, 2002.

[3] A.E. Hirsh and H.B. Fraser, Protein dispensability and rate of evolution. *Nature*, 411: 1046-1049, 2001.

[4] H.B. Fraser, A.E. Hirsh, L.M. Steinmetz, C. Scharfe and M.W. Feldman, Evolutionary rate in the protein interaction network. *Science*, 296: 750-752, 2002.

[5] L. Duret and D. Mouchiroud, Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol. Biol. Evol.*, 17: 68-74, 2000.

[6] E.P. Rocha, The quest for the universals of protein evolution. *Trends Genet.*, 22: 412-416, 2006.

[7] N. Lartillot and R. Poujol, A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Mol. Biol. Evol.*, 28: 729-744, 2011.

[8] A.L. Koch, Oligotrophs versus copiotrophs. *Bioessays*, 23: 657-661, 2001.

[9] H. Bremer and P.P. Dennis (1996) Modulation of cell parameters by growth rate. *Escherichia coli and Salmonella: cellular and molecular biology.* Washington DC: ASM Press. pp. 1553-1569.

[10] S. Vieira-Silva and E.P.C. Rocha, The systemic imprint of growth and its uses in ecological (meta)genomics. *PLoS Genet.*, 6: e1000808, 2010.

Horizontal Gene Transfer of a Chloroplast Protein to Thaumarchaeota: The Unique Case of a Ferredoxin and a J-domain Fusion

Céline PETITJEAN[1], Céline BROCHIER-ARMANET [1], Purificación LÓPEZ-GARCIA[2] and David MOREIRA[2]

[1] LCB, UPR9043 CNRS IFR88 LCB, 31 chemin Joseph Aiguier, 13402, Marseille, Cedex 20, France
{celine.petitjean, celine.brochier}@ifr88.cnrs-mrs.fr
[2] ESE, UMR8079 Université Paris-sud, Bâtiment 360, 91405, Orsay, Cedex, France
{puri.lopez, david.moreira}@u-psud.fr

Abstract *An unusual protein composed of a DnaJ domain and a ferredoxin domain has been found to be present only in Viridiplantae and Thaumarchaeota. To understand this unexpected repartition, we have carried out a phylogenetic analysis of this protein, and highlighted a transfer from plants to the Thaumarchaeota. We have also studied the evolution of the archaeal chaperone DnaK and its cochaperones, which interact with DnaJ domain containing proteins. Our results suggest a more complex evolutionary history than previously thought, involving multiple horizontal gene transfers from diverse donors. This has implications for our knowledge of the adaptation to mesophilic lifestyle in Archaea.*

Keywords DnaJ, Archaea, Horizontal Gene Transfer, DnaK.

In 2004, the study of a genome fragment from an uncultured thaumarchaeon highlighted an unusual protein composed of a N-ter DnaJ domain fused with a C-ter ferredoxin (Fer) domain [1]. A first analysis showed homologues of this DnaJ/Fer protein only in Thaumarchaeota and Viridipantae (green algae and land plants), a surprising taxonomic repartition. In fact, Thaumarchaeota are a widespread lineage of Archaea, formerly classified as mesophilic crenarchaeota group I but recently proposed to be a third phylum within the Archaea, together with the Euryarchaeota and Crenarchaeota [2]. Thaumarchaeota are therefore extremely distant from the Viridiplantae. Moreover, the plant homologues contain a N-terminal chloroplast-targeting signal region. This has been studied in detail in the green algae *Chlamydomonas reinhardtii*, showing the chloroplast localization of two of the three homologues found in this species. In addition, they have been shown to interact with the chloroplast Hsp70B (Heat Shock Protein 70 also called DnaK) [3]. Two hypotheses can be proposed to explain this particular taxonomic distribution of such a rare association between DnaJ and Fer domains: it results either of two independent fusions in Thaumarchaeota and in Viridiplantae or of a single fusion event followed by a horizontal gene transfer (HGT) between these two distant groups. To clarify its evolutionary history we have carried out a phylogenomic analysis of this protein, including the study of each domain independently.

We have also studied the evolutionary histories of the chaperone DnaK and its co-chaperone GrpE, because DnaK interacts specifically with DnaJ domains. In fact, DnaK recognizes unfolded or misfolded proteins through the intermediate of DnaJ-domain containing proteins. From an evolutionary point of view, DnaK and GrpE homologues are present in bacteria, eukaryotes and some archaea, and genes coding for DnaK are often located close to *grpE* and *dnaJ* (a protein different from the DnaJ/Fer protein) in many bacterial genomes. Previous studies have proposed that the archaeal DnaK had been acquired by HGT from bacteria [4,5].

The phylogenomic analysis of the DnaJ/Fer protein strongly supported the sistership of the Viridiplantae and Thaumarchaeota homologues indicating that the hypothesis of an HGT between these two lineages was the most parsimonious explanation. More precisely, our analyses suggested that the HGT occurred from Viridiplantae to Thaumarchaeota. The absence of any DnaJ/Fer homologue in the recently published genome sequence of '*Candidatus* Caldiarchaeum subterraneum' (representative of a deep branching archaeal lineage related to Thaumarchaeota) suggested that Thaumarchaeaota have acquired the DnaJ/Fer protein after the divergence of '*Candidatus* C. subterraneum'.

The evolutionary study of the three proteins DnaK and GrpE and DnaJ is particularly interesting in the context of the adaptation to mesophily in Archaea. In fact, their taxonomic repartition is clearly correlated to a non-hyperthermophilic lifestyle, as they are encoded only in (even if not in all) the genomes of mesophilic or thermophilic archaea, but not in hyperthermophilic species. Although not always in the same order, the three genes coding for these proteins are next to each other in the genomes of Archaea, except for some Halobacteriales. The phylogeny of each of these three proteins showed a very complex history, including at least two HGTs from bacteria to archaea (one to Halobacteriales and the other likely to Methanomicrobia), followed by many HGT among different archaeal lineages, including Thaumarchaeota. In fact, the latter appear to have acquired these three genes from euryarchaeota, as suggested by their robust placement among the euryarchaeotal sequences, although the limited resolution of our trees did not allow identifying precisely the euryarchaeal donor.

Until now, the presence of DnaK in Archaea has been proposed to be linked to the adaptation to mesophily. Our results suggested that this adaptation may have been more complex than previously thought, probably including multiple adaptations in different archaeal groups. In conclusion, we have highlighted the complex evolutionary history of a set of interacting chaperone proteins in Archaea, and more precisely in Thaumarchaeota. All of them appeared to have been acquired by independent HGTs from diverse donors, including Viridiplantae (in the case of the DnaJ/Fer protein), bacteria and other archaea (in the case DnaK, GrpE and DnaJ).

Acknowledgements

This work was supported by the ANR EvolDeep.

References

[1] P. López-García, C. Brochier, D. Moreira and F. Rodríguez-Valera, Comparative analysis of a genome fragment of an uncultivated mesopelagic crenarchaeote reveals multiple horizontal gene transfers. *Environ Microbiol.*, Jan;6(1):19-34, 2004.

[2] C. Brochier-Armanet, B. Boussau, S. Gribaldo and P. Forterre, Mesophilic crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. *Nature Reviews Microbiology.*, 6:245-252, 2008.

[3] KV. Dorn, F. Willmund, C. Schwarz, C. Henselmann, T. Pohl, B. Hess, D. Veyel, B. Usadel, T. Friedrich, J. Nickelsen and M. Schroda, Chloroplast DnaJ-like proteins 3 and 4 (CDJ3/4) from Chlamydomonas reinhardtii contain redox-active Fe-S clusters and interact with stromal HSP70B. *Biochem J.*, Mar 29;427(2):205-15, 2010.

[4] S. Gribaldo, V. Lumia, R. Creti, EC. de Macario, A. Sanangelantoni, and P. Cammarano, Discontinuous Occurrence of the hsp70 (dnaK) Gene among Archaea and Sequence Features of HSP70 Suggest a Novel Outlook on Phylogenies Inferred from This Protein. *J Bacteriol.*, January; 181(2): 434–443, 1999.

[5] AJ. Macario, L. Brocchieri, AR. Shenoy and E. Conway de Macario, Evolution of a Protein-Folding Machine: Genomic and Evolutionary Analyses Reveal Three Lineages of the Archaeal hsp70(dnaK) Gene. *J Mol Evol.*, Jul;63(1):74-86, 2006.

Tpms: a Tree Pattern-matching Utility for Querying Gene Trees Collections

Thomas BIGOT, Vincent DAUBIN and Guy PERRIÈRE

Laboratoire de Biométrie et Biologie Évolutive, UMR 5558 CNRS, Université Claude Bernard – Lyon 1, 43 bd. du 11 Novembre 1918, 69622, Villeurbanne Cedex, France

{thomas.bigot,vincent.daubin,guy.perriere}@univ-lyon1.fr

Abstract *Tpms is a portable C++ program allowing to retrieve gene trees from large collections, this according to tree patterns defined by the users. It can be used for different purposes such as orthologs search or horizontal gene transfers identification. Documentation, source code, as well as Linux and MacOSX binaries can be freely downloaded at ftp://pbil.univ-lyon1.fr/pub/mol_phylogeny/tpms/.*

Keywords phylogenetic trees, tree pattern-matching, orthologs, horizontal gene transfers.

1 Introduction

Comparative genomics is a common approach in sequence analysis, and many biological results have been obtained through its use. Among the different programs and packages developed for comparative genomics, those using the information contained in phylogenetic trees are of special interest. Indeed, orthology detection methods using phylogenetic trees usually perform better than the simpler (and easier to use) methods based on best reciprocal hits of sequence similarity scores [1]. In that context we developed tpms, a program allowing to retrieve gene trees from a tree collection, this according to patterns defined by the users. Those patterns usually include some kind of constraints, such as node nature (duplication, speciation), or subtree content. Therefore this program can be used for orthologs search, but also for any studies that require to retrieve sets of gene families matching constraints in their corresponding phylogenetic tree (*e.g.*, gene duplications identification or horizontal gene transfers prediction).

2 System and methods

The tree pattern-matching algorithm used in tpms is a C++ version of the one from the RAP program implemented in FamFetch [2]. It requires the Bio++ [3, 4] and Boost [5] libraries to be run. This new implementation consists in a command-line standalone binary and is not embedded into a graphical interface. Moreover, it is also no longer dependant on the use of the HOVERGEN, HOGENOM and HOMOLENS gene families databases [6], and it can be used on collections build by the users. Binaries of the program are provided for Linux and MacOSX (Intel architectures only), as well as the source code at ftp://pbil.univ-lyon1.fr/pub/mol_phylogeny/tpms/.

3 Program use

At first, the user needs to build a gene trees collection in the RAP format [7]. This collection will be then accessed by the program when performing pattern searches. Collection construction can be done easily through the use of tpms_mkdb program, distributed with tpms. This tool uses a reference species tree, also in RAP format, and a set of individual gene trees in standard Newick format. The species tree can contain unresolved nodes (multifurcations), but not the individual gene trees.

Tree patterns have to be written in an extended Newick format. The simplest constraints that can be introduced are represented by the taxa found on the leaves of the pattern. The labels can stand for a given species (*e.g.*, *Homo sapiens*) or larger taxonomic groups (*e.g.*, Primates). For instance, the pattern:

```
((Homo sapiens,Pan troglodytes),Rodentia)
```

will allow to find all the gene trees in which a subtree with sequences from *H. sapiens* and *Pan troglodytes* species are grouped, while sequences from any rodents are located outside of this group.

The programs also allows to specify constraints on subtrees: one can ask that a subtree from the entered pattern will only contains genes from a defined set of species. This set can be one species, one taxon, or a list with a combination of species and taxa. This makes it possible to exclude a subgroup of a larger group. Elements of the list are bracketed, and indication of addition or removal is done through + or − operators. For example, the pattern:

```
((Homo sapiens,Pan troglodytes),Mammalia{-Primates})
```

will allow to find all the gene trees in which a subtree with sequences from *H. sapiens* and *P. troglodytes* species are grouped, while any sequences from mammals – excluding primates – is located outside this group.

A third kind of constraint can be set on nodes if the program is running on a reconciled trees collection: it is possible to search specifically for speciation or duplication nodes. This kind of nodes can be specified in the pattern by the use of letters S or D. In the following example:

```
(Homo sapiens,Mus musculus){D}
```

will allow to find all the gene trees in which a subtree with sequences from *H. sapiens* and *M. musculu*s species are grouped, while the node that groups them is a duplication node.

In order to search for orthologs, the program can perform queries in the gene trees collection with a pattern extracted from the reference species tree. Orthologs can be then identified from the subparts in the gene trees that match that pattern. Another possibility is to search for all gene trees in which there is a subtree containing a list of taxa defined by the user, this whatever the topology of the subtree is, and giving the fact that all taxa are present in single copy. This later approach is suited only for the detection of 1:1 orthology relationships, but it is extremely fast and can be used on very large collections containing thousands of trees.

Search for horizontal gene transfers is also straightforward, as it will only require to enter anomalous patterns, this relatively to the reference species tree.

References

[1] T. Gabaldón, Large-scale assignment of orthology: back to phylogenetics? *Genome Biol.*, 9:235, 2008.
[2] J.F. Dufayard, L. Duret, S. Penel, M. Gouy, F. Rechenmann and G. Perrière, Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics*, 21:2596-2603, 2005.
[3] J. Dutheil, S. Gaillard, E. Bazin, S. Glemin, V. Ranwez, N. Galtier and K. Belkhir, Bio++: A set of C++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics. *BMC Bioinformatics*, 7:188, 2006.
[4] http://biopp.univ-montp2.fr/
[5] http://www.boost.org/
[6] S. Penel, A.M. Arigon, J.F. Dufayard, A.S. Sertier, V. Daubin, L. Duret, M. Gouy and G. Perrière, Databases of homologous gene families for comparative genomics. *BMC Bioinformatics*, 10(S6):S3, 2009.
[7] http://pbil.univ-lyon1.fr/software/RAP/RAP.htm

Session 6.B : Annotation

Combined Approach for Transposable Elements Detection

Olivier Inizan[1], Véronique Jamilloux[1], Sandie Arnoux[1], Thimothée Flutre[2] and Hadi Quesneville[1]

[1] Unité de Recheche en Génomique-Info, UR 1164 INRA, route de Saint Cyr, Centre de Versailles-Grignon, 78026, Versailles, France

[2] University of Chicago, Department of Human Genetics, Chicago, IL 60605, USA

olivier.inizan@versailles.inra.fr

Abstract *Transposable elements (TEs) play a major role in genome evolution and their dynamics are particular. Although many algorithms are designed for repeats detection, but none is able to take into account all TEs dynamics specificities. In this study, we present a combined approach for TEs detection. We evaluated three programs based on different algorithms: GROUPER, RECON and PILER. Our results show that a combined approach provide a better TEs detection than using a program alone.*

Keywords transposable elements, detection, annotation.

Une Approche Combinée pour la Détection d'Eléments Transposables

Résumé *Les éléments transposables (ETs) jouent un rôle important dans l'évolution des génomes. Leur dynamique d'évolution présente des caractéristiques particulières. Comme aucun algorithme ne couvre l'ensemble de ces caractéristiques, nous présentons une approche combinée pour la détection d' ETs. Nous avons évalués 3 programmes basés sur différents algorithmes : GROUPER, RECON et PILER. Nos résultats montrent que cette approche est plus efficace pour la détection d'ETs.*

Mots-clés éléments transposables, détection, annotation.

1 Introduction

Les éléments transposables (ET) sont des séquences d'ADN mobiles présentes dans pratiquement tous les génomes eukaryotes. Longtemps considérés comme des parasites, les ETs sont maintenant reconnus comme jouant un rôle majeur au cours de l'évolution. Ils ont un impact sur la structure et l'organisation des génomes [1]. Nous développons une suite d'outils permettant d'identifier, caractériser et annoter les ET dans un génome. Pour cela, nos outils reposent sur une compréhension fine de leur dynamique évolutive dans les génomes. La dynamique d'évolution des éléments transposables est très particulière. Leur mobilité et leur nature répétée dans les génomes en font des objets biologiques évoluant de façon très spécifique. La compréhension de celle-ci est au coeur d'une annotation efficace.

2 Un Modèle de Dynamique des ET dans les génomes

Schématiquement, la dynamique d'évolution des ET comprend une phase d'amplification et de dégénérescence. La phase d'amplification suit un transfert horizontal ou apparaît en réaction à une mutation ou un stress. L'ET envahit alors rapidement le génome en se multipliant. On parle souvent de « burst » de transpositions. Suit alors une phase de dégénérescence où chacune des copies créées lors de la première phase évolue au cours du temps par des événements d'insertion, de délétion, de substitution et de recombinaison. Pendant cette phase, il n'est pas rare que ces copies subissent l'insertion d'une autre famille d'ET qui est alors en phase du «burst ». Il en résulte des copies dégénérées et d'âges d'insertion variables dans le génome. On ne retrouve alors que des copies plus ou moins fragmentées traces des éléments ancestraux.

3 Intérêt d'une Approche Combinée

Cette dynamique spécifique des éléments transposables n'est prise en compte que partiellement par les algorithmes actuels de détection d'ET.

L'algorithme de GROUPER [2,3] procède en deux étapes: (i) Les copies fragmentées issues d'alignements

2 à 2 sont connectées par programmation dynamique. (ii) Les chaînes obtenues sont ensuite clusterisées par un algorithme simple lien avec contrainte de recouvrement à 95%. Ainsi ne clusterisent que les copies ayant des tailles identiques, c'est à dire celles supposées être peu dégénérées. On obtient alors des copies proches des copies fonctionnelles, ce qui en fait une méthode très spécifique. RECON [4] repose aussi sur du clustering simple lien, mais exploite des informations issues d'alignements multiples afin d'identifier (i) les extrémités des copies et (ii) les familles d'éléments homologues. L'approche est plus sensible que celle de GROUPER mais perd en spécificité. PILER [5] exploite des profils d'alignements locaux (les « piles ») propres à certaines répétitions. L'approche est identique dans le principe à celle de GROUPER: ne regrouper que les copies ayant des tailles identiques, c'est à dire celles supposées être peu dégénérées. Cependant PILER ne réalise pas de connections des fragments comme GROUPER, il apparaît donc moins sensible que ce dernier, ne prenant pas en compte cette caractéristique de leur dynamique.

4 Résultats

Nous avons combiné ces trois programmes afin de tester la détection des ET sur les génomes de *D.melanogaster* et *A.thaliana.* Comme les ET de ces espèces sont connus par ailleurs, il est possible d'évaluer les performances de cette approche par la mise au point de mesures de sensibilité (Sn*) et de spécificité (Sp*) ainsi que par le taux récupération de copies complètes (Rcc), [3] et Table1.

Génome	Programmes	Sn*	Sp*	Rcc
D.melanogaster	GROUPER	80.34%	85.89%	66.20%
D.melanogaster	RECON	92.31%	73.17%	66.20%
D.melanogaster	GROUPER + RECON	93.16%	81.03%	79.40%

Table 1. Résultats partiels de l'approche combinée.

Nous avons aussi comparé les séquences retrouvées par GROUPER à celles retrouvées par RECON (données non fournies, cf [3]): les premières sont proches en taille et en identité alors que les secondes sont plus hétérogènes. Cela nous mène à penser que chaque algorithme puisse être associé à une phase particulière de la dynamique des ET: GROUPER identifie des évolutions récentes de type « burst » alors que RECON met en évidence des évolutions plus anciennes.

5 Conclusion

Nous avons implémenté cette approche combiné dans un pipeline le « TEdenovo ». Celui-ci est intégré au package REPET (http://urgi.versailles.inra.fr/Tools/REPET) de détection et d'annotation automatique de génomes. Plusieurs génomes eukaryotes ont été annotés avec ce package.

Références

[1] M. G. Kidwell and D. R. Lisch, Perspective: Transposable elements, parasitic dna and genome evolution. *International journal of organic evolution*, 55:1-24, 2001.

[2] H. Quesneville, C. M. Bergman, O. Andrieu, D. Autard, D. Nouaud, M. Ashburner and D. Anxolabehere, Combined Evidence Annotation of Transposable Elements in Genome Sequences. *PLOS Comput Biol,* 1:166-175, 2005.

[3] T. Flutre, E. Duprat, C. Feuillet and H. Quesneville, Considering Transposable Element Diversification in De Novo Annotation Approaches. *PLOS One*, 6:e16526-e16526, 2011.

[4] Z. Bao and S. R. Eddy, Automated de novo identification of repeat sequence families in sequenced genome. *Genome Res*. 12:1269-1276, 2002.R.

[5] C. Edgar and E. W. Meyers, PILER: identification and classification of genomics repeats. *Bioinformatics* 21:i152-i158, 2005.

Integrated Gene Prediction for Prokaryotic Genomes using EuGene

Erika SALLET[1], Jérôme GOUZY[1], Brice ROUX[1], Delphine CAPELA[1], Laurent SAUVIAC[1], Claude BRUAND[1], Pascal GAMAS[1] and Thomas SCHIEX[2]

[1] Laboratoire Interactions Plantes Micro-organismes (LIPM) UMR441/2594, INRA/CNRS
{Erika.Sallet,Jerome.Gouzy}@toulouse.inra.fr
[2] Unité de Biométrie et d'Intelligence Artificielle UR 875, INRA, F-31320 Castanet Tolosan, France
Thomas.Schiex@toulouse.inra.fr

Abstract *With the advent of new generation sequencing, the annotation of new prokaryotic genomic sequences will occur in a data-rich context, including a variety of libraries of short reads of transcriptomic sequences. This rich context creates new potentialities in annotation. In this paper, we describe the new prokaryotic variant of the integrative gene prediction software EuGene. By leveraging RNA-Seq data, EuGene becomes capable of predicting new functional structures, including RNA genes and untranslated transcribed regions inside operons.*

Keywords Gene prediction, RNA gene, NGS, RNA-Seq.

1 Introduction

Following the initial development of gene prediction tools for prokaryotic genomes, the complexity of eukaryotic gene prediction led to the development of highly integrative gene prediction tools. Very few, if any, prokaryotic gene prediction tools have evolved along the same line, mostly because prokaryotic protein gene structures are simple and defined by open reading frames. Other dedicated tools have however been designed for the prediction of other functional (transcribed) elements such as RNA genes.

Through RNA-Seq short reads, new generation sequencing gives unprecedented access to transcriptomic data. In genomes with low gene density, such as plant or animal genomes, the availability of such transcribed sequences sampling is extremely useful to delineate gene structures. In bacterial genomes, most if not all the genome is transcribed, making such data much less easy to exploit. However, NGS technology is able to produce oriented read for which the strand of transcription is known. Such data facilitates the automatic prediction of a variety of transcribed elements, including protein genes, (possibly antisense) RNA genes and operon structures.

2 Changing EuGene Gene Model

EuGene is an eukaryotic gene finder [1,2] that can be described as a Conditional Random Field (CRF) predictor [3], a variant of random Markov fields capturing the conditional probability of structural annotations given available evidence. The default gene model of EuGene includes intergenic regions, coding exons, introns, 5'/3' untranslated terminal regions and introns within UTRs.

To be able to predict new functional elements in prokaryotes, the gene model underlying EuGene has been extensively modified. In the absence of splicing, intronic and spliced exonic states have been removed (overall 34 states removed). Conversely, new states have been introduced to capture:

- overlapping protein gene regions (on the same strand or not) on any of the 6 different coding frames.
- untranslated transcribed internal regions (UIR) between non overlapping gene appearing in the same operon on either strand. These new states complete the existing 5' and 3' UTR (untranslated terminal regions) defining operon extremities.
- and finally RNA genes on either strand.

Overall, the new prokaryotic variant of EuGene includes 30 states, compared to the 45 origianl states.

This work has been partially funded by the french "Agence nationale de la Recherche" SYMbiMICS project.

3 Integrating Evidence

In EuGene, each "feature", representing a specific type of evidence used for prediction is weighted in the CRF model and integrated through independent software plugins. We just integrated the prokaryotic translation start predictor of FrameD [4,5], based on RBS/ribosomal RNA hybridation energy, as a new plugin to get a fully functional prokaryotic gene finder capable of predicting protein genes, RNA genes and operons.

We are experimenting with the integration of oriented RNA-Seq data through existing generic plugins, either directly or following a segmentation based on the level of transcription. In the simplest variant, partial transcripts defined by oriented pair-end short reads are mapped to the genome. Their abundance at a given position is used as a weighted feature that indicates that the current region is transcribed on the corresponding strand. By integrating translation/transcription start and stop prediction, statistical models of different regions (especially coding regions) and RNA-Seq data inside a unique tool, EuGene becomes capable of discriminating protein genes (which are transcribed and follow a coding region statistical model) from RNA genes (which are transcribed but do not follow a coding model). In some sense, this is related to the QRNA [6] comparative RNA gene predictor which relies on a stochastic context free grammar model for RNA genes and a usual 3-periodic Markov model for coding regions. In a non comparative settings, we use a simple homogeneous Markov model for RNA regions instead but the integration of oriented RNA-Seq restricts the discrimination between coding and RNA genes to transcribed regions.

Similarly, a "stable" expression level inside a region, in several different conditions, identified through prior segmentation, should help delineate operons. This information can be directly injected inside EuGene as a feature informative about transcription start/stop but has not been evaluated yet.

Most, if not all, eukaryotic gene finders assume that only one strand is transcribed at a given position. To overcome this limitation, EuGene has been slightly modified to allow to perform independent gene prediction on each strand. Together with oriented RNA-Seq data, this allows to perform an automatic annotation that includes protein genes, RNA genes but also anti-sense RNA (RNA gene predicted on one strand overlapping a gene predicted on the other strand).

We are currently applying this new strand-independent prokaryotic variant of EuGene to the genome of *Sinorhizobium meliloti* using oriented RNA-seq data (representing 48Gb of reads). The results we have obtained closely match the existing genome annotation (with 6483 genes predicted compared to the 6235 annotated) and show that EuGene correctly identifies ribosomal and transfer RNA genes and many potentially new RNA genes (2040 ncRNA predicted compared to the 64 annotated ones). These new genes, predicted without any specific RNA related information (except for RNA-Seq and 3-periodic Markov coding models), needs to be experimentally evaluated.

References

[1] T. Schiex, A. Moisan, and P. Rouzé, Eugène: an eukaryotic gene finder that combines several sources of evidence. In M. Sagot, editor, *Selected papers from JOBIM'2000*, volume 2066 of *LNCS*, pages 118–133. Springer Verlag, 2001.

[2] S. Foissac, J. Gouzy, S. Rombauts, C. Mathe, J. Amselem, L. Sterck, Y. de Peer, P. Rouzé, and T. Schiex, Genome annotation in plants and fungi: EuGene as a model platform. *Current Bioinformatics*, 3(2):87–97, 2008.

[3] J. Lafferty, A. McCallum, and F. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of the Machine Learning International Workshop*, pages 282–289, 2001.

[4] T. Schiex, P. Thébault, and D. Khan, Recherche des génes et des erreurs de séquençage dans les génomes bactériens GC-riches. In O. Gascuel and M.-F. Sagot, editors, *Proc. of JOBIM'2000 (Journées Ouvertes Biologie Informatique Mathématiques)*, Montpellier, France, 2000.

[5] T. Schiex, J. Gouzy, A. Moisan, and Y. de Oliveira, FrameD: A flexible program for quality check and gene prediction in prokaryotic genomes and noisy matured eukaryotic sequences. *Nucleic Acids Res*, 31(13):3738–41, 2003.

[6] E. Rivas and S. R. Eddy, Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, 2:8, 2001.

URGI Genome Annotation System: an Integrated System for Structural and Functional Genome Annotation

Joelle AMSELEM[1,4], Michael ALAUX[1], Nathalie CHOISNE[1], Nicolas LAPALU[1,4], Baptiste BRAULT[1,4], Aminah KELIET[1], Erik KIMMEL[1], Françoise ALFAMA[1], Sandie ARNOUX[1], Marc BRAS[1], Laetitia BRIGITTE[1], Olivier INIZAN[1], Véronique JAMILLOUX[1], Jonathan KREPLAK[1], Fabrice LEGEAI[2], Isabelle LUYTEN[1], Cyril POMMIER[1], Sébastien REBOUX[1], Stéphanie SIDIBE-BOCS[3], Marc-Henri LEBRUN[4], Delphine STEINBACH[1] and Hadi QUESNEVILLE

[1] Unité de Recherche en Génomique-Info, UR1164 INRA, Route de Saint-Cyr, 78026, Versailles cedex, France
joelle.amselem@versailles.inra.fr

[2] UMR BIO3P INRA IRISA, Domaine de la Motte, BP35327, 35653, Le Rheu cedex , France

[3] UMR DAP CIRAD, TA A-96 / 03, Av Agropolis, 34398, Montpellier cedex 5, France

[4] Biologie et gestion des Risques en agriculture, UMR1290 INRA Agro-ParisTech, Av Lucien Brétignières, BP01, 78850, Thiverval-Grignon, France

Abstract *The URGI platform (http://urgi.versailles.inra.fr) develops a genome annotation system dedicated to plants and their pathogens. This Integrated System relies on: (i) pipelines for Transposable Elements annotation (REPET) and gene structural and functional predictions (ii) databases and user-friendly interfaces to browse and query the data (URGI Information System GnpIS, Genome Report System GRS), (iii) A distributed annotation system for curation of gene structure.*

Keywords genome annotation, pipelines, databases, interfaces, genes, transposable elements, plants, fungi.

1 Introduction

The INRA URGI (Unité de Recherche en Génomique-Info) develops and maintains an information system for plant and pathogens genomes. This system is used in number of national and international collaborative projects involving biologists and bioinformaticians. Nowadays, the recent development of new generations of sequencing tools leads to a spectacular increase of the number of sequenced genomes. But, genome annotation has difficulties to follows this pace, introducing a lack between the release of genome sequences and their annotations. To face this problem, the URGI develops and provids tools to annotate entirely sequenced genome (pipeline, databases, and interfaces).

2 The URGI Annotation System

The URGI annotation system relies on three components: pipelines, databases and interfaces.

2.1 Pipelines

- A transposable element detection and annotation package, called REPET [1,2] is composed of two pipelines: TEdenovo and TEannot. Thanks to their high level of automation and accuracy, they were used within many international genome projects concerning plants, fungi and insects.

- A gene prediction pipeline, based on *ab initio* and similarity gene finding softwares. It uses the EuGene program to integrate all sources of information [3].

- A functional annotation pipeline, based on (i) various methods of patterns matching and motifs recognition, (ii) intracellular targeting prediction methods, and (iii) comparative genomics with other fungal genomes.

2.2 Databases

Our database component relies on the well known schemas from the GMOD consortium (http://gmod.org). All annotation features and analysis results are primarily stored in the Chado or Bio ::SeqFeature ::Store schema according to the need (speed access or genericity). Data can then be searched through GnpIS QuickSearch (http://urgi.versailles.inra.fr/gnpis) and Biomart (GMOD).The GnpIS QuickSearch is based on the Apache Lucene™ full-featured text search engine library. Indexes are generated to query structural or functional data stored in same or separate DBs. Query results are returned according to significance with terms, and linked to GnpIS modules and/or Genome Report System (GRS). BioMart based datamarts were used as an advance search tool. Results of complex search criteria could be exported in different formats or directly send to Galaxy (http://main.g2.bx.psu.edu/) for further bioinformatic analysis.

2.3 Interfaces

We provide textual or graphical interfaces over the databases. We use GBrowse as graphical interface to display sequence annotations. Apollo or Artemis are used for gene structure curation shared by a community, as they are committed in the database using "pure JDBC" direct communication protocol between Apollo (or Artemis) and Chado. The Genome Report System GRS was developed (in Java) in the frame of the ANR GnpAnnot project. It provides comprehensive categories of reports through a user-friendly textual interface over structural and functional genomic data stored in Chado databases.GRS also proposes a Gene Ontology browser and an editing module (GRE) to allow manual functional curations.

2.4 Conclusions and Perspectives

The platform was chosen by the international grapevine consortium (IGGP) to manage grapevine genomic annotations and to help the community to perform the manual gene annotation. It also hosts wheat genomic and genetic data for the International wheat scientific community (IWGSC). It is used for the annotation of the first wheat chromosome (3B) sequence. The integrated genome annotation system was also successfully used for fungal genomes as *Botrytis cinerea* T4 (grey mould disease) and *Leptosphaeria maculans* (stem canker) [5] in the frame of their genome consortium for sequencing and annotation. Portal for the different plant and fungi species are available at http://urgi.versailles.inra.fr/index.php/Species.

Data integration of sequences from the next generation sequencing technologies is a new scientific challenge in bioinformatics. To face this challenge, evolution of GnpIS architecture is in progress: evolution of DB schemas and interfaces, new datamarts and galaxy workflow manager based pipelines to mine data.

Acknowledgements

We ackowledge :
- URGI Information System, pipelines and data "Agile" development teams.
- ANR for the funding of GnpAnnot and GnpInteGr projects.

References

[1] T. Flutre, E. Duprat, C. Feuillet and H. Quesneville, Considering transposable element diversification in de novo annotation approaches. *PLoS ONE* 6(1):e16526, 2011

[2] H. Quesneville, C.M. Bergman, O. Andrieu, D. Autard and D. Nouaud, Combined evidence annotation of transposable elements in genome sequences. *PLoS comput. Biol.,* 1(2):e22, 2005

[3] S. Foissac, J. Gouzy, S. Rombauts, C. Mathé, J. Amselem, Y. Van de Peer, P. Rouzé and T. Schiex, Genome Annotation in Plants and fungi : Eugene as a model platform. *Current Bioinformatics*, 3(2):87-97, 2008

[4] T. Rouxel, J. Grandaubert, J. Hane, C. Hoede, A. van de Wouw, A. Couloux, V. Dominguez, V. Anthouard, P. Bally, S. Bourras, A. Cozijnsen, L. Ciuffetti, A. Degrave, A. Dilmaghani, L. Duret, I. Fudal, S. Goodwin, L. Gout, N. Glaser, J. Linglin, G.H. Kema, N. Lapalu, C. Lawrence, K. May, M. Meyer, B. Ollivier, J. Poulain, C. Schoch, A. Simon, J. Spatafora, A. Stachowiak, B.G. Turgeon, B. Tyler, D. Vincent, J. Weissenbach, J. Amselem, H. Quesneville, R. Oliver, P. Wincker, M.H. Balesdent and B. Howlett, Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by Repeat-Induced Point mutations. *Nat. Comms.*, 2:202, 2011

NGS: Neglected Genome Sequencing

Assembly and Annotation Challenges in a Highly Divergent Protozoan Genome

Sandrine MOREIRA[1], Marcel TURCOTTE[2] and Gertraud BURGER[1]

[1] Robert Cedergren Center for Bioinformatics and Genomics, Université de Montréal, 2900 Edouard-Montpetit, H3T 1J4, Montréal, Québec, Canada
{sandrine.moreira, gertraud.burger}@umontreal.ca

[2] School of Information Technology and Engineering, University of Ottawa, 800 King Edward, K1N 6N5, Ottawa, Ontario, Canada
turcotte@site.uottawa.ca

Keywords 454-sequencing, Assembly, Annotation, Protozoan, RNA processing.

1 Introduction

Our laboratory focusses on novel gene expression mechanisms and genome structures. The organisms of choice are a group of poorly investigated unicellular eukaryotes, the Diplonemids (Euglenozoa). Euglenozoa are thought to have emerged as one of the earliest diverging eukaryotic groups, much before the emergence of the well-studied animals, fungi and plants.

Diplonemids possess a highly unusual mitochondrial genome. Not only is this genome composed of a hundred or so circular chromosomes, but also mitochondrial genes are all split into several pieces (modules) each of which is located on a different chromosome. Gene modules are transcribed separately into RNA and then joined to a complete messenger RNA by some sort of trans-splicing. At least one mitochondrial pre-mRNA is modified in sequence post-transcriptionally, and this RNA editing proceeds by addition of uridines exactly at the junction of two modules [1,2]. Uridine-based RNA editing is known from the Diplonemids' sister group, the Kinetoplastids (addition and deletion), but editing being interlinked with trans-splicing is unheard of.

RNA editing and trans-splicing are likely performed by a multifunctional molecular machine. My project consists in the identification of genes involved in this machinery using bioinformatics methods. These genes must be encoded in the nucleus, yet, nothing is known about the nuclear genome, except the size estimated experimentally.

2 454 Sequencing and Assembly

Sequencing of the nuclear genome from the Diplonemid species *Diplonema papillatum* is well underway at the IMG in Prague, using the 454 massive parallel pyrosequencing technology of Roche Life Science, which produces reads of about 300 nt in length. A preliminary sequence assembly with the Newbler assembler using 900 Mb of reads including paired-end libraries (3 kb, 8 kb, and 20 kb), led to 66 Mb in more than 70000 contigs. The large number of contigs is a known problem. No eukaryotic genome, sequenced by 454 solely, and assembled without a reference genome, has ever been completed (and published). To address this problem, we built smaller datasets, benchmarked a selection of 454 sequence assemblers [3], and tested different ways of removing sequencing errors. Our tests are still underway. I will present the strategy that we plan to adopt to reach an assembly that is well suited for gene annotation.

3 Annotation of Highly Divergent Species

The annotation of poorly investigated protozoan genomes such as *Diplonema* is the second challenge. Experimental data are scarce and sequences are highly divergent from others so that similarity-based annotation (*e. g.* Blast) is of limited success. Similar difficulties were encountered by others when annotating the nuclear genomes of the divergent Kinetoplastid and Apicomplexan genomes [4,5], but a general strategy has not yet been designed. We aim at using *Diplonema* as a model

for developing an effective annotation approach for divergent genomes. A powerful and affordable asset for genome annotation is the availability of EST data. This allows defining an effective gene model for the first, syntactic, annotation phase. For *Diplonema*, a dataset of about 4000 EST clusters has been generated previously in the context of the pan-Canadian Protist EST Program [6]. By mapping ESTs to genome sequence, we addressed the following questions: (i) are nuclear genes of *Diplonema* discontinuous as its mitochondrial genes or rather orthodoxically contiguous? (ii) Do they contain introns, and if yes, what is their frequency, length and type? I will report my preliminary results on the nuclear gene structure of *D. papillatum* and use the identification of a key composent of the trans-splicing machinery to examplify challenges typically encountered in annotating genes of highly divergent species.

Acknowledgements

This work was supported by the CIHR (Canadian Institutes of Health Research) and the FESP (Faculté des Etudes Supérieures et Postdoctorales) of the Université de Montréal.

References

[1] W. Marande and G. Burger, Mitochondrial DNA as a genomic jigsaw puzzle. *Science,* 318:415, 2007.

[2] C. Vlcek, W. Marande, S. Teijeiro, J. Lukes, and G. Burger, Systematically fragmented genes in a multipartite mitochondrial genome. *Nucleic Acids Res.,* 39:979-88, 2011.

[3] S. Kumar and M. L. Blaxter, Comparing de novo assemblers for 454 transcriptome data. *BMC Genomics,* 11:571, 2010.

[4] E. A. Worthey and P. J. Myler, Protozoan genomes: gene identification and annotation. *Int J Parasitol,* 35:495-512, 2005.

[5] R. Salavati and H. S. Najafabadi, Sequence-based functional annotation: what if most of the genes are unique to a genome? *Trends Parasitol,* 26:225-9, 2010.

[6] P. Keeling, G. Burger, D. Durnford, B. Lang, R. Lee, R. Pearlman, A. Roger, and M. Gray, The tree of eukaryotes. *Trends Ecol Evol (Amst),* 20:670-676, 2005.

Session 6.C : Software Tools

ecoQuery: a Semantic Module to Query Biodiversity Data on the Web

Julie CHABALIER, Olivier COULLET, Amandine SAHL and Olivier ROVELLOTTI
NATURAL SOLUTIONS, 68 rue Sainte, 13001 Marseille, France
{julie_chabalier, olivier_coullet, amandine_sahl, olivier_rovellotti}@natural-solutions.eu

Abstract *The rapid growth of collected biodiversity data, their heterogeneity and their dissemination can be the bottleneck of safeguarding our environment. Semantic Web technologies aim to address these issues. Based on these technologies, we extended the ecoRelevé platform which is dedicated to biodiversity data management with the ecoQuery module. This module is able to query the Web of data in order to move up scale and provide an answer to complex issues.*

Keywords Biodiversity, linked data, Web of data, semantic Web.

ecoQuery : un Module Sémantique pour Interroger les Données de Biodiversité sur le Web

Résumé *L'accroissement rapide du nombre de données de biodiversité collectées, leur hétérogénéité et leur dispersion sur le Web peuvent être un frein pour la préservation du monde qui nous entoure. Les technologies du Web sémantique ont pour objectifs de pallier ces problèmes. Basé sur ces technologies, nous avons étendu la plateforme ecoRelevé dédiée à la gestion des données de biodiversité avec le module ecoQuery. Ce module est capable d'interroger le Web de données pour apporter un changement d'échelle et une réponse à des questions complexes.*

Mots-clés Biodiversité, linked data, Web de données, Web sémantique.

1 Introduction

L'érosion accélérée de la biodiversité, diagnostiquée lors de l'Evaluation des écosystèmes du Millénaire [1], a été confirmée par les résultats de l'étude des coûts de l'inaction en matière de biodiversité en 2008 [2]. Ces études ont eu pour conséquence d'accentuer, ces dernières années, la prise de conscience internationale sur le rôle vital de la biodiversité.

Cette prise de conscience s'est naturellement accompagnée d'un besoin généralisé d'approfondir les connaissances du monde qui nous entoure dans l'objectif de mieux le préserver. Ce besoin se traduit par une augmentation de la collecte des données de biodiversité. Leur analyse vise à proposer des zones de protection spéciale ou de conservation i.e. dans le cadre du réseau Natura 2000 [3], faire des études d'évaluation du risque pour une espèce ou un écosystème, quantifier le nombre d'individus d'une espèce dans une zone géographique donnée, etc. Répondre à ces objectifs nécessite d'intégrer un grand nombre de données de biodiversité. La nature complexe, l'hétérogénéité et la dispersion sur le Web de ces données rendent leur intégration et, en conséquence, leur analyse longues et fastidieuses.

Le Web sémantique désigne un ensemble de technologies dont l'objectif est de favoriser l'interopérabilité des données en vue notamment de leur intégration. Utilisant la famille de langages développés par le W3C, le contenu des ressources du Web devient accessible et utilisable par les programmes et agents logiciels, grâce à un système de métadonnées formelles [4]. Dans ce cadre, le « Linked Data » désigne un ensemble de bonnes pratiques pour publier et connecter les données structurées sur le Web [5]. Ces données constituent le Web de données. Elles peuvent être découvertes et collectées automatiquement par des machines sans intervention humaine. L'utilité du « Linked Data » pour représenter la sémantique des observations relatives à la biodiversité et favoriser ainsi l'interopérabilité des données a été décrite dans [6]. Les auteurs soulignent également l'importance de construire des outils offrant aux utilisateurs la possibilité de fouiller le Web de données et facilitant l'intégration automatique des données.

Récemment, nos travaux se sont concentrés sur la construction de la plateforme open source ecoRelevé[1] dédiée à la gestion des données de biodiversité [7]. Cette plateforme a été initialement conçue pour stocker et visualiser les données d'observation. Le module ecoQuery, présenté ici, est une extension de cette plateforme, permettant de consommer les données publiées au sein du Web de données rendant ainsi transparentes leur recherche et leur intégration. Agrémentée du module ecoQuery, la plateforme ecoRelevé favorise des études de données à plus large échelle et a pour objectif de répondre à des questions plus complexes.

Ce papier s'organise de la manière suivante : la section 2 présente le cas d'utilisation sur lequel nous avons basé la construction du module ecoQuery, la section 3 traite du Web de données et de son application à la biodiversité, la section 4 présente l'extension de la plateforme ecoRelevé avec le module ecoQuery.

2 Définition d'un Cas d'Utilisation

Pour spécifier les exigences fonctionnelles du module ecoQuery, nous avons défini un cas d'utilisation concret dans le domaine de la biodiversité. Ce cas d'utilisation devait répondre à un problème courant et transposable et nécessiter l'intégration de données pour être résolu.

Suivant ces critères, nous nous sommes intéressés à la réalisation d'une étude environnementale dans le cadre du projet de restructuration des dispositifs de protection contre les crues entre les Communes de Sénas et de Cheval-Blanc dans le département du Vaucluse. Plusieurs études ont en effet démontré qu'il existe un risque majeur de débordement de la Durance lors de crues pouvant conduire à l'inondation des communes situées sur la rive droite de la rivière. Les communes concernées ont décidé d'établir une ligne de protection empêchant tout risque de débordement (digue insubmersible, élévation du remblai de la voie ferrée, etc.). La construction de ce type d'ouvrages nécessite la réalisation d'un inventaire de la zone. Nous avons restreint ce cas d'utilisation à l'inventaire de l'avifaune.

Pour chaque inventaire naturaliste, la première étape consiste généralement à établir une liste de taxons déjà observés sur le site. Ces informations sont, tout d'abord, recherchées sur le Web et utilisées, dans un deuxième temps sur le terrain, en tant que référence. Les naturalistes notent cependant avec attention les nouveaux taxons observés. Différentes informations peuvent être ensuite collectées de manière à évaluer l'incidence de la construction d'ouvrages sur ces taxons. Dans le cadre des oiseaux, il est, par exemple, important de distinguer pour le site donné, les oiseaux nicheurs des oiseaux de passage et des oiseaux hivernants. Le statut de protection des espèces est également important lors des études d'impacts.

Répondre à cette étude nécessite donc d'intégrer un ensemble de données. A minima, ces données doivent être décrites par trois attributs spécifiques, i.e. le nom scientifique du taxon, ses coordonnées géographiques et la date d'observation. La section suivante présente les données disponibles dans le Web de données et comment y accéder.

3 Le Web de Données et la Biodiversité

Le « Linked Open Data Cloud » désigne l'ensemble des données en accès libre publiées suivant les principes du « Linked Data ». En septembre 2010, 203 jeux de données dont 40 en sciences de la vie et 15 en géographie[2] étaient présents dans le Web de données libres. Parmi ces jeux de données, nous avons recensé six jeux impliquant des données ouvertes relatives à la biodiversité (voir table 1).

Les données minimales nécessaires à la mise en place de notre cas d'utilisation concernent essentiellement des données d'observations (taxons, coordonnées, dates d'observation). Seul le jeu de données « TaxonConcept » propose des données d'observations. Cependant, bien qu'il y ait actuellement pratiquement 100000 concepts relatifs aux taxons, ce jeu de données ne contient qu'un nombre relativement faible d'observations. Ce nombre est amené à évoluer, de nombreux travaux vont dans ce sens (notamment dans le cadre du Global Biodiversity Information Facility, GBIF [8]).

En attendant l'enrichissement du Web de données, nous avons choisi d'appliquer les principes du « Linked Data » sur une source qui, nativement, ne les respecte pas. L'agrégateur de données aviaires le plus

[1] https://code.google.com/p/ecoreleve/ (stable release)
[2] http://richard.cyganiak.de/2007/10/lod/lod-datasets_2010-09-22_colored.html

important est l'« Avian Knowledge Network (AKN) » [9]. L'AKN contient actuellement environ 89 millions d'observations sur plus de 9500 taxons. Nous avons extrait de l'AKN les données correspondant aux observations d'oiseaux en France et transformé ces données en RDF (Resource Description Framework) en utilisant le standard de données de biodiversité Darwin Core[1]. Les données obtenues ont été stockées dans un entrepôt RDF, la version open source de Virtuoso (VOS). Elles sont accessibles par le langage de requête SPARQL, spécifique au données RDF, via un « endpoint[2] ». La section suivante présente les outils mis en place pour consommer ces données de manière à répondre aux besoins du cas d'utilisation.

Jeu de données	Description
TaxonConcept	Fournit des URI informatives afin d'améliorer la qualité et la stabilité des liens entre une espèce et les données relatives à cette espèce.
Geospecies	Information sur les ordres, familles, espèce.
European Nature Information System	Informations sur les espèces et sur les sites européens où elles sont attendues.
DPpedia	Extraction d'informations structurées de Wikipedia. Les données sont liées à un ensemble de données présentes sur le Web.
Fishes of Texas	Standardisation et géoréférencement des informations scientifiques connues sur les poissons d'eau douce du Texas.
Geonames	Fournit des données géographiques à partir de différentes sources telles que les noms de lieux, l'altitude, la population, etc. Les coordonnées (latitude et longitude) sont en WGS84 (World Geodetic System 1984).

Table 1. Jeux de données du Web de données relatif à la biodiversité.

4 Les Requêtes Sémantiques et l'Intégration de Données

Cette section décrit nos travaux en cours, à savoir la plateforme ecoRelevé, un système de gestion de données de biodiversité, et son nouveau module ecoQuery interrogeant, de manière transparente pour l'utilisateur, le Web de données.

4.1 La Plateforme ecoRelevé

La plateforme ecoRelevé comprend actuellement trois modules : 1) ecoRelevé Core dédié au stockage des données, 2) ecoRelevé Data pour l'import des données de terrain à partir du logiciel mobile pocket eRelevé [10] et 3) ecoRelevé Explorer permettant de visualiser les données sur une carte. Cette application internet riche (RIA) est basée sur Adobe Flex/AIR et OpenScales pour la couche de présentation (module Explorer et Data), JAVA/Hibernate pour la manipulation/persistance des objets biologiques (i.e. relevés, taxons), geoserver pour la manipulation des objets cartographiques, et PosgreSQL/PosGIS pour le stockage des données (module Core).

Le module Explorer propose, via un filtre, une visualisation cartographique des données suivant trois dimensions : la taxonomie, le temps et la localisation géographique. Le filtre traduit les contraintes sur ces dimensions en un service Web interrogeant la base de données sous-jacente au module ecoRelevé Core. Les filtres ainsi créés sont gérés au sein d'un gestionnaire. Afin d'obtenir de la plateforme ecoRelevé qu'elle réponde aux exigences fonctionnelles spécifiées lors de l'élaboration du cas d'utilisation, il s'agissait d'étendre l'application d'un filtre au Web de données, d'intégrer les données obtenues avec celles stockées dans le module Core et de visualiser les résultats. La section suivante présente le module ecoQuery répondant à ces exigences.

[1] http://rs.tdwg.org/dwc/
[2] http://natural01.gn-noc.com:8890/sparql (URI du graphe : urn:rdf.Occurences_AKN_dwc)

4.2 Le Module ecoQuery

Le module ecoQuery transforme à la volée les contraintes sur les trois dimensions de chaque filtre établi par l'utilisateur dans le langage SPARQL. La requête ainsi obtenue s'exécute sur le « endpoint » spécifique que nous avons mis en place pour pallier le manque actuel de données d'observations au sein du Web de données. Le XML résultant de la requête est ensuite traité afin de visualiser les résultats au sein du module Explorer. Le gestionnaire a également été modifié de manière à étendre un filtre à une autre source de données sans modifier la requête sous-jacente.

La Figure 1 présente la requête SPARQL sur laquelle sont automatiquement appliquées les contraintes définies par l'utilisateur dans le module Explorer (sous forme de filtres). La construction de la requête est transparente pour l'utilisateur. A l'instar du service Web interrogeant la base de données sous-jacente au module ecoRelevé Core, la requête SPARQL est construite automatiquement à partir de l'interface cartographique, de la liste de taxons et du diagramme de temps présents au sein de l'interface graphique du module Explorer.

Actuellement, le standard Darwin Core autorise seulement des requêtes basiques, d'expressivité limitée, nous verrons dans la discussion les efforts effectués pour pallier ce problème.

```
PREFIX dwc: <http://rs.tdwg.org/dwc/terms/#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX dcterms: <http://purl.org/dc/terms/>
SELECT distinct ?scientificName ?lat ?long ?date
FROM <urn:rdf:Occurences_AvianKnowledgeNetwork>
WHERE{
?occurrence rdf:type dwc:Occurrence;
dwc:Identification ?id;
geo:lat ?lat;
geo:long ?long;
dcterms:date ?date.
?id dwc:scientificName ?scientificName.
}
```

Figure 1. Requête SPARQL générique.

Associé à la plateforme ecoRelevé[1], le module ecoQuery permet désormais de réaliser une requête sémantique sur le Web de données et d'intégrer les résultats avec les données stockées dans le module Core de la plateforme (voir Figure 2 pour l'architecture du système).

[1] https://code.google.com/p/ecoreleve/ (latest release)

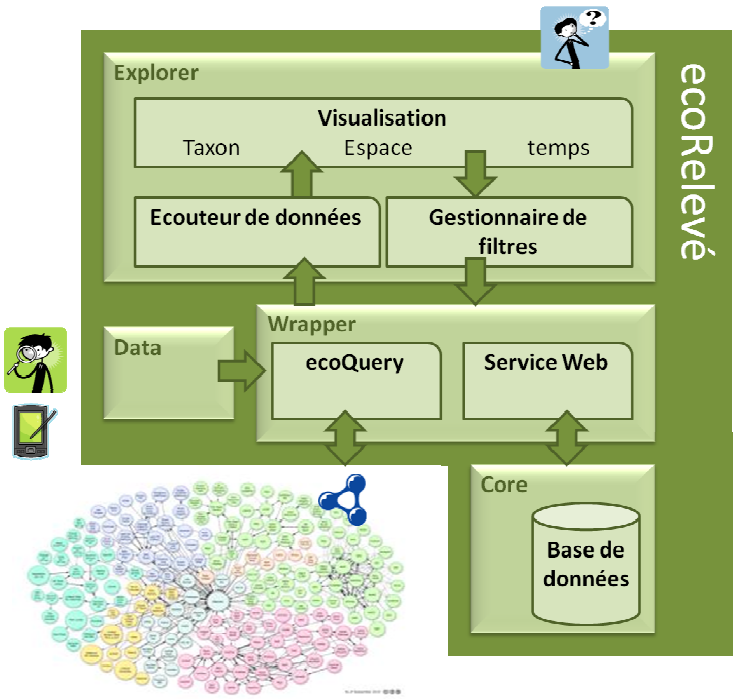


Figure 2. Architecture de la plateforme ecoRelevé agrémentée du module ecoQuery.

4.3 L'Inventaire de l'Avifaune

Afin de valider l'utilisation de la plateforme ecoRelevé agrémentée du module ecoQuery, nous avons réalisé l'étude définie par le cas d'utilisation présenté dans la section 2.

L'utilisation du module ecoRelevé Explorer débute par la sélection d'une emprise géographique. Cette sélection peut se faire à partir de la carte ou via une recherche textuelle utilisant le service Web associé à l'ontologie Geonames [11]. Dans ce dernier cas, l'échelle de la portion de carte affichée dépend de l'unité administrative associée à la localisation (Pays, Région, Département, Ville, etc.). Nous avons ensuite établi un filtre en fixant comme contraintes la commune de Cheval-Blanc, la classe des oiseaux (taxon Aves) et une plage de dates (2000 – 2011). Le module ecoQuery permet d'exécuter le filtre sur le jeu de données présent dans le VOS, à savoir les données de l'AKN.

Seules deux stations (association des coordonnées géographiques, d'une date, d'un observateur) sont affichées sur la carte, couvrant six taxons : *Columba livia* (pigeon biset), *Garrulus glandarius* (geai des chênes), *Phoenicurus ochruros* (Rougequeue noir), *Erithacus rubecula* (Rouge-gorge familier), *Sitta europaea* (Sittelle torchepot) et *Fringilla coelebs* (Pinson des arbres). Avec deux stations, l'effort de prospection est trop faible pour tirer des conclusions sur ces données. Cette première utilisation de la plateforme enrichie du module ecoQuery permet cependant de dégager une liste de taxons déjà observés sur le site.

Pour pallier le manque de données, nous avons effectué une mission d'inventaire de l'avifaune muni du logiciel mobile pocket eRelevé. La liste de taxons précédente a été utilisée comme référence pour cette étude de terrain. Après deux jours de prospection sur la commune de Cheval-Blanc, tous les oiseaux rencontrés ont été identifiés et les données correspondantes ont été enregistrées sur l'appareil mobile, i.e. la date, le nom scientifique du taxon et les coordonnées GPS. De retour de mission, les données ont été importées à l'aide du module ecoRelevé Data au sein de la base de données du module Core. Les données ainsi stockées dans le Core, le gestionnaire de filtres est utilisé pour les intégrer avec les données du Web. L'ensemble des stations est ensuite visualisée sur une carte. En supplément des six taxons de la liste de référence, 17 taxons sont désormais affichés sur la carte. Parmi ces taxons, certains, tels que *Coracias garrulus* (Rollier d'Europe), sont quasi-menacés (statut de conservation de l'IUCN).

Pour finir, les données de l'AKN ont permis d'étendre cette étude à tout le département du Vaucluse sans nouvel effort de prospection. Nous obtenons 49 taxons dont 34 nouveaux par rapport au jeu de données

précédent (voir Figure 3). Ces nouveaux taxons pourraient éventuellement être associés au rapport d'inventaire en tant que taxons potentiels.

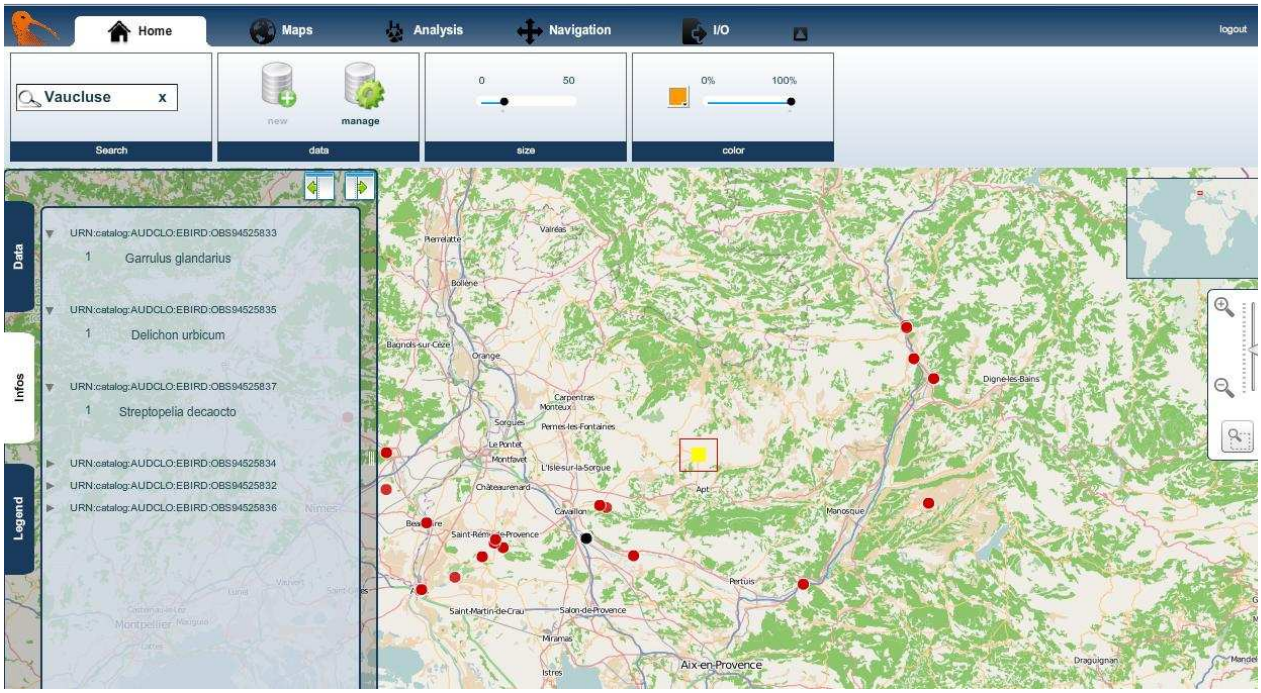


Figure 3. Présentation des stations sur l'emprise du Vaucluse.

5 Discussion

ecoQuery est un module sémantique permettant d'interroger le Web de données de manière transparente et de visualiser les résultats intégrés avec des données locales au sein de la plateforme ecoRelevé, une plateforme dédiée à la gestion des données de biodiversité. Bien qu'actuellement les observations naturalistes soient peu présentes dans le Web de données, ce premier prototype démontre la faisabilité de l'approche sur un jeu de données aviaires transformé en RDF et stocké dans un entrepôt accessible via un « SPARQL endpoint ».

Dans cette approche, l'avantage d'utiliser le RDF réside essentiellement dans l'interopérabilité qu'il procure. Contrairement aux données contenues dans des bases de données classiques, difficilement accessibles et généralement non compatibles entre elles, les jeux de données, décrits en RDF et présents dans le Web de données sont désormais accessibles, liés et exploitables par tous. L'application sur un cas d'utilisation concret a démontré que l'accès aux données de biodiversité via le Web de données est un gain de temps pour le naturaliste par rapport au travail long et fastidieux de recherche de données, notamment bibliographiques, en amont des études environnementales. En effet, la plateforme ecoRelevé agrémentée du module ecoQuery propose un accès centralisé et transparent à des données unifiées.

Récemment, un jeu de données d'observation a été mis à disposition au sein du « endpoint » du créateur de TaxonConcept[1]. Il s'agit des données obtenues dans le cadre d'un bioblitz[2] organisé lors de la conférence TDWG 2010 à Woods Hole, Massachussetts[3]. ecoQuery a donc été utilisé pour interroger ce nouveau jeu de données et obtenir une visualisation intégrée avec les données de l'AKN concernant le Massachussetts dans la plateforme ecoRelevé. Cette dernière expérimentation a permis de valider notre étude sur un jeu de données distant, RDF-natif et a démontré que notre système était techniquement prêt à accueillir les données d'observation qui viendront prochainement enrichir le « Linked Open Data Cloud ».

[1] http://lsd.taxonconcept.org/sparql
[2] Bioblitz : inventaire biologique intensif sur une portion bien précise de terrain.
[3] http://bioblitz.tdwg.org/

Actuellement, la version RDF du standard Darwin Core utilisé pour structurer les données de biodiversité n'exploite pas pleinement les capacités du Web sémantique pour définir des données. Bien qu'offrant un vocabulaire commun pour partager les données, ce vocabulaire ne permet pas de définir formellement les termes utilisés ni de raisonner sur les données de manière à produire automatiquement de nouvelles connaissances. Le niveau d'expressivité des requêtes qui en découle reste donc assez bas. Pour pallier ce problème, nous travaillons sur l'élaboration d'une ontologie de la biodiversité dans le cadre du projet collaboratif « ecoOnto » [12]. Basée sur une extension de l'ontologie OBOE [13], les classes de cette ontologie seront mises en relation avec les standards actuels de la biodiversité, notamment Darwin Core, de manière à obtenir une meilleure structuration des données d'observation et ainsi résoudre des problèmes complexes de biodiversité via les mécanismes de raisonnement associés aux logiques de description. Cette ontologie sera notamment utilisée dans la plateforme ecoRelevé pour élaborer des requêtes complexes sur le Web de données et fournir ainsi une aide à la décision pour l'utilisateur. Dans cet objectif, la plateforme ecoRelevé sera prochainement étendue de manière à accueillir des données de biodiversité autres que les observations, e.g. les statuts de protection des espèces, les habitats, etc.

Remerciements

Ce travail est soutenu par OSEO INNOVATION.

Références

[1] Millennium Ecosystem Assessment (MEA). Current state and trends assessment, *Washington D.C., Island Press*, 2005.

[2] P. Sukhdev, The Economics of Ecosystems and Biodiversity (TEEB) interim Report, 2008.

[3] O. P. Ostermann, The need for management of nature conservation sites designated under Natura 2000. *Journal of Applied Ecology*, 35, 968-97, 1998.

[4] T. Berners-Lee, J. Hendler and O. Lassila. *The semantic web. Scientific American*, p. 29-37, 2001.

[5] C. Bizer, T. Heath and T. Berners-Lee, Linked Data – The Story So Far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5 (3). pp. 1-22. 2009.

[6] O. J. Reichman, M. B. Jones and M. P. Schilldhauer, Challenges and Opportunities of Open Data in Ecology. *Science*, 331,6018:703-705, 2011.

[7] O. Coullet, A. Sahl, J. Chabalier and O. Rovellotti, ecoRelevé: an open source response to the biodiversity crisis. *Proceedings of the conference for Open Source Geospatial Software (FOSS4G)*, 2010.

[8] T. Catapano, D. Hobern, H. Lapp, R.a. Morris, N. Morrison, N. Noy, M. Schildauer and D. Thau, Recommendations for the Use of Knowledge Organisation System by GBIF, *GBIF white paper*, February 2011.

[9] M. Iliff, L. Salas, E. Ruelas Inzunza, G. Ballard, D. Lepage and S. Kelling, The avian knowledge network: a partnership to organize, analyse, and visualize bird observation data for education, conservation, research, and land management. *Proceedings of the fourth Inrternational Partners in Filght Conference: Tundra to Tropics*, pp. 365-376.

[10] R. Kipré, O. Coullet, A. Sahl, J. Chabalier, C. Duval, O. Assunçao, O. Rovellotti, Pocket eRelevé : nouvelle approche de collecte de données sur le terrain. *Géomatique Expert*, 69, 24 – 27, 2009.

[11] GeoNames ontology. http://www.geonames.org/ontology/ontology_v2.2.1.rdf.

[12] J. Chabalier, EcoOnto : une ontologie pour la biodiversité. *Acte du colloque national d'écologie scientifique (Ecologie 2010)*, 2010.

[13] J. Madin, S. Bowers, M. Shildhauer, S. Krivov, D. Pennington and F. Villa. An ontology for describing and synthesizing ecological observation data. *Ecological Informatics*, 2: 279-296, 2007.

Integrated Bioinformatics Solutions for Microbial Genome, Proteome and Metabolome Comparative Analysis

Agnès ILTIS[1], Pierre-Emmanuel CIRON[1] and François RECHENMANN[1,2]
[1] Genostar, 60 rue Lavoisier, 38330 Montbonnot, France
[2] INRIA Grenoble Rhône-Alpes, 655 avenue de l'Europe, 38334 Saint Ismier Cedex, France
Francois.Rechenmann@inria.fr, rechenmann@genostar.com

Solutions Bioinformatiques Intégrées pour l'Analyse et la Comparaison de Génomes, Protéomes et Métabolomes Microbiens

Avec l'objectif de concevoir et développer un environnement bioinformatique pour l'annotation de génomes bactériens, le consortium Genostar rassemble en 1999 l'Institut Pasteur de Paris, l'INRIA (Institut National de Recherche en Informatique et en Automatique) et les sociétés de biotechnologie Hybrigenics (Paris) et GENOME Express (Grenoble).

Financé par ses partenaires et par le Ministère de la Recherche (Direction de la Technologie et Direction de la Recherche), le consortium fait une présentation publique en mai 2002 à l'Institut Pasteur de la première version opérationnelle du logiciel. Deux ans plus tard, la société Genostar est créée pour valoriser ce développement et ceux réalisés au sein de l'équipe Helix de l'INRIA Rhône-Alpes.

La société Genostar propose des solutions bioinformatiques pour l'annotation, la comparaison et la gestion de génomes, protéomes et métabolomes microbiens (virus, bactéries, levures). Le logiciel Metabolic Pathway Builder permet de sélectionner et enchaîner un vaste ensemble de méthodes d'analyse, et de visualiser, évaluer et exporter les résultats d'annotation et de comparaison. À l'issue de ces processus d'analyse, les génomes peuvent rejoindre une base de données structurée pour accueillir les informations génomiques, protéiques et biochimiques adéquatement connectées et consolider ainsi les données de référence utilisées pour les analyses comparatives ultérieures.

Les solutions Genostar évoluent avec les progrès technologiques (NGS) et méthodologiques (biologie des systèmes) qui marquent ses domaines d'application. Caractérisation de génomes et protéomes viraux, mise en contexte de données métabolomiques, analyses métagénomiques sont ainsi les projets majeurs actuellement menés au sein de la société. Dans ce contexte de R&D, Genostar est ouvert aux partenariats, en particulier avec des laboratoires publics, tant sur des projets de bioanalyse que de développements méthodologiques et logiciels.

SynTView: a Dynamic Genome Browser for Microbial Synteny and Polymorphism Information

Pierre LECHAT[1], Erika SOUCHE[1] and Ivan MOSZER[1]

[1] Institut Pasteur, Bio-informatique pour l'Analyse Génomique, 28 rue du Docteur Roux, 75015 Paris, France
{pierre.lechat, erika.souche, ivan.moszer}@pasteur.fr

Keywords Synteny, NGS, SNP, Genome Browser, Flash.

1 Introduction

Dynamic visualization interfaces are required to explore data obtained by "next-generation sequencing" technologies (NGS). SynTView offers such functionalities. The program foundation is a generic genome browser with sub-maps holding information about genomic objects (in a broad meaning). The main features of the software are the presentation of the syntenic [1] organization of microbial genomes (prokaryotes and lower eukaryotes) and the interactive visualization of polymorphism data along these genomes.

2 Implementation and Access

SynTView is a Flash software (AS3 language) [2], a technology increasingly used in comparative genomics [3,4]. The user can access the application through a web interface, in combination with the GenoList environment [5], thus taking advantage of comparative genome data (see 3.2). Alternatively, a stand-alone client is available (multi-platform AIR), allowing the user to work with its own locally stored data. Currently, the accepted file formats are ptt (GenBank) for genome annotation, tab-delimited files for other information (protein correspondences, Single Nucleotide Polymorphisms – SNPs) and the Newick format for phylogenetic trees. Web access can also be implemented using local user flat files.

3 Functional Modules

3.1 Genome Browser

SynTView is built as a generic genome browser which allows the user to visually explore genomes by genomic location, or to directly access genes by names. Several genomic maps can be stacked on top of each other. Users can dynamically change their respective order, adjust the scale of the maps to take advantage of the entire screen area, zoom in and out. Contextual menus are associated to genes, allowing the user to compute local views around a given gene, get sequence information, access the GenoList gene card, or add genes to a gene basket that can subsequently be used for various operations on gene lists.

3.2 Synteny Viewer

Besides its local implementation (see Section 2), SynTView is embedded in GenoList [5], an integrated environment dedicated to the analysis of microbial genomes, where comparative genomics data are pre-computed. There are two ways to access SynTView in this environment: either from one given gene card or through a direct access to the 750 organisms stored in GenoList. The synteny information is computed from the correspondence between proteins of different organisms (Bi-Directional Best Hits – BDBH) and the conserved order of the corresponding genes along the genomes.

A color is randomly assigned to every gene of the reference organism. By construction, BDBH genes get the same color as the reference gene and orphan genes remain black. When the user clicks on a gene to show its local synteny, the other genomes shift to be aligned with the main sequence and the non-syntenic genes fade away. In the GenoList-associated implementation, clicking on the links between corresponding genes redirects to a protein multiple alignment performed in GenoList.

In addition to the local view described above, four global views are available. (i) The Dot Plot shows the synteny ruptures and the chromosome re-arrangements. From there, BDBH can be graphically selected and exported to a file. (ii) The Line Plot shows the organization of syntenic groups at the chromosome level. (iii) The phylogenetic profile consists in a heat map of BLASTP hits sorted according to a phylogenetic tree. (iv) Finally the user can browse the pivot genome or the content of user-defined gene baskets through gene tables with sorting functionalities and backward access to the local view.

3.3 SNP Viewer

The SNP map allows the user to navigate through SNP data sets (*e.g.*, obtained by NGS), which are colored according to the mutation type and the gene/intergenic location (Fig. 1). SNP types can be dynamically hidden and are mutually linked to cognate genes (*e.g.*, when the mouse is over a gene, this triggers an animation of the enclosed and surrounding SNPs). The user can build groups of strains (according to epidemiological, phylogenetic or other criteria) to obtain a SNP density map, *i.e.* a histogram where the size of the bars represents the number of SNPs per gene. Sequence variations (both at the nucleotide and protein levels) can be obtained in a dedicated view. Finally, an artificial sequence can be determined across all genomes or for a group of genes, in order to compute phylogenetic distances (Fig. 1).

Figure 1. SNP visualization in the SynTView application (upper left), dynamically linked to a reconstructed artificial sequence from site sequence variations (bottom left) and to a SNP density map per gene and per strain (right).

Both in synteny and in SNP visualizations, the various views that can be generated are interactively linked together: a user selection (genes, genome region) in one panel dynamically triggers either a related selection or a view modification in other panels.

4 Conclusion

The most important asset of SynTView is the interactivity inherent to the use of the Flash technology. In addition, there is a tight integration between SynTView and the GenoList environment, and dynamic interactions between the various views. Further developments will take into account the management of incomplete genomes and multiploid organisms. SynTView is freely available for download at the URL: http://genopole.pasteur.fr/SynTView. Documentation, tutorials and demonstration sites are also provided.

References

[1] E.V. Koonin, L. Aravind and A.S. Kondrashov, The impact of comparative genomics on our understanding of evolution. *Cell*, 101:573-576, 2000.

[2] Flare library: http://flare.prefuse.org/.

[3] C.T. Lopes, M. Franz, F. Kazi, S.L. Donaldson, Q. Morris and G.D. Bader, Cytoscape Web: an interactive web-based network browser. *Bioinformatics*, 26:2347-2348, 2010.

[4] MEDEA, Broad Institute: http://www.broadinstitute.org/annotation/medea/.

[5] P. Lechat, L. Hummel, S. Rousseau and I. Moszer, GenoList: an integrated environment for comparative analysis of microbial genomes. *Nucleic Acids Res.*, 36:D469-474, 2008.

Mining Databanks to Analyse Functional and Taxonomic Diversity of Sequences

Alain MEIL, Goulven KERBELLEC and Patrick DURAND
Korilog Sarl, 4 rue Gustave Eiffel, 56230 Questembert, France
pdurand@korilog.com

The process of exploring banks of biological sequences responds to an important issue: given a new set of sequences to study, it is a question of locating the most similar sequences in a bank of known sequences. It is by recognizing these homologs and the associated information that possible functions and origins are usually predicted.

As well as sequence banks, biologists use additional databases (functional annotations, taxonomic bases, etc.), which are essential for in silico analyses of sequence diversity. The quantity of data available for studying genomes has become such that laboratories are faced with significant remote access issues to data providers. Transfer times, execution times for analysis tools on remote servers, and the user interfaces for exploring the results of these tools no longer live up to the needs of laboratories for studies involving huge batches of sequences.

Laboratories now have the ability to rapidly create vast repositories of sequences. The appearance of specialized databanks for a specific biological theme or even for a laboratory can be imagined. Biologists using these data will have a real need for Tools capable of effectively managing and analyzing their databanks.

KoriBlast is a software platform especially designed to gather within a single application the tools needed to explore sequence diversity, either functional or taxonomic. It provides the biologists with tools to prepare reference annotated databanks, to run databank search jobs in batch mode (either locally or on remote clusters), to query and to visualize the results. In connection with Pathway Explorer (provided by our partner Genostar), users can interactively explore metabolic pathways out of databank search results. During the presentation, some use cases will illustrate the capabilities of the KoriBlast Platform.

Session 7 : Algorithms and Evolution

Conférence invitée

Marie-France Sagot

INRIA Grenoble Rhône-Alpes & Université Claude Bernard, Lyon, France.

Towards an Algorithmic and Mathematical Exploration of Symbiosis

Symbiosis is described as a close relationship between different biological species, often of a long term nature. It is a pervasive phenomenon. It has for instance been estimated that 50% of all known species are parasites, that is maintain a symbiotic relation with another species from which they benefit while the partner in the relation is harmed. And it is believed that close to a 100% of all plants and animals are parasitised as individuals, in general by more than one species. Indeed, there are thought to be 10 times more bacterial cells in a human body than human cells (Savage, *Annual Review of Microbiology*, 1977). The idea of humans, and other animals or plants, as "superorganisms with an internal ecosystem of diverse symbiotic microbiota and parasites" has thus been advanced (Nicholson, *Nature Biotechnology*, 2004) and raises the issue of what is an individual, and what is species identity.

Symbiosis, or at least its extent, role and precise nature are controversial but symbiosis appears also essential to understand some of the most fundamental evolutionary and functional questions related to living organisms. The enormous variety in the observed types of pair- and multi-wise symbiotic relations, and the fact that these relationships touch upon almost every aspect of biology, from molecular to ecological, raise formidable mathematical and computational issues that should keep a computational biologist busy for decades.

This talk will survey part of the work we have done on this issue, and some of the questions we wish to address in the coming years.

Influence Function for Robust Phylogenetic Reconstruction

Mahendra MARIADASSOU[1] and Avner BAR-HEN[2]

[1] Laboratoire MIG- INRA, Bât. 233, Domaine de Vilvert, 78352 Jouy-en-Josas, Cedex , France
mahendra.mariadassou@jouy.inra.fr

[2] Laboratoire MAP5, UMR8145 CNRS, 45 rue des saints-pères, 75270 Paris, Cedex 06, France
avner.bar-hen@mi.parisdescartes.fr

Abstract *Phylogenies in short, are the most convenient way to describe the relationship between different species and are widely used in several fields of biology: comparative genomics, epidemiology, conservation biology, etc. However, most inferences drawn from phylogenies are accurate only if the reconstructed phylogeny itself is accurate. For a given reconstruction bias, robust phylogenies are preferred to non robust ones. We are concerned here with the loss of robustness induced by outliers. One way to mitigate this loss is to detect and remove outliers from the dataset.*

We advocate the use of empirical influence functions to detect influent characters and taxa, which are prone to be outliers, and their removal from the data set to build robust phylogenies. Three data sets (Zygomycetes, placental mammals, T-box gene family) show that maximum likelihood phylogenies are not robust and that removing as few as a handful of outliers can significantly increase the robustness of a tree, as measured by average bootstrap values.

Keywords Biostatistics, Phylogeny, Influence Function, Outliers, Robustness.

1 Introduction

Phylogenies are an essential tool in many fields of biology and it is thus crucial to reconstruct accurate phylogenies and moreover to assess the uncertainty associated with these phylogenies. The most frequent way of doing so is to use bootstrap replicates of the alignment and to compute bootstrap values [1] of inner branches. This approach produces a global index of uncertainty that captures, among others, the variability induced by sampling of characters. However, bootstrap probabilities should be handled with caution as they do not have a clear-cut statistical interpretation [2]. Moreover, the sampling of character is not the only cause for uncertainty in the inferred phylogeny: taxon sampling is also known to impact the accuracy of phylogenetic analysis[3]. Outlying characters resulting for example from alignment artifacts as well as rogue taxa can introduce bias in the reconstruction process which leads. If the bias is strong enough, measures of variability based on random resampling, such as bootstrap values, can be blind to the influence of these characters. Here, we use influence function to systematically investigate the influence of a given character and/or a given taxon on the inferred phylogeny.

2 Methods

We work with the maximum likelihood (ML) framework under which all characters $\mathbf{X} = (X_1, \ldots, X_n)$ of an alignment are considered as random variables independently drawn from the same distribution Q on a sample space $\mathcal{A}$ (for an alignment made of s taxa and nucleotide characters, $\mathcal{A} = \{A, C, G, T\}^s$). To each topology T with branch lengths $\mathbf{b}_T$, we can associate a probability distribution $P(.; T, \mathbf{b}_T)$ on $\mathcal{A}$. The goal of ML phylogenetic inference is to find the tree $(\hat{T}, \hat{\mathbf{b}}_{\hat{T}})$ that minimizes the Kullback-Leibler divergence between Q (unknown and replaced by the empirical distribution Q_n of the X_i) and $P(.; T, \mathbf{b}_T)$ or similarly that maximizes the per-character log-likelihood $L(X_1, \ldots, X_n; T, \mathbf{b}_T)$ of the alignment under tree $(T, \mathbf{b}_T)$:

$$L(\mathbf{X}; T, \mathbf{b}_T) = L(X_1, \ldots, X_n; T, \mathbf{b}_T) = \frac{1}{n} \sum_{i=1}^{n} \log P(X_i; T, \mathbf{b}_T).$$

Using definitions first introduced in the robustness literature [4], we define the influence $IF(X_i)$ of character X_i as the normalized shift in per-character log-likelihood induced by the removal of that character:

$$IF(X_i) = (n-1)[L(\mathbf{X}; \hat{T}, \hat{\mathbf{b}}_{\hat{T}}) - L(\mathbf{X}_{-i}; \hat{T}_{-i}, \hat{\mathbf{b}}_{\hat{T}_{-i}})]$$

where $\mathbf{X}_{-i}$ is the alignment deprived of character X_i and $(\hat{T}_{-i}, \hat{\mathbf{b}}_{\hat{T}_{-i}})$ is the tree reconstructed from $\mathbf{X}_{-i}$. Characters with a high positive $IF(X_i)$ have a phylogenetic signal that strongly conflicts the signal coming from the rest of the alignment and are potential outliers.

Similarly, we define the influence $TII(T_j)$ of taxon T_j as

$$TII(T_j) = d(\hat{T}^{-j}, \widehat{T^{-j}})$$

where d is a distance between trees, $\hat{T}^{-j}$ is the tree reconstructed on the complete alignment and then pruned of taxon T_j and finally $\widehat{T^{-j}}$ is the tree reconstructed on the alignment deprived of taxon T_j from the start. Taxa with a high $TII(T_j)$ strongly change the topology when included in the alignment and are potential rogue taxa.

3 Results

We applied our method to three datasets to detect outliers and propose alternative phylogenies: 16S rRNA from fungi (Zygomycetes and Chytridiomycetes), mtDNA from placental mammals and the T-Box transcription factor gene family in bilaterians. Our results on Zygomycetes [5] show that outliers have a strong effect on the inferred phylogeny. The two most influential characters affect the topology in no less than 20 inner branches (out of 155) and reduce the log-likelihood of the ML tree by more than 100 units. Excluding these two characters leads to a robust topology, which has higher bootstrap values and reduced influence values for the remaining characters. Our results on placental mammals [6] show that rogue taxa also have a strong impact on the resulting topology and confirms the status of the guinea-pig as a rogue taxa for this dataset. Our results on the T-Box gene family enable us to identify a subset of taxa for which the phylogeny can be reconstructed with greater confidence (higher bootstrap values) for both recent and old branches than in the initial alignment. The resulting phylogeny is then used to ascertain the position of a new T-Box gene within the family.

References

[1] J. Felsenstein, Confidence Limits on Phylogenies: An Approach Using the Bootstrap, *Evolution*, 39:783-791, 1985.

[2] E. Susko, Bootstrap support is not first-order correct, *Systematic Biology*, 58, 211–233, 2009.

[3] T. A. Heath, S. M. Hedtke and D. M. Hillis, Taxon sampling, the accuracy of phylogenetic analyses, *J. Mol. Evol.*, 46:239-257, 2008.

[4] F. R. Hampel, The influence curve and its role in robust estimation, *JASA*, 69:383–393, 1974.

[5] A. Bar-Hen, M. Mariadassou, M.-A. Poursat and P. Vandenkoornhuyse, Influence function for robust phylogenetic reconstructions, *Molecular Biology and Evolution*, 25:869–873, 2008

[6] M. Mariadassou, A. Bar-Hen and H. Kishino, Taxon Influence Index, *Systematic Biology*, in press, 2011.

Character Trimming Algorithms to Build a Compositionally Homogeneous Character Subset from a Multiple Sequence Alignment

Application to Phylogenetic Tree Inference

Alexis CRISCUOLO

INSTITUT PASTEUR,
Unité de Biologie Moléculaire du Gène chez les Extrêmophiles, Département de Microbiologie,
25 rue du Dr Roux, 75015 Paris, France
alexis.criscuolo@pasteur.fr

Abstract *When a multiple sequence alignment suffers from a strong compositional heterogeneity across sequences, some biasing effect may alter the phylogenetic analysis of this dataset. Indeed, many phylogenetic tree inference methods artificially group together sequences with strong compositional similarities. To circumvent this problem, this paper presents new algorithms that search for a compositionally homogeneous character subset from a multiple sequence alignment. These polynomial algorithms progressively remove the characters that are involved in the heterogeneous composition across sequences. The benefit of this approach is illustrated by simulation results, as well as two real-case phylogenetic studies.*

Keywords Compositional heterogeneity across sequences, matched-pairs Stuart test of marginal homogeneity, polynomial algorithms, phylogenetic tree.

1 Introduction

Most of the phylogenetic tree reconstruction methods are based on an alignment of homologous sequences. Therefore, the quality of a multiple sequence alignment can have a strong impact on the accuracy of the inferred phylogenetic tree, some characters (i.e. columns or sites inside the multiple sequence alignment) being ambiguously aligned or too variable [25,28,29]. In order to minimize errors in the phylogenetic tree, a current approach is to detect and remove these problematic characters prior to the phylogenetic analysis [7,8,10,12,30,38]. However, even if the homology is correctly depicted by a multiple sequence alignment, a strong heterogeneous composition of character states (e.g. nucleotides, amino acids) across sequences may cause systematic errors during the phylogenetic analysis. Indeed, sequences with very similar composition are often inaccurately grouped together by some phylogenetic tree inference methods [17,24,31,36]. This bias is often corrected by two distinct techniques: the first method performs a character state recoding[10,11,21,31,37], whereas the second approach allows using phylogenetic tree reconstruction methods that do not invoke the stationarity assumption (i.e. considering that the character state composition do not remains constant over all lineages) [4,5,15,17,18,19,26,27,39]. Knowing that each of these two techniques sometimes leads to incorrect results (e.g. [16,31,36,40]), this paper suggests using a third approach by removing the problematic characters (i.e. involved in the heterogeneous composition across sequences) in order to produce compositionally homogeneous data that can be analyzed with traditional phylogenetic tree inference methods.

Among the numerous statistical tests that assess whether aligned sequences are compositionally homogeneous (see [24]), the matched-pairs Stuart test of marginal homogeneity [35] is used throughout this paper. This test, described in subsection 2.1, allows assessing the marginal symmetry in the character state frequency table built from a pair of aligned sequences. If the Stuart test does not assess homogeneous composition, the subsection 2.2 describes an algorithm that allows building a compositionally homogeneous character subset from the two aligned sequences. This polynomial algorithm iteratively removes one character from the pairwise alignment as long as the Stuart test does not assess the homogeneous composition between the two aligned sequences. In subsection 2.3, this algorithm is extended to the case of multiple sequence alignments. Simulation results are described in section 3, as well as two real-case phylogenetic studies, in order to illustrate the usefulness of this approach for nucleotide sequences.

2 Methods

2.1 Assessing Compositional Heterogeneity Between Two Aligned Sequences

Given an alphabet Σ of character states (e.g. nucleotides, amino acids), and the pairwise alignment of length ℓ of two sequences i and j over the alphabet Σ, the matched-pairs Stuart test of marginal homogeneity can be used to assess whether the two aligned sequences have similar composition [1,35]. This test requires the preliminary building of the frequency matrix $(\mathbf{F}_{xy})$, where $\mathbf{F}_{xy}$ is the number of times the character state x of the sequence i is aligned with the character state y of the sequence j. The matrix $\mathbf{F}$ is then a square matrix of size $|\Sigma|\times|\Sigma|$. The Stuart test then verifies the null hypothesis of marginal symmetry in the matrix $\mathbf{F}$ [35]:

$$\mathbf{F}_{x.} = \mathbf{F}_{.x}, \qquad \forall x \in \Sigma, \tag{1}$$

where $\mathbf{F}_{x.}$ and $\mathbf{F}_{.x}$ are the sum of $\mathbf{F}_{xy}$ and $\mathbf{F}_{yx}$ over y, respectively. The test statistic is computed by using some simple matrix operations. Given the column vector $(\mathbf{V}_x)$ containing any $\nu_\Sigma = |\Sigma| - 1$ of the values $\mathbf{V}_x = \mathbf{F}_{x.} - \mathbf{F}_{.x}$, the $\nu_\Sigma \times \nu_\Sigma$ variance/covariance matrix $(\mathbf{S}_{xy})$ of the elements of $\mathbf{V}$ is computed with the two formulae $\mathbf{S}_{xx} = \mathbf{F}_{x.} + \mathbf{F}_{.x} - 2\mathbf{F}_{xx}$ and $\mathbf{S}_{xy} = -\mathbf{F}_{xy} - \mathbf{F}_{yx}$. After computing the transpose $\mathbf{V}^{\mathrm{T}}$ of $\mathbf{V}$ and the inverse $\mathbf{S}^{-1}$ of the matrix $\mathbf{S}$, the Stuart statistic

$$X_{ij} := \mathbf{V}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{V} \tag{2}$$

is interpreted as a χ^2 value with $\nu_\Sigma = |\Sigma| - 1$ degrees of freedom [35]. The null hypothesis (1) is then verified if $X_{ij} \leq X_\Sigma$, where X_Σ is a known constant value. With a 10% critical value, $X_\Sigma :\approx 6.251$ for the nucleotide alphabet Σ (i.e. $\nu_\Sigma = 3$), and $X_\Sigma :\approx 27.203$ for the amino acid alphabet Σ (i.e. $\nu_\Sigma = 19$). Verifying the null hypothesis (1) from the pairwise alignment of length ℓ of two sequences i and j over the alphabet Σ then requires $O(\ell)$ time complexity to compute the contingengy matrix $\mathbf{F}$, $O(|\Sigma|^2)$ to compute $\mathbf{V}$ and $\mathbf{S}$, and $O(|\Sigma|^3)$ to perform the matrix operations in formula (2). Therefore, assessing the compositional homogeneity between two aligned sequences with the Stuart test requires $O(\ell + |\Sigma|^3)$ time complexity.

2.2 Compositionally Homogeneous Character Subset from Two Aligned Sequences

The formula (2) allows the level of heterogeneous composition to be quantified: the value X_{ij} is as large as the two aligned sequences i and j are compositionally heterogeneous. Based on this observation, this subsection describes a method that allows selecting a compositionally homogeneous character subset from two aligned sequences i and j (i.e. a character subset implying $X_{ij} \leq X_\Sigma$). If a character $c = (c_i, c_j)$ is removed from the two aligned sequences i and j, the two character states c_i and c_j are removed from the sequences i and j, respectively. Consequently, the entry $\mathbf{F}_{c_i c_j}$ is decremented by 1, which modifies the value computed by formula (2). Let $X_{ij}^{(c_i,c_j)}$ denote this value, and $\gamma_{ij}(c_i, c_j)$ the criterion defined as

$$\gamma_{ij}(c_i, c_j) := \begin{cases} X_{ij} \left(X_{ij} - X_{ij}^{(c_i,c_j)}\right) & \text{if } c_i, c_j \in \Sigma, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

If $\gamma_{ij}(c_i, c_j) < 0$, then removing the character (c_i, c_j) leads to a worsening of the heterogeneous composition between sequences i and j, i.e. $X_{ij} < X_{ij}^{(c_i,c_j)}$; reciprocally, if $\gamma_{ij}(c_i, c_j) > 0$, removing (c_i, c_j) leads to an improvement of the compositional homogeneity. Moreover, as the difference $X_{ij} - X_{ij}^{(c_i,c_j)}$ is multiplied by the factor X_{ij}, the criterion $\gamma_{ij}(c_i, c_j)$ allows quantifying how the character (c_i, c_j) is involved in the compositional heterogeneity between the two aligned sequences i and j. Note that $\gamma_{ij}(c_i, c_j) := 0$ if $c_i \notin \Sigma$ or $c_j \notin \Sigma$, which could occur when c_i or c_j is a gap or an unknown character state.

As X_{ij} is already known, formula (3) only requires the computation of $X_{ij}^{(c_i,c_j)}$. When (c_i, c_j) is removed, decrementing $\mathbf{F}_{x_i y_j}$ by 1 is performed in $O(1)$ computing time. Therefore, computing $X_{ij}^{(c_i,c_j)}$ only requires the computation of formula (2), with $O(|\Sigma|^3)$ time complexity (see subsection 2.1). However, the two aligned sequences i and j being defined over the alphabet Σ, there exist only $|\Sigma|^2$ distinct characters $c = (x, y)$. Precomputing all possible values $X_{ij}^{(x,y)}$ — then every values $\gamma_{ij}(x, y)$ — requires $O(|\Sigma|^5)$ time complexity. This computation time is acceptable, given that standard alphabets are small (i.e. $|\Sigma| = 4$ for nucleotides, $|\Sigma| = 20$ for amino acids). However, it should be stressed that precomputing the $|\Sigma|$ values $\gamma_{ij}(x, x)$ is unnecessary, because constant characters (i.e. diagonal entries in $\mathbf{F}$) are not taken into account by the Stuart test. Moreover, precomputing $\gamma_{ij}(x, y)$ is only required for the characters (x, y) that exist in the pairwise alignment,

i.e. $\mathbf{F}_{xy} > 0$. The precomputing step of required values $\gamma(x,y)$ is then performed in $O(\eta|\Sigma|^3)$ time complexity, where η is the number of non-diagonal and non-zero entries of the matrix $\mathbf{F}$.

By iteratively removing (one of) the character $\tilde{c} = (\tilde{c}_i, \tilde{c}_j)$ that maximizes the criterion γ_{ij}, building a compositionally homogeneous character subset from two aligned sequences i and j can be easily performed. The $O(|\Sigma|^2)$ required values of $\gamma_{ij}(x,y)$ being precomputed, searching for one of the character $\tilde{c} = \text{argmax}_{c=1,2,\cdots,\ell}(\gamma_{ij}(c_i,c_j))$ is performed with $O(\ell)$ time complexity. After the removal of $\tilde{c}$, the update of $\mathbf{F}$ (i.e. $\mathbf{F}_{\tilde{c}_i\tilde{c}_j} := \mathbf{F}_{\tilde{c}_i\tilde{c}_j} - 1$) and X_{ij} (i.e. replaced by $X_{ij}^{(\tilde{c}_i,\tilde{c}_j)} = X_{ij} - \gamma_{ij}(\tilde{c}_i,\tilde{c}_j)/X_{ij}$) is done with $O(1)$ time complexity. This algorithm iteratively performs these different polynomial steps (i.e. precomputing $\gamma_{ij}(x,y)$, removing the character $\tilde{c}$, updating $\mathbf{F}$ and X_{ij}) until $X_{ij} \leq X_\Sigma$.

2.3 Compositionally Homogeneous Character Subset from a Multiple Sequence Alignment

The algorithm described in subsection 2.2 can be easily extended for more than two aligned sequences. Let $(\mathbf{F}_{ij,xy})$ denote the contingency matrix $(\mathbf{F}_{xy})$ for each pair of sequences i,j from an alignment of n sequences of length ℓ over the alphabet Σ. From $(\mathbf{F}_{ij,xy})$, the $n(n-1)/2$ different values X_{ij} are computed with formula (2). From $\mathbf{F}$ and X_{ij}, all the required values $\gamma_{ij}(x,y)$ are precomputed with formula (3) for each pair of sequences i,j and each pair of character states (x,y) such that $x \neq y$ and $\mathbf{F}_{ij,xy} > 0$. Let $\sigma(c)$ denote a criterion for each character $c = 1,2,\cdots,\ell$ in the multiple sequence alignment defined as:

$$\sigma(c) := \sum_{i<j} \gamma_{ij}(c_i,c_j). \tag{4}$$

If $\sigma(c) > 0$, then the removal of the character c from the multiple sequence alignment leads to the decrease of the $n(n-1)/2$ values X_{ij} on average; reciprocally, if $\sigma(c) < 0$, then removing c leads to an overall worsening of the compositional heterogeneity across sequences. Consequently, removing a character that maximizes the criterion (4) will produce a character subset with an improved compositional homogeneity.

The character trimming algorithm that builds a compositionally homogeneous character subset from a multiple sequence alignment iteratively removes one of the characters $\tilde{c}$ that maximize σ as computed by the formula (4). After the removal of $\tilde{c}$, the matrix $(\mathbf{F}_{ij,xy})$ and the $n(n-1)/2$ values X_{ij} are updated for each pair of sequence i,j, and this procedure is iteratively performed until the remaining character subset is compositionally homogeneous, i.e. $X_{ij} \leq X_\Sigma$ for each pair of sequences i,j. This algorithm is summerized below:

Algorithm 1

(a) • `For each` pair of sequences i,j
(b) • Computing the matrix $(\mathbf{F}_{ij,xy})$;
(c) • Computing X_{ij} with formula (2) ;
(d) • `While` $\exists\, i,j$ such that $X_{ij} > X_\Sigma$
(e) • `For each` pair of sequences i,j
(f) • `For each` pair of character states (x,y) such that $x \neq y$ and $\mathbf{F}_{ij,xy} > 0$
(g) • Computing $\gamma_{ij}(x,y)$ with formula (3) ;
(h) • Removing the character $\tilde{c}$ that maximizes the criterion (4) ;
(i) • `For each` pair of sequences i,j
(j) • $\mathbf{F}_{ij,\tilde{c}_i\tilde{c}_j} := \mathbf{F}_{ij,\tilde{c}_i\tilde{c}_j} - 1$;
(k) • $X_{ij} := X_{ij} - \gamma_{ij}(\tilde{c}_i,\tilde{c}_j)/X_{ij}$;

Steps (a-c) in Algorithm 1 correspond to the calculations described in subsection 2.1 performed for each pair of sequences; they are then performed in time $O(n^2(\ell + |\Sigma|^3))$. Step (g) requiring $O(|\Sigma|^3)$ time complexity (see subsection 2.2), steps (e-g) are performed in time $O(n^2|\Sigma|^5)$. Step (h) requires the computation in time $O(n^2)$ of the criterion (4) for each of the $O(\ell)$ characters in the multiple sequence alignment; this step then requires $O(n^2\ell)$ time complexity. Finally, steps (i-k) perform constant time update operations for each pair of sequences, then requiring $O(n^2)$ time complexity. Note that the end condition can be verified during step (k). Steps (e-k) then require $O(n^2(|\Sigma|^5 + \ell))$ time complexity, and the Algorithm 1 runs in time $O(n^2\ell(|\Sigma|^5 + \ell))$.

The Algorithm 1 being based on the different Stuart statistics X_{ij}, it should be recalled that these values are interpreted as a χ^2 value with $\nu_\Sigma = |\Sigma| - 1$ degrees of freedom (subsection 2.1). It should also be stressed that a large sampling is required to verify $X_{ij} \sim \chi^2(\nu_\Sigma)$, especially when ν_Σ is large (e.g. [42]). Therefore, the Algorithm 1 is expected to fulfill its purpose (i.e. building a character subset that is effectively compositionally

homogeneous) when used with a large number of characters (e.g. $\ell \geq$ 1,000 non-constant characters). Unfortunately, its use with large sets of characters involves important running times (see section 3). However, faster running times are expected by simultaneously removing more than one character from the multiple sequence alignment at each iteration, i.e. steps (e-k) in Algorithm 1. Indeed, if the characters c are sorted according to their $\sigma(c)$ values, an alternative is to simultaneously remove the character set $\tilde{C}$ containing the $m > 1$ characters $\tilde{c}$ that maximize the criterion (4). This new algorithm is summerized below:

Algorithm 2

(a) • `For each` pair of sequences i, j
(b) • Computing the matrix $(\mathbf{F}_{ij,xy})$;
(c) • `While` $\ell > 0$
(d) • `For each` pair of sequences i, j
(e) • Computing X_{ij} with formula (2) ;
(f) • `For each` pair of character states (x, y) such that $x \neq y$ and $\mathbf{F}_{ij,xy} > 0$
(g) • Computing $\gamma_{ij}(x, y)$ with formula (3) ;
(h) • `If` $X_{ij} \leq X_{\Sigma}$ for all pair of sequences i, j `Then STOP` ;
(i) • `For each` character $c = 1, 2, ..., \ell$ of the multiple sequence alignment
(j) • Computing $\sigma(c)$ with formula (4) ;
(k) • Computing the set $\tilde{C}$ such that $|\tilde{C}| = m$ and $\forall c \notin \tilde{C}, \forall \tilde{c} \in \tilde{C}, \sigma(c) \leq \sigma(\tilde{c})$;
(l) • `For each` character $\tilde{c} \in \tilde{C}$
(m) • Removing $\tilde{c}$ from the multiple sequence alignment ;
(n) • `For each` pair of sequences i, j
(o) • $\mathbf{F}_{ij,\tilde{c}_i\tilde{c}_j} := \mathbf{F}_{ij,\tilde{c}_i\tilde{c}_j} - 1$;
(p) • $\ell := \ell - m$;

As steps (d-h), (i-k), and (l-p) in Algorithm 2 require $O(n^2|\Sigma|^5)$, $O(\ell(n^2+\log m))$, and $O(n^2 m)$ time complexity, respectively, the Algorithm 2 runs in time $O\left(n^2\ell(|\Sigma|^5+\ell) + \ell^2 \log m\right)$. However, it runs faster than the Algorithm 1 (see subsection 3.1). Indeed, one iteration (d-p) in Algorithm 2 removes m characters in time $O\left(n^2(|\Sigma|^5+\ell)+\ell \log m\right)$, whereas the Algorithm 1 removes the same number of characters by performing m iterations (e-k) in time $O\left(n^2(|\Sigma|^5+\ell)\right)$. The main difference between the two algorithms is the precomputing of the different values $\gamma_{ij}(x, y)$, which is performed after the removal of $m = 1$ character by the Algorithm 1, and $m > 1$ characters by the Algorithm 2. Consequently, the Algorithm 2 is expected to stop after the removal of more characters than the Algorithm 1. However, by setting m to $\lfloor \ell/1{,}000 \rfloor$ in the provided implementation of the Algorithm 2, very close results are observed between the two algorithms during simulations.

3 Results

3.1 Simulation Results

A method searching for compositionally homogeneous character subsets from multiple sequence alignments was previously described in [10]. First, this method, named the Stationary-based Character Trimming (SCT), progressively removes the characters ranked in function of their decreasing entropy values in order to obtain a first compositionally homogeneous character subset; secondly, SCT completes this subset by adding the remaining characters sorted following a criterion closely related to (4) (see [10] for more details). The efficiency of SCT was assessed from artificially generated nucleotide sequences with heterogeneous base compositions. From a 4-taxon tree $uv|xy$ (external and internal branches of lengths 0.475 and 0.025, respectively), the evolution of sequences of length ℓ = 10,000 nucleotide character states was simulated with the evolutionary model F81 [13]. For each value p = 0, 10, ..., 50, p% of the sequence length was simulated with 80% GC-content for the external branches corresponding to the leaves u and x, and 80% AT-content for the external branches corresponding to the leaves v and y. For the other characters (i.e. $100-p$% of the sequence length), equal relative character state frequencies were used to generate compositionally homogeneous regions. For each value p, 200 alignments of four sequences u ,v, x and y were simulated following this protocol. From each of these intial data, a *Maximum Likelihood* (ML) phylogenetic tree was inferred with the software PhyML [20] (evolutionary model F81). Knowing that base composition is as heterogeneous as p is large, the Table 1 clearly shows that the quartet tree $uv|xy$ is almost never inferred from the sequences generated with $p > 20$% (see [24] for similar findings). However, when the compositionally heterogeneous data are trimmed by the method SCT, the quartet tree $uv|xy$ is almost always revovered (Table 1; see [10] for more details).

	Proportion p of characters with heterogeneous composition					
	0%	10%	20%	30%	40%	50%
Average running time						
SCT	2.6 s	9.3 s	10.7 s	16.8 s	21.4 s	25.1 s
Algorithm 1	1.6 s	12.3 s	27.3 s	46.1 s	70.2 s	88.3 s
Algorithm 2	1.3 s	2.9 s	4.9 s	7.5 s	10.6 s	13.7 s
Average proportion of removed characters						
SCT	2.96%	6.04%	10.82%	15.88%	21.09%	26.26%
Algorithm 1	0.08%	2.56%	6.12%	10.77%	16.45%	22.13%
Algorithm 2	0.10%	2.60%	6.18%	10.91%	16.57%	22.27%
Overlapping rate between removed character subsets						
Algorithms 1 and 2	64.67%	68.89%	90.08%	80.20%	75.88%	81.53%
Proportion of correctly inferred trees						
Initial data	100%	95.0%	50.5%	2.0%	0.5%	0%
SCT	100%	98.0%	93.5%	96.0%	91.0%	91.0%
Algorithm 1	100%	100%	100%	100%	96.0%	94.5%
Algorithm 2	100%	100%	100%	100%	96.0%	94.5%

Table 1. Simulation results. The average running times and proportions of removed characters are reported for the Algorithms 1 and 2, as well as for the Stationary-based Character Trimming method (SCT [10]). Given the subsets R_1 and R_2 of characters removed by the Algorithms 1 and 2, respectively, the overlapping rate is estimated by $|R_1 \cap R_2|/|R_1 \cup R_2|$. The proportion of correctly inferred trees is reported for the initial (non-trimmed) data, as well as for the character subsets returned by SCT, and Algorithms 1 and 2.

The same simulation protocol was used to compare the respective performance of the Algorithms 1 and 2, and the observed results were reported in Table 1. This shows the Algorithm 1 allows building the largest compositionally homogeneous character subsets, but at the cost of very important running times. However, as expected, the Algorithm 2 leads to character subsets of similar size, with large overlapping rates (i.e. $> 60\%$), but with faster running times. Interestingly, Algorithms 1 and 2 remove less characters than SCT (Table 1). This shows that directly removing characters from a multiple sequence alignment (i.e. Algorithms 1 and 2) allows building larger compositionally homogeneous character subsets than the inverse procedure (i.e. adding characters in an initial compositionally homogeneous character subset) performed by SCT. Consequently, Algorithms 1 and 2 lead to more correctly inferred quartet trees than the use of SCT (Table 1).

3.2 The Rokas, Williams, King and Carroll (2003) Dataset of Yeasts

A 106-gene dataset of nucleotide sequences gathered from eight yeast genomes (seven *Saccharomyces* taxa, and *Candida albicans* used as ougroup taxon [34]) was observed to suffer from a heterogeneous GC-content bias [31]. Indeed, when these 106 phylogenetic markers are concatenated, they form a supermatrix of 127,026 characters that leads to a monophyletic relationship between the two taxa *S. bayanus* and *S. kudriavzevii* in the phylogenetic tree (Fig. 1A) inferred by optimizing the *Minimum Evolution* (ME) criterion from the pairwise GTR [33,43] and LogDet [26,27] evolutionary distance estimates (see [31] for more details; see also [9] for

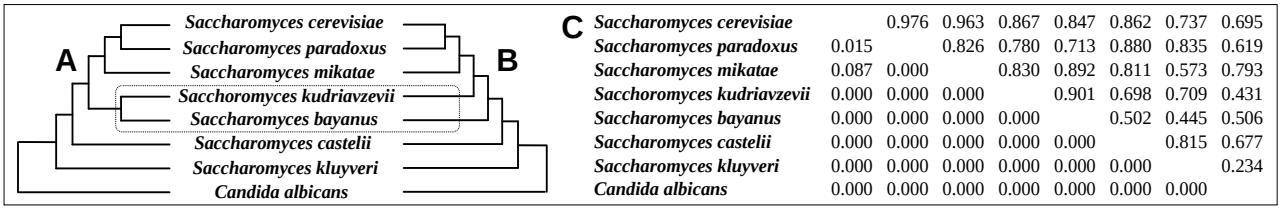
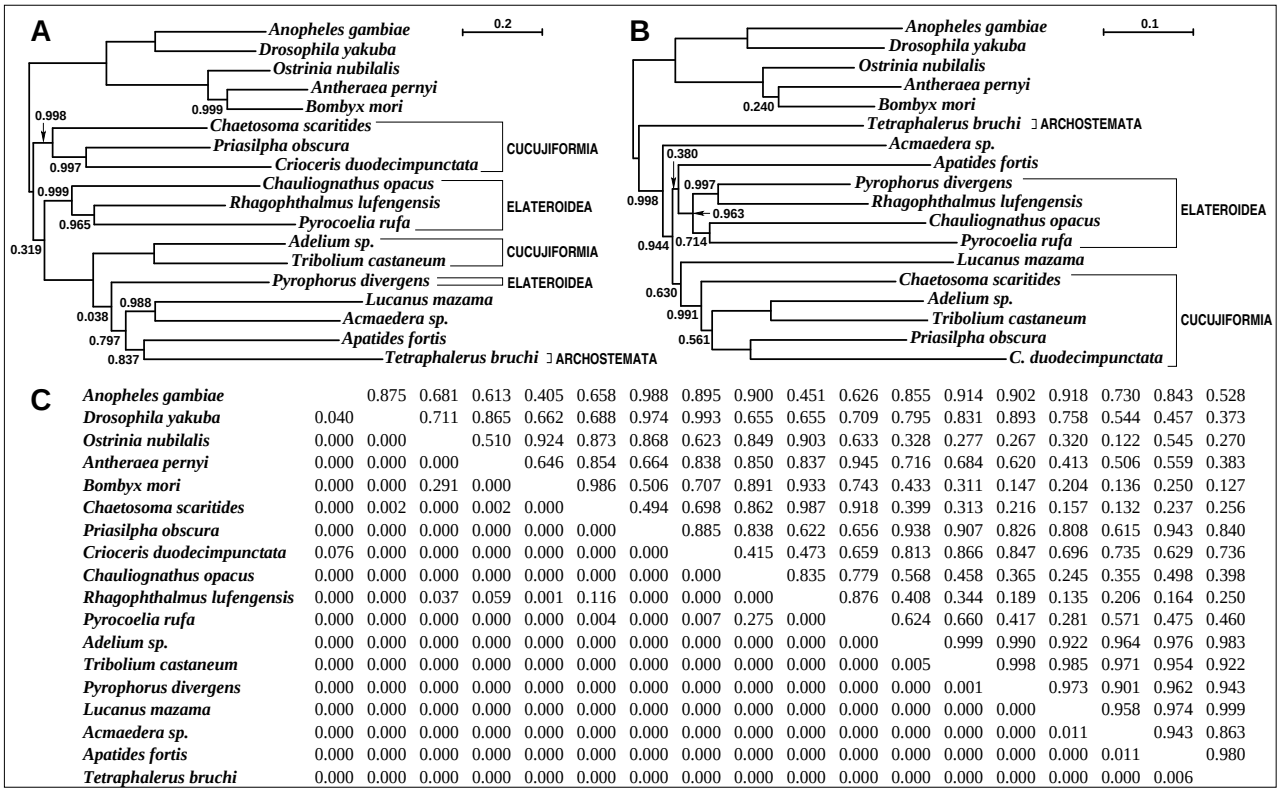


Figure 1. ME trees inferred from a GC-content heterogeneous dataset [34]. Evolutionary distances: GTR [33,43] and LogDet [26,27]. A: initial data ($\ell = 127{,}026$); B: compositionally homogeneous character subset ($\ell = 116{,}826$). For each character set and each evolutionary distance estimate, all branches are 100%-supported by a bootstrap analysis [14]. C: Stuart test $\chi^2(X_{ij}|\nu_\Sigma)$ values estimated from the character sets A (below the diagonal) and B (above); if $X_{ij} \leq X_\Sigma$, then $\chi^2(X_{ij}|\nu_\Sigma) \geq 0.1$.



C

Anopheles gambiae		0.875	0.681	0.613	0.405	0.658	0.988	0.895	0.900	0.451	0.626	0.855	0.914	0.902	0.918	0.730	0.843	0.528
Drosophila yakuba	0.040		0.711	0.865	0.662	0.688	0.974	0.993	0.655	0.655	0.709	0.795	0.831	0.893	0.758	0.544	0.457	0.373
Ostrinia nubilalis	0.000	0.000		0.510	0.924	0.873	0.868	0.623	0.849	0.903	0.633	0.328	0.277	0.267	0.320	0.122	0.545	0.270
Antheraea pernyi	0.000	0.000	0.000		0.646	0.854	0.664	0.838	0.850	0.837	0.945	0.716	0.684	0.620	0.413	0.506	0.559	0.383
Bombyx mori	0.000	0.000	0.291	0.000		0.986	0.506	0.707	0.891	0.933	0.743	0.433	0.311	0.147	0.204	0.136	0.250	0.127
Chaetosoma scaritides	0.000	0.002	0.000	0.002	0.000		0.494	0.698	0.862	0.987	0.918	0.399	0.313	0.216	0.157	0.132	0.237	0.256
Priasilpha obscura	0.000	0.000	0.000	0.000	0.000	0.000		0.885	0.838	0.622	0.656	0.938	0.907	0.826	0.808	0.615	0.943	0.840
Crioceris duodecimpunctata	0.076	0.000	0.000	0.000	0.000	0.000	0.000		0.415	0.473	0.659	0.813	0.866	0.847	0.696	0.735	0.629	0.736
Chauliognathus opacus	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000		0.835	0.779	0.568	0.458	0.365	0.245	0.355	0.498	0.398
Rhagophthalmus lufengensis	0.000	0.000	0.037	0.059	0.001	0.116	0.000	0.000	0.000		0.876	0.408	0.344	0.189	0.135	0.206	0.164	0.250
Pyrocoelia rufa	0.000	0.000	0.000	0.000	0.000	0.004	0.000	0.007	0.275	0.000		0.624	0.660	0.417	0.281	0.571	0.475	0.460
Adelium sp.	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000		0.999	0.990	0.922	0.964	0.976	0.983
Tribolium castaneum	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.005		0.998	0.985	0.971	0.954	0.922
Pyrophorus divergens	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001		0.973	0.901	0.962	0.943
Lucanus mazama	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000		0.958	0.974	0.999
Acmaedera sp.	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.011		0.943	0.863
Apatides fortis	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.011		0.980
Tetraphalerus bruchi	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.006	

Figure 2. ML trees inferred from a GC-content heterogeneous dataset [36]. ML evolutionary model: GTR+ Γ_8+I+F [33,43]. A: initial data (ℓ = 11,633, $-\log lk/\ell$ = 11.80, ML parameters Γ_8 = 1.030, I = 0.246, base frequencies A: 0.360, C: 0.123, G: 0.096, T: 0.420); B: compositionally homogeneous character subset (ℓ = 7,975, $-\log lk/\ell$ = 8.02, ML parameters Γ_8 = 0.860, I = 0.319, base frequencies A: 0.336, C: 0.113, G: 0.128, T: 0.422). Confidence values at each branch represent approximate likelihood ratio test values (aLRT [3]) as estimated by PhyML. No confidence value corresponds to the maximum aLRT value (i.e. 1.0). C: see Fig. 1C.

similar findings). However, numerous phylogenetic analyses of the same dataset (e.g. [2,6,9,10,31,32,34,41]) lead to a different tree (Fig. 1B). It was shown that the incorrect grouping between *S. bayanus* et *S. kudriavzevii* is due to the compositional heterogeneity across sequences that is sufficiently important to mislead the ME criterion [31]. This was corroborated by a recent re-analysis [10]: a compositionally homogeneous subset of 114,105 characters was built with the method SCT from this dataset, and the likely correct ME yeast tree (Fig. 1B) was recovered from this character subset. From the same dataset, the Algorithm 2 has allowed building a compositionally homogeneous subset of 116,825 characters in ∼13 minutes with a 2-GHz Intel(R) Core(TM)2 Duo with 2.0 Gb RAM. In agreement with the previous simulation results (subsection 3.1), the Algorithm 2 allows building larger character subset than SCT. As expected, these 116,825 compositionally homogeneous characters (Fig. 1C) allow inferring the likely correct ME tree (Fig. 1B).

3.3 The Sheffield, Song, Cameron and Whiting (2009) Dataset of Beetles

A 13-gene dataset was recently assessed to suffer from a strong compositional bias [36]. This dataset was built from the coding regions of 18 mitochondrial genomes of beetles (13 coleopterans, as well as 3 lepidopterans and 2 dipterans used as outgroup). The concatenation of the 13 multiple sequence alignments leads to a supermatrix of 11,655 nucleotide characters. This was analyzed with various phylogenetic tree reconstruction methods, and the inferred trees were evaluated using the three following criteria: (i) the taxon *Tetraphalerus bruchi* (suborder *Archostemata*) must emerge first within coleopterans, and each taxon group (ii) *Cucujiformia* and (iii) *Elateroidea* must make up a monophyletic subtree (e.g. [22,23]; see [36] for details). It has been observed [36] that phylogenetic tree reconstruction methods invoking the stationarity assumption infer incorrect trees (i.e. that do not verify the above three criteria). Moreover, incorrect trees were also inferred by several methods using non-stationary models of sequence evolution (i.e. [4,5]). However, several other non-stationary tree inference approaches (i.e. [15,19]) have led to trees in agreement with the criteria (i-iii).

A subset of 7,975 characters with homogeneous compositions was built with the Algorithm 2 in ∼17 minutes with a 2-GHz Intel(R) Core(TM)2 Duo with 2.0 Gb RAM. This character subset as well as the initial dataset were analyzed with the script morePhyML[1] (stationary model; see Fig. 2) to infer ML phylogenetic trees (Fig. 2). In agreement with previous results (see [36]), the use of an evolutionary model invoking the stationary assumption leads to a ML phylogenetic tree (Fig. 2A) that does not verify the three previous criteria. Indeed, the taxon *Tetraphalerus bruchi* does not emerge first within coleopterans, contrary to the criterion (i). Moreover, the tree in Fig. 2A does not verify the monophyly of *Cucujiformia* and *Elateroidea*, contrary to criteria (ii) and (iii). These errors appear to be due to compositional heterogeneity across sequences in the dataset (Fig. 2C). Indeed, when applied on the compositionally homogeneous character subset, the same ML approach leads to a phylogenetic tree (Fig. 2B) verifying the three criteria. Moreover, despite the removal of ∼31% characters, the criteria (i-iii) are strongly supported (i.e. corresponding confidence value at branches >90%; Fig. 2B).

4 Conclusion

This paper introduces novel polynomial algorithms to build compositionally homogeneous character subsets from sequence alignments. When applied on datasets suffering from a strong heterogeneity of character state composition across sequences, these algorithms build character subsets that allow minimizing compositional biases when analyzed with standard phylogenetic tree inference methods. Therefore, they represent an alternative approach to other existing methods that reduce compositional biases during phylogenetic inference (character state recoding, non-stationary models of sequence evolution). These two new algorithms are implemented in the software BMGE[2], replacing the less efficient Stationary-based Character Trimming method [10].

These two algorithms can be easily modified to use other statististal tests that assess the marginal symmetry in a contingency matrix (see e.g. [42]), or other useful symmetry properties (see [1]). It will also be interesting to adapt these algorithms to build compositionally homogeneous datasets by removing character states inside compositionally biased sequences instead of performing character trimming. Finally, despite evidence of heterogeneous composition across sequences in several studied amino acid datasets (not shown), the character subsets selected by the two algorithms do not lead to different phylogenetic trees. Therefore, it would also be interesting to assess the level of resistance of current phylogenetic tree reconstruction methods against the biasing effect due to compositional heterogeneity across amino acid sequences.

Acknowledgements

I thank Simonetta Gribaldo and the BMGE team of the French Pasteur Institute for support. Many thanks to Corrine Maufrais for providing ftp home pages. This research was supported by the PhyloCyano project of the Agence Nationale de la Recherche (ANR; 07-JCJC-0094-01).

References

[1] F. Ababneh, L.S. Jermiin, C. Ma and J. Robinson, Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences, *Bioinformatics*, 22:1225-31, 2006.
[2] C. Ané, B. Larget, D.A. Baum, S.D. Smith and A. Rokas, Bayesian estimation of concordance among gene trees, *Mol. Biol. Evol.*, 24:412-26, 2007.
[3] M. Anisimova and O. Gascuel, Approximate Likelihood-ratio test for branches: a fast, accurate, and powerful alternative, *Syst. Biol.*, 55:539-52, 2006.
[4] S. Blanquart and N. Lartillot, A Bayesian compound stochastic process for modelling nonstationary and nonhomogeneous sequence evolution, *Mol. Biol. Evol.*, 23:2058-71, 2006.
[5] B. Bousseau and M. Gouy, Efficient likelihood computations with non-reversible models of evolution, *Syst. Biol.*, 55:756-68, 2006.
[6] J.G. Burleigh, A.C. Driskell and M.J. Sanderson, Supertree bootstrapping methods for assessing phylogenetic variation among genes in genome-scale data sets, *Syst. Biol.*, 55:426-40, 2006.
[7] S. Capella-Gutiérez, J.M. Silla-Martínez and T. Gabaldón, trimAl: a tool for automated alignment triming in large-scale phylogenetic analyses, *Bioinformatics*, 25:1972-3, 2009.
[8] J. Castresana, Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis, *Mol. Biol. Evol.*, 17:540-52, 2000.

[1] `ftp://ftp.pasteur.fr/pub/gensoft/projects/morePhyML/`
[2] `ftp://ftp.pasteur.fr/pub/gensoft/projects/BMGE/`

[9] A. Criscuolo and C.J. Michel, Phylogenetic inference with weighted codon evolutionary distances, *J. Mol. Evol.*, 68:377-92, 2009.
[10] A. Criscuolo and S. Gribaldo, BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments, *BMC Evol. Biol.*, 10:210, 2010.
[11] T.M. Embley, M. van der Giezen, D.S. Horner, P.L. Dyal and P.G. Foster, Mitochondria and hydrogenosomes are two forms of the same fundamental organelle, *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, 358:191-203, 2003.
[12] A.W.M. Dress, C. Flamm, G. Fritzsch, S. Grünewald, M. Kruspe, S.J Prohaska and P.F. Stadler, Noisy: Identification of problematic columns in multiple sequence alignments. *Algorithms Mol. Biol.*, 3:7, 2008.
[13] J.Felsenstein,Evolutionary tree from DNA sequences:a maximum likelihood approach,*J.Mol.Evol.*,17:368-76,1981.
[14] J. Felsenstein, Confidence limits on phylogenies: an approach using the bootstrap, *Evolution*, 39:783-91, 1985.
[15] P.G. Foster, Modeling compositional heterogeneity, *Syst. Biol.*, 53:485-95, 2004.
[16] P.G. Foster and D.A. Hyckey, Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions, *J. Mol. Evol.*, 48:284-90, 1999.
[17] N. Galtier and M. Gouy, Inferring phylogenies from DNA sequences of unequal base composition, *Proc. Natl. Acad. Sci. USA*, 92:11317-21, 1995.
[18] N. Galtier and M. Gouy, Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis, *Mol. Biol. Evol.*, 15:871-9, 1998.
[19] V. Gowri-Shankar and M. Rattray, A reversible jump method for Bayesian phylogenetic inference with a nonhomogeneous substitution model, *Mol. Biol. Evol*, 24:1286-99, 2007.
[20] S. Guindon and O. Gascuel, A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood, *Syst. Biol.*, 52:696-704, 2003.
[21] I. Hrdy, R.P. Hirt, P. Dolezal, L. Bardonova, P.G. Foster, J. Tachezy and T.M. Embley, Trichomonas hydrogenosomes contain the NADH deshydrogenase module of mitochondria complex I, *Nature*, 432:618-22, 2004.
[22] J. Hughes, S.J. Longhorn, A. Papadopoulou, K. Theodorides, A. de Riva, M. Mejia-Chang, P.G. Foster and A.P. Vogler, Dense taxonomic EST sampling and its applications for molecular systematics of the Coleoptera (beetles), *Mol. Biol. Evol.*, 23:268-78, 2006.
[23] T. Hunt, J. Bergsten, Z. Levkanicova, A. Papadopoulou, O.S. John, R. Wild, P.M. Hammond, D. Ahrens, M. Balke and M.S. Caterino, A comprehensive phylogeny of beetles reveals the evolutionary origins of a superradiation, *Science*, 318:1913-6, 2007.
[24] L.S. Jermiin, S.Y.W. Ho, F. Ababneh, J. Robinson and A.W.D. Larkum, The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated, *Syst. Biol.*, 53:638-43, 2004.
[25] J.A. Lake, The order of sequence alignment can bias the selection of tree topology, *Mol.Biol.Evol.*,8:378-85, 1991.
[26] J.A. Lake, Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances, *Proc. Natl. Acad. Sci. USA*, 91:1455-9, 1994.
[27] P.J. Lockhart, M.A. Steel, M.D. Hendy and D. Penny, Recovering evolutionary trees under a more realistic model of sequence evolution, *Mol. Biol. Evol.*, 11:605-12, 1994.
[28] D.A. Morrison and J.T. Ellis, Effects of nucleotide sequence alignment on phylogeny estimation: a case study on 18S rDNAs of Apicomplexa, *Mol. Biol. Evol.*, 14:428-41, 1997.
[29] T.H. Ogden and M.S. Rosenberg, Multiple sequence alignment accuracy and phylogenetic inference, *Syst. Biol.*, 55:314-28, 2006.
[30] O. Penn, E. Privman, G. Landan, D. Graur and T. Pupko, An alignment confidence score capturing robustness to guide tree uncertainty, *Mol. Biol. Evol.*, 27:1759-67, 2010.
[31] M.J. Phillips, F. Delsuc and D. Penny, Genome-scale phylogeny and the detection of systematic biases, *Mol. Biol. Evol.*, 21:1455-8, 2004.
[32] F. Ren, H. Tanaka and Z. Yang, An empirical examination of the utility of codon-substitution models in phylogeny reconstruction, *Syst. Biol.*, 54:808-18, 2005.
[33] R. Rodríguez, J.L. Oliver, A. Marin and J.R. Medina, The general stochastic model of nucleotide substitution, *J. Theor. Biol.*, 142:485-501, 1990.
[34] A. Rokas, B.L. Williams, N. King and S.B. Carroll, Genome-scale approaches to resolving incongruence in molecular phylogenies, *Nature*, 425:798-804, 2003.
[35] A.Stuart,A test for homogeneity of the marginal distributions in a two way classification.*Biometrika*,42:412-6,1955.
[36] N.C. Sheffield, H. Song, S.L. Cameron and M.F. Whiting, Nonstationary evolution and compositional heterogeneity in beetle mitochondrial phylogenomics, *Syst. Biol.*, 58:381-94, 2009.
[37] E. Susko E and A.J. Roger, On reduced amino acid alphabets for phylogenetic inference, *Mol. Biol. Evol.*, 24:2139-50, 2007.
[38] G. Talavera and J. Castresana, Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments, *Syst. Biol.*, 56:564-77, 2007.
[39] K. Tamura and S. Kumar, Evolutionary distance estimation under heterogeneous substitution pattern among lineages, *Mol. Biol. Evol.*, 19:1727-36, 2002.
[40] R. Tarrío, F. Rodríguez-Trelles and F.J. Ayala, Shared nucleotide composition biases among species and their impact on phylogenetic reconstructions of the Drosophilidae, *Mol. Biol. Evol.*, 18:1464-73, 2001.
[41] D.J. Taylor and W.H. Piel, An assessment of accuracy, error, and conflict with support values from genome-scale phylogenetic data, *Mol. Biol. Evol.*, 21:1534-7, 2004.
[42] H. Uesaka, Validity and applicability of several tests for comparing marginal distributions of a square table with ordered categories, *Behaviormetrika*, 30:65-78, 1991.
[43] Z. Yang, Estimating the pattern of nucleotide substitution, *J. Mol. Evol.*, 39:105-11, 1994.

A Comparison of Regression Models for Testing QTL/eQTL co Location

Xiaoqiang WANG[1,2], Jean Michel ELSEN[3], Olivier FILANGI[1,2] and Pascale LE ROY[1,2]
[1]INRA, UMR0598, 65 rue de St-Brieuc, 35042, Rennes, France
[2]Agrocampus Ouest, UMR0598, 65 rue de St-Brieuc, 35042, Rennes, France
{xiaoqiang.wang, olivier.filangi, pascale.leroy}@rennes.inra.fr
[3]INRA, UMR0631, Chemin de Borde Rouge, BP27, 31326, Castanet-Tolosan, France
jean-michel.elsen@toulouse.inra.fr

Abstract- *Exploring the genetic architecture in biology relies on the accurate identification of quantitative trait loci (QTL) and gene expression quantitative loci (eQTL). Currently, as more and more eQTL have been mapped in various species, the importance of searching the eQTL which co locate with a given QTL is highlighted. A number of statistical approaches are proposed to look for the eQTL of interest, in particular test whether a same QTL affects two different traits or two different linked QTL affect separately two traits. All the methods were useful and widely applied in eQTL studies. However, the required conditions when using these methods hinder their applications for all the circumstances. This study firstly discusses the suitable conditions for the methods existing in the literature. Then we propose two methods, making use of the trait and QTL information, in order to look for the eQTL co located with the QTL. In addition, the parameters influencing the power for above eQTL mapping are discussed. Finally, through simulated data in an outbreed population, we verify the required conditions for the existed method and validate our methods in co location studies.*

Keywords QTL, eQTL, regression, co location.

1 Introduction

Since the concept of genetical genomics emerged and the heritability for gene expression is proved [1,2], numerous expression quantitative trait loci (eQTL) have been found in different species such as yeast [2], mice [3], human [4,5], maize [6], chicken [7], porcine [8], rainbow trout [9]. In the meantime, the test of the co location of these eQTL with quantitative trait loci (QTL) which affect complex traits is likely to be more and more important. Indeed, one of the eQTL mapping purposes is to identify the causative mutations (or polymorphisms) of QTL effects. The objective is to improve the efficiency of genomic selection [10]. In that framework, eQTL detected in the neighborhood of one QTL are used to refine the QTL location. Moreover, exploring causal phenotype networks has become a hot topic in recent years. However, for these both purposes, it would be necessary to decide which genes, among those analyzed, have an eQTL co located with one given QTL.

In livestock, the linkage analysis is currently used to detect QTL and several statistical strategies based on this technique have been developed for eQTL mapping [11,12,13]. This kind of approaches takes into account the huge dimensionality of transcriptomic dataset but, generally, does not provide much attention to inaccuracy of the estimated eQTL locations. Indeed, the transcriptome analysis is generally carried out on a limited number of animals, i.e. between 50 and 300 animals. In such designs, the wide confidence intervals of eQTL locations remain a difficult and challenging problem [14,15]. Consequently, after the eQTL have been mapped applying linkage analysis, further investigations are necessary to focus on the eQTL detected which are co located with a given QTL.

Standard scenario to test the eQTL and QTL co location would be the test of the null hypothesis of one pleiotropic QTL *versus* the general hypothesis of two close linked QTL in multitrait analyses [13,16,17,18,19,20]. This statistical technique was applied to a broad range of QTL co location studies and it is powerful, under some circumstances. In particular, Jiang and Zeng [16] or Gilbert and Le Roy [19] pointed out the favorable situation for joint analysis, i.e. when the product of the residual correlation and the effects of the two QTL is negative. In that case, the power of the joint analysis is always higher than that of the single trait analyses and the accuracy of QTL locations is better. Furthermore, the application of these

techniques, especially in outbreed populations, is possible only for some variables jointly analyzed because of the CPU time consuming to estimate all the parameters of these multivariate models [19].

A strategy thus consists to preliminary test, gene by gene, the co location of eQTL and QTL to reduce as much as possible the eQTL list. For that purpose, "fast" algorithms are necessary and thus regression models appear as good opportunities. To the best of our knowledge, several regression methods have appeared in the literature for testing the co location of eQTL and QTL. These linear regression methods are available, and easily implemented in practice, whereas little attention was paid to the required conditions which guarantee their efficiency. The goals of this paper are to 1) explore the concrete situations such that the regression methods appeared in the literature can be successfully applied; 2) develop new methods to generalize these regression methods to any situations; 3) demonstrate the feasibility of the proposed methods in an outbreed population context through simulations. In this framework, we will firstly describe the relevant statistical model in a backcross population. There, some statistical properties of QTL detection test statistics will be discussed. Secondly, the required conditions for the existing methods will be searched. Then, approaches will be proposed to overcome the limitations due to the assumptions in regression method for detected eQTL co located with QTL. Finally, using simulated data set, we will verify if the analytical results obtained on a backcross population are also valuable for an outbreed population.

2 Materials and Methods

2.1 Model and Notations

Traditional linkage analysis methodologies for mapping QTL are mainly interval mapping methods [23,24]. In a first step, we will consider the statistical model constructed in a backcross population.

Now let's assume that a true QTL is located at t_0 and a true eQTL at t. Let T_i denotes the trait and G_i the gene expression for individual i. Let g_{it0} denotes the genotype of i at the t_0 location, i.e. the genotype at the QTL, and g_{it} the genotype of i at the t location, i.e. the genotype at the eQTL. We will suppose here that g can take two values, 1 if i is homozygous qq or -1 if i is heterozygous Qq. Then the relationship among T and G and the genotypes at the QTL and eQTL can be explained by the following system of linear equations:

$$\begin{cases} T_i = \mu + \frac{a}{2} g_{it0} + \varepsilon_t, \\ G_i = \nu + \frac{b}{2} g_{it} + \varepsilon_g, \end{cases} \quad (1)$$

where μ and ν are the overall means of the T and G random variables respectively, a and b are respectively the allelic substitution effects of the QTL and of the eQTL, $(\varepsilon_t, \varepsilon_g)$ follows a binormal distribution with a mean 0 and a covariance matrix $\begin{pmatrix} \sigma_t^2 & \rho\sigma_t\sigma_g \\ \rho\sigma_t\sigma_g & \sigma_g^2 \end{pmatrix}$.

It should be noted that the trait and the gene expression may share other loci or be influenced by the same environmental factors. This fact implies that the correlation among the residual errors cannot be neglected in co location studies.

We can also define the part of the T variance explained by the QTL (H_t^2) and the part of the G variance explained by the eQTL (H_g^2) as two coefficients of heritability:

$$H_t^2 = \frac{a^2}{a^2+4\sigma_t^2}; \qquad H_g^2 = \frac{b^2}{b^2+4\sigma_g^2}. \quad (2)$$

Naturally, as the heritability increases, the power of QTL detection and the accuracy of QTL location are higher [14,15].

2.2 Regression Methods in the Literature

Several regression methods have been proposed to test the QTL/eQTL co location.

2.2.1 Method I

One of ideas consists of performing linkage analysis on the residual trait value corrected by the transcript expression value [8]. This approach comprises two steps. The first step consists in predicting the trait *T* using the linear regression by expression *G*, $\hat{T} = \hat{\beta}_0 + \hat{\beta}_1 G$. Next, the residual error, $Z = T - \hat{T}$, is submitted to a linkage analysis. If a QTL is yet detected analyzing the new trait *Z*, the conclusion is that the eQTL and the QTL are not at the same location.

Following the model (1), *Z* may be written:

$$Z = m + \frac{a}{2} g_{t0} - \frac{b}{2} \widehat{\beta}_1 g_t + \varepsilon_z,$$

where *m* is the overall mean for trait *Z*, $\hat{\beta}_1 = cov(G,T)/var(G)$, $\varepsilon_z \sim \mathcal{N}\left(0,\ \sigma_t^2 + \widehat{\beta}_1^2 \sigma_g^2 - 2\widehat{\beta}_1 \rho \sigma_t \sigma_g\right)$. Under the null hypothesis ($g_{it0} = g_{it}$),

$$Z = m + \frac{1}{2} \frac{a\sigma_g^2 - b\rho\sigma_t\sigma_g}{\frac{b^2}{4} + \sigma_g^2} g_t + \varepsilon_z.$$

Thus, this approach would be efficient only under the condition:

$$a\sigma_g^2 - b\rho\sigma_t\sigma_g = 0. \tag{3}$$

2.2.2 Method II

A similar approach is to consider the expression trait as a covariable in QTL detection procedures [21] following the linear model:

$$T = \mu' + \frac{a'}{2} g_{t0} + \beta G + \varepsilon_t'.$$

Similarly, if a significant QTL is detected applying this model, the eQTL is not considered as co located with the QTL. For simplicity, we use the Haley-Knott regression method to analyze the efficiency of this approach. In this case, the genotypes of QTL g_{t0} will be replaced by their expectation conditional on the information of markers, say $\tilde{g}_{t0} = \mathbb{E}(g_{t0}|M)$. The distribution of $\tilde{g}_{t0}$ is depicted in the table below. The estimated partial regression coefficients can be written as

$$\widehat{a'} = 2\,\frac{\mathbb{E}G^2\mathbb{E}T\tilde{g}_{t0} - \mathbb{E}G\tilde{g}_{t0}\mathbb{E}TG}{\mathbb{E}\tilde{g}_{t0}{}^2\mathbb{E}G^2 - (\mathbb{E}G\tilde{g}_{t0})^2}, \qquad \hat{\beta} = \frac{\mathbb{E}\tilde{g}_{t0}{}^2\mathbb{E}TG - \mathbb{E}G\tilde{g}_{t0}\mathbb{E}T\tilde{g}_{t0}}{\mathbb{E}\tilde{g}_{t0}{}^2\mathbb{E}G^2 - (\mathbb{E}G\tilde{g}_{t0})^2}. \tag{4}$$

Thus, it could be demonstrate that this approach had the same condition (3) to be efficient because a' should be 0 under the null hypothesis ($g_{it0} = g_{it}$).

In conclusion, these two similar strategies may be not efficient because the statistical tests applied are biased, i.e. the null hypothesis is not properly defined.

2.2.3 Method III

To avoid this problem, an alternative method was proposed by Li et al. [22]. In this approach, the trait *T* was analyzed two times, i.e. two log of odds ratio (LOD) scores were calculated, taking into account or not the gene expression *G* as a covariable in the model. Then, the difference between these two scores was used as test statistics: if the ΔLOD is large in absolute value, the authors suggested that the variable *G* is causally connected to the trait *T* and, thus, to the QTL. The threshold to determine if ΔLOD is large could be obtained by simulation under the null hypothesis, i.e. when the gene expression *G* was unrelated to the trait.

So, the QTL was mapped two times successively with the two linear models:

$$T = \mu + \frac{a}{2} g_{t0} + \varepsilon_t, \qquad T = \mu' + \frac{a'}{2} g_{t0} + \beta G + \varepsilon_t'$$

and the maximum LOD scores were recorded. Next, the comparison of these two LOD scores was used to test whether QTL and eQTL were located at the same location. Comparing to the preceding strategies, the power of the ΔLOD test statistics is here the capacity of the test statistics to conclude $t = t0$. However, when $t = t0$, note that if the estimation of β is 0, then ΔLOD is also small. Hence, we obtain from the expression (4) a necessary but not sufficient condition to guarantee the efficiency of this approach, that is

$$\rho \neq 0. \tag{5}$$

It means that this method will be efficient when *G* and *T* share same loci or when *G* and *T* are affected by the same environmental factors.

2.3 New Regression Methods

As shown above, some restricted conditions hinder the application of the previous methods to all the situations. In order to overcome this barrier, we proposed in this section two new methods. If their principle remains the same than in the preceding methods, i.e. adjust one trait by the other one, we tried here to improve the efficiency exploiting the information available after the QTL and eQTL primo detection, respectively on *T* and *G* variables. Indeed, several estimates of the parameters influencing the power are available after this initial step: $\hat{a}$ the QTL effect, $\hat{b}$ the eQTL effect and $t0$ the QTL location. Generally, the t eQTL location is less accurate than $t0$ because the eQTL detection was performed on a reduced design compared to QTL detection. Consequently, in the following propositions, we chose to rather adjust *G* by *T* than *T* by *G*.

2.3.1 Method IV

In this approach, a QTL detection is carried out on the variable *Z*:

$$Z = G - \frac{\hat{b}}{\hat{a}}T. \tag{6}$$

If the LRT score exceeds a rejection threshold, then this suggests that the QTL and the eQTL are not at the same location. The rejection threshold can be obtained by simulation under the null hypothesis where the trait *Z* is supposed to be unrelated to any QTL.

Under the model (6), the part of the *Z* trait variance due to the QTL is:

$$H^2 = \frac{1}{1+\frac{1}{\theta_{t,t0}}\left(\frac{\sigma_g^2}{b^2}+\frac{\sigma_t^2}{a^2}-\frac{2\rho\sigma_t\sigma_g}{ab}\right)}. \tag{7}$$

From the expression (7), we can see that the power of QTL detection on *Z* depends on the following parameters: the recombinant rate between the locations of the QTL ($t0$) and of the eQTL (t), the part of the *T* variance explained by the QTL, the part of the *G* variance explained by the eQTL and the correlation between the residual errors. In particular, when $t0 = t$, i.e. when $\theta_{t,t0} = 0$, H^2 will tend towards 0 as the power of QTL detection on *Z*. In the other cases, when the distance between the QTL and the eQTL increases, the power of QTL detection on *Z* increases.

2.3.2 Method V

In this approach, we assume that the QTL position ($t0$) is known. Then, we build a new trait as:

$$Z = G - \frac{\hat{b}}{2}\mathbb{E}(g_{t0}|M), \tag{8}$$

where *M* denote the genotypes at the two flanking markers of the t_0 location, $\mathbb{E}(g_{t0}|M)$ is the expectation of QTL genotype conditional to *M* for each individual. In a backcross population of $AAQQBB \times AaQqBb$, let l and r denote the positions of the left and of the right markers respectively, then the distribution of $\mathbb{E}(g_{t0}|M)$ is:

$\mathbb{E}(g_{t0}\|M)$	Probability	Markers genotype
$(\theta_{l,t0} + \theta_{t0,r} - 1)/(1-\theta_{l,r})$	$(1-\theta_{l,r})/2$	AABB

$(\theta_{l,t0} - \theta_{t0,r})/\theta_{l,r}$	$\theta_{l,r}/2$	AABb
$-(\theta_{l,t0} + \theta_{t0,r} - 1)/(1-\theta_{l,r})$	$(1-\theta_{l,r})/2$	AaBb
$-(\theta_{l,t0} - \theta_{t0,r})/\theta_{l,r}$	$\theta_{l,r}/2$	AaBB

where $\theta_{x,y}$ denotes the recombination rate between the two locations x, y using Haldane distance.

As above, if the LRT score does not exceed the rejection threshold, then the eQTL will be considered to be co located with the QTL.

The part of the Z variance explained by the QTL is:

$$H^2 = \frac{1}{1+\frac{1}{\mathbb{E}[\mathbb{E}(g_{t0}|M)]^2}\frac{1}{\theta_{t,t0}}\frac{\sigma_g^2}{b^2}}, \quad (9)$$

where $\mathbb{E}[\mathbb{E}(g_{t0}|M)]^2$ is equal to the variance of the variable $\mathbb{E}(g_{t0}|M)$ which is a constant when the density of markers is fixed and the position t_0 is known. According to the expression (9), it can be seen that, as in the preceding method, the power of QTL detection is null when $\theta_{t,t0}$ is null and that the power increases when the distance between the QTL and the eQTL increases. However, the dependency of the power to the QTL effect a or to the residual correlation ρ is solving with this strategy.

2.4 Application in Outbreed Populations

We have shown some required conditions, for the existing regression methods or the proposed two new methods, to be efficient to declare the QTL/eQTL co location in a backcross population. In this section, we will investigate the more complex situation of an outbreed population, through simulated data set using the QTLMap software [25,26,27,28,29,30,31,32], to validate or not the preceding results in that other context. For the purpose of the comparison of the methods mentioned above, we perform simulations under the one situation of population and genetic map. We simulated one QTL and one eQTL in a linkage group of 60cM where 13 markers are equally distributed, i.e. with a distance of 5cM between two successive markers. The population is composed of 5 independent sire families, with 2 unrelated dams per sire and 30 progeny per dam. The location of the QTL is fixed at 7cM and its effect is always assumed to be 1 phenotypic standard deviation of the trait T (σ). In order to see if the results obtained in the backcross population are also valid in the outbreed population, 2 different values (σ and $-\sigma$) for the eQTL effect and 3 different eQTL locations (7cM, 32cM and 57cM) are envisaged. The rejection threshold to QTL detection, which depends on the family structure and markers information, is gotten with the help of simulations of the T trait, with a polygenic heritability of 0.25, and under the null hypothesis of no QTL on the studied linkage group.

The boxplots in Figure 1 show the distribution of LRT scores for each method. Because the distribution of LRT score for method II is very similar as that for method I, we omitted it here. From these distributions, it can be seen that 1) for Method I: only when $a\sigma_g^2 - b\rho\sigma_t\sigma_g$ is small, this method is useful, 2) for Method III: when $\rho = 0$, this method loses efficiency, 3) for Method IV: this method is applicable to any circumstance. The power of the QTL detection depends still on the effect of eQTL (b) and the residual correlation (ρ), 4) for Method V: this method is also applicable to any circumstance. The power of the QTL detection does not depend on the sign of the eQTL effect or on the residual correlation (ρ). We observe that all these results from the simulation studies in an outbreed population are consistent with the analytical results in a backcross population. Hence, our propositions for the test of the QTL/eQTL co location in that context seem to be also valuable.

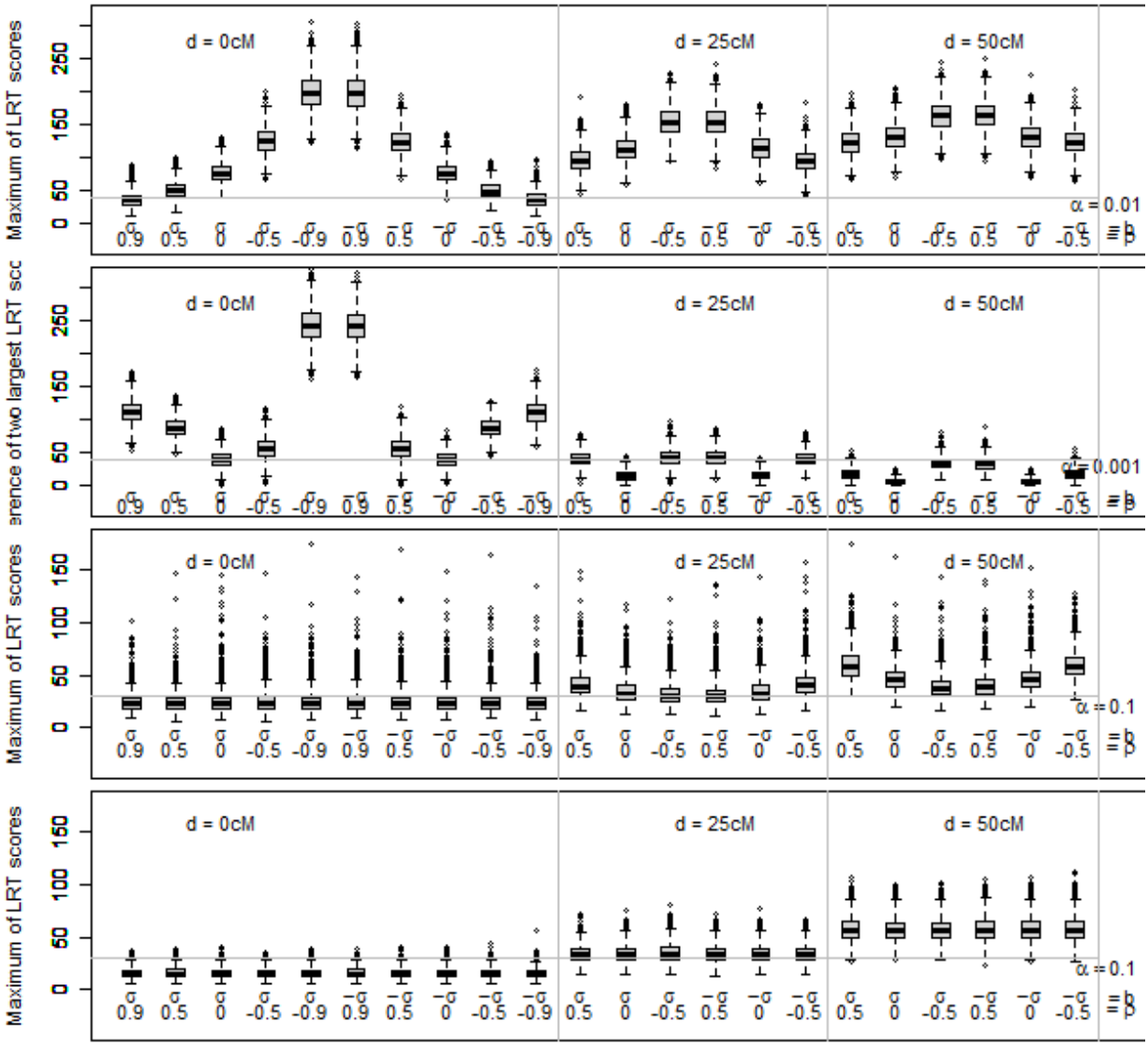


Figure 1. Boxplots of LRT scores depending on the parameters: the distance between the QTL and the eQTL (d), the eQTL effect (b) and the residual correlation (ρ).

3 Discussion

Understanding the genetic architecture in biology relies on the accurate identification of QTL and eQTL. In particular, the eQTL studies have contributed to the identification of causative mutations and the inference of the phenotypic networks. These applications highlight the importance of the estimation of eQTL location. So far, traditional linkage analysis methodologies were always applied to map eQTL in a broad range. Nevertheless, the location confidence intervals given by this kind of methods remains an important and challenging problem to find a precise location of eQTL. In this framework, we addressed to search the eQTL which co locate with a given QTL in this paper.

Many research studies have been carried out in the field of testing whether the same QTL could be affecting two traits or whether different QTL explain the observations. Generally, the methodologies for the statistical hypothesis testing were investigated in two ways: 1) making use of the linear regression between the trait and the gene expression, 2) using the multivariate analysis to test the hypothesis of one pleiotropic QTL versus two close linked QTL. These methods have been widely applied in practice; however, their required assumptions are either not given or ignored. In this context, we attempted to provide the situations in which these regression methods are efficient and the analytical results in the backcross population. A simulation study in an outbreed population demonstrated that the given circumstances for each method were feasible. Next, in order to eliminate the limitations caused by those required conditions, we developed two approaches based on the prior information on the traits and QTL. The analytical results and the simulation

study demonstrated the feasibility of these proposed methods. In particular, when the location of QTL is known, we suggest the last method, i.e. method V, to test whether an eQTL co locates truly at the position of the QTL.

Our studies concentrated on the comparison of regression models for testing QTL/eQTL co location. Further research should be investigated to reduce the dimension of eQTL co localizing with QTL by multiple testing when we use the method IV and V.

Acknowledgements

This work was funded and supported by SABRE ('cutting edge genomics for sustainable animal breeding') and the department of animal genetics INRA France. We thank the group of QTLMap software for assistance in simulation studies.

References

[1] R.C. Jansen and J.P. Nap, Genetical genomics: the added value from segregation. *TRENDS in Genetics* 17: 388-391, 2001.

[2] R.B. Brem, G. Yvert, R. Clinto and L. Kruglyak, Genetic Dissection of Transcriptional Regulation in Budding Yeast. *Science* 296: 752-755, 2002.

[3] E.E. Schadt, S.A. Monks, T.A. Drake, A.J. Lusis, N. Che, *et al.*, Genetics of gene expression surveyed in maize, mouse and man. *Nature* 422:297-302, 2003.

[4] S.A. Monks, A. Leonardson, H. Zhu, P. Cundiff, P. Pietrusiak, *et al.*, Genetic inheritance of gene expression in human cell lines. *Am. J. Hum. Genet.*, 75:1094-1105, 2004.

[5] M. Morley, C.M. Molony, T.M. Weber, J.L. Devlin, K.G. Ewens, R.S. Spielman and V.G. Cheung, Genetic analysis of genome-wide variation in human gene expression. *Nature* 430:743-747, 2004.

[6] C. Shi, A. Uzarowska, M. Ouzunova, M. Landbeck, *et al.*, Identification of candidate genes associated with cell wall digestibility and eQTL (expression quantitative trait loci) analysis in a Flint × Flint maize recombinant inbred line population. *BMC Genomics*, 8:22, 2007.

[7] G. Le Mignon, C. Desert, F. Pitel, S. Leroux, O. Demeure, *et al.*, Using transcriptome profiling to characterize QTL regions on chicken chromosome 5. *BMC Genomics* 10: 575, 2009.

[8] S. Ponsuksili, E. Murani, M. Schwerin, K. Schellander and K. Wimmers, Identification of expression QTL (eQTL) of genes expressed in porcine M. longissimus dorsi and associated with meat quality traits. *BMC Genomics* 8:22, 2010.

[9] Y. Le Bras, N. Dechamp, J. Montfort, A. Le Cam, F. Krieg, E. Quillet, P. Prunet and P. Le Roy, Acclimation to seawater in rainbow trout: QTL/eQTL approach for plasmatic ions and gill tissue. *Proc. 9th WCGALP*, Leipzig, PP4-120, 2010.

[10] G. Le Mignon, Y. Blum, O. Demeure, C. Diot, E. Le Bihan-Duval, P. Le Roy and S. Lagarrigue, Apports de la génomique fonctionnelle à la cartographie fine de QTL. *INRA Prod. Anim.*, 23(4): 343-358, 2010.

[11] J.D. Storey, J.M. Akey and L. Kruglyak, Multiple locus linkage analysis of genomewide in yeast. *PLoS Biol* 3(8): e267, 2005.

[12] P. Wang, J.A. Dawson, M.P. Keller, B.S. Yandell, N.A. Thornberry, *et al*., A model approach for expression quantitative trait loci (eQTL) mapping. *Genetics* 187: 611-621, 2010.

[13] G.Y. Sun and P. Schliekelman, 2010 A genetical genomics approach to genome scans increases power for QTL mapping. *Genetics* 110.123968

[14] G.A. Walling, C.S. Haley, M. Perez-Enciso, R. Thompson and P.M. Visscher, On the mapping of quantitative trait loci at marker and non-marker locations. *Genetical Research*, 79: 97-106, 2002.

[15] X.Q. Wang, J.M. Elsen, H. Gilbert, C. Moreno, O. Filangi and P. Le Roy The repercussions of statistical properties of interval mapping methods on eQTL detection. *Submission*.

[16] C.J. Jiang and Z.B. Zeng, Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics* 140: 1111-1127, 1995.

[17] S.A. Knott and C.S. Haley, Multitrait least squares for quantitative trait loci detection. *Genetics* 156: 899-911, 2000.

[18] E. E. SCHADT, J. Lamb, X. Yang, J. Zhu, S. Edwards et *al.*, An integrative genomics approach to infer causal associations between gene expression and disease. *Nature genetics* 37: 710-717, 2005.

[19] H. Gilbert and P. Le Roy, Comparison of three multitrait methods for QTL detection. *Genet. Sel. Evol.* 35: 281-304, 2003.

[20] H. Gilbert and P. Le Roy, Methods for the detection of multiple linked QTL applied to a mixture of full and half sib families. *Genet. Sel. Evol.* 39: 139–158, 2007.

[21] D.C. Kulp and M. Jagalur, Causal inference of regulator-target pairs by gene mapping of phenotypes. *BMC genomics* 7: 125, 2006.

[22] R.H. Li, S.W. Tsaih, K. Shockley, I.M. Stylianou, J. Wergedal *et al.*, Structural model analysis of multiple quantitative traits. *PLos Genetics* 2: e114, 2006.
[23] E.S. Lander. and D. Botstein, Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* 121: 185-191, 1989.
[24] C.S. Haley and S.A. Knott, A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* 69: 315-324, 1992.
[25] A. Legarra and R.L. Fernando, Linear models for joint association and linkage QTL mapping. *Genet Sel Evol*., 41:43, 2009.
[26] J.M. Elsen, O. Filangi, H. Gilbert, P. Le Roy and C. Moreno, A fast algorithm for estimating transmission probabilities in QTL detection designs with dense maps. *Genet Sel Evol*., 41:50, 2009.
[27] H. Gilbert, P. Le Roy, C. Moreno, D. Robelin and J. M. Elsen, QTLMAP, a software for QTL detection in outbred population. *Annals of Human Genetics*, 72(5): 694, 2008.
[28] H. Gilbert and P. Le Roy, Methods for the detection of multiple linked QTL applied to a mixture of full and half sib families. *Genet Sel Evol.*, 39(2):139-58, 2007.
[29] C.R. Moreno, J.M. Elsen, P. Le Roy and V. Ducrocq, Interval mapping methods for detecting QTL affecting survival and time–to–event phenotypes. *Genet. Res. Camb.*, 85 : 139-149, 2005.
[30] B. Goffinet, P. Le Roy, D. Boichard, J.M. Elsen and B. Mangin, Alternative models for QTL detection in livestock. III. Heteroskedastic model and models corresponding to several distributions of the QTL effect.. *Genet. Sel. Evol.*, 31, 341-350, 1999.
[31] B. Mangin, B. Goffinet, P. Le Roy, D. Boichard and J.M. Elsen, Alternative models for QTL detection in livestock. II. Likelihood approximations and sire marker genotype estimations. *Genet. Sel. Evol.*, 31, 225-237, 1999.
[32] J.M. Elsen, B. Mangin, B. Goffinet, D. Boichard and P. Le Roy, Alternative models for QTL detection in livestock. I. General introduction. *Genet. Sel. Evol*., 31, 213-224, 1999.

Session 8 : Genome Analysis

Conférence invitée

Mathieu Blanchette

McGill University, Montréal, Canada

Ancestral Mammalian Genome Reconstruction and its Uses Toward Annotating the Human Genome

With the number of sequenced mammalian genomes rapidly growing, the exciting prospect of being able to accurately infer ancestral genomes becomes within reach. In this presentation, I will discuss how ancestral DNA sequences can be inferred and how they can be then used to help addressing some key questions in genomics. Reconstructing ancestral sequences poses a number of algorithmic challenges. I will first describe some of our work on aligning orthologous sequence and inferring ancestral sequences, focusing on the accurate identification of insertions and deletions. Next, I will discuss how one can take advantage of the availability of inferred ancestral sequences to help at three important tasks: (i) identify non-coding sites under selection in the human genome; (ii) improve the detection of transcription factor binding sites; and (iii) determine the target gene(s) of long-range enhancers. Evolution has been conducting site-specific functionality assays for hundreds of millions of years. The ability to decipher the results of these experiments has and will continue to provide us with a wealth of information about our genome and the impact of mutations.

Rearrangements Occur Mostly Neutrally in Eukaryotic Genomes

Camille Berthelot[1], Matthieu Muffato[1] and Hugues Roest Crollius[1]

[1] DYOGEN Lab, ECOLE NORMALE SUPERIEURE, UMR8197 CNRS, INSERM U1024,
46 rue d'Ulm, 75230, Paris, Cedex 05, France
{cberthel, muffato, hrc}@biologie.ens.fr

Keywords Evolutionary rearrangements, genome evolution, multivariate analysis.

1 Introduction

The mechanisms underlying evolutionary genomic rearrangements and their fixation remain much debated. It is especially unclear if the occurrence of rearrangements is reflected by the current distribution of breakpoints in extant genomes, or if the latter is the result of a selective process where only a subset of non-deleterious rearrangements are maintained, or a combination of both. The original model of random breakage [1] has been challenged in recent years. Indeed, breakpoints are more clustered [2][3] and occur in regions richer in genes, GC, CpG islands, segmental duplications than expected at random [4][5]. Here we infer ancestral genomic characteristics to model breakage occurrence, and our results suggest that this process is mostly neutral and mechanistic, with a measurable but small contribution from negative selection.

2 Results and Discussion

The Boreoeutheria ancestor is the last common ancestor of primates, rodents and laurasiatherians (e.g. dog, cow, horse). Its gene content and gene order was reconstructed using AGORA [6], a new parsimony method relying on protein-coding gene annotations of its 28 sequenced descendant species available in ENSEMBL v.57. We compared this ancestral gene order to the gene orders of human, mouse, dog, horse and cow to identify 798 breakpoints (adjacent boreoeutherian genes separated in at least one modern species).

The aim of this work is to model the distribution of breakpoints in intergenes to identify the genomic parameters that underlie breakage. Candidate explanatory variables were chosen from parameters previously correlated with breakage, and an estimation of their ancestral value (i.e. before breakage occurred) was carried out. Lengths and GC contents of ancestral intergenes were estimated as the median of modern values in species where these intergenes still exist. To test for negative selection between genes and their regulatory sequences, we used computational predictions of target genes of long-range regulation and differentiated intergenes that flank a target gene from those that do not.

Using these three ancestral characteristics, breakage rates in intergenes were modeled using classical multivariate Poisson regression. The random model predicts that breakpoints will be distributed in intergenes following a Poisson law (Figure 1a, grey line). The regression equation obtained for observed breakpoints is strikingly different from the random expectations, as it involves a root of the intergene length rather than intergene length itself (slope < 1 in a log-log plot; Figure 1a, black line). However intergene length is the major predictor of breakage, explaining 74% of the observed deviation (McFadden's pseudo R^2). Long-range regulation also has a small but significant negative effect on breakage (Figure 1b), explaining an additional 6% of the deviation. GC content, on the other hand, is not significantly contributing to the breakage rate.

To evaluate if our new model based on intergene length and regulatory status can explain previous observations, we simulated breakpoints according to the regression equation obtained above. Results show that simulated breakpoints occur in regions that are significantly richer in GC, genes, CpG islands, SINEs and segmental duplications than expected at random, and therefore recapitulate observations that appear recurrently in the literature [5][6]. Although these simulations do not strictly rule out that those characteristics may be the underlying determinants of the rearrangement process, they show that a determination of breakpoints by intergene length might be sufficient to yield those previously observed correlations.

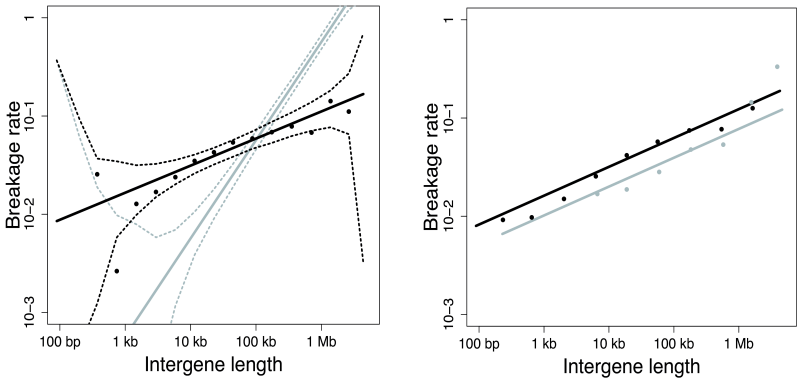


Figure 1.A. Breakage rate as a function of intergene length (Poisson regression). Black dots: observations; grey lines: expectations of the random model with confidence interval; black lines: regression model with CI. B. Intergenes that flank target genes of long-range regulation (grey) have lower breakage rates than those that do not (black).

While intergene length appears as a strong predictor of breakage, it is legitimate to ask if the true determinant of breakpoints might not be another genomic characteristic for which intergene length would be a proxy. We therefore searched for candidate characteristics that are tightly correlated with a root of intergene length in modern genomes. Transposable elements and conserved non-coding elements were ruled out as plausible explanations. Predicted replication origins [7], on the other hand, display a genomic distribution that makes them good candidates but they are in fact independent of observed breakpoints. This suggests that the distributions of both breakpoints and replication origins arise from the same cause. We hypothesize that the 3D nuclear organization of the genome may play an important mechanistic role in these processes, although the exact mechanism at work remains unknown.

3 Conclusions

The results presented in this work show that an intergene length alone is highly predictive of its breakage probability, suggesting that except for a minority of regions with evolutionarily constrained gene topologies due to regulatory domains, breakpoints occur mostly neutrally and mechanistically. This is in apparent contradiction with the fact that the distribution of breakpoints is not strictly random, and predicts that an unknown property of intergenes, perhaps related to the 3D structure of chromatin, influences the breakage probability. Our model also provides a new framework to investigate evolutionary breakpoints. With more breakpoint data available, this model will be able to make fine predictions of expected breakage rates for specific intergenes, and to compare them with the observed rates to identify regions under negative selective pressure against rearrangements, highlighting functional interactions.

References

[1] J. H. Nadeau and B. A. Taylor, Lengths pf chromosomal segments conserved since divergence of man and mouse. *Proc Natl Acad Sci U S A*, 81:814-818, 1984.

[2] P. Pevzner and G. Tesler, Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proc Natl Acad Sci U S A*, 100:7672-7677, 2003.

[3] Q. Peng, P. A. Pevzner and G. Tesler, The fragile breakage model versus random breakage models of chromosome evolution. *PloS Comput Biol*, 2:el4, 2006.

[4] D. M. Larkin, G. Pape, R. Donthu, L. Auvil, M. Wedge and H. A. Lewin, Breakpoint regions and homologous synteny blocks in chromosomes have different evolutionary histories. *Genome Res*, 19:770-777, 2009.

[5] C. Lemaitre, L. Zaghloul, M.-F. Sagot, C. Gautier, A. Arneodo, E. Tannier and B. Audit, Analysis of fine-scale mammalian evolutionary breakpoints provides new insight into their relation to genome organisation. *BMC Genomics*, 10:335-346, 2009.

[6] M. Muffato, A. Louis and H. Roest Crollius, AGORA: an Ancestral Gene Order Reconstruction Algorithm. *In prep.*

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

Exploration of the Genetic Diversity of the *Lachancea kluyveri* Yeast Species

Anne FRIEDRICH, Cyrielle REISSER, Paul JUNG and Joseph SCHACHERER

DEPARTEMENT DE GENETIQUE MOLECULAIRE, GENOMIQUE, MICROBIOLOGIE, UMR7156, 28 rue Goethe, 67083, Strasbourg, France

{anne.friedrich, creisser, pauljung, schacherer}@unistra.fr

Keywords Intraspecific genomic diversity, genotype-phenotype relationship, evolution.

The genetic variation that occurs naturally in a population represents a unique resource for both studying the basis of phenotypic differences between individuals and elucidating the evolutionary history of the species. In this context, microbial models such as yeast are of particular interest, as phenotypic diversity among isolates is significant and variation is apparent among the natural strains at different levels.

To date, yeast population genomics focused on two species: *Saccharomyces cerevisiae* and *Saccharomyces paradoxus* [1,2]. The *Saccharomyces* genus underwent whole-genome duplication followed by a massive loss of genes that may have an important impact on the structure of these genomes and on the phenotypic evolution of these species.

In this context, we launched a comprehensive survey of genomic variations among isolates within a protoploid yeast species that did not undergo whole genome duplication, *Lachancea kluyveri*. Comparative population genomics might allow us to assess the influence of genome duplication on genetic and phenotypic variation within a species. The *L. kluyveri* CBS3082 strain has already been fully sequenced and annotated, providing access to a high quality reference genome [3]. With the ambition to explore the genetic diversity and compare the patterns and levels of DNA sequence variation between *L. kluyveri* isolates, we sought to combine high-throughput sequencing, phenotyping and computational methods. We have consequently launched a high-coverage sequencing project of 38 *L. kluyveri* strains representative of the whole population. Based on an approach that combines commonly used software such as SOAP2 [4], SOAPsnp [5], STRUCTURE [6] and in-house PYTHON scripts, we are currently generating high coverage SNPs (Single Nucleotide Polymorphisms) maps. These data are being exploited to measure the intraspecific diversity as well as to assess the population structure, linkage disequilibrium and identify the selection patterns.

Our results will lay the foundation for phenotype-genotype linkage mapping and will next be recovered to compare the genotype-phenotype map of different natural populations of yeast: *S. cerevisiae* and *L. kluyveri.*

Acknowledgements

This work is supported by the ANR grant 2010 BLANC 1606 05 (GB-3G). AF was supported by a 2010 grant of the Scientific Council of the University of Strasbourg. CR is supported by a grant from the CNRS and Region Alsace.

References

[1] J. Schacherer, J.A. Shapiro, D.M. Ruderfer and L. Kruglyak, Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae*. *Nature*, 458(7236):342-345, 2009.

[2] G. Liti G, D.M. Carter, A.M. Moses, J. Warringer, L. Parts, S.A. James, R.P. Davey, I.N. Roberts, A. Burt, V. Koufopanou , I.J. Tsai, C.M. Bergman, D. Bensasson, M.J. O'Kelly, A. van Oudenaarden, D.B. Barton, E. Bailes, A.N. Nguyen, M. Jones, M.A. Quail, I. Goodhead, S. Sims, F. Smith, A. Blomberg, R. Durbin and E.J. Louis, Population genomics of domestic and wild yeasts. *Nature*, 458(7236):337-341, 2009

[3] The Génolevures Consortium, Comparative genomics of protoploid *Saccharomycetaceae*. *Genome Res*., 19:1696-1709, 2009.

[4] R. Li, C. Yu, Y. Li, T.W. Lam, S.M. Yiu, K. Kristiansen and J. Wang, SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15):1966-1967, 2009.

[5] R. Li, Y. Li, X. Fang, H. Yang, J. Wang, K. Kristiansen and J. Wang, SNP detection for massively parallel whole-genome resequencing. *Genome Res*., 19(6):1124-1132, 2009.

[6] J.K. Pritchard, M. Stephens and P. Donnelly, Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959, 2000.

An Evolutionary Analysis of the Type III Secretion System

Sophie S. ABBY[1] and Eduardo P.C. ROCHA[1]

[1] MICROBIAL EVOLUTIONARY GENOMICS, Institut Pasteur, CNRS URA2171, F-75724, Paris, France
{sabby, erocha}@pasteur.fr

Abstract *The type III secretion system (T3SS), or "injectisome", is a bacterial machinery that allows the injection of protein effectors directly into eukaryotic cells. This system is often used in the establishment of pathogenic or symbiotic relationships with both animals and plants. The T3SS is related to the bacterial flagellum and some of its core genes are homologous to conserved flagellar genes. We present the first systematic search of T3SS in thousands of prokaryotic genomes. We identified 203 putative T3SS in 147 complete bacterial genomes. T3SS are found in specific clades of alpha-, beta- and gamma- proteobacteria, are present in all intracellular pathogens chlamydiales, are rare in deltaproteobacteria and absent from epsilonproteobacteria sequenced to date. Among the nine core genes used for the characterization of the T3SS, eight are homologous to core genes of the flagellum. We detected more than 170 potential flagellar systems in our bacterial genome dataset, and perform phylogenetic analyses of the eight genes conserved between the two systems, in order to give insights on the evolutionary history of these two bacterial systems. Accordingly with previous studies, the taxonomic distribution of the T3SS is patchy and phylogenetic analyses suggest a complex evolutionary history of this system, punctuated by lateral gene transfers.*

Keywords Injectisome, Type III secretion system, T3SS evolution, Phylogenetic congruence, Lateral gene transfers.

1 Introduction

The type III secretory system (T3SS) or "injectisome" is a cellular machinery that spans the inner and outer membrane of some groups of diderm (typically gram−) bacteria [1]. The general function of the T3SS is the exportation of bacterial proteic effectors to a eukaryotic host. In several pathogens, the T3SS is required for virulence, and allows via a molecular syringe the injection of virulence factors directly into eukaryotic cells (e.g. in *Yersinia pestis*). Some T3SS were shown to be necessary to the establishment of symbiotic relationships (e.g. in plant symbionts Rhizobiales). Bacteria use T3SS in both plant and animal hosts. Some authors proposed the use of this system by *Pseudomonas aeruginosa* as a weapon to survive predation by amoebae in biofilms [2]. T3SS is thus an important player in relationships between bacteria and eukaryotes. Several families of T3SS were previously described on phylogenetic grounds, and T3SS scarce taxonomic distribution led some authors to invoke lateral gene transfers along its evolution. T3SS is related to the bacterial flagellum in two ways. First, advances on the structural characterization of both systems demonstrated a similar basal structure. Second, sequence analyses gave evidence on the common origin of several components of the T3SS and of the flagellum. T3SS is in average constituted by 20 proteins encoded by genes gathered in a genomic cluster (except in chlamydiae). Nine of these genes are conserved among all described T3SS. Eight out of nine core genes show a clear homology with core genes of the flagellum. The remaining gene, the secretin, has no flagellar homolog, but belongs to a super-family of proteins of the general secretion pathway [3]. In this study we propose i) a computational method for T3SS detection in bacterial genomes, ii) the study of the taxonomic distribution of T3SS iii) a phylogenetic analysis of the T3SS along with the related flagellar system.

2 Results and Discussion

2.1 Mining Bacterial Genomes for the T3SS

We took advantage of the existence of nine conserved core genes in all T3SS described to date, and built HMM profiles to search homologs in prokaryotic genomes. 464 replicons (428 genomes) out of 1342

analysed presented at least one hit for each of the nine core genes. Another conserved feature is the genome architecture of the system since genes encoding proteins of T3SS structure and functioning colocalized in one or few gene clusters. We thus searched for clusters of T3SS core genes homologs. A potential source of false positives detection is the homology of T3SS core genes with some flagellar core genes. We could sort out flagella from T3SS by segregating clusters containing a secretin homolog (no homolog in flagellum), from clusters containing conserved flagellar genes with no homolog in T3SS. Finally, 159 bacterial replicons distributed in 147 genomes presented at least one T3SS. A total of 203 putative T3SS were detected. 17 systems, found in the chlamydiales genomes, were situated on several loci on bacterial chromosome, whereas other systems clustered at a single genomic location. Most of the systems (180) are on chromosomes, but 23 are on plasmids. 173 flagellar clusters were also detected.

2.2 Taxonomic Distribution of the T3SS

T3SS are present in gammaproteobacteria (enterobacteriales, pseudomonadales, vibrionales...), betaproteobacteria (burkholderiales), alphaproteobacteria (rhizobiales), deltaproteobacteria (desulfovibrionales, myxococcales) and chlamydiales. These bacteria are mainly plant and animal pathogens or symbionts. Few occurences of T3SS are reported in free-living bacteria. This reinforces the general role of T3SS in the establishment of close relationships with eukaryotic hosts.

2.3 Phylogenetic Analyses of T3SS Core Genes and their Flagellar Homologs

The experimental and structural characterization of the T3SS in different species gave insight on the function, shape and relative positioning of T3SS proteins. For example, the core genes *sctR, sctS, sctT, sctU* and *sctV* encode for inner-membrane proteins known to be part of the T3SS basal structure and to closely interact. SctN is an ATPase suspected to interact with SctQ, whereas the secretin SctC is an outer-membrane protein. Moreover, the genomic location of the nine genes suggests that those genes could share a same evolutionary history. Based on patchy taxonomic distribution of the T3SS compared to that of the flagellum, widespread in bacterial phyla, previous phylogenetic analyses proposed that T3SS derived from a flagellar ancestor and spread through lateral gene transfers. We first studied the congruence of phylogenetic signal of the nine core genes. Then we could systematically test the hypothesis of *en bloc* transfers of T3SS using more data than previously included in such analysis. Reconstructed core genes phylogenies showed evidence of lateral gene transfers. Moreover, plasmidic type III systems alternatively branch with chromosomal systems, suggesting that conjugation may be one of the modes of T3SS lateral transmission.

3 Conclusions

Our study allowed the systematic search of the T3SS, system crucial in bacteria-plant symbioses, the virulence of many well-known pathogens, and also potentially in the emergence of new animal and plant pathogens. This search provided the material for a phylogenetic analysis of T3SS core genes, most of them being related to flagellar core genes. As suggested by its scarce taxonomic distribution, the T3SS presents a complex evolutionary history relying on multiple lateral gene transfers of the whole system.

References

[1] G.R. Cornelis, The type III secretion injectisome. *Nature Review Microbiology*. 4(11):811-25, 2006.
[2] C. Matz, A.M. Moreno, M. Alhede, M. Manefield, A.R. Hauser, M. Givskov, S. Kjelleberg. Pseudomonas aeruginosa uses type III secretion system to kill biofilm-associated amoebae. *The ISME Journal*. 2(8):843-52, 2008.
[3] S.A. Clock, P.J. Planet, B.A. Perez, D.H. Figurski. Outer membrane components of the Tad (tight adherence) secreton of Aggregatibacter actinomycetemcomitans. *Journal of Bacteriology.* 190(3):980-90, 2008.

Identification of Putative Parasitism Genes in Plant-Parasitic Nematodes

In silico Screening of Whole Genomes and Transcriptomes

Amandine Campan-Fournier[1], Laetitia Perfus-Barbeoch[1], Marie-Noëlle Rosso[1], Marie-Jeanne Arguel[1], Corinne Da Silva[2], Celine Vens[3], Nathalie Marteu[1], Karine Labadie[2], François Artiguenave[2], Pierre Abad[1] and Etienne G.J. Danchin[1]

[1] Plant-Nematode Interactions, UMR 1301 INRA - UNS - CNRS, 400 route des Chappes, 06903, Sophia-Antipolis Cedex, France

amandine.fournier@sophia.inra.fr

[2] Institut de Génomique, Genoscope, CEA, 2 rue Gaston Crémieux, 91000, Evry, France

[3] Department of Computer Science, Katholieke Universiteit Leuven, Celestijnenlaan, 3001, Leuven, Belgium

Abstract *Plant-parasitic nematodes (PPN) are microscopic roundworms that cause disease on nearly all economically important crop plants. They are responsible for estimated losses of several billion Euros/year. Currently, nematicides are the most important means of controlling nematodes. However, current nematicides are non-specific, notoriously toxic and pose a threat to soil ecosystem, ground water and human health. Therefore, novel and specific targets are needed to develop new strategies directed against plant-parasitic nematode species.*

In this context, our project aims at identifying putative parasitism genes in plant-parasitic nematodes. A multi-disciplinary approach combining bioinformatics and functional genomics based on large-scale screening of genomic and proteomic data from nematodes showing different modes of plant parasitism is proposed. Candidate targets are identified by bioinformatics methods and the most promising candidates will be selected for further functional analyses.

Here we report the semi-automated bioinformatics pipeline developed for that purpose. We have undertaken a comparative analysis of the sets of predicted proteins in Meloidogyne incognita and Meloidogyne hapla (two fully sequenced plant-parasitic nematodes) with a large dataset of whole genomes and transcriptomes. As our objective is to identify druggable parasitism genes we have searched for proteins conserved in other parasitic or plant-associated species, but absent from species that could be negatively affected by newly developed drugs or control means. We also have undertaken bioinformatics annotations of these proteins, including but not limited to: detection of signal peptide and Pfam domains, assignment of gene ontology terms and identification of specific motifs.

Keywords Plant-parasitic nematodes (PPN), parasitism genes, *in silico* screening, automatic functional annotation, comparative genomics.

1 Introduction

Plant-parasitic nematodes (PPN) are microscopic roundworms. Their strategy to infest plants and their host range depend on the species. Most of them feed on root tissue and damage their host mainly by altering root growth (resulting in reduced water uptake), by promoting microbial infections through wound sites or by serving as vectors for pathogenic viruses. They cause disease on nearly all economically important crop plants, including corn, soybean, cotton, rice, tomato, carrots and tobacco. They are thus responsible for estimated losses of several billion Euros/year. The most economically impacting plant-parasitic nematodes are root-knot nematodes and cyst nematodes that both pertain to the phylum Tylenchida (or clade 12) [1]. During their life cycle, these plant-parasitic nematodes penetrate the root and migrate within plant tissue. They induce the development of a specialized feeding structure from root plant cells and settle sedentary at this feeding site. It has been shown that they secrete proteins called “effectors” in plant tissue. Several of these effectors have been shown to be involved in degradation of the plant cell wall or potentially implicated in modulation of plant defenses. Many effectors remain uncharacterized and are suspected to be involved in the development of feeding structures or other processes related to successful parasitism.

Measures such as growing resistant crop varieties and the use of nematicides are extensively employed to control plant-parasitic nematodes. However, it happens that some nematodes overcome resistance genes and become "virulent" (able to infect varieties that were previously resistant to these nematodes). Moreover, current nematicides are costly, non-specific, notoriously toxic and pose a threat to the soil ecosystem, ground water and human health. This has lead to the banning of the most efficient chemicals that were previously commonly used. Therefore, novel and specific targets are needed to develop new strategies directed against plant-parasitic nematode species.

In 2008, the careful analysis of the first sequenced genome of a plant-parasitic animal, the root-knot nematode *Meloidogyne incognita* [2], highlighted new potential targets for anti-parasitic strategies. To confirm the relevance of these genes as good candidate targets, efforts are needed to produce high-throughput data on additional plant-parasitic nematode species. Indeed, very few genomic and transcriptomic resources were available so far: only two genomes of plant-parasitic nematodes (*M. incognita* and *Meloidogyne hapla* [3]) are fully sequenced and annotated, and most EST come from species from the Meloidogyne genus (preventing comparative studies).

That is why we propose: (i) an in-depth search for potential new targets by comparaison of the *M. incognita* and *M. hapla* sets of predicted proteins with proteins from other organisms (parasitic or not) ; (ii) the generation and analysis of large-scale transcriptomic data (RNA-seq) from four other plant-parasitic nematodes representing diverse parasitic strategies (*Pratylenchus coffeae*, *Ditylenchus dipsaci*, *Bursaphelencus xylophilus* and *Xiphinema index*). The rationale of our analysis is that the more a protein is broadly conserved across parasitic or plant-associated species yet restricted to them, the more it is likely to be involved in the parasitism process. Thus, conservation of a protein in a parasitic or a plant-associated species is considered like a "bonus". In contrast, we call "forbidden species" hereinafter the species that are neither plant-parasitic, nor plant-associated, and that could be negatively affected by the development of novel drugs or control means. In concrete terms, by "forbidden species", we mean species like plants, mammals, fishes, pollinating insects… Indeed, the long-term application of our project is to manage parasitic nematode infestations, without affecting crop plants or being toxic to ecosystem and human health. For example, a novel chemical developed against nematodes should not kill honeybees. Some species are neither "forbidden" nor "bonus" and are therefore considered as "neutral" (bacteria and viruses that are neither plant-parasitic nor plant-associated for example). After identification and annotation of candidate targets by bioinformatics methods, the most promising candidates will be selected for further functional analyses.

Here we report the semi-automated bioinformatics pipeline and the data management system developed for the identification of genes involved in plant-parasitism. To date, it has not been possible to develop such a comparative pipeline in the context of plant-parasitic nematodes because of the scarcity of genomics and transcriptomics data available for these species.

2 Material and Methods

The bioinformatics pipeline begins with two steps of screening, based on sequence similarity (Fig. 1). After the screening steps, the remaining proteins from *M. incognita* and *M. hapla* are analysed in terms of transcription evidence and automatic functional annotation. Lastly, all the data produced are stored into a relational database dedicated to this project.

Throughout the process, all the scripts necessary to parse the results or format the data files have been written with the Perl language with use of some BioPerl modules.

2.1 *In silico* Screenings of Potential Targets

The first screening step consists in a comparative analysis of the sets of predicted proteins from *M. incognita* and *M. hapla* with sets of predicted proteins from twenty-three other fully sequenced species. The dataset includes putative proteomes from 1 human-parasitic nematode, 2 plant-pathogenic fungi and 2 plant-eating insects. They are considered as "bonus species". In addition, the dataset includes putative proteomes from 3 nematodes, 5 mammals, 1 bird, 1 amphibian, 2 fishes, 3 insects and 1 plant. They are neither parasitic, nor plant-associated, and are considered as "forbidden species". To perform the comparative analysis, the OrthoMCL tool [4] was run with default parameters. The OrthoMCL procedure starts with all-against-all BLASTp comparisons of protein sequences from the submitted genomes. Putative orthologous relationships are identified between pairs of genomes by reciprocal best similarity pairs.

« Recent » paralogs (or in-paralogs) are identified as sequences within the same genome that are (reciprocally) more similar to each other than to any sequence from other species. Then, putative orthologous relationships are converted into a graph, to which the MCL (Markov Clustering) algorithm [5] is applied. The final output consists in clusters of putative orthologs and « recent » paralogs.

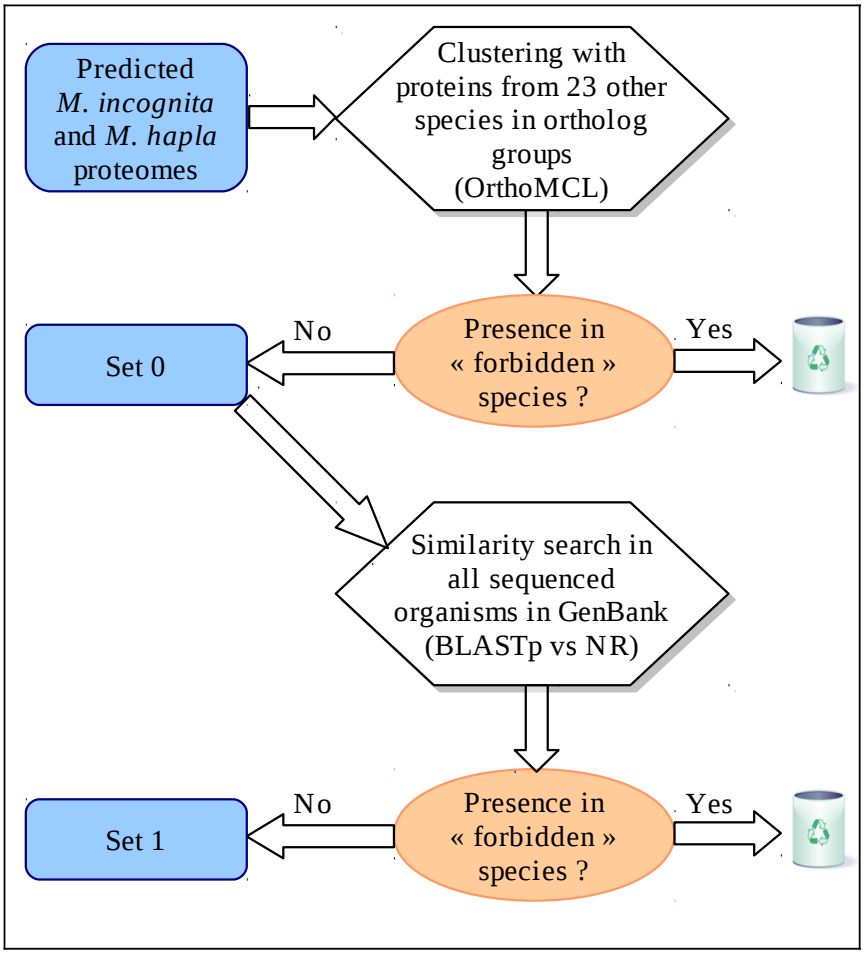


Figure 1. Schematic overview of the screening pipeline to identify potential targets in *Meloidogyne incognita* and *Meloidogyne hapla* whole sets of predicted proteins.

The results were parsed to exclude proteins conserved in forbidden species. The “remaining” proteins constitute the ***set 0***. We also assigned a “bonus tag” to proteins passing this filter that presented a potential ortholog in a known parasitic or plant-pathogenic species.

The second screening step consists in a BLAST search [6] of the ***set 0*** against GenBank (NR database, blastp, evalue max = 0.01, no filter for low complexity regions). This second step is perfectly complementary to the first one. Indeed, several proteomes included in the OrthoMCL run are absent from NR. In addition, most species in NR can not be included in the OrthoMCL run, because we do not have their complete putative proteomes (species not fully sequenced).

For each protein, all BLAST hits were analysed sequentially. We excluded proteins showing significant similarity (at least 40 % identity and 70 % of query length covered by the alignment) with one or more forbidden species. The remaining proteins presenting a significant similarity (at least 30 % identity and 50 % of query length covered by the alignment) with proteins from known plant-parasitic or plant-pathogenic species were assigned a “bonus tag”. (Fig. 2) The criteria are more stringent for exclusion than tag assignment to avoid considering hits in forbidden species that are not true orthologs, since only one hit in a forbidden species lead to the exclusion of the protein (whereas bonus tags are rather informative).

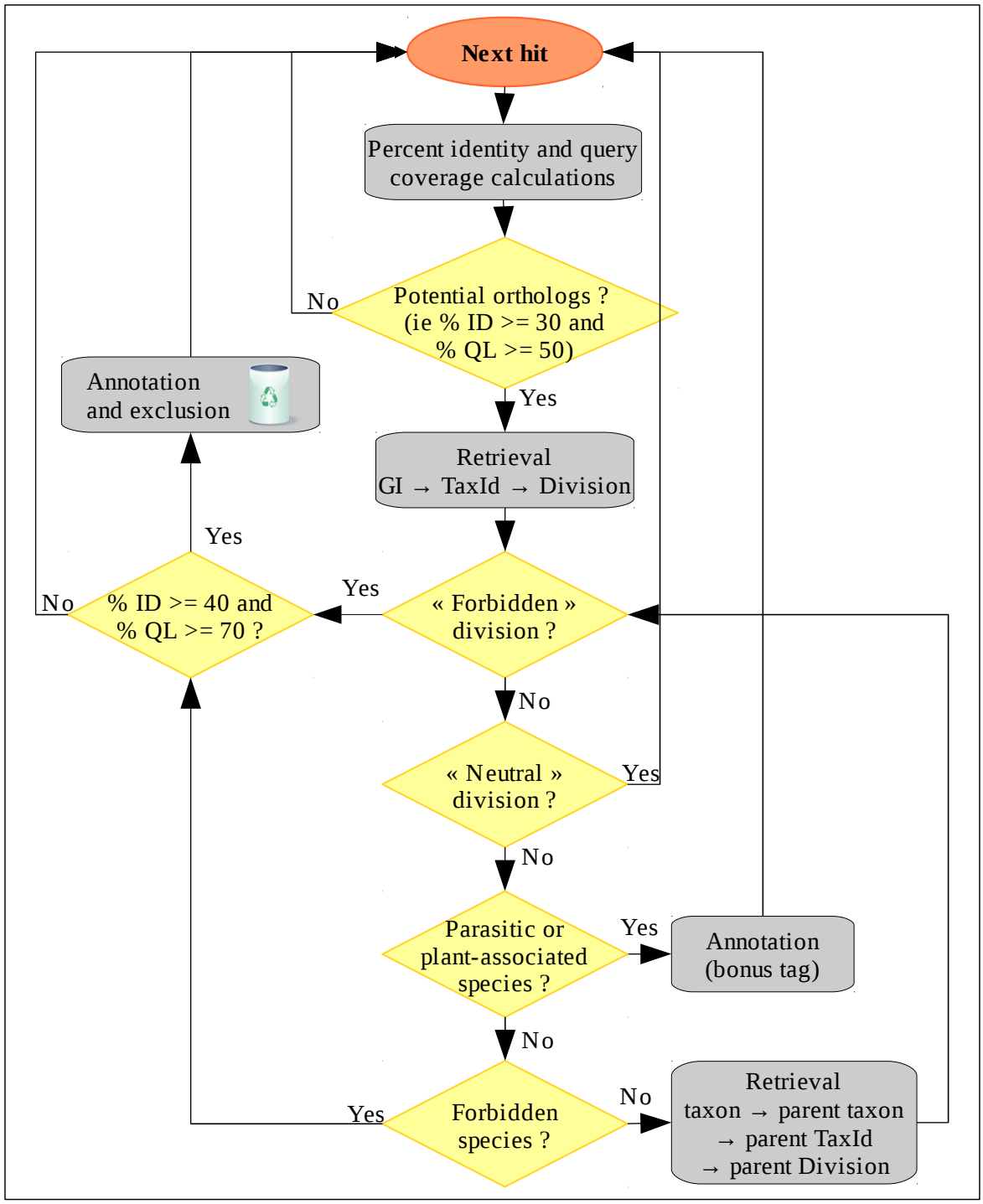


Figure 2. Schematic overview of the algorithm implemented in the BLAST parser. Remaining proteins after the first screening step based on an OrthoMCL run underwent a second screening step based on BLAST searches. All BLAST hits were sequentially analysed. When there is an hit in a species that is neither a "bonus" one, nor a "forbidden" one, we get the parent node in the taxonomy and test this parent node in the same way (until reaching the root if necessary), because the taxonomy identifiers (TaxId) we have listed sometimes correspond to clades (a higher level than a species). Moreover, to reduce computations, the parser first sorts the hits on the "division" criteria, as a division is assigned to each taxon node by the NCBI taxonomy. Forbidden divisions are Mammals, Primates, Rodents and Vertebrates. Neutral divisions are Phages, Synthetic, Unassigned and Environmental samples. In the other cases, the parser has to check if the TaxId is in the "bonus" list or in the "forbidden" list. On the figure, "% ID" means percent identity and "% QL" means percentage of query length covered by the alignment.

We are aware that this methodology would overlook genes involved in parasitism that are duplicates or mutated versions of existing genes shared with forbidden species. As our aim is to identify druggable targets, we can not take the risk to select genes that would be too similar to genes conserved in such species.

As there is no large-scale database that propose an inventory of plant-associated species, we collected information from the bibliography, from plant-pathologists and from two partial databases. The first one is the Comprehensive Phytopathogen Genomics Resource (http://cpgr.plantbiology.msu.edu/). It consists in a data warehouse of finished, draft and in progress genome and EST sequencing projects for viral, bacterial, oomycete, fungal, and nematode plant pathogens. The second one is the Pathogen Hosts Interactions base (http://www.phi-base.org/, [7]). It contains expertly curated information on experimentally verified pathogenicity, virulence and effector genes from fungal, oomycete and bacterial pathogens. Moreover, according to bibliography, we considered that four clades of nematodes are plant-associated: Tylenchida, Nordiidae, Longidoroidea and Trichodoroidea. In the end, we derived a list of 834 NCBI's taxonomy identifiers (TaxId) corresponding to species and clades known to be involved in parasitic or pathogenic interactions with plants. This is not an exhaustive list, but it represents more than 28000 species in total (as numerous species pertain to a clade) of nematodes, oomycetes, fungi, bacteria, trypanosomes, insects, virus and viroids.

To parse the BLAST results, we also needed to download the NCBI taxonomy and to list clades that we consider as “forbidden” (four clades: Chordata, Annelida, Mollusca, Viridiplantae).

The proteins kept at the end of the pipeline constitute ***set 1***. As computation requires huge memory, BLAST search and parsing have been computed on a cloud: the ProActive PACA Grid (http://proactive.inria.fr/pacagrid/).

2.2 Evidence of Expression at the Transcriptional Level

Evidence of the transcription of a gene coding for a putative protein supports the existence of this putatively expressed gene.

We already had accumulated data about EST evidence for the *M. incognita* set of proteins. They come from the NCBI dbEST database and from “in-house” *M. incognita*-specific EST clusters. The latter provides information about the stage(s) of the life cycle during which the gene is expressed. In our case, we are particularly interested in genes expressed during the free-living stage (as nematodes are more reachable by control means), but expression during plant-nematode interaction can provide insights into the mechanisms of parasitism.

We also downloaded datasets of protein predictions derived from clustered EST of seventeen plant-parasitic nematodes (including *M. incognita* and *M. hapla*) from the NEMBASE4 resource [8]. These collections are publicly available from http://nematodes.org/downloads/databases/NEMBASE4/index.shtml. We performed BLAST searches of our ***set 1***, using the polypeptides from NEMBASE4 as subject sequences (blastp, evalue max = 0.01, no filter for low complexity regions). Data were then parsed and criteria (identity and alignment percentages) were fit according to the phylogenetic distance between the subject species and the Meloidogyne phylum. Here, data provide not only expression evidence, but also information about conservation of the gene in plant-parasitic nematodes (as seen before, the more a gene is conserved in plant-parasitic species, the more it is likely to be involved in parasitism).

However, the amount of available transcriptomic data for plant-parasitic nematodes is relatively limited and most information is restricted to root-knot nematodes and to a lesser extent to cyst nematodes. This limits the possibility of comparing various different plant-parasitic nematodes that have adopted different strategies to feed on plant material. Hence, we have performed the RNA-seq transcriptome sequencing of four plant-parasitic nematode species presenting diverse parasitic strategies (*Pratylenchus coffeae*, *Ditylenchus dipsaci*, *Bursaphelencus xylophilus* and *Xiphinema index*). We have also generated the RNA-seq of different developmental stages of *M. incognita* in order to bring additional transcription support to the identified genes. Bioinformatics analyses are currently *in progress*.

2.3 Automatic Functional Annotation

As we are working on proteins predicted from the genomes, most of them have currently unknown functions. We have therefore undertaken bioinformatics annotations of these proteins.

- Functional regions (commonly termed *domains*) have been identified into our proteins, by using the PfamScan tool with the Pfam-A database (the part of Pfam containing high quality, manually curated families) and default parameters [9].
- "Standard" gene ontology (GO) terms have then been assigned to the proteins, based on the correspondence between the Pfam domains and the GO terms. "Slim" terms associated to the "standard" terms have also been subsequently assigned to the proteins. GO slims are cut-down versions of the GO ontologies containing a subset of the terms in the whole GO. They give a broad overview of the ontology content, without the detail of the specific fine grained terms. [10]
- The presence and location of signal peptide cleavage sites have been detected with the SignalP v3.0 tool [11], using both available methods (artificial neural networks and hidden Markov models).
- Prediction of transmembrane protein topology has been performed by searching for transmembrane helices in protein with the TMHMM v2.0 tool [12].
- Motifs specific to effectors of root-knot nematodes have been identified in the whole proteome of *M. incognita* using the MERCI software [13].

2.4 Database Development

A relational database has been developed using MySQL and phpMyAdmin in order to store all the data generated during the project. It allows to make data integration and analyses easier. Complex queries that combine results obtained at different steps of the pipeline (screening and/or annotation) can be launched. Outputs can be saved as simple tables or spreadsheets that can readily be used by the biologists.

3 Results and Discussion

The relational database contains all the data generated so far: results from the screening steps (which proteins have been "excluded" or "kept to go further" at each step, description of the hits) and the different types of annotations (as previously described: domains, gene ontology terms, signal peptides, transmembrane topology, specific motifs). Some more general information are included too, as the description of each step (who carried out this step, when, which tool and parameters have been used, what was the previous step...). This database allows us to compute a wide range of queries.

First of all, our ***set 1*** contains 16320 proteins. These root-knot nematode proteins are not present in species that could be negatively affected by the development of novel drugs and part of these are conserved in other plant-parasites. Among them:

– 5497 proteins (~ 34%) are shared with at least one other parasitic or plant-associated species,

– 3462 proteins (~ 21%) are supported by a transcription evidence from the same species and 4203 proteins (~ 26 %) are supported by a transcription evidence whatever the species (itself or another plant-parasitic nematode),

– less than a quarter of proteins have been associated with a functional annotation. Indeed, 3835 proteins (~ 24 % of ***set 1***) are annotated with one Pfam domain or more, and 2255 proteins (~ 14 %) have a GO term assigned.

A more detailed analysis of GO slim terms associated with the proteins shows that some terms seem to be under- or over-represented in the ***set 1*** compared to the whole putative proteomes of *M. incognita* and

M. hapla. In particular, we notice that the terms *nucleus* (GO:0005634 from the 'cellular component' ontology), *transcription factor activity* (GO:0003700 from the 'molecular function' ontology) and *transcription* (GO:0006350 from the 'biological process' ontology) seem to be over-represented ; suggesting that we may have identified specific transcription factors. By combining criteria on GO slim terms, transcription evidence and conservation in other species, we observed that twelve proteins are shared between several parasitic or plant-associated species, are annotated with the three GO slim terms related to transcription mentioned above and are supported by a transcriptional evidence. These proteins are of particular interest and are currently under experimental investigation.

Moreover, it is possible to identify putative effectors by combining the following criteria: conservation in other parasitic or plant-associated species, transcriptional evidence, no transmembrane helix, presence of a signal peptide and presence of one of the effector-specific motifs previously identified [13]. We obtain a list of 158 proteins. Among them, 18 are known to be carbohydrate-active enzymes (CAZymes: http://www.cazy.org/) involved in plant cell wall degradation, which is one of the known function of effectors [2]. They are able to degrade carbohydrates like cellulose, hemicellulose or pectin. Presence of known effectors within the reduced set of predicted effectors validates the screening method. Proteins of as yet unknown function constitute a set of interest to identify and characterize new effectors.

4 Conclusion and Perspectives

To date, the bioinformatics pipeline has generated a number of data which are all stored in a relational database. By combining several criteria, the database allows identification of sets of target genes restricted to parasitic or plant-associated species (such as putative transcription factors or effectors) for the design of durable new strategies to manage parasitic nematode infestations. As a control, we could align the identified genes back to the genomes of some neither parasitic nor plant-associated species to check that their successful outcome through filters is not due to annotation problems. But it would be quite surprising that these genes would have been missed in all the proteomes of the 18 forbidden species included in the OrthoMCL run.

In near future, it is planned to implement a graphical user interface, probably by using BioMart [14], as it is described as a simple and robust data integration system for large scale data querying. This interface would allow users to query the database more easily (without the need of writing SQL instructions). It is also planned to include more data, according to the users needs (such as bibliography or comments).

Furthermore, to partially overcome the scarcity of omics data available for plant-parasitic nematodes, we have performed the RNA-seq transcriptome sequencing of four plant-parasitic nematode species presenting diverse parasitic strategies as well as the RNA-seq of different developmental stages of *M. incognita*. Bioinformatics analyses are currently *in progress* and will soon provide additional information about the transcriptional support of the identified genes and their conservation across the four plant-parasitic nematodes sequenced.

In the end, the most promising candidates will be selected for further functional analyses in the plant-parasitic nematode model *Meloidogyne incognita*. This will include expression analysis, tissue localization of gene expression and gene inactivation by RNA interference assays.

Acknowledgements

This work is supported by the French National Research Agency (“Nematargets” project).

Celine Vens is a Postdoctoral Fellow of the Research Foundation - Flanders (FWO).

The authors are grateful to Franck Panabières and René Feyereisen for their explanation about parasitic and plant-associated oomycetes and insects (respectively).

References

[1] H. van Megen, S. van den Elsen, M. Holterman, G. Karssen, P. Mooyman, T. Bongers, O. Holovachov, J. Bakker and J. Helder, A phylogenetic tree of nematodes based on about 1200 full-length small subunit ribosomal DNA sequences. *Nematology*, 11: 927-950, 2009.

[2] P. Abad, J. Gouzy, J. Aury, P. Castagnone-Sereno, E.G.J. Danchin, E. Deleury, L. Perfus-Barbeoch, V. Anthouard, F. Artiguenave, V.C. Blok, M. Caillaud, P.M. Coutinho, C. Dasilva, F. De Luca, F. Deau, M. Esquibet, T. Flutre, J.V. Goldstone, N. Hamamouch, T. Hewezi, O. Jaillon, C. Jubin, P. Leonetti, M. Magliano, T.R. Maier, G.V. Markov, P. McVeigh, G. Pesole, J. Poulain, M. Robinson-Rechavi, E. Sallet, B. Ségurens, D. Steinbach, T. Tytgat, E. Ugarte, C. van Ghelder, P. Veronico, T.J. Baum, M. Blaxter, T. Bleve-Zacheo, E.L. Davis, J.J. Ewbank, B. Favery, E. Grenier, B. Henrissat, J.T. Jones, V. Laudet, A.G. Maule, H. Quesneville, M. Rosso, T. Schiex, G. Smant, J. Weissenbach and P. Wincker, Genome sequence of the metazoan plant-parasitic nematode Meloidogyne incognita. *Nat. Biotechnol.*, 26: 909-915, 2008.

[3] C.H. Opperman, D.M. Bird, V.M. Williamson, D.S. Rokhsar, M. Burke, J. Cohn, J. Cromer, S. Diener, J. Gajan, S. Graham, T.D. Houfek, Q. Liu, T. Mitros, J. Schaff, R. Schaffer, E. Scholl, B.R. Sosinski, V.P. Thomas and E. Windham, Sequence and genetic map of Meloidogyne hapla: A compact nematode genome for plant parasitism. *Proc. Natl. Acad. Sci. U.S.A.*, 105: 14802-14807, 2008.

[4] L. Li, C.J.J. Stoeckert and D.S. Roos, OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.*, 13: 2178-2189, 2003.

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

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

[7] R. Winnenburg, M. Urban, A. Beacham, T.K. Baldwin, S. Holland, M. Lindeberg, H. Hansen, C. Rawlings, K.E. Hammond-Kosack and J. Köhler, PHI-base update: additions to the pathogen host interaction database. *Nucleic Acids Res.*, 36: D572-6, 2008.

[8] J. Parkinson, C. Whitton, R. Schmid, M. Thomson and M. Blaxter, NEMBASE: a resource for parasitic nematode ESTs. *Nucleic Acids Res.*, 32: D427-30, 2004.

[9] R.D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J.E. Pollington, O.L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E.L.L. Sonnhammer, S.R. Eddy and A. Bateman, The Pfam protein families database. *Nucleic Acids Res.*, 38: D211-22, 2010.

[10] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin and G. Sherlock, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, 25: 25-29, 2000.

[11] O. Emanuelsson, S. Brunak, G. von Heijne and H. Nielsen, Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc*, 2: 953-971, 2007.

[12] A. Krogh, B. Larsson, G. von Heijne and E.L. Sonnhammer, Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, 305: 567-580, 2001.

[13] C. Vens, M. Rosso and E.G.J. Danchin, Identifying discriminative classification-based motifs in biological sequences. *Bioinformatics*, 27: 1231-1238, 2011.

[14] S. Haider, B. Ballester, D. Smedley, J. Zhang, P. Rice and A. Kasprzyk, BioMart Central Portal--unified access to biological data. *Nucleic Acids Res.*, 37: W23-7, 2009.

Session 9 : RNA and Transcription

Conférence invitée

Peter Stadler

Universität Leipzig, Leipzig, Germany.

The RNA Zoo: Diversity and Complexity of Transcriptomes

ENCODE and FANTOM showed that nearly the mammalian genomes are almost completely described, putting an end to idea of Junk DNA. Since then, we have learned that transcription is more extensive and more complex also in other eukaryotes and even in prokaryotes. In contrast to the common organizational principles governing the protein-coding minority, the collection of transcripts forms a surprisingly heterogenous zoo of RNAs differing in processing, transport, and function. Complex hierarchical processing pathways, furthermore, generate multiple RNA species from the same genomic information that can act in ways that are unrelated in both biochemical mechanism and biological function.

For bioinformaticians, new challenges keep popping up, ranging from the technicalities of analysing huge amounts of high throughput sequencing data to ncRNA annotation and the quest for a sensible taxonomy of ncRNA classes. In the main part of my presentation I will focus on novel approaches to analysing long ncRNAs and on the relation of short RNAs to their precursors, including the following topics. Prediction of novel long RNAs from comparative genomics data; giant transcript triggered in some signal pathways; production of coherent short fragments from well known structured house-keeping RNAs; unusual processing of 3' ends; and chemical modifications.

Five Vertebrate ChIP-seq Reveals the Evolutionary Dynamics of Transcription Factor Binding

Dominic SCHMIDT[1,2+], Michael WILSON[1,2+], Benoît BALLESTER[3+], Petra C. SCHWALIE[3], Gordon D. BROWN[1], Aileen MARSHALL[1,4], Claudia KUTTER[1], Stephen WATT[1], Celia P. MARTINEZ-JIMENEZ[5], Sarah MACKAY[6], Iannis TALIANIDIS[5], Paul FLICEK[3,7]and Duncan T. ODOM[1,2]

[+] Co-first authors

[1] CANCER RESEARCH UK, Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge, CB2 0RE, UK

{Dominic.Schmidt, Michael.Wilson, Gordon.Brown, Aileen.Marshall, Claudia.Kutter, Stephen.Watt, Sarah.Mackay, Duncan.Odom}@cancer.org.uk

[2] UNIVERSITY OF CAMBRIDGE, Department of Oncology, Hutchison/MRC Research Centre, Hills Road, Cambridge, CB2 0XZ, UK

[3] EUROPEAN BIOINFORMATICS INSTITUTE (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

{benoit, schwalie, flicek}@ebi.ac.uk

[4] CAMBRIDGE HEPATOBILIARY SERVICE, Addenbrooke's Hospital, Hills Road, Cambridge CB2 2QQ, UK

[5] BIOMEDICAL SCIENCES RESEARCH Center Al. Fleming, 16672, Vari, Greece

{celia, talianidis}@fleming.gr

[6] INTEGRATIVE AND SYSTEMS BIOLOGY, Faculty of Biomedical and Life Sciences, University of Glasgow, G128QQ, UK

[7] WELLCOME TRUST SANGER INSTITUTE, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

Abstract *Transcription factors (TFs) direct gene expression by binding to DNA regulatory regions. To explore the evolution of gene regulation, we used chromatin immunoprecipitation with high-throughput sequencing (ChIP-seq) to determine experimentally the genome-wide occupancy of two TFs, CCAAT/enhancer-binding protein alpha and hepatocyte nuclear factor 4 alpha, in the livers of five vertebrates. Although each TF displays highly conserved DNA binding preferences, most binding is species-specific, and aligned binding events present in all five species are rare. Regions near genes with expression levels that are dependent on a TF are often bound by the TF in multiple species yet show no enhanced DNA sequence constraint. Binding divergence between species can be largely explained by sequence changes to the bound motifs. Among the binding events lost in one lineage, only half are recovered by another binding event within 10 kilobases. Our results reveal large interspecies differences in transcriptional regulation and provide insight into regulatory evolution.*

Keywords Transcription Factor Binding, ChIP-seq, Evolution.

1 Summary

The relationship between genetic sequence and transcriptional regulation is central to understanding species-specific biology, disease, and evolution. Identifying the divergence and conservation among functional regulatory elements is an important goal of comparative genomic research, and this is often done via DNA sequence comparisons using distant and closely related species. Although both approaches have successfully identified conserved regulatory regions, the majority of transcription factor (TF) binding events change rapidly between closely related species, making them difficult to detect using DNA sequence alone. For instance, the experimentally-determined binding events for homologous TFs found in mouse and human livers are unlikely to align with each other, despite conservation of their functional targets and global liver transcription. The evolution of mammalian transcriptional regulation remains largely unexplored beyond limited mouse-human comparisons.

We therefore identified the genome-wide binding of two transcription factors: (i) CEBPA, in livers of species representing five vertebrate orders: human (primate), mouse (rodent), dog (carnivora), short-tailed

opossum (didelphimorphia), and chicken (galliformes), and (ii) HNF4A, in livers from human, mouse, and dog. Chromatin immunoprecipitation experiments were combined with high-throughput sequencing (ChIP-seq) using healthy, nutritionally unstressed adult liver from the heterogametic sex as a functionally and transcriptionally conserved homologous tissue type.

For these two liver-specific transcription factors, binding events appear to be shared 10%-22% of the time between mammals from any two of the three placental lineages we profiled, separated by approximately 80 million years of evolution. This reveals a rapid rate of evolution in transcriptional regulation among closely related vertebrates. Nevertheless, the number of CEBPA and HNF4A transcription factor binding events shared between any two of our five study species is far greater than could have occurred by chance.

Understanding the evolutionary dynamics of transcription factor binding is essential to understanding the evolution of gene regulation. Our analysis of experimentally determined in vivo occupancy of two TFs in multiple vertebrates revealed apparent limitations to this model and a number of other insights about the complex relationship between genetic sequence, transcription factor binding, and genome regulation.

First, the vast majority of ChIP-identified transcription factor binding events are unique to each species; in mammals, the binding events that occur within species-specific, repetitive DNA are more common than conserved binding events. Second, ultrashared TF binding events, which are the functional counterpart of ultraconserved sequences, appear rarely in vivo among all five vertebrates. Third, only approximately half of binding events that are lost in one placental mammal yet present in at least two others are potentially recovered by nearby turnover events. Fourth, neither motif nor strength of TF binding correlate with conservation of a transcription factor's genomic occupancy. Alterations in the DNA binding specificity of CEBPA and HNF4A cannot account for rapid binding divergence, nor can species-specific environmental differences.

Nevertheless, comparing binding events within 10 kb of the transcription start site (TSS) of experimentally determined target genes of CEBPA and HNF4A has shown that binding events near these genes are more likely to be shared with other species, although this does not correspond to an increase in sequence constraint. In fact, the set of the ultra-shared, five-way binding events is entirely disjoint from the set of genes directly dependent on CEBPA in adult liver. For HNF4A, only 6% of binding events shared across three placental mammals are near the highest-quality functional target genes, namely, those genes that depend on HNF4A for proper expression in both mouse and human. Given that most TFs are active in multiple cell types, it is possible that the remaining shared sites are active in other tissues or other developmental stages. Indeed, the ultra-shared CEBPA binding events are uniformly found near liver-specific genes that would be expected to be upregulated upon liver organogenesis. Conversely, those binding events near functional targets in adult liver that are neither shared nor show signs of sequence constraint may represent lineage-specific regulatory interactions.

The preponderance of specific-specific binding and the rapid lineage-specific loss of binding events suggests that a sizeable majority of specific TF-DNA interactions could be evolving neutrally. Liver-specific TFs and subsequent gene expression are both highly conserved, the rapid gain and loss of binding events may be indicative of compensatory changes that maintain local concentrations of TF binding near functional targets. Indeed, a recent computational approach which uses a high concentration of TF binding motifs, regardless of their alignment, showed improved ability to predict regulatory interactions.

Despite the rapid gain and loss of TF binding events in mammals, tissue-specific gene regulation seems to be maintained by identifiable regulatory architectures that can be independent of sequence constraint.

Session 10 : Protein Sequence Analysis

AuPosSOM: New Approach for the Identification of Active Compounds in the Set of Docked Molecules

Alexey B. Mantsyzov[1], Guillaume Bouvier[1], Nathalie Evrard-Todeschi[1] and Gildas Bertho[1]

[1] Laboratoire de Chimie et de Biochimie Pharmacologiques et Toxicologiques, UMR 8601 CNRS, Université Paris Descartes, 45 rue des Saints-Pères, 75006 Paris, France

{alexey.mantsyzov, guillaume.bouvier, nathalie.evrard-todeschi, gildas.bertho}@parisdescartes.fr

Abstract *Docking techniques on therapeutic targets are widely used for the investigation of the protein-ligand interaction and virtual screening. The very important problem is to distinguish biologically active compounds from inactive ones in the large set of the docked molecules. AuPosSOM is a new software for the evaluation of the docking results. The approach is based on the clustering of the docked molecule by the similarity of their contacts with the target.*

Keywords Docking, scoring function, virtual screening, self organizing map, contact, activity, clustering.

AuPosSOM: Nouvelle Approche pour l'Identification de Composés Actifs dans un Ensemble de Molécules Dockées

Résumé *Les techniques de Docking sur des cibles thérapeutiques sont largement utilisées pour l'étude des interactions protéine-ligand et pour le criblage virtuel. Le plus gros problème concerne la distinction des composés biologiquement actifs de ceux inactifs dans un large ensemble de molécules dockées. AuPosSOM est un nouveau programme pour évaluer les résultats de docking. L'approche est basée sur le regroupement (clustering) des molécules par leur similarité de contacts avec la cible.*

Mots-clés Docking, fonction de score, criblage virtuel, self organizing map. contact, activité, clustering.

1 Introduction

Evaluation of the docking results is one of the most important problems of the virtual screening and *in silico* drug design. Modern approaches for the identification of active compounds in the large data set of the docked molecules are based on the scoring functions. Scoring function estimation is governed by the calculation of the ligand binding energies. The general and the most significant limitations of scoring function methods are dealt with the inaccurate binding energy estimation.

AuPosSOM (Automatic analysis of Poses using SOM) [1] represents the new approach that utilizes contact fingerprint similarity conception for the virtual screening. This tool is available on-line: www.aupossom.com. Kohonen self-organizing maps (SOM) method [2] is applied for the unsupervised clustering of docked compounds [3]. Ligands and decoys are arranged in the hierarchal tree with respect to the similarity of binding modes. The problem of the correct pose selection is solved by the statistical analysis of the contact information over all poses for the ligand. The results of the clustering may be presented as a tree where leaves contain compounds with the similar binding modes (Fig.1). Benchmark tests of AuPosSOM for the several targets have been performed. It revealed that the approach is as efficient as conventional energy-based scoring functions or gives better results.

2 Materials and Methods

In order to evaluate the efficiency of the AuPosSOM approach, docking tests were performed for the datasets from the DUD (Database of the Useful Decoys) [4]. 9 targets were selected for the evaluation of the AuPosSOM clustering efficiency (CDK2, COX1, DHFR, HIV protease, HIV RT, HSP90, PR, thrombin and trypsin).

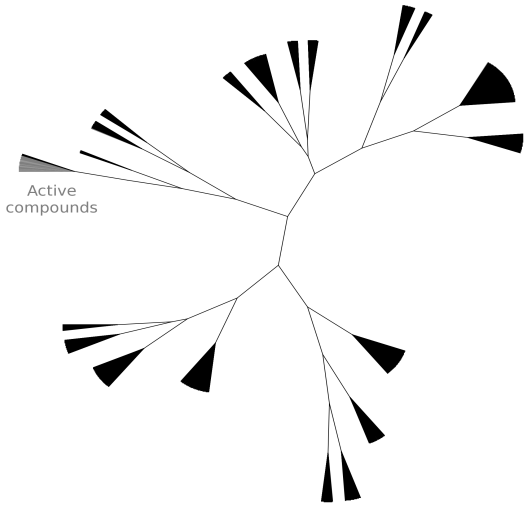


Figure 1. Example of the automatic clustering performed with AuPosSOM of ligands and decoys of DHFR (DUD database [4]) by the similarity of contacts with the target. 90.5% of active compounds were identified in the data set, which contains 201 active and 3318 inactive compounds.

Docking was performed by Surflex-Dock 2.0 from Sybyl 8.1.1 [5] package using mol2 files of targets and ligands. Protomol was generated with default parameters (threshold of 0.50 and bloat equal to 0). Each docking experiment was performed 20 times yielding 20 docked poses. Ligand energy minimization prior to docking and all-atom in-pocket minimization after docking was accomplished. Virtual screening with docking was performed on one Linux PC (quadricore Intel 2.66 GHz, 2 GB RAM).

The results of the docking were evaluated with AuPosSOM. From docking results mean vectors contacts involved in protein/molecule interactions were computed. Incremental subensemble are then used to train random SOM. Clustering was repeated 10 times to obtain representative results. For each trained SOM, all vectors were calibrated on it and then clustered according to the SOM. Each clustering can be visualized as a tree.

Four CScore scoring functions were utilized to compare AuPosSOM results with conventional scoring function approach: Chem score, PMF, G-score and D-score. The quality of the docking results evaluation was estimated by ROC curves (Receiver Operating Characteristic curves).

3 Results and Discussions

We applied AuPosSOM for the evaluation of nine datasets of the challenging database. Obtained results revealed that AuPosSOM clustering depends on the given data set and docking quality. For 6 out of 9 datasets clustering method gave better ROC curves than the best scoring functions; for the rest 3 datasets efficiency was approximately the same.

The important difference of the AuPosSOM approach from the scoring one is that it takes information for the contacts of all poses of the docked compound simultaneously. This allows to average imperfections of the docking and avoid errors related to the best pose search. The weak point of this approach might be inability to evaluate the results correctly when the number of the poses with correct set of contacts is low. In this case scoring function based approach might be possible to extract the right pose by energy estimation.

Meanwhile, in accordance with our results, scoring functions used in the tests were not better for difficult targets. The another important idea is that contact based approach does not take the conformation of the pose into consideration. It greatly simplifies analysis as the main requirement for the successful clustering is only the presence of the unique set of contacts for the active compounds but not the correct overall conformation of the pose. The last one is often hard to obtain especially for the ligands that were not extracted from the receptor's crystal structure used for docking.

Our results evidently demonstrate that AuPosSOM contact analysis may be more efficient than classical scoring function approach. Moreover, the clustering of compounds according to their contacts with the target will give information to identify key contacts which are specific and relevant for active compounds. This new approach provides an opportunity for the contact-activity relationship (CAR) analysis for the set of docked molecules.

Acknowledgements

This work was supported by the ANR, CNRS and University Paris Descartes.

References

[1] G. Bouvier, N. Evrard-Todeschi, J.-P. Girault and G. Bertho, Automatic clustering of docking poses in virtual screening process using self-organizing map. *Bioinformatics*, 26:53-60, 2010.

[2] T. Kohonen, *Self-Organizing Maps.* Springer Series in Information Sciences, Heidelberg, Germany, 2001.

[3] E.V. Samsonova, J.N. Kok and A.P. Ijzerman, TreeSOM: Cluster analysis in the self-organizing map. *Neural Netw.*, 19:935-949, 2006.

[4] N. Huang, B.K. Shoichet and J.J. Irwin, Benchmarking sets for molecular docking. *J. Med. Chem.,* 49:6789-6801, 2006.

[5] SYBYL-8.1.1., *Tripos International*, 1699 South Hanley Rd., St. Louis, Missouri, 63144, USA.

Ancestral HMMs and their Use to Detect Distant Homologs

Jean-Baka DOMELEVO ENTFELLNER and Olivier GASCUEL

Méthodes et Algorithmes pour la Bioinformatique, LIRMM, CNRS, Université Montpellier 2, France
{domelevo,gascuel}@lirmm.fr

Keywords Hidden Markov Models (HMMs), phylogeny, tree-HMM, remote homology.

Profile Hidden Markov Models are popular tools to model a family of aligned sequences. They have a canonical linear structure, Plan7, e.g. used in HMMER [1] (Fig. 1). These models emit residues on *Match* and *Insert* states, while *Delete* states allow to skip consensus residues ; all emissions and transitions are probabilistic. The score of a sequence through the profile HMM is the product of all the emission and transition probabilities it triggers along its best-scoring (Viterbi) path in the model. Homologous sequences are expected to yield high scores, but the standard HMM-based approach often fails to detect remote homologs. Following Mitchison & Durbin [2] and Qian & Goldstein [3], we propose to build a series of HMMs from a single multiple alignment, provided that we know the underlying phylogeny. We build one HMM on each of the nodes of the phylogenetic tree, with all the parameters of the resulting HMMs being the products of an ancestral reconstruction. As far as we know, our approach is novel regarding steps 1, 4 and 5 described in the following.

1 Method

Throughout this section, we will describe the process of building the HMM for some node n of a phylogeny $\mathcal{T}$. Obs is the observed data on the leaves of $\mathcal{T}$.

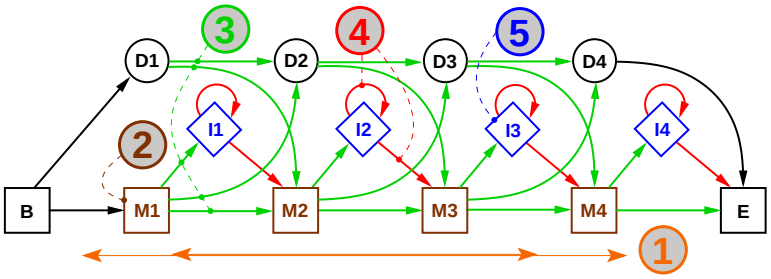


Figure 1. The traditional Plan7 structure of a profile HMM. Building an HMM for each node of the phylogeny breaks up in 5 steps, described below. Step 1 refers to the number of *Match* states and their mapping to columns of the alignment.

Step 1 shown in Fig. 1 consists in *selecting some columns* of the alignment to model them afterwards *as Match states*. Such a step is usually done with a heuristic (e.g. select the column iff it displays a fraction of gaps $\leq 50\%$). Here we proceed as follows: (1) transform the alignment of amino acids into an alignment on a binary alphabet, replacing any letter with 'N' and any gap with 'G' ; (2) learn a General Time-Reversible process on the resulting alignment ; (3) reconstruct the ancestral distribution on node n, and decide that a column is a *Match* iff the ancestral probability for 'N' is above a threshold (0.5 here, a value subject to optimisation).

Step 2 consists in determining the probability distributions for *residues emitted by Match states*, from the contents Obs of the relevant column (see [3] for a similar approach). We calculate the ancestral likelihood for each a.a. α on node n to use it as an emission probability: $\Pr(\alpha, n) \propto \Pr(\alpha)\mathrm{Lk}(n = \alpha, \mathcal{T}, Q|Obs)$. In this study, we use the LG model for Q.

Step 3 has also been treated by [3]. It concerns the setup of *phylogenetic-based transitions* leaving *Match* (M) or *Delete* (D) states. Through the alignment of transitions followed by the training sequences in the HMM, we build a phylogeny upon characters representing these transitions (e.g. $M \to D$ stands for a transition from a *Match* to a *Deletion*). It yields, e.g. for *Match* states: $\Pr(X, n) \propto \Pr(X)\mathrm{Lk}(n = X, \mathcal{T}, Q_M|Obs)$, where $X \in \{M \to M, M \to D, M \to I\}$. These probabilities are then used to weigh the corresponding transitions.

Step 4 is a second feature we introduce: while previous authors discarded the phylogenetic information contained in inserted regions, we choose to build a *phylogeny on the insert lengths* observed in sequences passing through any given *Insert* state. Such a phylogeny on characters drawn from $\mathbb{N}^+$ is analysed through an ad-hoc Birth-and-Death Markovian evolution model. With such a model, an ancestral distribution is calculated on node n. Identifying its first moment with the one of the geometric law yielded by the I state in the Plan7 architecture (Fig. 1, step 4), we deduce the appropriate value for the looping probability on the considered state.

Step 5 brings *phylogeny-based emissions from Insert states*. While the contents of insert columns are commonly disregarded to be replaced by a standard, hydrophilic-biased distribution, we take them into account by considering phylogenies where leaves do not carry a single residue, but a collection of them: if a certain sequence inserts AALV between two *Match* states, we consider the corresponding leaf in the phylogeny tying up insert contents to bear a *composite* character made out of $\frac{1}{2}$ A, $\frac{1}{4}$ L and $\frac{1}{4}$ V. This requires only a tweak in Felsenstein's pruning algorithm to calculate likelihoods and then deduce emission probabilities.

2 Data and Results

The score of a sequence against the $(2N-2)$ HMMs obtained from a phylogeny on N taxa is simply the best of its $(2N-2)$ scores. We test our approach on the same benchmark as employed by [3]: on each of 39 ASTRAL/SCOP superfamilies [4], we train the models on the sequences from all but one of the families belonging to the superfamily and test them on the entire database. Sequences from the left-out family form the set of true positives (214 in total over the 39 cases). The 4383 sequences in the database share $\leq 40\%$ sequence identity, which makes it a difficult issue. Results are shown in the form of a ROC curve in Fig. 2. Illustrated here on proteins, the same methodology should apply to DNA sequences with only minor and straightforward changes.

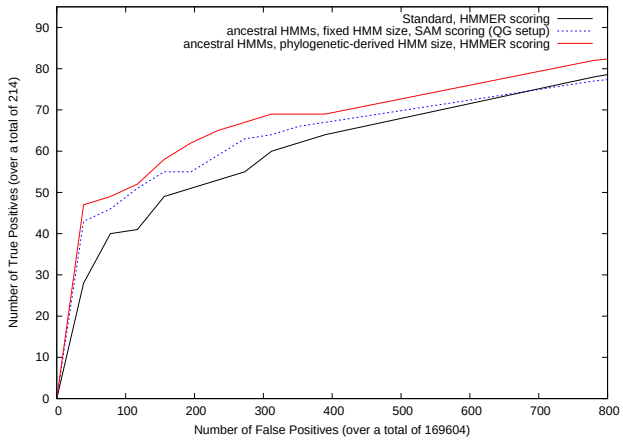


Figure 2. ROC curve showing the benefit of our models on 39 difficult cases of remote homology detection. The method implementing all the tools described herein is displayed in red. The blue dashed line corresponds to the models in [3].

Acknowledgements

We thank L. Bréhélin, J. Dutheil, A. Jean-Marie and V. Ranwez for fruitful discussions and help.

References

[1] SR Eddy, HMMER: Profile hidden Markov models for biological sequence analysis. *Washington University School of Medicine, St Louis, MO (http://hmmer.janelia.org/)*, 2000.

[2] G. Mitchison and R. Durbin, Tree-based maximal likelihood substitution matrices and hidden Markov models. *Journal of Molecular Evolution*, 41(6):1139–1151, 1995.

[3] B. Qian and R.A. Goldstein, Detecting distant homologs using phylogenetic tree-based HMMs. *Proteins: Structure, Function, and Bioinformatics*, 52(3):446–453, 2003.

[4] S.E. Brenner, P. Koehl, and M. Levitt, The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Research*, 28(1):254, 2000.

Fitting Hidden Markov Models of Protein Domains to a Target Species

Nicolas TERRAPON[1,2], Olivier GASCUEL[1] and Laurent BRÉHÉLIN[1]

[1] Méthodes et Algorithmes pour la Bioinformatique, LIRMM, CNRS - Université Montpellier 2, France
{laurent.brehelin, olivier.gascuel}@lirmm.fr

[2] Evolutionary Bioinformatics, Institute for Evolution and Biodiversity, Universität Münster, Deutschland
n.terrapon@uni-muenster.de

Keywords Protein domains, HMM correction, Malaria.

1 Introduction

Hidden Markov Models (HMMs) are a powerful tool for protein domain detection. The Pfam database notably provides a large library of HMMs for the annotation of sequenced organisms. When analyzing a new protein sequence, each Pfam HMM is used to compute a score that measures the similarity between the sequence and the domain family. If this score is above a recommended threshold, provided by Pfam, then the occurrence of the domain is *asserted* in the protein. However, when applied to highly divergent organisms, this strategy can miss numerous domains. For example, no Pfam domain is detected in 50% of the proteins of *Plasmodium falciparum* (the main causal agent of human malaria) and only 1 420 distinct families are identified among the 10 340 families defined in Pfam 23.0. In contrast, in *Saccharomyces cerevisiae*, 76% of all proteins contain at least one Pfam domain and 2 370 distinct Pfam families are involved. Although these observations can be explained by the existence of domains that are unique to a parasitic life style, it is likely further exacerbated by the A+T richness of the *P. falciparum* genome (76% on average in coding regions). This induces a compositional bias in the protein contents, as only 3 amino-acids (NIK) account for more than 35% of *P. falciparum* residues. This in turn makes homology detection particularly difficult. Two previous studies address the issue of HMM correction to enhance domain detection in a target species (hereafter denoted S). First, an *a posteriori* correction of domain scores has been proposed [1]. This correction takes the prior probability of each domain family in S into account. Prior probabilities are estimated using asserted domain occurrences in the closest relative of S (this species set, including S, is denoted R). A second approach builds taxon-specific models that integrate already asserted domain occurrences from R into the alignments used to learn the HMMs (*e.g.* [2]). However, these two approaches only allow the discovery of new occurences for the domain families already asserted in R.

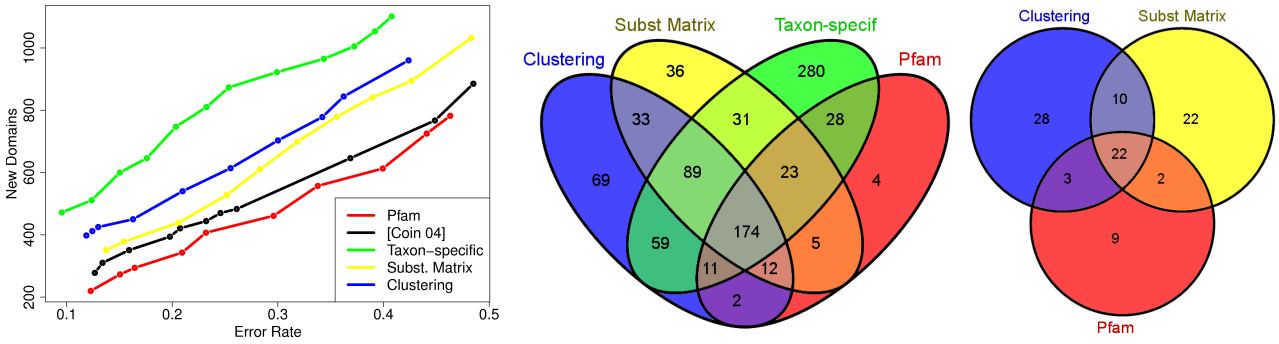
2 Methods

We propose two new approaches to circumvent this limitation by correcting the whole HMM library. These corrections involve learning general correction rules which are applied to the emission probabilities of the match states of all HMMs (match states represent the expected probability of observing a given amino acid at the corresponding position of the multiple sequence alignment defining the family). Our first approach requires the estimation of a substitution rate matrix for S. This matrix is built by combining the amino-acid frequencies observed in S, with the exchangeability coefficients of the LG matrix [3]. Using this substitution matrix, we then simulate the evolutionary drift of match states from standard species towards amino-acid composition of S. Our second approach involves partitioning all the match states of the Pfam library into clusters having similar amino-acid probability distributions, *i.e.* modeling common physico-chemical constraints. We apply a *K-means* procedure using the χ^2 distance to partition the match states into 100 clusters. Then, we use the asserted domain occurrences in R to align these amino-acid sequences with the successive match states of the models using the Viterbi algorithm. Each match state is thus associated with a set of specific amino-acids from sequences in R. Since each match state of the Pfam library is assigned to a unique cluster, a probability distribution is computed for each cluster from the amino-acids associated with all match states of the cluster. This distribution represents

the expected amino-acid distribution in R (and by extension in S) for this cluster and the corresponding physico-chemical constraints. Finally, the HMM library is corrected by combining, for each match state, the original emission probabilities with the computed distribution of the corresponding cluster. To assess the accuracy of the HMM libraries, we estimate the error rate of the *new* domains discovered by each library under various score thresholds. New domains are all predicted hits that are not asserted by the standard Pfam library. This error rate is estimated with a resampling procedure that exploits the well known tendency of domains to appear preferentially in proteins with a small set of favorite domains [4].

3 Results

We apply the four correction methods, taking *P. falciparum* as S and the *Alveolata* superphylum as R. Figure 1(a) shows the number of new domains according to the estimated error rates for each corrected library. For comparison purpose, we also estimate the error rate of the standard Pfam library when loosening the recommended threshold. We observe that all correction procedures improve the results achieved by the original Pfam library at equivalent error rates. The taxon-specific library, with 2 465 reconstructed HMMs, achieves the best results (green curve). The correction by clusters of states (blue) benefits from the combination of domain knowledge in R with the clustering of similar physico-chemical constraints over the whole library. The correction using the substitution matrix (yellow) may suffer from the employed standard substitution schema; exchangeability matrices model "universal" evolutionary mechanics, while *P. falciparum* is constrained by more extreme evolutionary circumstances. Finally, the taxonomic correction [1] (black) slightly improves the results of the original Pfam library (red). We illustrate by Venn diagrams (Fig. 1(b)) the common and specific new domains obtained by the different methods at 20% error rate. Each correction method discovers specific domains not discovered by the others. Moreover, the corrected libraries include most of the domains found by the standard Pfam library. To conclude, Figure 1(c) shows the new domains previously thought to be absent in alveolates and discovered by our two methods at 20% error rate. These new families, that cannot be identified by former approaches [1,2], reveal several new GO annotations that was previously unknown in *P. falciparum* and bring new insights into the biology of this complex organism.



(a) (b) (c)

Figure 1. Comparison of the HMM correction procedures. (a): Number of new domains (y-line) according to the estimated error rate (x-line). (b): Number of common and specific new domains at 20% error rate. (c): Number of new domains (20% error rate) previously thought to be absent in alveolates.

References

[1] L. Coin, A. Bateman and R. Durbin, Enhanced protein domain discovery using taxonomy, *BMC Bioinf.*, 5:56, 2004.

[2] I. Alam, S. Hubbard, S. Olivier and M. Rattray, A kingdom-specific protein domain HMM library for improved annotation of fungal genomes, *BMC Genomics*, 8:97, 2007.

[3] S. Le and O. Gascuel, An improved general amino acid replacement matrix, *MBE*, 25(7):1307-1320, 2008.

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

Communications affichées (revues par le CP)

Les contributions contenues dans cette section correspondent aux communications affichées reçues lors de l'appel à communications initial et ont été revues par le comité de programme (CP) de JOBIM.

Large-Scale Phylogenomic Analyses Indicate a Deep Origin of Primary Plastids within Cyanobacteria

Alexis CRISCUOLO and Simonetta GRIBALDO
INSTITUT PASTEUR, Unité de Biologie Moléculaire du Gène chez les Extrêmophiles, Département de Microbiologie, 25 rue du Dr Roux, 75015 Paris, France
{alexis.criscuolo, simonetta.gribaldo}@pasteur.fr

Keywords Origin of primary plastids, cyanobacteria, phylogenomics, ancestral sequences.

Des Analyses Phylogénomiques à Grande Échelle Montrent une Origine Ancienne des Plastes au Sein des Cyanobactéries

Mots-clés Origine des plastes primaires, cyanobactéries, phylogénomique, séquences ancestrales.

1 Introduction

L'apparition des eucaryotes photosynthétiques constitue un évènement majeur au cours de l'évolution, leur émergence ayant significativement modifié l'écologie de la planète. Il est largement reconnu qu'une unique endosymbiose entre un eucaryote hétérotrophe et une souche cyanobactérienne est à l'origine des plastes primaires. Toutefois, malgré un grand nombre d'analyses phylogénétiques mené depuis 25 ans, la lignée cyanobactérienne la plus proche de celle étant à l'origine des plastes primaires reste mal caractérisée. Toutefois, ces analyses s'appuient soit sur un large échantillonnage taxonomique mais peu de marqueurs phylogénétiques (e.g. [1]), soit sur un grand nombre de marqueurs mais peu de taxons (e.g. [2]). Cette question demeure encore ouverte car plusieurs singularités évolutives provoquent de nombreux biais durant l'analyse phylogénétique (i.e. transferts horizontaux de gènes entre cyanobactéries, forte hétérogénéité de composition entre génomes cyanobactériens et plastidiques, transferts de gènes entre plastes et le noyau de leurs hôtes, radiation évolutive importante chez les eucaryotes après l'endosymbiose primaire). Afin de minimiser ces sources de biais, un nombre important de marqueurs phylogénétiques et de taxons doit être considéré, ainsi que des outils méthodologiques adaptés, pour mener une analyse solide. En s'appuyant sur différents jeux de données de grandes tailles, nous avons conduit différentes analyses phylogénomiques afin de déterminer avec précision l'origine des eucaryotes photosynthétiques au sein de l'arbre des cyanobactéries.

2 Réalisations

Afin de conduire une analyse phylogénomique sur un jeu de données de taille importante, un grand nombre de génomes a été collecté : 61 de cyanobactéries, 22 de plastes (i.e. 1 glaucophyte, 5 algues rouges, 11 algues vertes, 5 plantes vertes), 11 de noyaux eucaryotes (i.e. 1 algue rouge, 5 algues vertes, 5 plantes vertes), ainsi que de nombreuses séquences EST issues de glaucophytes et d'algues rouges. Nous avons sélectionné 250 gènes cyanobactériens ayant au moins une séquence homologue au sein des eucaryotes considérés (i.e. plastes et/ou noyaux). Le développement d'une nouvelle méthode de recherche d'incongruence (basée sur des tests SH [3]) a ensuite permis d'identifier 191 marqueurs phylogénétiques orthologues. Quatre sous-ensembles de jeux de données ont été étudiés: restriction aux seules séquences de cyanobactéries (cyanobactéries ; 191 marqueurs phylogénétiques), séquences de cyanobactéries et leurs seuls homologues plastidiques (cyanobactéries+plastes ; 127 marqueurs) et nucléaires (cyanobactéries+noyaux ; 134 marqueurs), ainsi que le jeu de données complet (cyanobactéries+plastes+noyaux ; 191 marqueurs). Pour chacun de ces 4 jeux de données, les séquences de chaque marqueur ont été alignées et les caractères au sein de ces alignements contenant un signal phylogénétique pertinent ont été sélectionnés par un nouveau logiciel, BMGE [4]. Ayant montré que ces différentes séquences d'acides aminés souffraient d'un fort biais de composition, les états de caractère de ces 4 jeux de données ont été recodés : recodage en 4 classes d'états de caractère homogènes ('4-bin recoding' [5]), et recodage en codons dégénérés [4]. Grâce à l'utilisation de ces différentes méthodes, nous avons ainsi construit 4 jeux de données (i.e. cyanobactéries, 32193 caractères ; cyanobactéries+plastes, 18934 caractères ;

cyanobactéries+noyaux, 22019 caractères; cyanobactéries+plastes+noyaux, 30149 caractères), chacun disponible dans 3 versions différentes (i.e. acides aminés originaux, recodages '4-bin' et codons dégénérés). Les différents jeux de données ont été analysés phylogénétiquement en optimisant le critère du maximum de vraisemblance (ML). Dans le but d'obtenir les arbres phylogénétiques les plus fiables possibles, nous avons développé et utilisé un script informatique, nommé morePhyML, permettant d'améliorer les performances du logiciel PhyML [6] en implémentant la technique du 'ratchet' (i.e. bruitage des jeux de données afin d'échapper aux optima locaux au sein de l'espace de recherche de l'arbre ML [7]). En utilisant morePhyML sur les différentes versions de nos 4 jeux de données, nous avons donc inféré des arbres phylogénétiques robustes modélisant l'histoire évolutive des cyanobactéries et l'origine des plastes primaires. Les nouveaux résultats établis s'appuient sur le plus grand jeu de données cyanobactéries+plastes+noyaux jamais construit (i.e. 83 taxa, 191 marqueurs, >30000 caractères). De plus, afin d'utiliser la plus grande quantité possible de signal phylogénétique, nous n'avons pas restreint nos analyses à un sous-ensemble de séquences eucaryotes d'origines cyanobactériennes, contrairement aux approches récentes basées uniquement soit sur les séquences provenant des plastes [2], soit sur celles transférées dans le noyau (e.g. [8]). Afin de renforcer nos résultats, nous avons utilisé une approche originale consistant à construire la séquence ancestrale aux eucaryotes photosynthétiques pour chaque gène de notre jeu de données en utilisant des modèles d'évolution différents selon que les séquences sont issues de plastes ou de noyaux. Le taxon artificiel correspondant à ces séquences ancestrales a été ensuite inséré sur toutes les branches de l'arbre des cyanobactéries, et la probabilité de chacun de ces scénarios évolutifs a été estimée à l'aide de tests AU [9] à partir des différents encodages de séquences. Cette approche, minimisant les biais de composition et les artefacts d'attraction de longues branches, confirme le point d'émergence des plastes primaires estimé dans nos différentes analyses phylogénomiques.

3 Conclusion

Les différents arbres que nous avons inférés (i.e. cyanobactéries, cyanobactéries+plastes, cyanobactéries+noyaux, cyanobactéries+plastes+noyaux, chacun sous 3 encodages différents) sont très similaires entre eux et montrent donc une information phylogénétique solide. Il ressort de nos analyses que l'apparition des plastes est un événement ancien dans l'histoire évolutive des cyanobactéries, juste avant la diversification de la plupart des souches séquencées à ce jour. Ce résultat s'oppose à de récentes hypothèses [10] énonçant que les plastes primaires sont issus plus tardivement de l'endosymbiose d'une souche cyanobactérienne classifiée dans la sous-section IV (i.e. forme filamenteuse hétérocystée [11]). Finalement, nos analyses suggèrent qu'un séquençage futur doit cibler les lignées cyanobactériennes basales si l'on souhaite déterminer avec précision celle qui a été impliquée dans l'évènement d'endosymbiose primaire.

Références

[1] S. Turner, K.M. Pryer, V.P.W. Miao and J.D. Palmer, Investigating deep phylogenetic relationships among cyanobacteria and plastids by small subunit rRNA sequence analysis, *J. Eukaryot. Microbiol.*, 46:327-38, 1999.

[2] N. Rodriguez-Ezpeleta, H. Brinkmann, S.C. Burey, B. Roure, G. Burger, W. Löffelhardt, H.J. Bohnert, H. Philippe and B.F. Lang, Monophyly of primary photosynthetic eukaryotes : green plants, red algae, and glaucophytes, *Curr. Biol.*, 25:1325-30, 2005.

[3] H. Shimodaira and M. Hasegawa, Multiple comparisons of log-likelihoods with applications to phylogenetic inference, *Mol. Biol. Evol.*, 16:1114-6, 1999.

[4] A. Criscuolo and S. Gribaldo, BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments, *BMC Evol. Biol.*, 10:210, 2010.

[5] E. Susko and A.J. Roger, On reduced amino acid alphabets for phylogenetic inference, *Mol. Biol. Evol.*, 24:2139-50, 2007.

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

[7] D.A. Morrison, Increasing the efficiency of searches for the Maximum Likelihood tree in a phylogenetic analysis of up to 150 nucleotide sequences, *Syst. Biol.*, 56:988-1010, 2007.

[8] P. Deschamps and D. Moreira, Signal conflicts in the phylogeny of the primary photosynthetic eukaryotes, *Mol. Biol. Evol.*, 26:2745-53, 2009.

[9] H. Shimodaira, An approximately unbiased test of phylogenetic tree selection, *Syst. Biol.*, 51:492-508, 2002.

[10] O. Deusch, G. Landan, M. Roettger, N. Gruenheit, K.V. Kowallik, J.F. Allen, W. Martin and T. Dagan, Genes of cyanobacterial origin in plant nuclear genomes point to a heterocyst-forming plastid ancestor. *Mol. Biol. Evol.*, 25:748-61, 2008.

[11] R. Rippka, J. Deruelles, J.B. Waterbury, M. Herdman and R.Y. Stanier, Generic assignments, strain histories and properties of pure cultures of Cyanobacteria, *J. Gen. Microbiol.*, 111:1-61, 1979.

Conformément au souhait des auteurs, cette contribution n'est pas reproduite dans la version en ligne des actes de JOBIM 2011.

Following the wishes of the authors, this paper is not included in the online version of the JOBIM 2011 proceedings.

Estimating Phylogenetic Correlations between Molecular data and Longevity in Mammals

Raphaël POUJOL[1] and Nicolas LARTILLOT[1]

Centre Robert Cedergren, UdeM, Montréal, Université de Montréal, Montréal, Québec, Canada,
raphael.poujol@umontreal.ca, nicolas.lartillot@umontreal.ca

Abstract *Studies on ageing have proposed several genes playing a role in prevention of cell degeneration. Assuming that these genes are subject to a more stringent selective pressure in long-lived species, our laboratory use Bayesian modeling to infer their involvement in senescence. We expose here an evolutionary approach for a genome-wide study linking genes and longevity in mammalian species. We describe a Bayesian phylogenetic model, where we investigate correlations between genes and senescence. Based on the sequences and the phylogeny of on the 36 fully sequenced mammalian species.*

Keywords molecular evolution, phylogeny, bayesian sampling, mammals, ageing.

1 Introduction

Studies on the biology of senescence suggest that it is due to the accumulation of biochemical damage in DNA, proteins and lipids. Many genes and pathways have been proposed to play a role in prevention of cell degeneration and premature ageing such as anti-oxidant pathway, DNA reparation or protein recycling system.

Some evolutionary theories of ageing as mutation accumulation theory [1] and disposable soma theory [2], assume a correlation between selection pressure and longevity. In a population where environmental pressure down regulates life expectancy, deleterious mutations occurs more frequently in late-acting genes. On the contrary, the gene will be under a high strength of stabilizing selection when ageing is only the consequence of somatic damage. According to that, we are assuming in this study that the correlation between gene specific selection pressure and longevity depend whether or not this gene is late acting.

The estimate of this correlation should be made in a phylogenetic framework, in order to dissociate the dependencies due to evolutionary inertia. Using a codon substitution model, we are able to estimate the variations of gene specific selective pressure using ω [3] the ratio of non-synonymous (dN) to synonymous (dS) substitution rates over time ($\omega = dN/dS$). Lower values of ω indicate a stronger selective pressure. Therefore, when ω is negatively correlated with longevity (i.e. this gene has been under more intense purifying selection in long-living species) the gene is likely involved in the regulation of ageing.

Numerous observations show a negative correlation between population size and longevity. It can be explained by the neutral theory of evolution [4], which state that the fixation rate of synonymous mutations strongly vary over time, depending on population size. Therefore, we expect a positive correlation between the genome wide longevity and ω. Because effective population size is hard esitmate, we decide to build a single model, computing simultaneously all the correlations (α_i) and then an estimate of $\bar{\alpha}$ the correlations of the coding sequences with longevity. Another problem, is that estimations of longevity in different species are neither accurate or very discussed, consequently we introduce different life traits such as female maturity, mass, and metabolism. Our model will estimate their history and the correlation between each pair of traits in order to have a more precise longevity history.

2 Implementation and results

The model is implemented in $C++$, in the Bayesian Monte Carlo framework from coevol software [5]. We use Metropolis Hastings algorithm to sample all the parameters from their posterior distribution. The method

allows then to have a good estimate of the posterior probability of each value according to the data and the model. Because of the great amount of data needed for a genome wide study in a single model, parallelisation was added to the framework using a message passing interface. The connected graph structure of the model made this parallelisation very challenging, but we were able to take advantage of the hierarchical form of our model. We model all the history of continuous data as a Brownian process. All histories of the life traits and the genomewide selection pressure have been linked in a covariance matrix. Branch lengths were also estimated using fossil calibrations. We marginalize the covariance between ω and longevity and compute its gene specific posterior probability to be greater than the genome wide coupling.

We apply the model to 36 species fully sequenced with multiple alignments from the Orthomam database [6] and life traits data from AnAge database [7]. Preliminary results showed a good power of the model. This model can also be used to address further question about interdependence between molecular evolution and phenotypic continuous values or environmental factors such as temperature or atmospheric oxygen level.

Acknowledgements

This work was supported by NSERC : Natural Sciences and Engineering Research Council of Canada

References

[1] P. Medawar, An Unsolved Problem of Biology, *H.K. Lewis & Co.*, London, 1952.
[2] T.B.L. Kirkwood, Evolution of ageing, *Nature*, 31:301-304, 1977.
[3] Y. Suzuki and T.F. Gojobori, A method for detecting positive selection at single amino acid sites, *Molecular Biology and Evolution*, 16: 1315-1328, 1999.
[4] M. Kimura, The neutreal theory of neutral evolution *Cambridge*, 1983.
[5] N. Lartillot and R. Poujol, A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters, *Molecular Biology and Evolution*, 28(1):729-44, 2011.
[6] V. Ranwez, F. Delsuc, S. Ranwez, K. Belkhir, M. Tilak and E. J. P. Douzery, A database of orthologous genomic markers for placental mammal phylogenetics, *BMC Evolutionary Biology*, 7 : 241, 2007.
[7] J. P. de Magalhaes, and J. Costa, A database of vertebrate longevity records and their relation to other life-history traits, *Journal of Evolutionary Biology*, 22(8):1770-1774, 2009.

Peptidergic Signaling Systems in Bilaterian Genomes

Olivier MIRABEAU[1] and Jean-Stéphane JOLY[1]
[1] Institut de Neurobiologie Alfred Fessard (INAF), CNRS, 1 Avenue de la Terrasse, 91198 Gif sur Yvette,
mirabeau@inaf.cnrs-gif.fr

Keywords molecular evolution, neuropeptide evolution, GPCR evolution, hidden Markov models, urbilateria.

1 Introduction

Peptides and their corresponding G protein-coupled receptor (GPCR) genes are essential for the physiology and behavior of animals. We have screened the publicly available genomes to ask whether invertebrate genomes possess vertebrate-type peptide genes and their associated GPCR genes. In our molecular evolution context we define a peptidergic system (PS) to be the association of a group of related peptides with a group of cognate receptors. A vertebrate PS was hypothesized to be present in a given non-vertebrate genome when either a peptide precursor gene or a GPCR was found that belonged to that class of peptide or receptor.

The main purpose of this study was to clarify the evolutionary link between deutorostomian and protostomian peptidergic systems in bilaterians. Recent studies indicate that several vertebrate peptidergic systems are present in invertebrates including the CCK/sulfakinin [1] and GNRH/adopokinetic hormone [2] signaling systems in *C. elegans*. However comprehensive characterization of both peptide and GPCR repertoires have been restricted to single genomes like *C. elegans* [3].

2 Results

2.1 Analysis of GPCR Phylogenetic Trees

To study the origin of peptide GPCRs of subtype Rhodopsin, according to the GRAFS classification [4] we adopted a standard phylogenomics pipeline. Briefly it consisted in the following:

1- All annotated human GPCR protein sequences were downloaded from the Swissprot database.

2- Protein sequences derived from complete sets of gene models were retrieved from public databases from the Joint Genome Institute, Ensembl project and the Baylor Genome Center. The organisms that were surveyed are the following: two vertebrates, *Homo sapiens* and *Takifugu rubripes*, two non-vertebrate chordates *Branchiostoma floridae* and *Ciona intestinalis*, two non chordate deuterostomes, the urchin *Strongylocentrotus purpuratus* and *Saccoglossus kowalevskii*, two lophotrochozoans, *Capitella telata* and *Lottia gigantea*, one nematode *Caenorhabditis elegans* and three arthropods *Daphnia pulex*, *Tribolium castaneum* and *Drosophila melanogaster*.

3- Reciprocal BLAST analysis of human genome vs. genome in consideration was performed to cluster large groups of related sequences.

4- These groups of sequences were aligned to derive phylogenetic trees using PhyML [5], a maximum likelihood-based method for gene tree reconstruction.

5- Presence of ancestral systems and phylum/species-specific history (large-scale duplications, losses) were inferred from the analysis of the trees.

We found several robust subtrees that have a topology congruent with the species tree. The structure of these subtrees strongly suggests that these receptors occurred before the split of deuterostomes and protostomes.

2.2 Analysis of Homologous Peptide Precursors

Conservation between homologous peptide precursor sequences from different phyla (e.g. nematode vs. arthropds) is usually restricted to very few amino acids in the peptide region, that are buried inside larger precursors. As a result, standard phylogenomics approaches are not applicable in this case. First we established a list of potential peptide precursors using a modified version of the hidden Markov model (HMM)-based program described in [6]. These candidates were then screened for the presence of short conserved motifs often found at the C-terminal end of peptides (e.g. RF-amides). Often the general structure of homologous peptide precursor genes was found to be conserved, including the position of the peptide and the overall length of precursor. In most cases the conclusion of our "ligand" and of our "receptor" analysis were consistent (e.g. absence of both vasopressin/oxytocin peptides and receptors in flies), indicating that our bioinformatics method is relevant for studying the history of this complex gene family.

2.3 Conclusion

Our results lend further support to the theory that urbilateria was an animal with a sophisticated physiology and nervous system, capable of integrating complex sensory information. It strengthens the case for using alternative models such as lophotrochozoans (e.g. capitella) to study how sensory information is integrated in animals.

Functional studies to test for the binding of putative peptides to their predicted receptors, as well as expression studies of those genes and protein products, are needed to further substantiate the homology hypotheses.

We believe that some of these newly established homologies will provide the evo-devo community with new markers for the study of ancestral cell types [7, 8], yield insights into the fundamental functions of vertebrate peptidergic systems and offer new molecular data for computational biologists interested in peptide-receptor coevolution studies.

Acknowledgements

Vincent Lefort for help with running PhyML on large sets of genes, Laurent Bréhélin and Olivier Gascuel for helpful discussions about HMMs and phylogenetic inference, Robert Freeman for the Saccoglossus gene models set and Vincent Laudet for advices on how to present the data. This work was supported by a postdoc grant from the Fondation pour la Recherche médicale.

References

[1] T. Janssen, E. Meelkop, M. Lindemans, K. Verstraelen, S. Husson, L. Temmerman, R. Nachman and L. Schoofs, Discovery of a cholecystokinin-gastrin-like signaling system in nematodes. *Endocrinology*, 149: 2826-2839, 2008.

[2] M. Lindemans, F. Liu, T. Janssen, S. Husson, I. Metrens, G. Gäde and L. Schoofs, Adipokinetic hormone signaling through the gonadotropin-releasing hormone receptor modulates egg-laying in Caenorhabditis elegans. *Proc Natl Acad Sci U S A.*, 106:1642-1647, 2009.

[3] A. Nathoo, R. Moeller, B. Westlund and A. Hart Identification of neuropeptide-like protein gene families in Caenorhabditis elegans and other species. *Proc Natl Acad Sci U S A.*, 98:14000-14005, 2001.

[4] R. Fredriksson, M. Lagerström, L. Lundin and H Schiöth, The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. *Mol Pharmacol.*, 63:1256-1272, 2003.

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

[6] O. Mirabeau, E. Perlas, C. Severini, E. Audero, R. Possenti, E. Birney, N. Rosenthal and C. Gross, Identification of novel peptide hormones in the human proteome by hidden Markov model screening. *Genome Res.*, 17:320-327, 2007.

[7] K. Tessmar-Raible, F. Raible, F. Christodoulou, K. Guy, M. Rembold, H. Hausen and D. Arendt, Conserved sensory-neurosecretory cell types in annelid and fish forebrain: insights into hypothalamus evolution. *Cell*, 129:1389-1400, 2007.

[8] O. Hobert, Regulatory logic of neuronal diversity: terminal selector genes and selector motifs. *Proc Natl Acad Sci U S A.*, 105:20067-20071, 2008.

Early Effect of Antiviral Therapy on HIV-1 Metapopulations

Sophie BROUILLET [1], Mary KEARNEY [2], Frank MALDARELLI [2], John COFFIN [2] and Guillaume ACHAZ [1]

[1] ATELIER DE BIOINFORMATIQUE & UMR7138, UPMC, 4 place Jussieu, 75252 Paris Cedex 05
{sophieb, achaz}@abi.snv.jussieu.fr

[2] HIV DRUG RESISTANCE PROGRAM, NCI, Frederick, MD, USA
{kearneym, fmalli, coffinj}@mail.nih.gov

Keywords HIV-1, population genetics, metapopulations.

A metapopulation can be defined as a population subdivided into several subpopulations. Individuals from the same subpopulation are more genetically related to each other than individuals from different subpopulations. We apply a statistical permutation test for subdivision to temporal data to assess the genetics differences between samples of the same population taken at different time points. A graphical implementation of the test is available online at http://wwwabi.snv.jussieu.fr/public/mpweb/.

We compared all pairs of 18 samples (278 sequences of 991-nt) taken from a single patient shortly before and after the start of the anti-retroviral therapy (hereafter ART). The sequences cover the whole protease region and the beginning of reverse-transcriptase. We found that the 10 samples before ART show an excess of differences to the 8 samples after ART; excess when compared to the differences we observe within the 10 samples before ART or within the 8 samples after ART. Importantly, this pattern was also observed, to a lesser extent, on another patient we looked at. These results highly suggest that the therapy has changed (directly or indirectly) the global composition of the viral population.

We then characterized the molecular basis of the differences. We found that reducing the 991nt to 2 selected ones leaves unchanged the observed differences. These 2 sites have a different frequency in the samples before or after ART. Further investigations of the phylogeny of the viruses revealed that a whole clade of viruses was only present before ART, suggesting that a type of virus has disappeared after the start of ART.

Out of the several interpretations that can be put forward, we favor two: (a) the ART has a different action on the different types of virus that inhabit the host or (b) the ART has an identical effect on all types of virus (i.e. it stops the viral replication) but the half-life of the infected cells differs from type to type. .

A Block Regression approach for Simultaneous Variables Clustering and Selection: Application to Genetic Data

Loïc YENGO[1,2], Julien JACQUES[1] and Christophe BIERNACKI[1]

[1] Laboratoire Paul Painleve, UMR8524 CNRS, UFR de Mathematiques, F-59655 Villeneuve d'Ascq, France
[2] Institut de Biologie de Lille, UMR8199 CNRS, 1 rue du Professeur Calmette, B.P 245, F-59019, Lille Cedex, France
loic.yengo@good.ibl.fr
Julien.Jacques@polytech-lille.fr
Christophe.Biernacki@math.univ-lille1.fr

Keywords dimension reduction, EM algorithm, variable clustering and selection.

1 Introduction

Genome Wide Association studies have uncovered the implication of numerous single nucleotides polymorphisms (SNP) in the aetiology of common diseases. Nevertheless, only a small part of the expected heritabiliy is explained by those variants. A large number of researches that have been lately investigating this missing heritability have considered interactions between genes and/or environmental factors as a plausible and promising explanation. Considering all if not a large number (thousands) of variants altogether, as underlain by the latter hypothesis, stresses the problem of the high dimensionality that most regression-based methods cannot afford. To solve this problem one either reduces the number of variants to be analyzed (Variable selection, LASSO or elastic net) or groups them according to a certain similarity (OSCAR [1] or PLS regression). We introduce here a regression-based method that simultaneously clusterizes the variants sharing close effect size while selecting the most informative clusters. Our approach differentiates from a method like group LASSO since no penalization is used and the clusters are not predefined. This method assumes a high level of sparsity and uses the EM algorithm (see [2]) to conduct maximum likelihood estimation (MLE) in the presence of unobserved grouping variables. Our approach offers a wider flexibility than preexistent methods since it can account for supplementary genetic information such as variants position and linkage desequilibrium. The results are presented in the context of linear regression models.

2 Block Regression Methodology

Notations Consider the usual linear regression model with observed data on n observations and p predictors. Let $\mathbf{y} = (y_1 \dots y_n)'$ to be a vector of responses (say a quantitative phenotypic trait), $\mathbf{x}_j = (x_{1j} \dots x_{nj})'$ denote the j^{th} predictor (say a SNP), $j = 1, \dots, p$ and ε_i a vector of independant error terms, each following a centered normal distribution of variance σ^2. The number of predictors being very large with respect to the number of observations, we cannot uniquely estimate the effect size of each predictor: β_j. We therefore assume the existence of g groups $(G_1 \dots G_g)$ of predictors such as all predictors in the group k $(k = 1, \dots g)$ have exactly the same effect size: b_k. We thus define z_{jk} as a random variable that indicates whether the predictor $\mathbf{x}_j$ belongs to the group G_k. $\mathbf{Z}$ is defined as the set of all the z_{jk} and $\mathbf{X}$ is the set of all the predictors.

Model for variables clustering The regression model given by the following equation

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij} \left(\sum_{k=1}^{g} b_k z_{jk} \right) + \varepsilon_i = \beta_0 + \sum_{k=1}^{g} b_k \left(\sum_{j=1}^{p} z_{jk} x_{ij} \right) + \varepsilon_i. \quad (1)$$

defines entirely the law $p(\mathbf{y}|\mathbf{Z}, \mathbf{X})$. The law $p(\mathbf{Z}|\mathbf{X})$ is chosen so as to integrate genetic information that could help infering the grouping structure. This very critical choice is expected to improve the performance of the model both in terms of prediction and biological relevance.

Maximum Likelihood Estimation As $\mathbf{Z}$ is totally unobserved, the calculation of the likelihood becomes intractable. MLE can therefore be obtained using the EM algorithm. The complete log-likelihood $p(\mathbf{y}, \mathbf{Z}|\mathbf{X})$ is expressed as following:

$$\begin{aligned}\log p(\mathbf{y}, \mathbf{Z}|\mathbf{X}) \quad = \quad & -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0)^2 + \frac{1}{2\sigma^2}\sum_{i,j,k}\left(2b_k x_{ij}(y_i - \beta_0) - (b_k x_{ij})^2\right) z_{jk} \\ & -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{j,k}\sum_{j\neq j', k\neq k'} 2b_k b_{k'}\left(x_{ij}x_{ij'}\right) z_{jk}z_{j'k'} + \log p(\mathbf{Z}|\mathbf{X}).\end{aligned}$$

The **E**-step requires the calculation of the following quantities $p(z_{jk}|\mathbf{X}, \mathbf{y})$, $p(z_{jk}z_{j'k'}|\mathbf{X}, \mathbf{y})$ and $\mathrm{E}_{\mathbf{Z}|\mathbf{X},\mathbf{y}}[\log p(\mathbf{Z}|$ that are analytically intractable.

Three strategies are then proposed to sidestep this limitation.

- We first suggest to maximize the complete likelihood alternatively with respect to $\mathbf{Z}$ and $\mathbf{b}$. This strategy also referred to as Classication EM (CEM) was already explored by Govaert and Nadif (see [3]).
- The second strategy suggests to approximate numerically the **E**-step using Markov Chain Monte Carlo algorithms such as the Gibbs sampling. The performances of this strategy known as Monte Carlo EM (MCEM) were extensively explored by Levine and Casella (see [4]).
- The third approach consists in the use of a variational approach (see [5]) which no longer maximizes the likelihood, but a lower bound well chosen so that to be as tight as possible.

Model for variables selection The selection purpose is a achieved by imposing a special class to have an exactly null coefficent. For instance, such constraint on G_1 translates as $b_1 = 0$. The equations underlying the three approaches presented above can still be easily derived under this constraint.

2.1 Numerical Experiments

The three strategies presented above will be compared in terms of quality of estimation and prediction. An application to GWA data will subsequently be achieved as well as a comparison to classical approaches. The influence of different choice of the law $p(\mathbf{Z}|\mathbf{X})$, including models published in the literature (for instance [6]), will also be explored.

References

[1] H. Bondell and B. Reich, Simultaneous regression shrinkage, variable selection and supervised clustering of predictors with OSCAR. *Biometrics*, 64:115-123, 2008.

[2] A. Dempster, N. Laird and D. Rubin, Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc.*, 39:1-38, 1977.

[3] G. Govaert and M. Nadif, An EM algorithm for the block mixture model. *IEEE Trans. Pattern Anal. Machine Intel.*, 27:643-647, 2005.

[4] R. Levine and G. Casella, Implementations of the Monte Carlo EM algorithm, *Journal of Computational and Graphical Statistics*, 10:422-439, 2001.

[5] M. Jordan, Z. Ghahramani, T. Jaakkola and L. Saul, An introduction to variational methods for graphical models. *Mach. Learn.*, 37:183-233, 1999.

[6] S. Kim and E. Xing, Feature Selection via Block-Regularized Regression. *Proceedings of the 24th Conference on Uncertainty in AI (UAI)*, 2008.

Differential Selection Profiles Using Statistical Phylogenetic Models For Understanding HIV Adaptation According To Host HLA

Sahar PARTO[1] and Nicolas LARTILLOT[1]
[1] ROBERT-CEDERGREN CENTER FOR BIOINFORMATICS AND GENOMICS, DEPARTMENT OF BIOCHEMISTRY, UNIVERSITÉ DE MONTRÉAL, 2900 Edouard-Montpetit, H3C 1J4, Montreal, QC, Canada
{sahar.parto, nicolas.lartillot}@umontreal.ca

Keywords selection, escape mutation, virus adaptation, Bayesian, phylogenetic, MCMC, HLA.

1 Introduction

Acquired Immunodeficiency Syndrome (AIDS), the disease caused by the progression of HIV (Human Immunodeficiency Virus), is one of the most challenging current diseases and does not have any definite cure or vaccine yet. The extensive rate of HIV mutation and adaptation makes the design of vaccine difficult, as it enables the virus to escape from the immune system (escape mutation) [1]. This HIV adaptation is due to its genetic diversity which is the result of its fast replication cycle and large population size, and its high mutation rate of $3x10^{-5}$ per nucleotide per cycle of replication [2]. Frequent recombination and the natural selection driven by immune system even intensify this diversity by creating an additional mechanism for virions to share beneficial mutations between individuals in a population.

So the first step for designing efficient vaccine is to identify consistent patterns in viral adaptation, as a function of the specific genetic background of the host. It has been shown that polymorphisms in HIV-I are associated with particular host HLA (Human Leukocyte Antigen) alleles [3, 4]. For example, HLA-*B57* and *B27* are associated with long-term HIV control and are likely to exert strong selection pressure on the virus. This association confirms the effect of HLA-restricted CTL (Cytotoxic T-Lymphocyte) response on HIV evolution.

2 Differential Mutation-Selection Model

Modeling the interplay between mutation and selection at the molecular level is one of the major goals in molecular evolutionary studies. Estimation of evolutionary patterns from homologous sequences is crucial for understanding the evolutionary processes like mutation rate and selection pressures. In recent years, codon-based evolutionary modeling efforts have increasingly been used to devise more realistic discription of the substitution process in protein coding genes [5-7].

In this study, a differential mutation-selection model is developed for HIV genes which parameterizes mutational and selective effects bearing on the overall substitution process. It is implemented in a phylogenetic Bayesian MCMC (Markov Chain Mont Carlo) object-oriented framework which allows us to tease out each parameter of the model from their joint posterior distribution and estimate differential selection profiles; one distinct selection profile is estimated for each host genetic background and specifies which amino acids are selected for or selected against at each position of the viral coding sequences. The different conditions are defined as *B-57* positive and *B-57* negative hosts which show different progression of the dissease.

This model is used to analyze the data of 445 gag sequences from 124 patients with identified genetic immune profile and HLA type. The phylogenetic tree of the sequences is shown in figure 1. The differential selection pressure is estimated between sequences in *B-57* $^{+}$ and *B-57* $^{-}$ hosts. It is also possible to estimate the mutation rate and codon usage bias of HIV and compare it with that of human in which the virus replicates.

By associating specific viral adaptation with specific host genetic background, it is possible to understand how HIV escapes from immune system, which in turn provides useful guideline to design an efficient vaccine against AIDS.

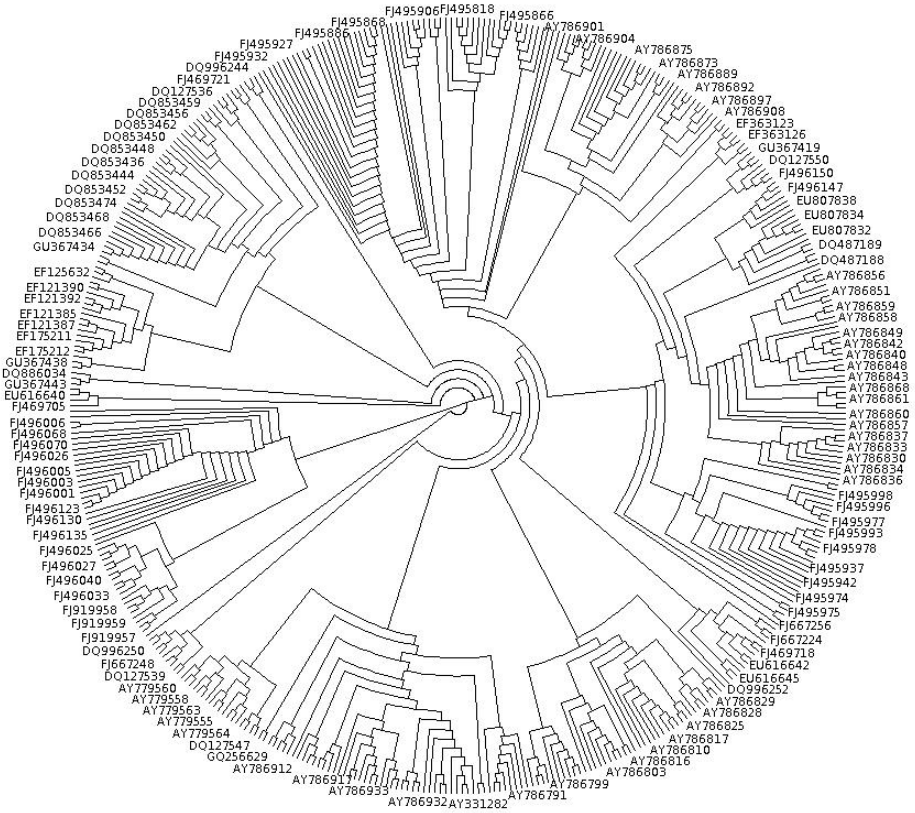


Figure 1. Phylogenetic tree of 445 HIV gag sequences.

Acknowledgements

This work was supported by National Sciences and Engineering Research Council of Canada (NSERC).

References

[1] P.J. Goulder and D.I. Watkins, HIV and SIV CTL escape: implications for vaccine design. *Nature reviews. Immunology*, 4:630-640, 2004.

[2] D.L. Robertson, B.H. Hahn and P.M. Sharp, Recombination in AIDS viruses. *Journal of molecular evolution*, 40:249-259, 1995.

[3] J.M. Carlson, Z.L. Brumme, C.M. Rousseau, C.J. Brumme, P. Matthews, C. Kadie, J.I. Mullins, B.D. Walker, P.R. Harrigan, P.J.R. Goulder and D. Heckerman, Phylogenetic dependency networks: inferring patterns of CTL escape and codon covariation in HIV-1 Gag. *PLoS computational biology*, 4:e1000225, 2008.

[4] C.B. Moore, M. John, I.R. James, F.T. Christiansen, C.S. Witt and S.A. Mallal, Evidence of HIV-1 Adaptation to HLA-Restricted Immune Responses at a Population Level. *Science*, 296:1439-1443, 2002.

[5] A.L. Halpern and W.J. Bruno, Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Molecular biology and evolution*, 15: 910-917, 1998.

[6] S.V. Muse and B.S. Gaut, A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular biology and evolution*, 11:715-724, 1994.

[7] N. Rodrigue, H. Philippe and N. Lartillot, Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles, *Proceedings of the National Academy of Sciences of the United States of America*, 107:4629-4634, 2010.

Cross-Species Comparison of *cis*-Regulatory Motifs: the Case Study of AP-1 Transcription Factors in Yeasts

Christel GOUDOT[1,2,3], Catherine ETCHEBEST[1,2,3], Frédéric DEVAUX[4] and Gaëlle LELANDAIS[1,2,3]

[1] Dynamique des Structures et Interactions des Macromolécules Biologiques (DSIMB), INSERM U665, Paris, F-75015

[2] Université Paris Diderot – Paris 7, UMR-S665, Paris, F-75015

[3] INTS, Paris, F-75015

{christel.goudot, catherine.etchebest, gaelle.lelandais}@univ-paris-diderot.fr

[4] Laboratoire de Génomique des Microorganismes, UMR7238 CNRS, Université Pierre et Marie Curie, Paris, F-75006

frederic.devaux@umpc.fr

Keywords Transcription factors; transcriptional modules, comparative functional genomics; *cis*-regulatory motifs, protein/DNA interactions.

1 Introduction

AP-1 proteins are transcription factors that belong to the basic leucine zipper family [1]. In yeasts, the AP-1 transcription factors Yap1p (in *Saccharomyces cerevisiae*), Cgap1p (in *Candida glabrata*) and Cap1p (in *Candida albicans*) play a central role in oxidative stress response and multidrug resistance [2,3,4] (Figure 1A). They are able to recognize DNA motifs (referred as YRE for Yap response element) that are palindromic or pseudo-palindromic sequences starting with a TTA or a TGA triplet with one or two central (C/G) base pairs [4,5,6]. To better understand the mechanisms that underlines the DNA binding specificity of yeast AP-1 transcription factors, we *(i)* combined transcriptomic and ChIP-chip data to infer species-specific transcriptional modules for Yap1p, Cgap1p and Cap1p (*i.e.* set of target genes) responding to the antifungal agent benomyl (Figure 1B), and *(ii)* performed a cross-species comparison of these modules, accurately inspected cis-regulatory motifs in promoter sequences of genes (Figure 1C).

2 Results

In each yeast species, *cis*-regulatory motif analyses revealed the presence of a conserved adenine in 5' position of the canonical YRE sites. Also, an impressive conservation was observed in the YRE consensus sequence (5'-MTKASTMA) over-represented in Yap1p (*S. cerevisiae*) and Cap1p (*C. albicans*) dependent genes. In Cgap1p (*C. glabrata*) dependant genes, two different YRE consensuses (5'-ATTACHAAW and 5'-MTTASSTAA) were identified and strongly suggested that Cgap1p, unlike Yap1p and Cap1p, tolerates YRE motifs with one or two central (C/G) base pairs. These findings were supported by structural data that show the interaction between the Schizosaccharomyces pombe Yap1p orthologue (Pap1p) and a DNA target sequence (Figure 1D).

3 Conclusions

We inferred condition-specific transcriptional modules associated to orthologous AP-1 proteins in three different yeast species, using an integrative framework that combined multiple sources of experimental data and multiple bioinformatics approaches. Exploitation of these modules in terms of predictions of the protein/DNA regulatory interactions considerably changed our vision of yeast AP-1 transcription factor evolution, and illustrated the complexity of the evolutionary pathways that lead to the modern transcriptional regulatory network architectures.

Acknowledgements

This work was supported by the ANR Jeunes Chercheurs and the programme Emergence of UPMC.

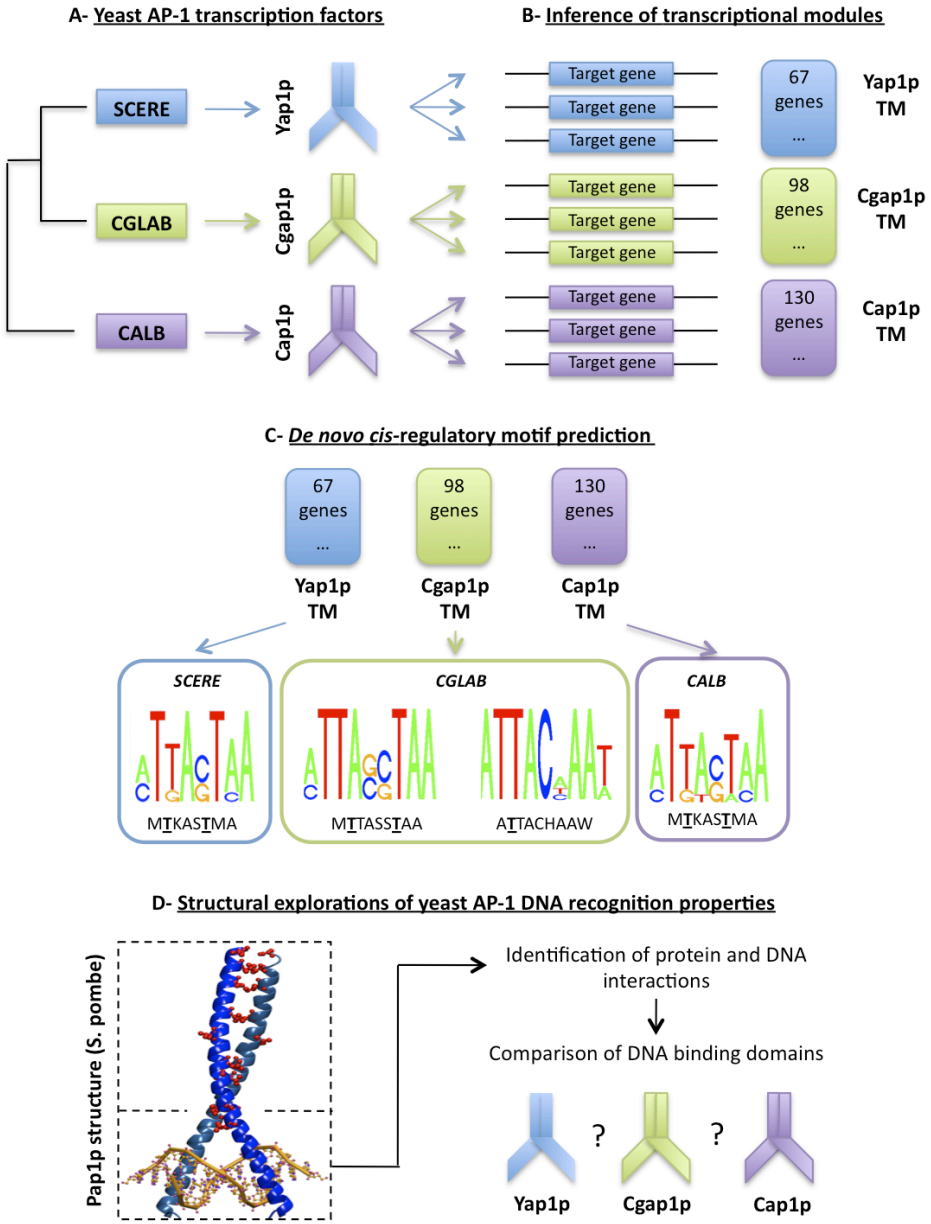


Figure 1. Global strategy to analyze the evolution of yeast AP-1 proteins.

References

[1] Y. Fujii, T. Shimizu, T. Toda, M. Yanagida and T. Hakoshima, Structural basis for the diversity of DNA recognition by bZIP transcription factors. *Nat Struct Biol* 7: 889-893, 2000.

[2] G. Lelandais, V. Tanty, C. Geneix, C. Etchebest, C. Jacq and F. Devaux, Genome adaptation to chemical stress: clues from comparative transcriptomics in Saccharomyces cerevisiae and Candida glabrata. *Genome Biol* 9: R164, 2008.

[3] A. Lucau-Danila, G. Lelandais, Z. Kozovska, V. Tanty, T. Delaveau, F. Devaux and C. Jacq, Early expression of yeast genes affected by chemical stress. *Mol Cell Biol* 25: 1860-1868, 2005.

[4] S. Znaidi, K.S. Barker, S. Weber, A.M. Alarco, T.T. Liu, G. Boucher, P.D. Rogers and M. Raymond, Identification of the Candida albicans Cap1p regulon. *Eukaryot Cell* 8: 806-820, 2009.

[5] D. Kuo, K. Licon, S. Bandyopadhyay, R. Chuang, C. Luo, J. Catalana, T. Ravasi, K. Tan and T. Ideker, Coevolution within a transcriptional network by compensatory trans and cis mutations. *Genome Res* 20: 1672-1678, 2010.

[6] D.T. Nguyen, A.M. Alarco, M. Raymond, Multiple Yap1p-binding sites mediate induction of the yeast major facilitator FLR1 gene in response to drugs, oxidants, and alkylating agents. *J Biol Chem* 276: 1138-1145, 2001.

MGCA: The Multiple Genome Cluster Analysor

A Flexible Tool for Phylogenomic Analyses of Prokaryotic Genomes

Kirsley Chennen[1], Pierre Lechat[1], Edouard Hirchaud[1], Romain Cahuzac[1], Pierre Dehoux[1] and Catherine Dauga[1]

[1] Institut Pasteur, Bio-informatique pour l'Analyse Génomique, 28 rue du Docteur Roux, 75015 Paris, France
{kirsley.chennen, pierre.lechat, edouard.hirchaud, romain.cahuzac, pierre.dehoux, catherine.dauga}@pasteur.fr

Keywords Comparative genomics, draft genomes, Inparanoid-Multiparanoid, RAST server, Chlamydiae.

1 Introduction

The advent of next generation sequencing approaches and the drop in the cost of sequencing have led to a sudden increase of the number of completely sequenced genomes and unfinished genomes. Consequently our ability to annotate genomes - identifies protein-encoding, rRNA and tRNA genes, assigns functions to the genes - and methods for data mining need to be adapted to this new volume of data. For this reason, we propose an automatic workflow MGCA "Multi-Genome Cluster Analysor", to integrate in a same resource, methods for functional annotation and genomic comparison, to help the researchers in the exploration of genomic information.

2 Implementation

MGCA is a package composed of a workflow of four pipelines coded in Perl, a MySQL database where all the results are stored and a user-friendly web interface designed with PHP to query and select data of interest.

3 Program Description and Potentialities

MGCA permits to compare data from one to about 20 genomes of medical or ecological interest. The workflow was designed to be flexible enough to take in input ENA-EMBL, Genbank or Tab-delimited genome files. In the case of draft genomes still in contigs, we used previously the RAST server, an automatic gene calls and annotation service which predicts RNA and CDS and provides DNA and protein sequences [3].

MGCA makes easier comparisons of gene repertoire and brings new insights on functional evolution of related genomes. A multi-genome clustering process using Inparanoid [6] and MultiParanoid [1] is done to obtain homologs (orthologs and in-paralogs) from a subset of genomes selected by the users. High-quality annotation and updated annotation, essential for understanding a genome, is obtained by using RPS-Blast [2] against the CDD database [5]. To make our system user-friendly, we developed a Web-based graphical interface to retrieve large amount of data by requests based on annotation and/or taxonomy criteria or through a list of gene names. The user can browse through out a refined list of clusters of homologs, retrieve all the corresponding annotations and export the corresponding DNA and protein sequences.

With MGCA, it is now feasible to analyze genomic data through a phylogenomic approach. MGCA gives the opportunity to compare closely related strains or species within a phylum as well as with bacteria from closely related phyla. The classification of CDSs as core genes, dispensable genes (present in some but not all compared genomes) and orphan genes (genes specific of one genome) can be obtained. Furthermore, MGCA allows to carry out in- depth phylogenetic analysis like detection of horizontal gene transfers (by tree topology tests) or measure of selection pressure within selected clusters.

MGCA gives the opportunity to compare bacteria from different phyla sharing a same lifestyle and/or different features. MGCA helps to predict genes conferring important phenotypes, which are present or absent depending on strains and to retrieve a short list of proteins for laboratory experiments.

4 Biological Application

To demonstrate a specific application case, we analyzed ten genomes of Chlamydiaceae including human pathogens, C. trachomatis & C. pneumoniae, through a comparison of genomes available in the superphylum PVC (Planctomycetes – Verrucomicrobia – Chlamydiae). The aim of the comparisons between Chlamydiae, obligated intracellular bacteria, infecting animals, humans, as well as Protozoa (amoebae) and closely related free-living environmental bacteria is to identify some proteins involved in intracellular survival mechanisms. We focused on protein involved in the Type III Secretion System, apparatus and pathways, for which we benefit from laboratory experiments [4], to test the efficiency of our workflow.

5 Conclusion

At end, MGCA would make easier the identification of genes involved in pathogen adaptation and in many other processes of biological interest and will be available for distribution on micro computers (Unix and Mac OS X).

References

[1] A. Alexeyenko, I. Tamas and E. Sonnhammer, Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics*, *22:9-15*, 2006.

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

[3] R. Aziz, D. Bartels, A. Best, M. DeJongh, V. Vonstein, A. Wilke and O. Zagnitko, RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics*, *9:75*, 2008.

[4] P. Dehoux, R. Flores, C. Dauga, G. Zhong and A. Subtil, Multi-genome identification and characterization of chlamydiae- specific type III secretion substrates: the Inc proteins. *BMC Genomics*, *12:109*, 2011.

[5] A. Marchler-Bauer, J. Anderson, F. Chitsaz, M. Derbushire and C. DeWeese-Scott, CDD : specific functional annotation with the Conserved Domain Database. *Nucleic Acid Research, 37:D205-10*, 2008.

[6] M. Remm, C. Storm and E. Sonnhammer, Automatic Clustering of Orthologs and In-paralogs from Pairwise Species Comparisons. *Journal of Molecular Biology*, 314:1041-1052, 2001.

Vibrios infecting Marine Invertebrates: new insights into Vibrio Virulence and their Adaptive Gene Reservoirs

David Goudenège[1], Evelyne Krin[2], Erwan Corre[1], Claudine Médigue[3], Didier Mazel[2] and Frédérique Le Roux[1]

[1] EQUIPE ÉMERGENTE IFREMER-UPMC "GÉNOMIQUE DES VIBRIO", FR2424 Station biologique de Roscoff
Place Georges Teissier, 29682, Roscoff
frederique.le-roux@sb-roscoff.fr

[2] UNITÉ "PLASTICITÉ DU GÉNOME BACTÉRIEN", CNRS URA2171, Département Génomes et Génétique
25 rue du Dr. Roux, 75724, Paris, cedex 15

[3] LABGeM, CNRS UMR8030 & CEA/DSV/IG/Genoscope, 2 rue Gaston Crémieux, 91057, Evry

Keywords Comparative and functional genomic, genome plasticity, pathogenicity.

1 Introduction

Vibrionaceae constitute a family of Gram-negative bacteria which belong to the ∝-group of proteobacteria and are ubiquitous in marine environments (for a review see [1]). They are coma shaped and are highly motile thanks to a polar flagellum appendage. They have two circular chromosomes and present a high genomic plasticity, especially at the level of chromosome 2, whereas chromosome 1 carries most of the genes involved in house keeping functions [2]. This high plasticity is reflected in the fact that frequently there is no good correlation between virulence, host specificity and taxonomy suggesting an evolution of virulence traits mostly through horizontal gene transfer [3]. Comparative genome analyses underscore a variety of genomic events such as chromosomal rearrangements, loss of genes by decay and/or deletion and gene acquisition through horizontal transfer as a source of plasticity [4,5]. Such phenomena have likely played an important role in the acquisition of virulence traits and pathogen emergence.

Vibrionaceae show a wide range of niche specialization, from estuarine to deep-sea habitats, from free-living forms to those attached to biotic and abiotic surfaces, and from symbiotic to pathogenic interactions [4]. It encompasses the ancient and well-studied human pathogen V. cholera. Perhaps less widely recognized are the consequences of vibrio infections in non-human species [6]. Vibrios have indeed been found to be pathogens of fish, coral, shellfish and shrimps and infections by these organisms have profound environmental and economic consequences. For instance V. splendidus has been associated with the abnormal mortality events that have plagued oysters in France over the last three years [7].

As virulence evolves by natural selection in order to increase the pathogen fitness [3], the emergence of virulence traits should result from genomic plasticity and adaptation to environmental constraints. Therefore evolutionary analysis of the vibrios may highlight adaptive gene reservoirs and identified new virulence mechanisms. Briefly this means finding correlations between the evolution of gene families and the evolution of the lifestyles and ecological niches of the different species/strains; and between niche/virulence and LGT. The data resources of such analysis are the subset of genes that is found on all the genomes (core genome), the set of genes that is found on more than one but not all genomes (accessory or distributed genome), and strain specific genes [8]. The existence of 17 complete genome sequences for closely related species from varied aquatic niches makes this group an excellent case study for genome comparison and research concerning the evolution-adaptation of these bacteria with emphasis on virulence. To date few studies on the Vibrio core and accessory genome have been published and they are based mainly on V. cholerae intra-specific genome comparison [5,9,10].

Finally sequences are of limited interest if there are no subsequent functional studies. In the last few years, new genetic strategies were developed to express or knock out genes in numerous vibrio species [13,14] and led us to demonstrate the role of specific genetic element such as plasmids, or genes such as metalloprotease genes in Vibrio virulence.

2 The Project

Our research groups aim at investigating the molecular mechanisms involved in the emergence of Vibrio pathogenic for marine invertebrates. This project is based on in silico approaches (phylogeny, comparative genomic analyses) combined to in vivo (functional genomic) and in vivo studies (experimental challenges).

The evolutionary analysis of the gene families amongst the vibrio will allow an estimation of the correlations between the evolution of gene families and the evolution of the lifestyles and ecological niches of the different species/strains. We are analyzing the evolution of functions known to be associated with pathogenicity such as secreted metalloproteases and secretion systems as well as genes and regulator screened by the comparative genomic.

Be it a gene annotated as a putative virulence gene, a gene of unknown function present in a putative pathogenic island or expressed specifically in virulent stage, this element will be subjected to mutagenesis for a functional demonstration of its role. As an example, the effect of metalloprotease (orthologs and paralogs) deletion on vibrio virulence will be presented.

The availability of multiple genome sequences, genetic tools and infection experimental systems allows us to propose the Vibrio genus as an especially suitable model to be at the interface of in silico and experimental approaches.

Acknowledgements

This work is supported by the Région Bretagne (SAD) and Ifremer.

References

[1] F.L. Thompson, T. Iida and J. Swings, Biodiversity of vibrios. *Microbiol Mol Biol Rev.*, 68(3):403-31, 2004.

[2] R. Dryselius, K. Kurokawa and T. Iida, Vibrionaceae, a versatile bacterial family with evolutionarily conserved variability. *Res Microbiol.*, 158(6):479-86, 2007.

[3] D. Mazel, Integrons: agents of bacterial evolution. *Nat Rev Microbiol*., 4(8):608-20, 2006.

[4] F.J. Reen, S. Almagro-Moreno, D. Ussery and E.F. Boyd, The genomic code: inferring Vibrionaceae niche specialization. *Nat Rev Microbiol*., 4(9):697-704, 2006.

[5] T. Vesth, T.M. Wassenar, P.F. Hallin, L. Snipen, K. Lagesen and D.W. Ussery, On the origins of a Vibrio species. *Microb Ecol*., 59(1):1-13, 2010.

[6] B. Austin, Vibrios as causal agents of zoonoses. *Vet Microbiol*., 140(3-4):310-7, 2010.

[7] F. Le Roux and B. Austin, Vibrio splendidus. Biology of Vibrio ASM Press, Washington DC, 2005.

[8] S. Bentley, Sequencing the species pan-genome. *Nat Rev Microbiol*., 7(4):258-9, 2009.

[9] D.J. Grimes, C.N. Johnson, K.S. Dillon, A.R. Flowers, N.F. Noriea and T. Berutti, What genomic sequence information has revealed about Vibrio ecology in the ocean--a review. *Microb Ecol.*, 58(3):447-60, 2009.

[10] C.C. Thompson, A.C. Vicente, R.C. Souza, A.T. Vasconcelos, T. Vesth, N.Jr. Alves, D.W. Ussery, T. Iida and F.L. Thompson, Genomic taxonomy of Vibrios. *BMC Evol Biol*., 9:258, 2009.

[11] P. Romby, F. Vandenesch and E.G. Wagner, The role of RNAs in the regulation of virulence-gene expression. *Curr Opin Microbiol*., 9(2):229-36, 2009.

[12] J.M. Liu, J. Livny, M.S. Lawrence, M.D. Kimball, M.K Waldor and A. Camilli, Experimental discovery of sRNAs in Vibrio cholerae by direct cloning, 5S/tRNA depletion and parallel sequencing. *Nucleic Acids Res*., 37(6):e46, 2009.

[13] F. Le Roux, J. Binesse, D. Saulnier and D. Mazel, Construction of a Vibrio splendidus mutant lacking the metalloprotease gene vsm by use of a novel counterselectable suicide vector. *Appl Environ Microbiol*., 73(3):777-84, 2007.

[14] F. Le Roux, B.M. Davis and M.K. Waldor, Conserved small RNAs govern replication and incompatibility of a diverse new plasmid family from marine bacteria. *Nucleic Acids Res.*, 39(3):1004-13, 2011.

TriAnnot: A High Performance Pipeline for the Automated Structural and Functional Annotation of Plant Genomes

Philippe LEROY[1], Aurélien BERNARD[1], Nicolas GUILHOT[1], Sébastien THEIL[1], Michael ALAUX[2], Sébastien REBOUX[2], Olivier INIZAN[2], Frédéric CHOULET[1], Hiroaki SAKAI[3], Tsuyoshi TANAKA[3], Takeshi ITOH[3], Hadi QUESNEVILLE[2] and Catherine FEUILLET[1]

[1] INRA-UBP, UMR 1095 Genetics, Diversity and Ecophysiology of Cereals, 234 Avenue du Brézet, F-63100 Clermont-Ferrand, France

Philippe.leroy@clermont.inra.fr

[2] INRA URGI, Route de Saint Cyr, F-78000, Versailles, France

[3] National Institute of Agrobiological Sciences, 2-1-2 Kannondai, Tsukuba, Ibaraki 305-8602, Japan

Keywords pipeline, computing cluster, plant genome, structural annotation, functional annotation, protein coding gene, transposable elements, repeats, wheat, chromosome 3B.

Genomics is the primary driver for unification in biological science, and now genome technologies are tools for hypothesis-driven research [1]. However, because of next-generation sequencing technologies and reduced costs, new genomes are being sequenced at a faster rate than they are being fully and correctly annotated [2]. Indeed, genome annotation is probably one of the most difficult tasks in genome sequencing projects [3]. However, structural and functional annotations are essential for connecting genome sequence to biology. To achieve a systematic and comprehensive annotation of the bread wheat (*Triticum aestivum* L.) genome (17 Gbp, 2n=6x=42, AABBDD), 45 time the rice genome, a pipeline called TriAnnot has been developed under the umbrella of the IWGSC (http://www.wheatgenome.org): **1.** to provide the international scientific community with an online user-friendly interface; **2.** To facilitate large scale analysis such as the ANR/FranceAgriMer 3BSEQ French flagship project which aims at annotating ~1 Gb of sequences from the wheat chromosome 3B which gene content is estimated to be around 8,000 [4]. At the end of the sequencing process 21,000 scaffold sequences should be delivered by the Génoscope. Therefore, a parallelized pipeline (V3.0) has been developed and installed on the INRA URGI cluster 'Sauron', Versailles. The pipeline has potentially access to 700 cores and 50 TB disk storage. The modular architecture of the TriAnnot pipeline allows the identification and annotation of repeats and Transposable Elements (TEs), protein-coding genes structural and functional annotation, RNA-coding genes and other biological features identifications. The pipeline is launched automatically using the FASTA files retrieved from the sequencing performed at Genoscope and after annotation, the output files are inserted automatically into a Chado database that is connected to an online GBrowse and Artemis graphical viewer to help further manual expertise. EMBL output files have been formatted to be used with GenomeView (http://genomeview.org/) as well. The manually expertise annotations will be automatically linked to the URGI Information System (http://urgi.versailles.inra.fr/gnpis/) that integrates into a single platform many biological data. Compared with three international pipelines (MIPS, RiceGAAS, Flowering Plant Gene Picker - FPGP), and based on 18 Mb of manually annotated wheat BAC sequences (148 genes) [4], the sensitivity (*Sn*) and specificity (*Sp*) of the TriAnnot pipeline, are suitable, therefore making TriAnnot more adequate for wheat genome annotation. TriAnnot can be easily applied to annotation efforts in other plant genomes with minor modifications. Full description of the TriAnnot pipeline is available at http://www.clermont.inra.fr/triannot.

References

[1] D.R. Cook and R.K Varshney, From genome studies to agricultural biotechnology: closing the gap between basic plant science and applied agriculture. *Current Opinion in Plant Biology*, 13:115-118, 2010.

[2] B.L. Cantarel, I. Korf, S.M.C. Robb, G. Parra, E. Ross, B. Moore, C. Holt, A.S. Alvarado and M. Yandell, MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research*, 18:188-196, 2008.

[3] C.G. Elsik, K.C. Worley, L. Zhang, N.V. Milshina, H. Jiang, J.T. Reese, K.L Childs, A. Anand Venkatraman, C.M. Dickens, G.M. Weinstock and R.A. Gibbs, Community annotation: Procedures, protocols, and supporting tools. *Genome Research*, 16:1329-1333, 2006.

[4] F. Choulet, T. Wicker, C. Rustenholz, E. Paux, J. Salse, P. Leroy, S. Schlub, M-C. Le Paslier, G. Magdelenat, C. Gonthier, A. Couloux, H. Budak, J. Breen, M. Pumphrey, S. Liu, X. Kong, J. Jia, M. Gut, D. Brunel, J.A. Anderson, B.S. Gill, R. Appels, B. Keller and C. Feuillet, Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. *Plant Cell* 22:1686-1701, 2010.

BiblioList

A Literature Manager for Annotated Organisms

Emmanuel Quevillon [1]

[1] Groupe Projets et développement en BioInformatique, 28 rue du Docteur Roux, 75015, Paris, France
emmanuel.quevillon@pasteur.fr

Abstract *While more and more genome sequencing projects have been initiated, scientists like and want to get rapid access to new informations about the sequenced organism. Most of the time, data from fresh genomes are first analysed to deliver the genome structure and thus to identify genes and their function(s). However, most of such websites don't provide informations about annotated genes literature. In order of filling this gap, a new web application has been developed : BiblioList. BiblioList helps scientists to enrich genes annotations with literature. To do so, first BiblioList uses an internal simple search engine that automatically links publications to each genes and reports it to researchers through its web interface. From this point, scientists can accurate gene identification with the associated publications. Once the literature for the gene annotation is complete, this bunch of publications can be directly exported to a remote organism annotation database with a simple click.*

Keywords Annotation, Organisms, PubMed, Literature, Search engine, Web application.

1 Introduction

In the last ten years, strong efforts have been lead to develop genome automatic annotation tools as well as genome assembly. Indeed, with new fast sequencing methods, also known as “next generation sequencing”, we can see pop up lots of new softwares to rapidly assemble DNA fragments. In the mean time, projects to automatically discover, store and display those genome structure annotations have been developed. For example, we can site tools from the GMOD consortium ([1]), such as the “Chado” database and the “Gbrowse” genome browser. However, all these development efforts only refer to genome annotations. Nothing really cares about genes associated literature. Such a tool might enhance the gene definition while referring directly to biological published work or books as a real proof for the gene discover and function. This is the idea BiblioList has been designed for.

2 Background

2.1 Functionalities

BiblioList is a web application that allow scientists to curate genome annotation by increasing gene informations with literature. This literature comes from the PubMed library[2] and is incorporated into its database to be filtered later by hand. For easiness, only title and abstract are fetched from PubMed. BiblioList is able to automatically update the associated literature for a particular organism and alert the user(s) that some new publications are ready to be curated (added to a gene or not). To do so, BiblioList uses a simple internal search engine which is able to link a publication and a gene based on its name and/or its description. Then the user is ready to curate associated literature. Validations can be accompanied with comments in order to keep track why the user decided to keep or not this link.

Once a gene is curated enough, the publications can be exported to a remote annotation database in a simple click.

Like most of other web applications, BiblioList provides an administration interface to manage organisms, publications, users, groups and genes. Finally, it is easy to create news feed as well as send email infos to users registered to the site.

2.2 Mapping and Export to Remote Databases

BiblioList is flexible and extensible enough that many remote annotation databases can be plugged with it. Indeed, nothing simpler than extending a Perl module and configure its database connection parameters and you're done.

2.3 Implementation

In an attempt to provide an intuitive and easy to understand tool, BiblioList has been designed as a web application, following the Model View Controller (MVC) architecture.

BiblioList is a pure Perl[3] MVC application. We used the Catalyst[4] framework to get a rapid and well structured running application. As a storage, we use the relational database management system (RDBMS) Sybase[5] in a first attempt and then switched later to PostGreSQL[6]. So SQL schemas are available for these two RDMS. The views are produced by the powerful Template Toolkit[7] and the the AJAX library jQuery[8].

2.4 Mechanism

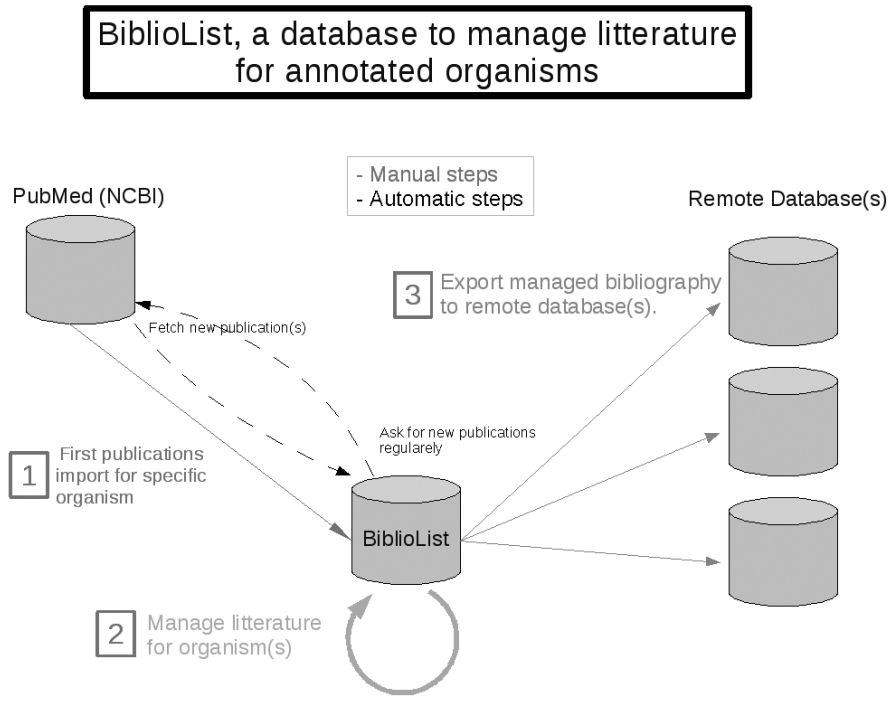


Figure 1. BiblioList data flow.

References

[1] Generic Model Organism Database, http://www.gmod.org/wiki/Main_Page

[2] PubMed, http://www.ncbi.nlm.nih.gov/pubmed

[3] The Perl Programming Language, http://www.perl.org

[4] Catalyst, http://www.catalystframework.org

[5] Sybase, http://www.sybase.com

[6] PostGreSQL, http://www.postgresql.org

[7] The Template Toolkit is a fast, flexible and highly extensible template processing system, http://template-toolkit.org

[8] jQuery: The Write Less, Do More, JavaScript Library, http://jquery.com

Protein Classification in the Case of Large and Many-Class Datasets

A Comparison with BLAST and BLAT

Rabie SAIDI[1,2], Wajdi DHIFLI[1,2], Sabeur ARIDHI[1,2], Marie AGIER[1,2], Gisèle BRONNER[3,4], Didier DEBROAS[3,4], Laurent D'ORAZIO[1,2], François ENAULT[3,4], Sylvie GUILLAUME[1,2], and Engelbert MEPHU NGUIFO[1,2]

[1] Clermont Université, Université Blaise Pascal, LIMOS, BP 10448, 63000, Clermont-Ferrand, France
[2] CNRS, UMR 6158, LIMOS, F-63173 Aubière
{saidi, dhifli, aridhi, dorazio, agier, guillaume mephu}@isima.fr
[3] Clermont Université, Université Blaise Pascal, LMGE, BP 80026, 63000, Clermont-Ferrand, France
[4] CNRS, UMR 6023, LMGE, F-63171 Aubière
{Gisele.bronner, francois.enault, didier.debroas}@univ-bpclermont.fr

Abstract *With the continuously increasing amounts of biological data, the need for automated, accurate and rapid classification is still challenging, especially when the number of classes is large. Here, we describe and compare the alignment-based approach (ABA) and the machine learning-based approach (MLBA). Then, we introduce a two-phase approach coupling hidden markov models (HMM) with standard classifiers and experimentally validate it.*

Keywords Protein classification, large scale learning, HMM, sequence alignment.

1 Alignment-based Classification

ABA refers just to assign to the query sequence the class of its reference sequence having the best hit score. Blast [1] is the most widely used program in bioinformatics. Its popularity is mainly due to its algorithm which focuses on sensitivity while being considerably speeder then its previous tools such as Fasta. This makes it practical on large scale database analysis. Blat [1] is much faster than Blast with less sensitivity. Its rapidity is mainly due to its indexing technique which is memory greedy since it charges all data indexes in RAM. ABA is class number independent, which represents its main advantage. Indeed, varying the number of classes (while keeping the same number of instances) has no effect on the alignment result in terms of speed and scoring. However, this advantage reveals an undeniable drawback i.e., ABA depends on the similarity to a unique sequence which yield both poor generalization and discrimination. This would lead to classification errors when varying the affiliations of sequences. In fact, a protein may belong to several classes of different nature e.g., functional, taxonomic, structural, etc.

2 Learning-based Classification

MLBA benefits from the panoply of developed classifiers that have shown high efficiency as decision aid tools in several fields such as finance, trade, medicine, etc, due to their strong discrimination and generalization. In general, MLBA requires data in relational format i.e., attribute/value table. Thus, two elements must be provided: a set of reliable attributes to be used as descriptors, and a reliable function of description e.g., frequency, incidence, etc. MLBA faces many problems when dealing with biological data classification. On one side, protein sequences are represented by strings of characters, which does not respect the format required by MLBA. On another side, the number of classes has an important impact on any learning task. Indeed, the discrimination ability of any classifier decreases with increasing number of classes especially in the case of unbalanced data. This problem has not been deeply investigated [2] whereas many efforts have been devoted to address large scale learning in term of number of instances [2].

3 Proposed Approach

3.1 Training

For each class we create an HMM profile then we build a binary model using a discriminative classifier. This model is trained on a dataset comprising that class and the other classes' consensuses, after being

encoded into relational format using a motif based approach [3]. This model discriminates between each class and the rest of the training set. A class consensus is a unique sequence which abstracts its class and is generated using the class HMM-profile. This allows us to bypass the unbalanced data and the many-class problems. Henceforth, each class is represented by a probabilistic model (its HMM-profile) and a discriminative model (its binary classifier model).

3.2 Prediction

Each query sequence is scanned against the HMM-profiles. Hence, some classes are suggested as potential targets sorted by their scores. The number of suggested classes is considerably below the total number of classes. At this level, we use the binary models corresponding to the suggested classes to confirm or refute the HMM results. The final sustained class is the one having the best score and confirmed by the binary model. The combination of the probabilistic and the discriminative aspects preserves an acceptable rapidity while enhancing the sensitivity of the prediction. Furthermore, the memory consumption in our approach is moderate compared to Blat since models are processed separately.

4 Experimental Comparison

To evaluate the above described methods, we utilized four protein datasets taken from the KEGG [4] (Table1). The datasets are characterized by a large number of sequences (from 12192 to 44572) and a large number of classes (from 25 to 100). Each class refers to an ortholog (functional) group [4] of less than 45% of identity. Experiments were conducted on a PC with a 3 Ghz duo core CPU / 3.25GB RAM. We used the hold-out technique to evaluate the classification approaches i.e., a third is reserved to test and the rest is used for training (for MLBA) or as reference base (for ABA). For our approach we use HMMER [5] as HMM tool, N-grams [3] as encoding method and SVM [3] as classifier. It is noteworthy that the functional groups in the KEGG base are built using many techniques including alignment. This explains the full accuracy reached by Blast. The experimental results between Blast and Blat confirm what we have mention in section 1. Blast is much more accurate and Blat is much faster. Our approach represents a tradeoff between Blast and Blat i.e., tradeoff between accuracy and speed with the ability to deal with other kinds of classification rather than the functional one e.g., taxonomic, structural.

Dataset	Sequence#	Classe#	Accuracy (%)			Time (mn)		
			Blast	Blat	Our approach	Blast	Blat	Our approach
DS1	12192	25	100	79	88	94	4	3
DS2	24301	50	100	90	92	187	6	8
DS3	33814	75	100	87	90	267	9	13
DS4	44572	100	100	87	90	392	15	18

Table 1. Experimental results.

Acknowledgements

This work is supported by PREFON_META project funded by the FEDER Auvergne program LifeGrid.

References

[1] W. J. Kent. BLAT—the BLAST-like alignment tool. *Genome research,* 12 (4): 656–664, 2002.
[2] O. Madani and M. Connor, Large-scale many-class learning, *SIAM Conf. on Data Mining (SDM)*, 2008.
[3] R. Saidi, M. Maddouri and E. M. Nguifo, Protein sequences classification by means of feature extraction with substitution matrices. *BMC Bioinformatics* 11:175, 2010.
[4] S. Mitra, P. Rupek, D. C. Richter, T. Urich, J. A. Gilbert, F. Meyer, A. Wilke and D. H. Huson, Functional analysis of metagenomes and metatranscriptomes using SEED and KEGG, *BMC Bioinformatics*, 12(Suppl 1):S21, 2011.
[5] L. S. Johnson, S. R. Eddy and E. Portugaly., Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* 11:431, 2010.

Predicting Copy Number Alterations and Structural Variants using Paired-end Sequencing data

Valentina BOEVA[1], Bruno ZEITOUNI[1], Kevin BLEAKLEY[2], Andrei ZINOVYEV[1], Jean-Philippe VERT[1], Isabelle JANOUEIX-LAROSEY[3], Olivier DELATTRE[3] and Emmanuel BARILLOT[1]

[1] Institut Curie, U900 Inserm, 26 rue d'Ulm, Paris, F-75248 France ; Mines ParisTech, Fontainebleau, F-77300 France
{valentina.boeva, bruno.zeitouni, andrei.zinovyev, jean-philippe.vert, emmanuel.barillot}@curie.fr

[2] INRIA Saclay Ile-de-France, 4 rue Jacques Monod, Orsay, F-91893 France
kevbleakley@gmail.com

[3] Institut Curie, U830 Inserm, 26 rue d'Ulm, Paris, F-75248 France
{isabelle.janoueix, olivier.delattre}@curie.fr

Keywords Cancer, Structural variants, Deep sequencing, Paired-ends, Copy number profiles.

1 Introduction

The detection of structural variants (SVs) in the human genome plays an important role in the understanding of many genetic diseases, including cancer. In cancer, tumor suppressor genes can be deleted or mutated, whereas oncogenes can be amplified or mutated with a gain of function. Translocations can result in cancer-causing fusion proteins (BCR/ABL fusion in CML, BCL1/IGH in multiple myeloma, EWS/FLI1 in Ewing sarcoma, etc.)

With the arrival of new high-throughput sequencing technologies, our current power to detect SVs has significantly improved. Genomic breakpoints of large structural variants (i.e., translocations or large duplications and deletions) can be identified using two complementary approaches: calculation of copy number profiles (CNPs) and analysis of 'discordant' mate-paired/paired-ends mappings (PEMs).

2 Results

The investigation of CNPs allows identification of genomic regions of gain and loss. There exist two frequent obstacles in the analysis of cancer genomes: absence of an appropriate control sample for normal tissue and possible polyploidy. We therefore developed a bioinformatics tool, called FREEC [1], able to automatically detect copy number alterations (CNAs) without use of a control dataset. FREEC normalizes copy number profiles using read mappability and GC-content and then applies a LASSO-based segmentation procedure to the normalized profiles to predict CNAs.

For PEM data, one can complement the information about CNAs (i.e., output of FREEC) with the predictions of structural variants (SVs) made by another tool that we developed, SVDetect [2]. SVDetect finds clusters of 'discordant' PEMs and uses all the characteristics of reads inside the clusters (orientation, order and clone insert size) to identify the SV type. SVDetect allows identification of a large spectrum of rearrangements including large insertions-deletions, duplications, inversions, insertions of genomic shards and balanced/unbalanced intra/inter-chromosomal translocations.

Here we present a package for automatic intersection of FREEC and SVDetect outputs that allows one to (1) refine coordinates of CNAs using PEM data and (2) improve confidence in calling true positive rearrangements (particularly, in ambiguous satellite/repetitive regions).

Both SVDetect and FREEC are compatible with the SAM alignment format and provide output files for graphical visualization of predicted genomic rearrangements.

Acknowledgements

This work was supported by The Ligue Nationale contre le Cancer (V.B., A.Z., E.B., I.J.-L. and O.D. are members of a labeled team).

References

[1] V. Boeva, A. Zinovyev, K. Bleakley, JP. Vert, I. Janoueix-Lerosey, O. Delattre and E. Barillot, Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics*, 27(2):268-9, 2011.

[2] B. Zeitouni, V. Boeva, I. Janoueix-Lerosey, S. Loeillet, P. Legoix-né, A. Nicolas, O. Delattre and E. Barillot, SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. *Bioinformatics*, 26:1895-1896, 2010.

De Novo Transcriptome Assembly in Non-Model Organisms from Next Generation Sequencing Data

Vincent Cahais[1], Philippe Gayral[1], Georgia Tsagkogeorga[1], Jose Melo-Ferreira[2], Ylenia Chiari[1], Khalid Belkhir[1], Vincent Ranwez[1] and Nicolas Galtier[1]

[1] ISEM, UMR5554 CNRS, Université de Montpellier II -CC 064, 34095, Montpellier, Cedex 05, France
{vincent.cahais, philippe.gayral02, georgia.tsagkogeorga, ylenia.chiari, khalid.belkhir, vincent.ranwez, nicolas.galtier}@univ-montp2.fr

[2] CIBIO, R. Monte-crasto, Campus Agrário de Vairão, 4485-661, Vairo, Portugal
jmeloferreira@mail.icav.up.pt

Keywords NGS, Transcriptome, Assembly, Non-model species.

Next-generation sequencing technologies give the opportunity for genomic study of non-model organisms sampled in the wild. The transcriptome is a convenient and popular target for such purposes. Assembling gene coding sequences out of short transcriptome reads, however, is a complex task, owing to gene duplications, genetic polymorphism, alternative splicing, and transcription noise. Typical assembling programs return thousand of predicted contigs, with unclear connection to species true gene content. This is especially problematic in taxa lacking a fully-sequenced, closely related genome. Here we use two animal species for which a reference genome is available to assess the potential for proper transcriptome assembly in absence of a reference. The transcriptome of Ciona intestinalis (Urochordata), Lepus granatensis (Mammalia) are assembled from newly-generated 454 and Illumina sequence reads. A new procedure is introduced to annotate each predicted contig as full length, partial, chimera, allele, paralogue, DNA or alien, based on the number and overlap of/between BLAST hits to appropriate reference transcriptomes and genomes. Transcriptoms of Emys orbicularis (Reptilia) and Ostrea edulis (Mollusca), from which no reference genomes are available, where de novo assembled with the same method. Analyses shows that (i) optimal assemblies are obtained when 454 and Illumina data are combined, (ii) existing assembling programs differ in their ability to correctly split paralogues and group alleles, (iii) typical de novo assemblies include a majority of irrelevant cDNA predictions, and (iv) assemblies can be appropriately cleaned by filtering contigs based on coverage and length. We conclude that robust, reference-free assembly of thousands of genes from transcriptomic next-generation sequence data is possible, which opens promising perspectives for transcriptome-based evolutionary genomics in animals.

Analyzing RNA-Seq Data within the MicroScope Web-based Platform

Béatrice CHANE-WOON-MING[1,3], Marion WEIMAN[1], David VALLENET[1], Véronique de BERARDINIS[2], Béatrice SEGURENS[2], Marcel SALANOUBAT[2], Maxime DUROT[1] and Claudine MEDIGUE[1]

[1] LABORATOIRE D'ANALYSES BIOINFORMATIQUES POUR LA GENOMIQUE ET LE METABOLISME, UMR 8030 CEA/Genoscope - CNRS - Université d'Evry, 2 rue Gaston Crémieux, 91057, Evry, Cedex, France

{mweiman, vallenet, mdurot, cmedigue}@genoscope.cns.fr

[2] LABORATOIRE DE GENOMIQUE ET BIOCHIMIE DU METABOLISME, UMR 8030 CEA/Genoscope - CNRS - Université d'Evry, 2 rue Gaston Crémieux, 91057, Evry, Cedex, France

{segurens, salanou, vberard}@genoscope.cns.fr

[3] Current address: ARCHITECTURE ET REACTIVITE DE L'ARN, UPR 9002 Institut de Biologie Moléculaire et Cellulaire - CNRS, 15 rue René Descartes, 67084, Strasbourg, Cedex, France

b.chanewoonming@ibmc-cnrs.unistra.fr

Keywords RNA-Seq, transcriptomics, high-throughput sequencing data, web platform, software tool, bioinformatics platform.

1 Introduction

MicroScope is a well-established web-based platform dedicated to microbial genome (re)annotation and comparative genomics [1]. It integrates several databases and software tools allowing advanced automated genome annotation to be performed and provides user-friendly web interfaces to query and curate gene annotations. Its annotation capabilities are widely used by the scientific community: 82 annotation projects including around 1000 organisms and involving more than 800 user accounts have been or are being performed since 2002. MicroScope also provides several layers of analytical tools focused on (i) comparative genomics, (ii) the reconstruction and analysis of metabolic networks, (iii) the integration of functional genomics data (e.g. mutant phenotypes [2],[3]), and, more recently, (iv) the analysis of bacterial polymorphism evolution from high-throughput sequencing data.

Among the applications of high-throughput sequencing (HTS) in functional genomics, RNA sequencing (RNA-Seq) is developing fast and offers significant improvements over microarray-based approaches [4]. RNA-Seq methods provide direct access to transcript structure, are not limited to a predefined list of transcripts, and cover a larger dynamic range of expression levels. As a growing number of MicroScope users make use of RNA-Seq data in their project, we recently integrated an RNA-Seq analysis and visualization module in the platform.

The current version of the RNA-Seq analysis module focuses on the analysis of differential expression. It is composed of two main components: a HTS data analysis pipeline combined with a dedicated database storing results, and a web-based visualization interface allowing MicroScope users to interact with RNA-Seq differential expression results.

2 RNA-Seq Data Analysis and Storage

Starting from raw sequencing reads, the analysis component processes in four successive steps (see Figure 1). First, raw reads are preprocessed by usual HTS tools to assess their quality, remove sequence adapters if needed, and prepare proper file formats for the subsequent analyses. Second, all reads are mapped to reference genomes. This step is performed using the software SSAHA2 [5] which is already used in MicroScope to analyze evolution projects from HTS data. This software provides good alignment results and supports small mismatches and insertions/deletions in sequencing reads while maintaining reasonable execution times on HTS datasets. Third, the coverage of transcripts is computed along genomes and expression levels are evaluated for each annotated genomic object that is stored in MicroScope annotation database (e.g. CDS, ncRNA). Expression levels are directly represented by the raw number of reads mapped in genomic objects, a representation that is needed by the statistical analysis method used in the following step to assess differential expression. Using raw read numbers actually allows the statistical package to better model expression variability occurring between experiments [6]. Finally, in a fourth step, differential

expression is tested between samples of distinct experimental conditions. This step relies on the R/Bioconductor package DESeq which normalizes expression levels across samples (to account for biases in sequencing depth) and makes use of an RNA-Seq specific statistical model to test for differential expression [6].

MicroScope RNA-Seq database stores information on experimental conditions, sequencing runs, transcript coverage along genomes, expression levels of all genomic objects, and results of statistical tests for differential expression. This design allows the analysis pipeline to support the process of several projects in parallel and integrates RNA-Seq data with all other MicroScope data. In addition, most parameters used to run the pipeline are stored in the database and can be used to run it again if needed (for instance if new genomic objects are annotated in the genome or if parameters of a given pipeline step need to be changed).

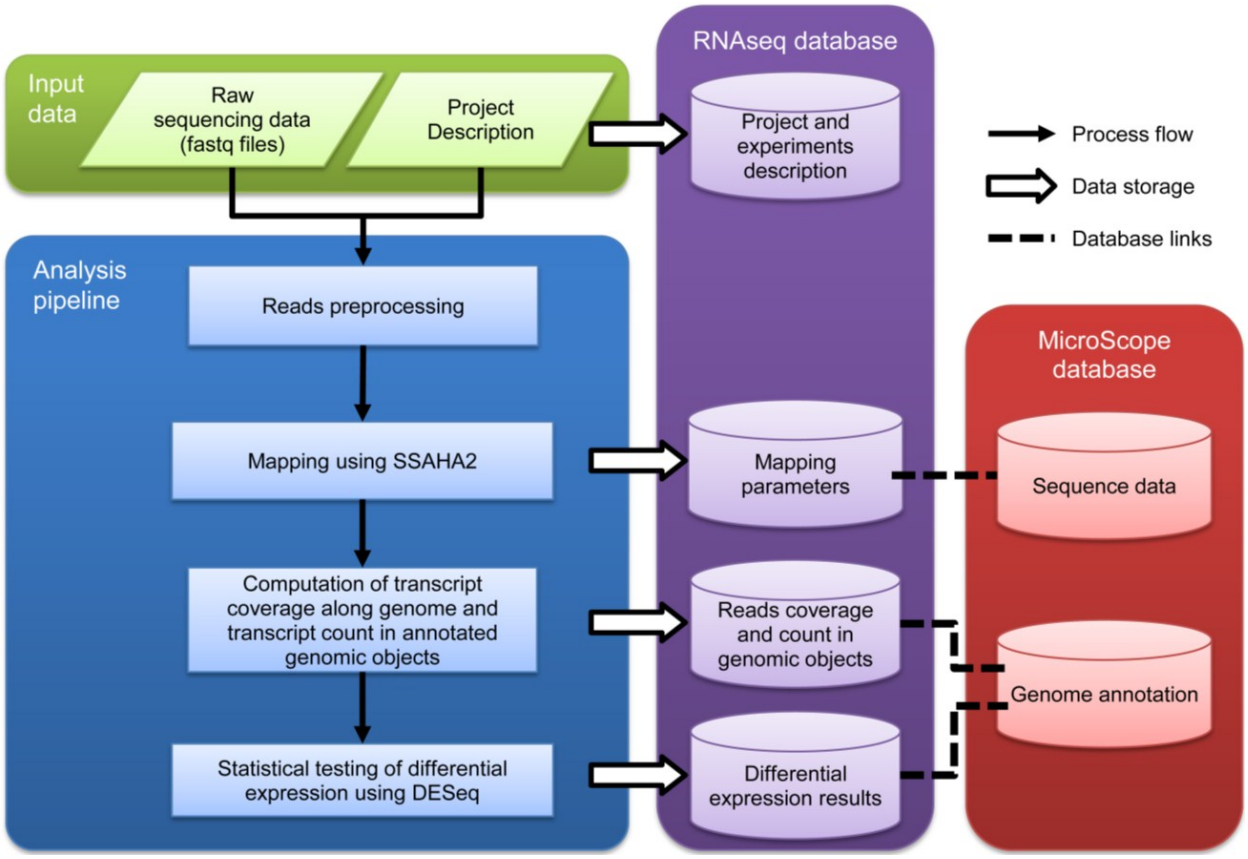
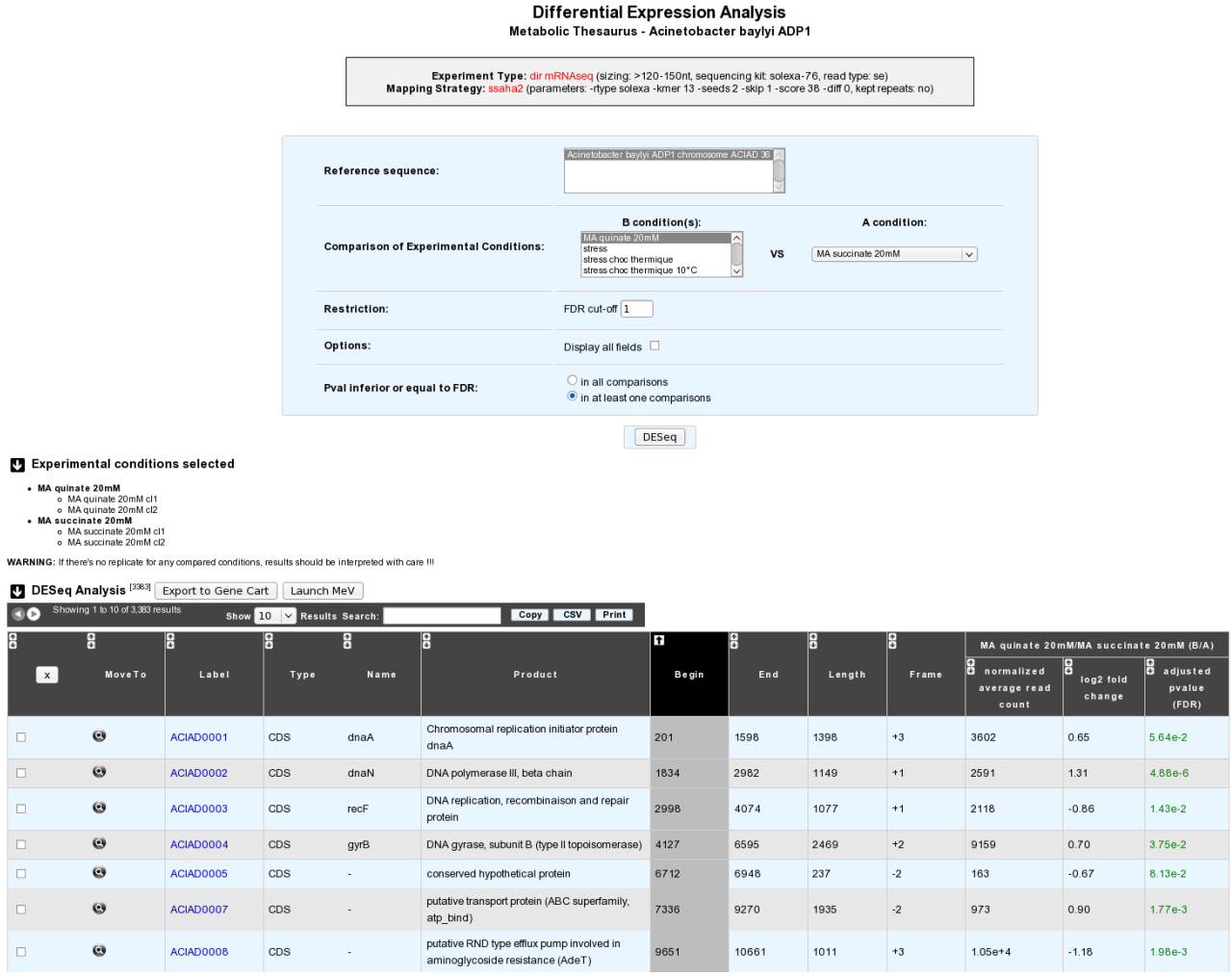


Figure 1. RNA-Seq data analysis pipeline and database. Relations to MicroScope genome annotation database are depicted on the right.

3 Visualization of Results

Similarly to other modules of MicroScope, we developed a web-based visualization component allowing users to explore most RNA-Seq results online, combine and analyze them using other tools from MicroScope (e.g. explore annotations, highlight metabolic pathways, search for orthologs of differentially expressed genes), and download results locally. More specifically, raw and normalized expression levels can be displayed for any genomic object on any experimental condition, and all appropriate pairwise comparisons of experimental conditions can be directly queried from the interface (see screenshot on Figure 2). In addition, transcript coverage over genomes are displayed together with genome annotations using the Integrative Genomics Viewer software [7] and expression levels can be automatically loaded into the Multiexperiment Viewer software [8] for further data analysis, such as clustering or gene-set enrichment analysis. The RNA-Seq visualization interface is available at the following URL: https://www.genoscope.cns.fr/agc/microscope/expdata/.



x	MoveTo	Label	Type	Name	Product	Begin	End	Length	Frame	MA quinate 20mM/MA succinate 20mM (B/A)		
										normalized average read count	log2 fold change	adjusted pvalue (FDR)
☐		ACIAD0001	CDS	dnaA	Chromosomal replication initiator protein dnaA	201	1598	1398	+3	3602	0.65	5.64e-2
☐		ACIAD0002	CDS	dnaN	DNA polymerase III, beta chain	1834	2982	1149	+1	2591	1.31	4.88e-6
☐		ACIAD0003	CDS	recF	DNA replication, recombinaison and repair protein	2998	4074	1077	+1	2118	-0.86	1.43e-2
☐		ACIAD0004	CDS	gyrB	DNA gyrase, subunit B (type II topoisomerase)	4127	6595	2469	+2	9159	0.70	3.75e-2
☐		ACIAD0005	CDS	-	conserved hypothetical protein	6712	6948	237	-2	163	-0.67	8.13e-2
☐		ACIAD0007	CDS	-	putative transport protein (ABC superfamily, atp_bind)	7336	9270	1935	-2	973	0.90	1.77e-3
☐		ACIAD0008	CDS	-	putative RND type efflux pump involved in aminoglycoside resistance (AdeT)	9651	10661	1011	+3	1.05e+4	-1.18	1.98e-3

Figure 2. Screenshot of the web interface retrieving differential expression analysis results. This interface is accessible from http://www.genoscope.cns.fr/agc/microscope/expdata/.

At the end of April 2011, MicroScope RNA-Seq module gathered data from four collaborative projects on five organisms and included 43 sequencing runs, mostly originating from Illumina Solexa GAIIx, for a total of $1{,}5.10^9$ reads. Relevant public RNA-Seq datasets are also being integrated. As an internal test case, we used the module to analyze RNA-Seq expression data for the bacterium *Acinetobacter baylyi* ADP1 on 9 distinct conditions – including pH, temperature and light stress conditions and two distinct carbon sources growth media, succinate and quinate – for a total a 17 sequencing runs. From raw expression levels, we found that 86% of all genes were expressed in at least one experiment and 47% were expressed in all experiments. Tests for differential expression confirmed the functions of several genes known to be involved in stress response or in the degradation of the carbon sources but also highlighted genes with unknown function. For instance, 65% of the 58 genes shown to be over-expressed at 42°C with respect to 30°C do not have any known function. RNA-Seq expression data therefore provides useful information on gene function which may prove to be powerful complements to other functional genomics datasets [9].

4 Perspectives

Because of its relative novelty, RNA-Seq data analysis is a rapidly evolving field. Experimental protocols are still under optimization and data processing methods are not completely settled. For instance, quantification and normalization of expression levels are still a matter of debate and some biases introduced by experimental protocols are currently not properly handled [10]. Part of improvements we will implement in MicroScope RNA-Seq module will therefore follow the state of the art in this field. In addition, we plan to extend the module to include analyses of the structure of transcripts. To this end, we currently design an

analysis pipeline aiming at locating transcription starting sites (TSS) from TSS specific RNA-Seq experiments. Results from this pipeline will then be combined with transcript coverage data to build global transcription maps.

Acknowledgements

This work is supported by the Agence Nationale de la Recherche (ANR) project EVOGENO.

References

[1] D. Vallenet, S. Engelen, D. Mornico, S. Cruveiller, L. Fleury, A. Lajus, Z. Rouy, D. Roche, G. Salvignol, C. Scarpelli, and C. Medigue, MicroScope: a platform for microbial genome annotation and comparative genomics, *Database*, 2009:bap021, 2009.

[2] E. Giraud, L. Moulin, D. Vallenet, V. Barbe, E. Cytryn, J. Avarre, M. Jaubert, D. Simon, F. Cartieaux, Y. Prin, G. Bena, L. Hannibal, J. Fardoux, M. Kojadinovic, L. Vuillet, A. Lajus, S. Cruveiller, Z. Rouy, S. Mangenot, B. Segurens, C. Dossat, W.L. Franck, W. Chang, E. Saunders, D. Bruce, P. Richardson, P. Normand, B. Dreyfus, D. Pignol, G. Stacey, D. Emerich, A. Verméglio, C. Médigue, and M. Sadowsky, Legumes symbioses: absence of Nod genes in photosynthetic bradyrhizobia, *Science*, 316:1307-1312, 2007.

[3] C. Rusniok, D. Vallenet, S. Floquet, H. Ewles, C. Mouzé-Soulama, D. Brown, A. Lajus, C. Buchrieser, C. Médigue, P. Glaser, and V. Pelicic, "NeMeSys: a biological resource for narrowing the gap between sequence and function in the human pathogen Neisseria meningitidis," *Genome Biology*, 10:R110, 2009.

[4] F. Ozsolak and P.M. Milos, "RNA sequencing: advances, challenges and opportunities," *Nature Reviews. Genetics*, 12:87-98, 2011.

[5] Z. Ning, A.J. Cox, and J.C. Mullikin, SSAHA: a fast search method for large DNA databases, *Genome Research*, 11:1725-1729, 2001.

[6] S. Anders and W. Huber, "Differential expression analysis for sequence count data," *Genome Biology*, 11:R106, 2010.

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

[8] A.I. Saeed, V. Sharov, J. White, J. Li, W. Liang, N. Bhagabati, J. Braisted, M. Klapa, T. Currier, M. Thiagarajan, A. Sturn, M. Snuffin, A. Rezantsev, D. Popov, A. Ryltsov, E. Kostukovich, I. Borisovsky, Z. Liu, A. Vinsavich, V. Trush, and J. Quackenbush, "TM4: a free, open-source system for microarray data management and analysis," *BioTechniques*, 34:374-378, 2003.

[9] V. de Berardinis, M. Durot, J. Weissenbach, and M. Salanoubat, Acinetobacter baylyi ADP1 as a model for metabolic system biology, *Current Opinion in Microbiology*, 12:568-576, 2009.

[10] J. Li, H. Jiang, and W.H. Wong, Modeling non-uniformity in short-read rates in RNA-Seq data, *Genome Biology*, 11:R50, 2010.

Conformément au souhait des auteurs, cette contribution n'est pas reproduite dans la version en ligne des actes de JOBIM 2011.

Following the wishes of the authors, this paper is not included in the online version of the JOBIM 2011 proceedings.

Conformément au souhait des auteurs, cette contribution n'est pas reproduite dans la version en ligne des actes de JOBIM 2011.

Following the wishes of the authors, this paper is not included in the online version of the JOBIM 2011 proceedings.

Microarray Image Segmentation using Parallel Spectral Clustering

Sandrine Mouysset, Joseph Noailles, Daniel Ruiz and Ronan Guivarch[1]

University of Toulouse, IRIT, UMR5505 CNRS, 2 rue Camichel, BP 7122, 31071 Toulouse, Cedex 7 France
{sandrine.mouysset,joseph.noailles,daniel.ruiz,ronan.guivarch}@enseeiht.fr

Abstract *Microarray technology generates large amounts of expression level of genes to be analysed simultaneously. The analysis of data crucially depends on the microarray image segmentation which extracts the quantitative information from spots. Spectral clustering is one of the most relevant unsupervised method able to gather data without a priori on its shapes. With a criterion for determining the number of clusters and by exploiting inherent properties of spectral clustering, a parallel strategy based on domain decomposition is proposed and tested on microarray images.*

Keywords Spectral clustering, domain decomposition, microarray image segmentation.

1 Introduction

Image segmentation in microarray analysis is a crucial step to extract quantitative information from the spots [1],[2]. Spectral methods, and in particular the spectral clustering algorithm introduced by Ng-Jordan-Weiss [3], are useful when considering no a priori shaped subsets of data. Spectral clustering uses eigenvectors of a Gaussian affinity matrix in order to define a low-dimensional space in which data points can be clustered. But when very large data set are considered, the extraction of the dominant eigenvectors become the most computational task in the algorithm [4],[5].In this paper, a parallel strategy based on domain decomposition is investigated. Two main problems still arise from the divide and conquer strategy : the difficulty to choose a Gaussian affinity parameter and the number of clusters k which may even vary from one subdomain to the other.

2 Parallel Spectral Clustering: justification and implementation

By exploiting the block structure of microarrays, clustering could be made on subdomains by breaking up the data set into data subsets with respect to their geometrical coordinates in a straightforward way. With an appropriate Gaussian affinity parameter and a method to determine the number of clusters, each processor applies independently the spectral clustering algorithm on subsets of data points and provide a local partition on these data subsets. Based on these local partitions, a gathering step ensures the connection between subsets of data and determines a global partition. We experiment this strategy which principle is represented in Fig. 1 (a) and a clustering result on a 4×2 greyscaled spotted microarray image is plotted in Fig. 1 (c).

2.1 Choice of the affinity parameter

The Gaussian affinity matrix is widely used and depends on a free parameter. It is known that this parameter affects the results in spectral clustering and spectral embedding. A global heuristic for this parameter was proposed in [6] in which both the dimension of the problem as well as the density of points in the given p-th dimensional data set are integrated. By considering an uniform distribution in which all pair of data points are separated by the same distance, a reference distance is defined. From this definition, clusters may exist if there are points that are at a distance no more than a fraction of this reference distance.

2.2 Choice of the number of clusters

After indexing data points per cluster for a value of k, we define the indexed affinity matrix L which diagonal affinity block represent the affinity among the cluster and the off-diagonal ones the affinity between clusters (Fig. 1 (b)). The ratios between the Frobenius norm of the off-diagonal blocks and that of the diagonal ones could be evaluated. By definition, the appropriate number of clusters, noted k, corresponds to a situation where points which belong to different clusters have low affinity between each other whereas points in same clusters have higher affinity. Among various values for k, the final number of cluster is defined so that the affinity between clusters is the lowest and the affinity within clusters is the highest. As numerical experiments, the computational cost (the time spent in the parallel Spectral Clustering part divided by the average number of points on each subdomain) is plotted in Fig. 1 (d) and shows good performance: this cost decreases drastically when we increase the number of processors.

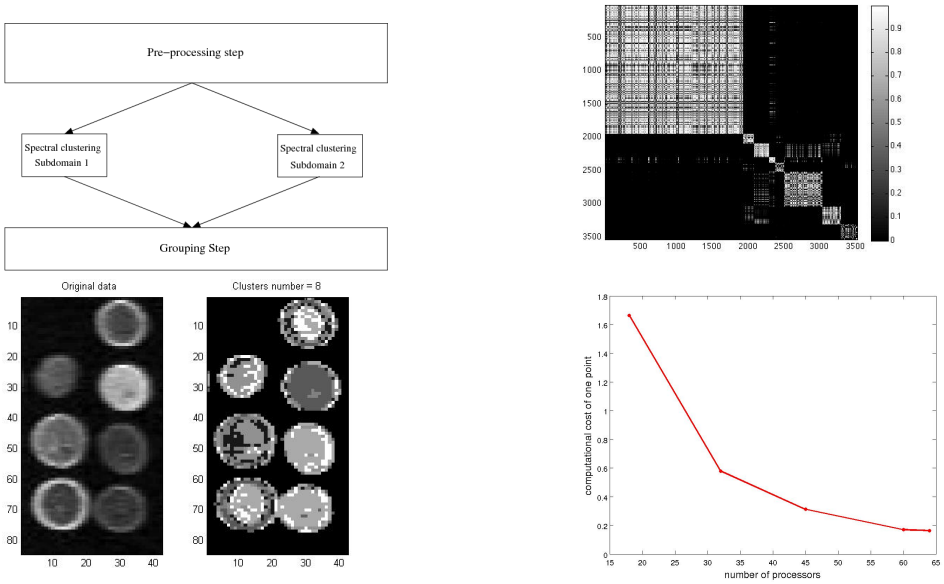


Figure 1. (a) Parallel Spectral clustering principle, (b) block structure of the indexed affinity matrix for $k = 8$, (c) Clustering on one sub-domain 4×2 greyscaled spotted microarray image (3500 pixels) and its clustering result, (d) Total computational costs.

3 Conclusion

With the domain decomposition strategy and heuristics for determining the choice of the Gaussian affinity parameter and the number of clusters, the parallel spectral clustering becomes robust for microarray image segmentation and combines intensity and shape features.

References

[1] L. Rueda and L. Qin, A new method for DNA microarray image segmentation. *Image Analysis and Recognition*, 886-893, 2005.

[2] N. Giannakeas, P.S. Karvelis and D.I. Fotiadis, A classification-based segmentation of cDNA microarray images using Support Vector machines, *Engineering in Medicine and Biology Society*, 875-878, 2008.

[3] A. Y. Ng, M. I. Jordan and Y. Weiss, On spectral clustering: analysis and an algorithm. *NIPS*, 849-856, 2002.

[4] E. Yom-Tov and N. Slonim, Parallel pairwise clustering *SIAM Int. Conf. on Data Mining*, 2009.

[5] W-Y. Chen, S. Yangqiu, H. Bai, C-J. Lin and E. Y. Chang, Parallel Spectral Clustering in Distributed Systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

[6] S. Mouysset, J. Noailles and D. Ruiz, Using a Global Parameter for Gaussian Affinity Matrices in Spectral Clustering, *High Performance Computing for Computational Science-VECPAR 2008*, 378-390, 2008.

Improving Biclustering for High Dimension Genomic Data using the Ensemble Methods

Blaise HANCZAR[1] and Mohamed NADIF[1]

Laboratoire d'Informatique Paris Descartes LIPADE, University Paris Descartes. 45 rue des saint-pres, 75006 Paris
hanczar_blaise@yahoo.fr mohamed.nadif@univ-paris5.fr

Abstract *One of the major tools of transcriptomics is the biclustering that simultaneously constructs a partition of both examples and genes. Several methods have been proposed for microarray data analysis that enables to identify groups of genes with similar expression profiles only under a subset of examples. We propose to improve the quality of these biclustering methods by using an ensemble approach. Our bagged biclustering method generates a collection of biclusters using the bootstrap samples of the original data and aggregate them into new biclusters. We show that our method improve the performance of biclustering on several public microarray datasets.*

Keywords Co-clustering, Biclustering, Microarray data, Gene expression, Bagging.

1 Introduction

The capacity of microarray to measure simultaneously the expression of a whole genome under different experimental condition, is of great interest for biologists. Biclustering methods allow the identification of relevant groups of genes and conditions that cannot be identified by classic clustering techniques. These kinds of methods consist in simultaneous clustering on rows and columns, to reorganize the data set into homogeneous blocks. Several biclustering algorithms have been proposed and used on microarray data [1]. In this paper, we try to improve the performance of biclustering algorithms in using the ensemble approach. The principle of ensemble methods is to construct a set of models, then to aggregate them into a single model, by using generally a voting scheme [2]. We propose in this paper a bagging approach for the biclustering of microarray data [3]. The experiments on public microarray data show that ensemble methods produce biclusters with lower residue than classic biclusters. Moreover the ensemble biclusters are also biologically more relevant with respect to the prior knowledge of data.

2 Methods

The principle of bagged biclustering consists in applying a biclustering method on multiple bootstrapped datasets and aggregate the results. Our method can be divided into 3 steps: construction of a collection of biclusters, identification of the meta-cluster, computation of the biclusters. In the first step we generate a high number of different biclusters. To do so, we generate bootstrap samples of the original data. We apply the bootstrap sampling only on the genes. On each of the R bootstrapped datasets, a biclustering algorithm,with the same parameters, is applied to produce K biclusters. We obtain a collection of KR biclusters noted B^b that are used to identify meta-clusters.The aim of the second step is to identify K meta-clusters merging the similar biclusters. The idea is that if two biclusters, generated from different bootstrapped data, are similar, it is likely that they represent the same bicluster. All bootstrapped biclusters representing the same bicluster should be grouped into a meta-cluster. The notion of similarity between two biclusters depends on the number of elements (genes and examples) they have in common. Here we use the Jaccard index to evaluate this similarity. From this similary we compute a distance matrix and a hierarchical clustering of the biclusters using the average linkage. From the obtained dendrogram we can identify K meta-clusters in cutting the dendrogram. Before deducing these meta-clusters in the third step, we estimate the probability of each gene and example to belong to the meta-clusters. The last step consists in computing the final biclusters of the original data. Each meta-cluster is assigned to a bicluster. Then each element can be assigned to biclusters depending on computed probabilities. We define a threshold t, if the probability is higher than t then the gene or example is associated to the bicluster.

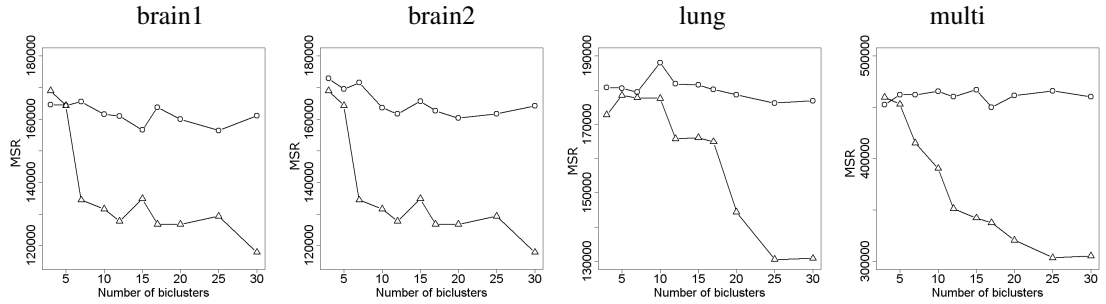


Figure 1. The results of single and bagged biclustering on four real datasets with the Cheng and Church algorithm.

3 Results and Discussion

In our simulation study, we evaluate the performance of bagged biclustering and compare it to single biclustering. We performed our experiments on four microarray datasets with five algorithms: Bimax, Cheng & Church, plaid model, spectral biclustering and Xmotifs. Because of the limitation of space we present here on ly the results with the Cheng and Church algorithm. Figure 1 shows the mean square residue (MSR) of the single (dot lines) and bagged (triangle lines) biclustering in function of the number of biclusters K on the four microarray datasets. We see that at a given number of biclusters the MSR of ensemble biclustering is much smaller than the MSR of single biclustering. These results show that the bagged biclusters is significantly better than single biclusters. We also compared single and bagged biclustering based on the biological coherence of the obtained gene partition. A good tool to check if there is an identified relation between two genes, is the pathway database of the Kyoto Encyclopedia of Genes and Genomes (KEGG). In the table 1, we report the average of the number of over-expressed pathways contained in the identified biclusters. It appears clearly that bagged biclusters contains much more over-expressed pathways and can be consider more biologically relevant than single biclusters.

	Brain1	Brain2	Lung	Multi
single	8	10	12	4
Bagged	18	26	20	15

Table 1. Number of over-represented pathways in the biclusters.

4 Conclusion

In this paper we have introduced the concept of ensemble methods for biclustering in the context of microarray data. On artificial data, we have shown that ensemble method enables to strongly decrease the biclustering error compared to classic methods. On real data, bagged biclustering provides biclusters more relevant than single biclustering according to their MSR value. The use of our approach is a new powerful tool for microarray analysis and should allow biologists to identify new relevant patterns in gene expression data.

Acknowledgements

This research was supported by the CLasSel ANR project ANR-08-EMER-002.

References

[1] S. C. Madeira and A. L. Oliveira, Biclustering algorithms for biological data analysis: a survey, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 1, no. 1, pp. 24–45, 2004.

[2] T. G. Dietterich, Ensemble methods in machine learning, *Lecture Notes in Computer Science*, vol. 1857, pp. 1–15, 2000.

[3] B. Hanczar and M. Nadif, Bagging for biclustering: application to microarray data, in *Proceedings of the 2010 European conference on Machine learning and knowledge discovery in databases: Part I*, (Berlin, Heidelberg), pp. 490–505, Springer-Verlag, 2010.

Validated Chip Annotation

A New Tool for Gene Annotation Quality Control

Gérôme JULES-CLEMENT[1], Jean-Claude HAW KING CHON[1], Jean-Philippe MEYNIEL[1], Philippe LA ROSA[2,3], Emmanuel BARILLOT[2,3] and Charles DECRAENE[1,4]

[1] Département de Recherche Translationnelle, Institut Curie, 26 rue d'Ulm, 75248, Paris, Cedex 05, France

charles.decraene@curie.fr

[2] Service de Bioinformatique, Institut Curie, 26 rue d'Ulm, 75248, Paris, Cedex 05, France

[3] U900 INSERM, Paris, France

[4] UMR144 CNRS, Paris, France

Keywords microarray, annotation, gene expression profiling, quality control.

Abstract

For more than a decade, the microarray technology is a powerful and widely used tool to explore the biological systems and is commonly based on the hybridization of fixed oligonucleotides (probes) with mRNA (or similar) in solution (targets) [1-4]. Probe sequences are designed to correspond to all or part of expressed genes in the cell to collect expression profiles at the gene, gene family, transcript or exon level. Each probe is linked to an annotation file to identify the corresponding gene at the analysis step. Using this technology, a large number of data are collected and the downstream processes developed to analyze the expression profiles are complex [5]. In this context it is crucial to have access to validated annotations for all interrogated genes in order to collect biologically relevant data.

Recent publications investigate microarray annotation based on updated probes alignment using available public sequence databank and show the relevance of well-annotated microarray data [6-10]. As an example, Dai *et al.* (BrainArray,[7]) and Liu *et al.* (AffyProbeMiner,[9]) purpose systematic solutions for Affymetrix microarray annotation building new Chip Definition Files (CDF) using RefSeq or EntrezGene NCBI databank. Ballester *et al.* [10] have recently developed a different approach where "original" probeset structure provided by Affymetrix is kept and the annotations are checked with updated probe alignment using ENSEMBL genome build.

According to these approaches, we have created a new microarray re-annotation protocol called Validated Chip Annotation (VCA) developed for all microarray technologies such as Affymetrix, Agilent, Illumina or Nimblegen. For each microarray, we have produced a specificity score in order to control which gene, transcript(s), exon(s) and CDS are targeted by the probeset. As a result, native Affymetrix CDF and custom CDF have been re-annotated including a quality score.

Thus, we propose a user-friendly annotation which is a common system allowing heterogeneous datasets comparison (intra- and inter-technologies; intra- and inter-species) by considering genes instead of probes or probe sets. This new microarray annotation tool gives the possibility to improve the quality of the data analysis and to explore the biology of the cell using the large number of available bioinformatics tools as IPA (Ingenuity), dedicated R packages, GO, Kegg, …

Acknowledgements

This work was supported by the Institut Curie, CNRS, INSERM, by the European Union under the auspices of the FP7 collaborative project TuMIC, contract no. HEALTH-F2-2008-201662.

References

[1] M. Schena, D. Shalon, R.W. Davis and P.O. Brown, Quantitative monitoring of gene expression patterns with acomplementary DNA microarray. *Science*, 270:467-470, 1995.

[2] T. Sorlie, C.M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M.B. Eisen, M. van de Rijn, S.S. Jeffrey, T. Thorsen, H. Quist, J.C. Matese, P.O. Brown, D. Botstein, P. Eystein Lonning and A.L. Borresen-Dale, Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A*, 98:10869-10874, 2001.

[3] C. Sotiriou, P. Wirapati, S. Loi, A. Harris, S. Fox, J. Smeds, H. Nordgren, P. Farmer, V. Praz, B. Haibe-Kains, C. Desmedt, D. Larsimont, F. Cardoso, H. Peterse, D. Nuyten, M. Buyse, M.J. Van de Vijver, J. Bergh, M. Piccart and M. Delorenzi, Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst*, 98:262-272, 2006.

[4] H. Liu, I. Bebu and X. Li, Microarray probes and probe sets. *Front Biosci (Elite Ed)*, 2:325-338, 2010.

[5] S. Imbeaud and C. Auffray, 'The 39 steps' in gene expression profiling: critical issues and proposed best practices for microarray experiments. *Drug Discov Today*, 10:1175-1182, 2005.

[6] L. Gautier, M. Moller, L. Friis-Hansen and S. Knudsen, Alternative mapping of probes to genes for Affymetrix chips. *BMC Bioinformatics*, 5:111, 2004.

[7] M. Dai, P. Wang, A.D. Boyd, G. Kostov, B. Athey, E.G. Jones, W.E. Bunney, R.M. Myers, T.P. Speed, H. Akil, S.J. Watson and F. Meng, Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res*, 33:e175, 2005.

[8] J. Harbig, R. Sprinkle, and S.A. Enkemann, A sequence-based identification of the genes detected by probesets on the Affymetrix U133 plus 2.0 array. *Nucleic Acids Res*, 33:e31, 2005.

[9] H. Liu, B.R. Zeeberg, G. Qu, A.G. Koru, A. Ferrucci, A. Kahn, M.C. Ryan, A. Nuhanovic, P.J. Munson, W.C. Reinhold, D.W. Kane and J.N. Weinstein, AffyProbeMiner: a web resource for computing or retrieving accurately redefined Affymetrix probe sets. *Bioinformatics*, 23:2385-2390, 2007.

[10] B. Ballester, N. Johnson, G. Proctor, and P. Flicek, Consistent annotation of gene expression arrays. *BMC Genomics*, 11:294, 2010.

Biomarkers Discovery in Breast Cancer by Interactome-Transcriptome Integration

Maxime GARCIA[1,2,3], Olivier STAHL[1,2,3], Pascal FINETTI[1,2,3],
François BERTUCCI[1,2,3], Daniel BIRNBAUM[1,2,3] and Ghislain BIDAUT[1,2,3]

[1] Centre de Recherche en Cancérologie de Marseille, UMR891 INSERM, 13009 MARSEILLE, France
[2] Institut Paoli-Calmettes, 13009 MARSEILLE , France
[3] Université de la Méditerranée, 13007 MARSEILLE , France
{maxime.garcia, olivier.stahl, ghislain.bidaut}@inserm.fr,
daniel.birnbaum@inserm.fr, {finettip, bertuccif}@marseille.fnclcc.fr

Keywords Breast cancer, large scale data integration, transcriptome, interactome, biomarkers.

1 Introduction

High-throughput gene-expression profiling technologies yield several genomic signatures to predict clinical condition or patient outcome. However, such signatures show dependency on training set, lack of generalization and instability. We are proposing an interactome-based algorithm ITI [1] to find a generalizable signature for prediction of breast cancer relapse by superimposition of a large scale protein-protein interaction data (human interactome) over several gene expression datasets. The algorithm extracts discriminative regions in the interactome (subnetworks) predicting 5 years relapse free survival in breast cancer. This method expands the algorithm proposed by Chuang et al [2] with the added capability to extract a genomic signature from several gene-expression data sets simultaneously. It was trained with four breast cancer DNA microarray data sets and allowed the discovery of a breast cancer relapse signature constituted by 58 subnetworks that was generalizable over independent data. Exploration of annotations has shown that this set of subnetworks reflects several biological processes linked to cancer and is a good candidate for establishing a subnetworks-based signature for prediction of 5 years relapse free survival in breast cancer.

2 Methods

Two data types are fed to the algorithm, large scale interaction data and gene expression profiles (GEP). To build our set of interaction data, we integrated five existing human protein-protein interaction (PPI) maps (HPRD[3], Ramani[4], MINT[5], IntAct[6] and DIP[7]). All PPI sets were integrated by uniqueness of NCBI EntrezGene identifiers, leading to a final set of 70,530 interactions among 13,202 proteins. We built a compendium of breast cancer tumors profiles by examining datasets available with clinical information on the NCBI GEO database. Each dataset was downloaded from GEO as raw data and normalized within Bioconductor using affy and gcrma packages. Tumors without relapse information were removed, leading to a final compendium of 5 datasets containing 787 tumors [8,9,10,11,12].

One dataset was left out for cross-validation purpose with independent testing. Pearson correlation is computed between GEPs and clinical information (Distant Metastasis Free Survival [DMFS] status) for each dataset. Interactome regions whose gene expression is highly correlated with DMFS status are then detected. Random distributions of score are drawn to assign p-values to the subnetworks and perform a statistical validation. Finally, the discriminative power of statistically significant subnetworks is tested against an independent dataset. We found a set of 58 subnetworks linked to 5 years relapse free survival in breast cancer. These were stored in our local resource, available from the ITI web site (http://bioinformatique.marseille.inserm.fr/iti).

Intrinsic biology of the 58 extracted subnetworks was examined using annotation information from the NCBI EntrezGene database and the Gene Ontology Consortium. We found that subnetworks formed complexes functionally supporting the studied disease for metabolism, cell cycle control, proliferation, cell-cell adhesion

and immunological response, which are known mechanisms of cancer and metastatic process. Several drivers genes were detected, including CDK1, NCK1 and PDGFB, some not previously linked to breast cancer relapse. Classification in Wang experiment [12] showed an accuracy of 0.59 over independent data within the same dataset. SVM classification based on a set of 85 subnetworks showed an accuracy of 0.78 (sensitivity of 0.25 and specificity of 0.88) over an independent dataset of 182 tumors [11].

3 Conclusion

We present an Interactome-Transcriptome Integration algorithm (ITI) to identify subnetwork-based prognostic signatures generalizable over multiple datasets of breast cancer. We performed large scale integration of 5 DNA microarray datasets to create a breast cancer compendium and constructed a large coverage human interactome by integrating 5 existing human protein-protein interaction datasets. These data, used conjointly with a discriminative subnetwork detection algorithm and significance scoring, allowed the identification of interactome regions linked with 5 years relapse free survival in breast cancer. Subnetworks found have been linked to biological functions related to metastasis and breast cancer, such as cell differentiation, cell cycle signaling, cell adhesion and proliferation, as well as functional links to immune response. This resource is the first of its kind to allow linking a human interactome to diseases or clinical situations. This resource can be mined for identification of potential drug targets to establish finer disease models.

Acknowledgements

Research is funded by the Institut National du Cancer and the Institut National de la Santé et de la Recherche Médicale (INSERM). Our Beowulf cluster was funded by a Fondation pour la Recherche Médicale grant. Maxime Garcia is funded by a fellowship from INSERM and the Provence-Alpes-Côte d'Azur Region.

References

[1] M. Garcia et al., Linking interactome to disease: a network-based analysis of metastatic relapse in breast cancer. *Handbook of Research on Computational and Systems Biology: Interdisciplinary Applications.* IGI Global, 406-427, 2011.

[2] H.Y. Chuang et al., Network-based classification of breast cancer metastasis. *Mol Syst Biol.*, 3:140, 2007.

[3] T.S.K. Prasad et al., Human protein reference database-2009 update. *Nucleic Acids Res.*, 37:767-772, 2009.

[4] A.K. Ramani et al., Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biol.*, 6(5):R40, 2005.

[5] A. Ceolet al., Mint, the molecular interaction database: 2009 update. *Nucleic Acids Res.*, 38:532-539, 2010.

[6] B. Aranda et al., The intact molecular interaction database in 2010. *Nucleic Acids Res.*, 38:525-531, 2010.

[7] L. Salwinski et al., The database of interacting proteins: 2004 update. *Nucleic Acids Res.*, 32:449-451, 2004.

[8] C. Desmedt et al., Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. *Clin Cancer Res.*, 14(16):5158-5165, 2008.

[9] S. Loi et al., Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen. *BMC Genomics*, 9:239, 2008.

[10] R. Sabatier et al., A gene expression signature identifies two prognostic subgroups of basal breast cancer. *Breast Cancer Res Treat*, 2010.

[11] M. Schmidt et al., The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer Res*, 68(13):5405-5413, 2008.

[12] Y. Wang et al., Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, 365(9460):671-679, 2005.

RNAspace: an Integrated Environment for the Prediction, Annotation and Analysis of non-coding RNA

Marie-Josée CROS[1], Antoine de MONTE[2], Jérôme MARIETTE[3], Philippe BARDOU[4], Daniel GAUTHERET[5], Hélène TOUZET[2] and Christine GASPIN[1,3]

[1] INRA, Unité de Biométrie et Intelligence Artificielle, UR 875, F-31320 Castanet, France
[2] LIFL, UMR CNRS 8022 Université Lille 1 and INRIA Lille Nord Europe, France
[3] INRA, Plateforme bioinformatique, F-31320 Castanet, France
[4] INRA, SIGENAE, UMR 444, F-31320 Castanet, France
[5] IGM UMR 8621 CNRS-U Paris sud, France

contact@rnaspace.org

Abstract *RNAspace is an environment that allows to create web sites dedicated to non-protein-coding RNA (ncRNA) prediction, annotation and analysis. The web sites allow users to run a variety of tools in an integrated and flexible way. RNAspace is focused on the integration of complementary ncRNA gene finders. It also offers a set of tools for the comparison, visualization, edition and export of ncRNAs candidates. Predictions can be filtered according to a large set of characteristics.*
A public web site http://rnaspace.org *has been created that allows for on line annotation of a complete bacterial genome or a small eukaryotic chromosome.*

Keywords non-protein-coding RNA, genome 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, led us to propose to biologists an *in silico* environment allowing structural and functional annotations of these molecules. For this purpose, an environment called RNAspace was developed that allows to install dedicated web sites just by adjusting various global parameters (gene finders to consider, maximal size for input genomic sequences ...).

A web site allows to (i) run a variety of ncRNA gene finders in an integrated environment, (ii) explore computed results with dedicated tools for comparison, visualization, alignment and edition and (iii) export them in various formats (FASTA, GFF, RNAML).
Gene finders are organized into three categories containing respectively:

1] known ncRNA based gene finders including (i) sequence homology search tools: BLAST [1], YASS [2] on ncRNA databases: Rfam [3], fRNAdb [4], miRBase [5], (ii) general purpose ncRNA motif search tools: Infernal [6], Darn [7], Erpin [8], (iii) specialized search tools: RNAmmer [9] for ribosomal RNAs, tRNAscan_SE [10] for transfer RNAs;
2] a comparative analysis gene finder: an *ad hoc* pipeline [11] has been implemented based on BLAST or YASS for similarities search and caRNAc [12] or RNAz [13] for consensus structure inference;
3] an *ab initio* gene finder based on detection of atypical GC% regions.

All gene finders can be run with default parameters values. However it is also possible for users, through a dedicated interface, to set some of these parameters to specific values according to the level of knowledge of biological data and user expertize. Once the execution of selected gene finders is achieved, combination of predictions is possible on demand. For example, predictions that have only tiny differences in positions on the input genomic sequence are merged into a single prediction. This avoids having a lot of redundant predictions for ncRNA families (*e.g.* tRNA) predicted by several gene finders. An overview of all putative

ncRNAs found on the genomic sequence is provided. 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 ...). Any putative ncRNA can be edited and deleted. It is also possible to visualize putative ncRNAs on the input genomic sequence with several genome browsers: JBrowse, CGview, ApolloRNA. Finally, a functionality allows the export of candidate ncRNAs in several formats for a future usage.

The environment relies on collaboratively-developed code using the Python language and the HTTP framework CherryPy (see http://cherrypy.org for more detail). The code is open source under GPL license and is available on Source Forge (https://sourceforge.net/projects/rnaspace/). It has been conceived to be as parameterizable and extensible as possible. This allows to configure web sites for special uses. Parametrization of a site includes declaration of available gene-finders, limits for process time execution, disk space, storage duration, execution on a connected computer cluster via a job scheduler. It is also worth to note that the environment can be used in command line and thus inserted in a pipeline. Using the RNAspace environment, a public site http://rnaspace.org has been created. It accepts genomic sequences up to 5Mb. Computations are executed on the computer cluster of the Genotoul bioinformatic platform. For example, it is possible to get an annotation for the *E. coli* genome (4.9M nucleotides) using a wide selection of gene finders and recovering the majority of known RNA genes in less than one hour.

In the near future, we plan to incorporate supplementary prediction approaches, to provide more advanced methods to eliminate redundant results, to include information on the genomic context, to define and compute a common normalized prediction score (indeed some gene finders now provide a score but these scores are not comparable). Furthermore with the huge quantity of high-throughput sequencing data obtained by transcriptome studies, it is also highly desirable to consider RNAseq and sRNAseq data for the annotation and the search for potential targets of regulatory RNA acting through RNA-RNA interactions.

Acknowledgements

This work was supported by IBiSA (Infrastructures en Biologie, Santé et Agronomie).

References

[1] S.F. Altschul, W. Gish, W. Miller, E.W. Myers and D.J. Lipman, Basic local alignment search tool. *J Mol Biol* 215 (3): 403–410, 1990.
[2] L. Noe, G. Kucherov, YASS: enhancing the sensitivity of DNA similarity search. *Nucleic Acids Res.*, 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 Res.*, 37(Database Issue), 2009.
[4] T. Kin, K. Yamada, G. Terai, H. Okida, Y. Yoshinari, Y. Ono, A. Kojima, Y. Kimura, T. Komori and K. Asai, fRNAdb: a platform for mining/annotating functional RNA candidates from non-coding RNA sequences. *Nucleic Acids Res*., 35 (Database Issue), 2007.
[5] S. Griffiths-Jones, H.K. Saini, S. van Dongen and A.J. Enright, miRBase: tools for microRNA genomics. *Nucleic Acids Res*., 36 (Database Issue), 2008.
[6] E.P. Nawrocki, D.L. Kolbe and S.R. Eddy, Infernal 1.0: Inference of RNA alignments. *Bioinformatics,* 25(10), 2009.
[7] M. Zytnicki, C. Gaspin and T. Schiex, DARN! A Weighted Constraint Solver for RNA Motif Localization. *Constraints*, Vol. 13, 2008.
[8] D. Gautheret and A. Lambert, Direct RNA Motif Definition and Identification from Multiple Sequence Alignments using Secondary Structure Profiles. *J Mol Biol*., 313:1003-11, 2001.
[9] K. Lagesen, P. Hallin, E.A. Rødland, H.-H. Stærfeldt, T. Rognes and D.W. Ussery, RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res*., 35(9), 2003.
[10] T.M. Lowe and S.R. Eddy, tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*., 25(5), 1997.
[11] B. Grenier-Boley, A. de Monte and H. Touzet, CG-seq: a toolbox for automatic annotation of genomes by comparative analysis. INRIA Research Report N°7428, available on HAL, Oct. 2010.
[12] H. Touzet and O. Perriquet, CARNAC: folding families of non coding RNAs. *Nucleic Acids Res*., 142(Web Server Issue), 2004.
[13] S. Washietl, I.L. Hofacker and P.F. Stadler, Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci*. U.S.A. 102, 2454-2459, Feb. 2005.

Dynamics of Small RNA-Directed DNA Methylation during the Arabidopsis Innate Immune Response

Anne-Laure ABRAHAM, Agnès YU, Gersende LEPÈRE and Lionel NAVARRO

Institut de Biologie de l'Ecole Normale Supérieure (IBENS), UMR8197 CNRS,
46 rue d'Ulm 75230 PARIS Cedex 05 France
{anne-laure.abraham, ayu, gersende.lepere, lionel.navarro}@ens.fr

Abstract *Small RNAs (sRNAs) are involved in the transcriptional and post-transcriptional control of gene expression. Whereas these small regulatory RNAs were originally characterized in plant and animal development they have recently emerged as key components of the innate immune response [1,2,3,4]. For example, the team has recently implicated the Arabidopsis small RNA-directed DNA methylation pathway in antibacterial resistance. However, the dynamic of DNA methylation mediated by short interfering RNA (siRNA) following bacterial detection remains elusive. In the present work, we will report the profiling of small RNAs and DNA methylation in the course of the Arabidopsis thaliana antibacterial defence response. We will also present the extent to which these siRNA-directed DNA methylation changes contribute to the regulation of immune-responsive gene expression.*

Keywords non-coding RNA, epigenetics and epigenomics, high throughput sequencing data analysis.

1 Introduction

In higher eukaryotes, the vast majority of the genome appears to be transcribed, leading to an extraordinary diversity of non-coding RNAs (ncRNAs). Whereas the functional significance of these ncRNAs is mostly unknown, increasing evidence suggests a role for these molecules in guiding chromatin modifications [5]. In plants, a large proportion of ncRNAs is processed by the RNA silencing machinery to produce short interfering RNAs. Some of them guide sequence specific DNA methylation through a phenomenon referred to as RNA-directed DNA methylation (RdDM) [6]. RdDM is usually associated with transcriptional silencing of transposons, retrotransposons and repeated sequences. By using a reverse genetic approach the team found that RdDM negatively regulates antimicrobial defence. Consistent with these findings, they found that the bacterial-derived elicitor flg22, a 22 amino acid peptide derived from the N-terminal part of bacterial flagellin, down-regulates the RdDM pathway in the course of the elicitation. The present study aims to unravel (1) the dynamics of Arabidopsis siRNA-directed methylation changes during the course of flg22 elicitation (2) the contribution of these changes in the control of immune-responsive gene expression.

2 Methods

Using Illumina sequencing we have deep sequenced sRNA libraries derived from Arabidopsis thaliana WT leaves treated for 3, 6 and 9 hours with active or inactive forms of flg22. We have introduced a spike internal control (artificial sRNA) in each library in order to normalize the data before sRNA mapping and data mining [7]. We will identify sRNAs that are differentially expressed in the course of the elicitation and retrieve their cognate targets. In parallel, we have deep sequenced mRNAs using Illumina sequencing technology to identify mRNAs differentially expressed in response to flg22. Finally, we have used bisulfite sequencing approach to determine the DNA methylation status of specific immune responsive genes in the course of flg22 elicitation. This method consists in converting unmethylated cytosine to thymine by bisulfite treatment followed by sequencing. We will compare changes in siRNA and DNA methylation levels and assess the contribution of such siRNA-mediated epigenetic modifications in the regulation of immune-responsive gene expression. Overall, these approaches will give us insights into the siRNA-directed transcriptional regulation of defence genes in the course of the elicitation.

References

[1] K.D. Taganov, M.P. Boldin, K.J. Chang and David Baltimore, NF- κB-dependent induction of microRNA miR-146, an inhibitor targeted to signaling proteins of innate immune responses. *Proc Natl Acad Sci U S A*, 103(33): 12481-12486, 2006.

[2] L. Navarro, P. Dunoyer, F. Jay, B. Arnold, N. Dharmasiri, M. Estelle, O. Voinnet and J.D.G. Jones, A plant miRNA contributes to antibacterial resistance by repressing auxin signalling. *Science*, 312(5772):436-9, 2006.

[3] L. Navarro,F. Jay, K. Nomura, S.Y. He and O. Voinnet , Suppression of the microRNA pathway by bacterial effector proteins. *Science*, 321 : 964-967, 2008.

[4] F. Jay, J-P. Renou, O. Voinnet and L. Navarro, Biotic stress-associated microRNAs : Identification, detection, regulation and functional analyses. *Methods in Molecular Biology*, Plant microRNAs. Humana Press, 592: 183-202, 2009.

[5] S. Katiyar-Agarwal and H. Jin, Role of small RNAs in host-microbe interactions. *Annu. Rev. Phytopathol.*, 48:225-46, 2010.

[6] M. A. Matzke and J. A. Birchler, RNAi-mediated pathways in the nucleus. *Nat. Rev. Genet.* 6(1):24-35, 2005.

[7] N. Fahlgren, C.M. Sullivan, K.D. Kasschau, E.J. Chapman, J.S. Cumbie, T.A. Montgomery, S.D. Gilbert, M. Dasenko, T.W. Backman, S.A. Givan and J.C. Carrington, Computational and analytical framework for small RNA profiling by high-throughput sequencing. *RNA*, 15(5):992-1002, 2009.

An Automatic Method for Identifying TE-derived Pre-miRNAs

Sébastien TEMPEL[1] and Fariza TAHI[1]

Laboratoire IBISC, Tour Evry 2, 523, Place des Terrasses, 91000 Evry, France
sebastien.tempel@ibisc.univ-evry.fr
fariza.tahi@ibisc.univ-evry.fr

Abstract *MicroRNAs and transposable elements (TEs) share numerous characteristics: size, stable secondary structure, maturation into shorter sequence. TE-derived pre-miRNAs are microRNAs that are generated from transposable elements. TE-derived miRNAs and TEs have a similar distribution of their occurrences that is distinct from 'classical' pre-miRNAs: a TE-derived pre-miRNA have many occurrences spread in many chromosomes, and a 'classical' pre-miRNA has generally one occurrence or few ones appearing in a cluster.*

We developed an automatic method called miRNAcheck to distinguish a TE-derived pre-miRNA from a 'classical' pre-miRNA. Given a pre-miRNA candidate, miRNAcheck calculates in a first step the number of occurrences of the candidate in the genome. The ten occurrences the most similar to the candidate sequence are then extended and a consensus sequence is created. Finally, the consensus sequence is compared to TE sequences in RepBase, a database of TEs.

From the 1048 human and 672 mouse pre-miRNAs of miRBase, we selected the pre-miRNAs that have at least 10 similar occurrences in the genome. We get 83 human and 84 mouse candidates. Among them, 60 human and 69 mouse pre-miRNAs are identified by miRNAcheck as TE-derived pre-miRNAs.

miRNAcheck is available at the Web site: http://EvryRNA.ibisc.univ-evry.fr/

Keywords pre-miRNA, Transposable Element, TE-derived pre-miRNA.

1 Introduction

Recent studies show the whole genomes of higher eukaryotes are transcribed [1] while the genes represent only few percentages of these genomes. Non-genic regions are mainly composed by non-coding RNAs (ncRNAs) and transposable elements (TEs) that represent a substantial fraction of many eukaryotic genomes. For example, about 50% of the human genome is derived from transposable element sequences [2].

Transposable elements (TE) are present in nearly all genomes that have been studied to date and in some cases represent most of the genome [3]. They move or are copied from one genomic location to another [4]. TEs are characterized and classified on the basis of terminal or sub-terminal remarkable structures or of their protein-coding capacity. TEs are conventionally divided into two classes [5]: Class I and Class II. Class I is represented by retrotransposons Long INterspersed Elements (LINEs), Short INterspersed Elements (SINEs), Long Terminal Retrotransposons (LTRs), and Endogenous RetroViruses (ERVs), all requiring reverse transcription from an RNA intermediate. Class II includes "cut-and-paste" DNA transposons, which are characterized by terminal inverted repeats (TIRs) and are mobilized by a transposase [4]. Many families of both classes do not show any coding capacity and are called non-autonomous transposable elements. They have cumulated so many mutations, insertions or deletions so they are generally solely defined by their extremities [6,7]. In Class I, SINEs are short sequences (100 to 500 nt) and present a stable secondary structure similar to the fusion of tRNA structure and hairpin structure [8,9]. In Class II, Miniature Inverted-repeat Transposable Elements (MITEs) are non-autonomous transposable elements characterized by a small size (80-500 nt), a stable secondary structure, generally an hairpin structure, and an insertion into A + T-rich regions [10]. MITEs could generate small interfering RNAs (22-24 nt) by a pathway similar to that required for TE-derived small interfering RNA (siRNA) biogenesis and by DICER-like proteins [11].

siRNAs are non-coding RNAs generated from a biological response to double-stranded RNAs (dsRNAs) called RNA interferences (RNAi) [12,13]. Long dsRNA molecules (for example TEs) initiate RNAi by being converted to smaller 21-23 nt siRNAs by the Dicer enzyme. Therefore, hairpin RNAs have been commonly used to induce RNAi [12].

MicroRNAs (miRNAs) are non-coding RNAs with only 21-25 nt in sequence length that are present in all sequenced higher eukaryotes [14,15]. They are involved as negative regulators of gene expression at the post-transcriptional level by binding to specific mRNA targets whose translations are inhibited or down-regulated [15,16]. According to the current understanding of miRNA biogenesis, miRNA genes are transcribed and then are cleaved into a 39-938 nt long precursor of miRNA sequences (pre-miRNAs) by the Drosha/Pasha complex. Pre-miRNAs, structured as hairpins, are transported into the cytoplasm by Exportin5 and cleaved by Dicer into mature miRNAs [14]. In the RISC complex, a miRNA binds with a specific mRNA transcript and leads to the cleavage or the degradation of the mRNA.

Non-autonomous TEs like SINEs and MITEs and pre-miRNAs share some characteristics (Fig. 1), especially the similarity of their biogenesis [17,18]. Moreover, some recent bioinformatic studies show that some pre-miRNAs share their sequences or an important part of their sequences with TEs [17,18,19]. These pre-miRNAs, annotated in miRBase [20], are called TE-derived pre-miRNAs and present a high number of occurrences in the genome [18]. Both classes of TEs could be involved in TE-derived pre-miRNAs [21].

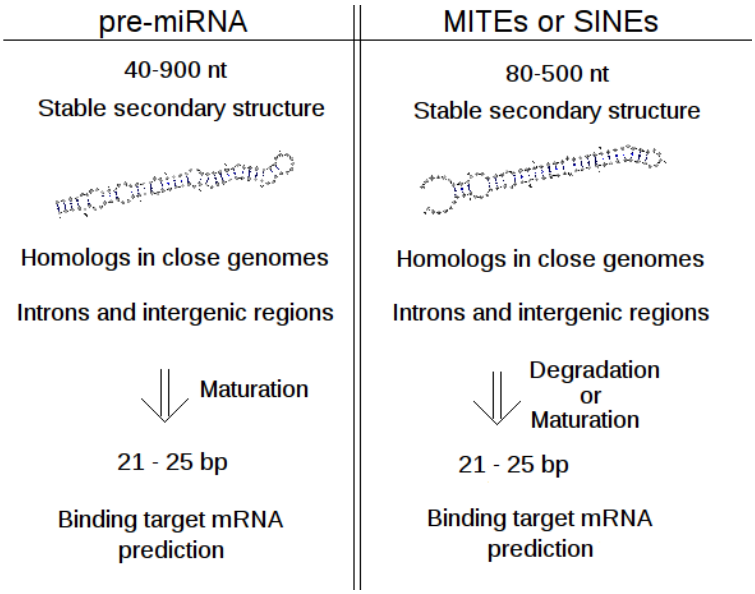


Figure 1. Bioinformatic characteristics of pre-miRNAs and transposable elements. The size and the secondary structure of pre-miRNAs and non-autonomous TEs (MITEs and SINEs) are similar. Mature miRNAs down regulate target binding genes [14] and siRNAs generated from TE regulate TE-genes [11].

In this article, we present an automatic method, called miRNAcheck, for identifying TE-derived pre-miRNAs.

2 Our approach

2.1 How to identify TE-derived pre-miRNAs

The main criteria that identifies TE-derived pre-miRNA candidates from other pre-miRNAs is the number and the distribution of the candidate occurrences. Pre-miRNAs do not have a transposition mechanism like TEs, and are not widespread in all chromosomes, not even widespread in one chromosome [22,23]. A mechanism that can copy pre-miRNAs is an error of chromosome replication that can give a cluster of miRNA genes [24]. We consider that a pre-miRNA candidate that has 10 or more copies in the genome has a strong probability to

be a TE-derived pre-miRNA. This difference in the copy mechanism changes the localization of occurrences and allows to distinguish TEs from satellites (tandem repeats) [4]. The part of TE sequences in pre-miRNAs are often too short (about 10 bp [19]) to be used directly as query with repeat identifier tools like Censor [25], RepeatMasker (www.repeatmasker.org) or Repet [26].

The study of a pre-miRNA occurrences distribution depends on:

1. The number of occurrences in the whole genome: excepted TE-derived pre-miRNAs, a pre-miRNA has few occurrences in the whole genome [22,23]; therefore, we can consider that a pre-miRNA candidate that occurs several times in the genome has a strong probability to be a TE-derived pre-miRNA.
2. The number of distinct chromosomes where appear the occurrences: the tandem repeat mechanism does not allow a sequence to jump to another chromosome [4]. Then, very few pre-miRNAs are found in two chromosomes. The presence on a second chromosome could be explained by chromosomic rearrangement during the evolution. Therefore, we can consider that a pre-miRNA candidate present in several chromosomes has a strong probability to be a TE-derived pre-miRNA.
3. The distance between the occurrences: some recent studies show that some similar miRNA genes are clustered in a small distance [24] and that the tandem repeat mechanism creates copies close to the original sequence [4]. For example, there is a cluster of 49 miRNA genes in human chromosome 19 spread on only 150 kb. Sewer *et al.* approximated the maximal distance of a miRNA gene cluster to 20kb [24]. Therefore, we can consider that if two or more similar occurrences are distant of more than 20 kb, there is a strong probability that the candidate is a TE-derived pre-miRNA.

2.2 Description of miRNAcheck method

In order to identify TE-derived pre-miRNAs, we developed an automatic method called miRNAcheck that works as follows.

Given a pre-miRNA candidate, the first step of our method consists in a study of the candidate occurrence distribution, using BLAT [27] of UCSC Genome Browser [28]. We chose BLAT instead of BLAST on NCBI or EBI because the results do not correspond to the chromosomes but to the scaffolds that do not allow a distribution study of the occurrences. We assume that two occurrences are in a same cluster if they are on two distinct chromosomes or are distant at least of 100000 nt. We calculate the number of occurrences, named ”hits”, of the candidate in the genome, and more particularly the number of ”similar hits”. Similar hits are hits whose similarity with the candidate is greater than 80% and whose size is between 80% and 120% of the candidate size. This definition is similar to the identification definition of transposable elements [21]. We calculate also the number of chromosomes containing the different similar hits.

After the study of the occurrences distribution, the second step of our method looks for a possible similarity with transposable elements. However, the size of human pre-miRNA candidates (11-186 nt in the last version of miRBase [20]) could be too short for an identification by Censor [25]. To extend the candidate sequence, our method extracts the ten best similar hits (or all similar hits if there is less than ten hits). Using UCSC genome browser [28], we get the surrounding sequence around each hit: 100 nt left to the hit and 100 nt right to the hit. These sequences are then aligned with ClustalW [29] and a consensus sequence is created. The nucleotide consensus at position *i* corresponds to the nucleotide present at least 5 times in the alignment at same position; otherwise there is the character N. We assume that ten hits are sufficient to create a consensus sequence since the hits have a similarity with the candidate greater than 80%.

Finally, we compare the consensus sequence to a TE database: RepBase [30]. For performing this comparison, we used Censor [25] (we choose Censor instead of Repet [26] and RepeatMasker (www.repeatmasker.org) because to our knowledge there is no Repet webserver; and we preferred Censor to RepeatMasker because it was easier to extract the data from Censor webserver). The candidate is a TE-derived pre-miRNA if the consensus is similar to a TE in RepBase.

2.3 miRNAcheck tool

Our method was implemented in JAVA. The obtained tool, called miRNACheck, is available on the Web site : http://EvryRNA.ibisc.univ-evry.fr.

Figure 2. JAVA interface of miRNAcheck. It shows here the results obtained for the human pre-miRNA HAS-MIR-1273e.

The interface of miRNAcheck (Fig. 2) works as follows: the user enters the sequence of a pre-miRNA candidate in STADEN format, enters a name and chooses the corresponding genome. miRNAcheck sends a request to BLAT at the UCSC Genome Browser and gets the hits of the sequence in the genome (2 in Fig. 2). The line above the hits table summarizes the BLAT result (the number of hits returned by BLAT, the number of chromosomes where appear the hits, and the number of similar hits, i.e. hits that have a size between 80% and 120% of the pre-miRNA size and that have a similarity greater than 80% with the candidate sequence). The user can check the hits obtained from BLAT with a link to the BLAT webpage that stored the results. miRNAcheck selects then the 10 most similar hits (or all if there are less than 10 similar hits) and extends the hits in the genome sequence. The extended hit sequences are then aligned by ClustalW and a consensus sequence is generated (3 in Fig. 2). Finally, the consensus is sent to RepBase database [30] in order to identify a TE candidate associated to the consensus sequence. The alignment between the consensus and the most similar TE is then shown (4 in Fig. 2). A pop-up summarizes the results and specifies if the candidate is a TE-derived pre-miRNA (5 in Fig. 2).

3 Results and discussion

For our tests, we considered the 1048 human pre-miRNAs and the 672 mouse pre-miRNAs present in MiRBase version 16 [20]. The first step was to calculate the number of hits of each pre-miRNA. Only 83 human pre-miRNAs and 84 mouse pre-miRNAs have more than 10 similar hits in the genome. These pre-miRNAs are listed in Fig. 3 and Fig. 4. Thanks to miRNAcheck, we found that among these 83 (84) human (mouse) pre-miRNAs, 60 (69) are TE-derived pre-miRNAs (see Fig. 3 and Fig. 4).

Pre-miRNA name	TE name	Hits	Chrom	Pre-miRNA name	TE name	Hits	Chrom
HSA-MIR-466	???	17	12	HSA-MIR-548y	MADE1	13	11
HSA-MIR-516a-1	???	10	1	HSA-MIR-548z	MADE1	145	26
HSA-MIR-518a-1	???	10	1	HSA-MIR-566	AluSx1	108	16
HSA-MIR-518c	???	10	1	HSA-MIR-570	MADE1	55	14
HSA-MIR-519a-1	???	10	1	HSA-MIR-603	MADE1	47	14
HSA-MIR-519b	???	11	1	HSA-MIR-620	???	12	8
HSA-MIR-519c	???	11	1	HSA-MIR-622	???	43	22
HSA-MIR-520b	???	12	1	HSA-MIR-649	MER8	15	9
HSA-MIR-520f	???	10	1	HSA-MIR-650	???	10	2
HSA-MIR-526b	???	12	1	HSA-MIR-526b	???	12	1
HSA-MIR-548a-1	MADE1	55	17	HSA-MIR-1233-1	???	27	6
HSA-MIR-548a-2	MADE1	91	19	HSA-MIR-1233-2	???	27	6
HSA-MIR-548a-3	MADE1	102	21	HSA-MIR-1244-1	???	22	20
HSA-MIR-548aa-1	MADE1	86	19	HSA-MIR-1244-2	???	22	20
HSA-MIR-548aa-2	MADE1	157	22	HSA-MIR-1244-3	???	22	20
HSA-MIR-548b	MADE1	14	10	HSA-MIR-1254	AluJr	32	12
HSA-MIR-548c	MADE1	145	26	HSA-MIR-1255b-1	TIGGER1	160	23
HSA-MIR-548d-1	MADE1	86	19	HSA-MIR-1261	TIGGER1	19	14
HSA-MIR-548d-2	MADE1	157	22	HSA-MIR-1268	AluSz	87	11
HSA-MIR-548e	MADE1	62	16	HSA-MIR-1273	AluS	148	23
HSA-MIR-548f-1	MADE1	89	24	HSA-MIR-1273d	AluSq	84	20
HSA-MIR-548f-2	MADE1	103	21	HSA-MIR-1273e	AluS	108	23
HSA-MIR-548f-3	MADE1	89	18	HSA-MIR-1285-2	Alu2_TS	36	14
HSA-MIR-548f-5	MADE1	47	12	HSA-MIR-1290	TIGGER4a	61	24
HSA-MIR-548g	MADE1	27	14	HSA-MIR-1299	CER	46	16
HSA-MIR-548h-1	MADE1	45	18	HSA-MIR-1302-1	MER53	54	15
HSA-MIR-548h-2	MADE1	179	24	HSA-MIR-1302-3	MER53 / HERV46I	10	9
HSA-MIR-548h-4	MADE1	34	14	HSA-MIR-1302-4	MER53	49	21
HSA-MIR-548i-1	MADE1	17	7	HSA-MIR-1324	???	10	9
HSA-MIR-548i-2	MADE1	16	7	HSA-MIR-1973	???	10	9
HSA-MIR-548i-3	MADE1	18	7	HSA-MIR-3118-1	L1PA13_5 / L1PA13_5	14	7
HSA-MIR-548i-4	MADE1	128	22	HSA-MIR-3118-2	L1PA13_5 / L1PA13_5	14	7
HSA-MIR-548l	MADE1	85	20	HSA-MIR-3118-3	L1PA13_5 / L1PA13_5	14	7
HSA-MIR-548m	MADE1	74	20	HSA-MIR-3118-4	L1PA13_5 / L1PA13_5	13	7
HSA-MIR-548n	MADE1	180	22	HSA-MIR-3118-5	L1PA13_5 / L1PA13_5	12	7
HSA-MIR-548o	HSMAR1	90	21	HSA-MIR-3118-6	L1PA13_5 / L1PA13_5	13	7
HSA-MIR-548p	MADE1	80	21	HSA-MIR-3135	AluJr	29	16
HSA-MIR-548t	MADE1	104	23	HSA-MIR-3669	???	16	6
HSA-MIR-548u	MADE1	91	19	HSA-MIR-3673	???	23	9
HSA-MIR-548v	MADE1	58	20	HSA-MIR-3674	LTR8A	94	21
HSA-MIR-548w	MADE1	132	24	HSA-MIR-3683	SATR1	14	8
HSA-MIR-548x	MADE1	89	16				

Figure 3. Human pre-miRNAs that have at least 10 similar hits in the genome. Pre-miRNAs that have "???" in 'TE name' column are not similar to known TEs listed in Repbase. The columns 'Hits' and 'Chrom' correspond respectively to the number of similar hits and to the number of chromosomes where appear these hits. There are 24 genomic chromosomes and 9 haplotype chromosomes in UCSC Genome Browser [27].

As shown in Fig. 3, 23 human and 15 mouse pre-miRNAs (that have more than 10 similar hits) do not correspond to a RepBase TE. Respectively, only 10 and 4 of these human and mouse pre-miRNAs (for example HSA-MIR-518c) occur in one chromosome. However, the hits of these pre-miRNAs are not close to each other (some of them are distant to more than two million nt from other hits). These 23 human and 15 mouse pre-miRNAs require therefore more study in order to know if they are TE-derived or not.

Fig. 3 and Fig. 4 show also that pre-miRNAs having a same name prefix (e.g. HSA-MIR-548A-1, HSA-MIR-548A-2, HSA-MIR-548B, etc.) correspond to a same TE, which is not surprising since these pre-miRNAs have similar sequences.

Pre-miRNA name	TE name	Hits	Chrom	Pre-miRNA name	TE name	Hits	Chrom
MMU-MIR-1194	MERVL_LTR	198	22	MMU-MIR-467a-4	ID_B1	22	6
MMU-MIR-1195	B1_Mur3	13	8	MMU-MIR-467a-5	ID_B1	22	6
MMU-MIR-1274a	ETNERV / ERVB7_4-LTR	15	12	MMU-MIR-467a-6	ID_B1 / CR1-79_HM	21	5
MMU-MIR-1935	B1_Mus2	125	19	MMU-MIR-467a-7	ID_B1 / CR1-79_HM	22	6
MMU-MIR-1937a	MMERGLN_I / RLTR1D2	10	3	MMU-MIR-467a-8	ID_B1	22	6
MMU-MIR-1937b-1	MuRRS4-int	34	13	MMU-MIR-467a-9	ID_B1	22	6
MMU-MIR-1937b-2	RLTR6C_Mm	142	22	MMU-MIR-467b	ID_B1	15	1
MMU-MIR-1937b-3	MuRRS4-int / MURVY_LTR	53	13	MMU-MIR-467c	ID_B1	17	3
MMU-MIR-1937b-4	RLTR6C_Mm	139	22	MMU-MIR-467d	ID_B1	15	1
MMU-MIR-1937b-5	RLTR6C_Mm	135	22	MMU-MIR-467e	ID_B1	19	4
MMU-MIR-1937c	MMERGLN_I / RLTR1D2	14	5	MMU-MIR-467g	???	24	12
MMU-MIR-1970	ORR1B2	33	16	MMU-MIR-467h	???	12	8
MMU-MIR-297a-2	???	18	10	MMU-MIR-669a-1	EnSpm-4_HM	19	3
MMU-MIR-297a-6	???	28	10	MMU-MIR-669a-10	EnSpm-4_HM	19	4
MMU-MIR-344e	???	10	1	MMU-MIR-669a-11	EnSpm-4_HM	19	4
MMU-MIR-3470a	B1_Mur3	56	16	MMU-MIR-669a-12	EnSpm-4_HM	19	4
MMU-MIR-3470b	B1_Mus1	174	19	MMU-MIR-669a-2	EnSpm-4_HM	17	2
MMU-MIR-3471-2	MTA_Mm_LTR	202	21	MMU-MIR-669a-3	EnSpm-4_HM	17	2
MMU-MIR-466a	Helitron-2_HM	16	1	MMU-MIR-669a-4	EnSpm-4_HM	19	4
MMU-MIR-466b-1	Kolobok-2_XT	14	1	MMU-MIR-669a-5	EnSpm-4_HM	19	4
MMU-MIR-466b-2	Kolobok-2_XT	14	1	MMU-MIR-669a-6	EnSpm-4_HM	19	4
MMU-MIR-466b-3	Kolobok-2_XT	16	1	MMU-MIR-669a-7	EnSpm-4_HM	19	4
MMU-MIR-466b-4	Kolobok-2_XT	14	1	MMU-MIR-669a-8	EnSpm-4_HM	19	4
MMU-MIR-466b-5	Kolobok-2_XT	14	1	MMU-MIR-669a-9	EnSpm-4_HM	19	4
MMU-MIR-466b-6	Kolobok-2_XT	14	1	MMU-MIR-669f	EnSpm-4_HM	16	3
MMU-MIR-466b-7	Kolobok-2_XT	14	1	MMU-MIR-669g	???	13	1
MMU-MIR-466b-8	Kolobok-2_XT	14	1	MMU-MIR-669j	???	13	1
MMU-MIR-466c-1	EnSpm-4_HM	14	1	MMU-MIR-669k	ID_B1 / EnSpm-2_HM	14	1
MMU-MIR-466c-2	???	14	1	MMU-MIR-669m-1	Chapaev-8_HM	16	5
MMU-MIR-466e	Kolobok-2_XT	14	1	MMU-MIR-669o	EnSpm-4_HM	18	3
MMU-MIR-466f-1	MuDr-3_HM	11	6	MMU-MIR-669p-1	EnSpm-4_HM	18	3
MMU-MIR-466f-3	EnSpm-1_HM	10	5	MMU-MIR-669p-2	EnSpm-4_HM	18	3
MMU-MIR-466f-4	???	14	9	MMU-MIR-680-1	ERVB4_1B-LTR_MM	194	22
MMU-MIR-466j	CR1-18_HM	20	11	MMU-MIR-680-2	ERVB4_1B-LTR_MM	167	21
MMU-MIR-466k	???	31	17	MMU-MIR-680-3	ERVB4_1B-LTR_MM	45	14
MMU-MIR-466o	Kolobok-2_XT	18	3	MMU-MIR-682	???	14	8
MMU-MIR-466p	EnSpm-4_HM	20	2	MMU-MIR-684-1	???	20	12
MMU-MIR-467a-1	ID_B1	15	1	MMU-MIR-684-2	???	17	10
MMU-MIR-467a-10	ID_B1 / CR1-79_HM	21	5	MMU-MIR-692-1	RTE-1_AG	18	11
MMU-MIR-467a-2	ID_B1 / CR1-79_HM	22	6	MMU-MIR-692-2	RTE-1_AG	18	10
MMU-MIR-467a-3	ID_B1 / CR1-79_HM	21	5	MMU-MIR-703	???	18	10
MMU-MIR-467a-4	ID_B1	22	6	MMU-MIR-713	???	10	1

Figure 4. Mouse pre-miRNAs that have at least 10 similar hits in the genome. Pre-miRNAs that have "???" in 'TE name' column are not similar to known TEs listed in Repbase. The columns 'Hits' and 'Chrom' correspond respectively to the number of similar hits and to the number of chromosomes where appear these hits. There are 22 genomic chromosomes in UCSC Genome Browser [27].

One important remark is that very few pre-miRNAs have several similar hits; only 83 (85) among 1048 human (mouse) pre-miRNAs have more than 10 similar hits. This observation confirms the fact that a pre-miRNA is normally unique or with very few and close similar hits.

King Jordan *et al.* have previously discussed the origin of pre-miRNAs and the possibility that they come from the evolution of MITEs [18,17]. Their hypothesis is supported by the similarity between their secondary structures and by the similarity between their targeting mechanism. Moreover, Smalheiser *et al.* shown that some mammal pre-miRNAs have a small fragment of L2 transposable element in their sequence [19]. If the hypothesis of pre-miRNAs with a TE-derived origin seems possible, some studies stipulate that pre-miRNAs derive from genomic loci distinct from any other recognized elements [14,31] and Yan *et al.* think for instance that mir4441 and mir4446 are misannotated as pre-miRNAs but are pre-siRNAs [32].

Our automatic method confirmed the previously result obtained manually by Jordan *et al.* which shown that 6 human pre-miRNAs 'HSA-MIR-548' are TE-derived [18]. Thanks to our tool miRNAcheck, we identified 64 new TE-derived human pre-miRNAs.

We planned to add somes features to the next version of miRNAcheck. One of them should be to choose the tools for identifying TE-derived pre-miRNAs. For example, RepeatMasker and Censor do not give always the same result and it is possible that Censor does not recognize a TE sequence in few cases while RepeatMasker

can do it. For the identification of the pre-miRNAs, we can also use the genome annotations, they could avoid to search for the similar hits of the pre-miRNA candidate obtained by Censor when the candidate corresponds already to known pre-miRNA or a TE.

4 Conclusion

In this paper, we present an automatic method called miRNAcheck for identifying TE-derived pre-miRNAs. TE-derived pre-miRNAs are pre-miRNAs that are derived from transposable elements (TEs).

Our method is based on the hypothesis that a pre-miRNA that has several occurrences widespread in the genome has a high probability to be TE-derived. The first step of miRNAcheck is to calculate the number of occurrences of the pre-miRNA candidate, the number of chromosomes where appear the different occurrences and the distance between the occurrences. The second step is then to calculate a consensus sequence to the ten occurrence sequences the more similar to the pre-miRNA sequence. Finally, the last step consists to check if the consensus sequence corresponds to a TE in RepBase database.

We tested our method on human and mouse pre-miRNAs of miRBase. There are a total of 1048 human and 672 mouse pre-miRNAs, and only 83 human and 84 mouse pre-miRNAs have more than 10 occurrences (with high similarity). Almost all these 83 human and 84 mouse pre-miRNAs are identified by miRNAcheck as TE-derived, i.e. corresponding to TEs in RepBase.

Thanks to miRNAcheck, one could check very quickly if a pre-miRNA candidate is a TE-derived pre-miRNA. It requires between 30 seconds to 1 minute to treat a pre-miRNA sequence (depending on the number of occurrences in UCSC and on the access to RepBase).

miRNAcheck is available at the Web site: http://EvryRNA.ibisc.univ-evry.fr/

Acknowledgements

This work was funded by the Council of Essonne Region (Pôle System@tic, OpenGPU project).

References

[1] P. Kapranov, A.T. Willingham, and T.R. Gingeras, Genome-wide transcription and the implications for genomic organization. *Nat Rev Genet*, 8:413-23, 2007.

[2] International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome. *Nature*, 409:860-921, 2001.

[3] M.G. Kidwell, and D.R. Lisch, Perspective: transposable elements and host genome evolution. *Trends Ecol. Evol*, 15:95-99, 2001.

[4] N.L. Craig, R. Gragie, M. Gellert and A.M. Lambowitz, Mobile DNA II Second Edition. *ASM Press*, 2002.

[5] T. Wicker, F. Sabot, A. Hua-Van, J.L. Bennetzen, P. Capy, B. Chalhoub, A. Flavell, P. Leroy, M. Morgante, O. Panaud, E. Paux, P. SanMiguel, and A.H. Schulman, A unified classification system for eukaryotic transposable elements. *Nat Rev Genet*, 8:973-82, 2007.

[6] S.R. Wessler, T.E. Bureau, and S.E. White, LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. *Genet. Dev.*, 5:814-821, 1995.

[7] C. Feschotte, and C. Mouches, Evidence that a family of miniature inverted-repeat transposable elements (MITEs) from the *Arabidopsis thaliana* genome has arisen from a pogo-like DNA transposon. *Mol. Biol. Evol*, 17:4051-730-737, 2000.

[8] H. Kawagoe-Takaki, N. Nameki, M. Kajikawa, and N. Okada, Probing the secondary structure of salmon SmaI SINE RNA. *Gene*, 365:67-73, 2006.

[9] J.D. Suntera, S.P. Patela, R.A. Skiltona, N. Githakaa, D.P. Knowlesb, G.A. Scolesb, V. Nened, E de Villiers, and R.P. Bishopa, A novel SINE family occurs frequently in both genomic DNA and transcribed sequences in ixodid ticks of the arthropod sub-phylum Chelicerata. *Genet. Dev.*, 415:13-22, 2008.

[10] Y. Chen, F. Zhou, G. Li, and Y. Xu, A recently active miniature inverted-repeat transposable element, Chunjie, inserted into an operon without disturbing the operon structure in Geobacter uraniireducens Rf4. *Genetics*, 179:2291-7, 2008.

[11] H. Kuang, C. Padmanabhan, F. Li, A. Kamei, P.B. Bhaskar, S. Ouyang Jiang, C. Robin Buell, and B. Baker, Identification of miniature inverted-repeat transposable elements (MITEs) and biogenesis of their siRNAs in the Solanaceae: New functional implications for MITEs. *Genome Res.*, 19:42-56, 2008.

[12] G.J. Hannon, and J.J. Rossi, Unlocking the potential of the human genome with RNA interference. *Nature*, 431:371-378, 2004.

[13] R. Rana, Illuminating the silence: understanding the structure and function of small RNAs. *Molecular Cell Biology*, 8:23-36, 2007.

[14] D. Bartel, MicroRNAs: genomics, biogenesis, mechanism and function. *Cell*, 116:281-197, 2004.

[15] L. He, and G. Hannon, microRNAs: small RNAs with a big role in gene regulation. *Nat. Rev. Genet*, 5:522-531, 2004.

[16] Y. Lee, M. Kim, J. Han, K. Yeom, K.S. Lee , S. Baek and V. Kim, microRNA genes are transcribed by RNA polymerase II. *EMBO J.*, 23:4051-4060, 2004.

[17] J. Piriyapongsa and I.K. Jordan, Dual coding of siRNAs and miRNAs by plant transposable element. *RNA*, 14:814-821, 2008.

[18] J. Piriyapongsa, and I.K. Jordan, A Family of Human MicroRNA Genes from Miniature Inverted-Repeat Transposable Elements. *PLoS ONE*, 2:e203, 2007.

[19] N.R. Smalheiser and V.I. Torvik, Mammalian microRNAs derived from genomic repeats. *Trends Genet.*, 21:322-326, 2005.

[20] S. Griffiths-Jones, H. Saini, S. van Dongen, and A. Enright, miRBase: tools for microRNA genomics. *Nucleic Acids Res.*, 36:D154-D158, 2008.

[21] J. Piriyapongsa, L. Marino-Ramirez and I.K. Jordan, Origin and Evolution of Human microRNAs From Transposable Elements. *Genetics*, 176:1323-1337, 2007.

[22] E. Berezikov, N. Robine, A. Samsonova, J.O. Westholm, A. Naqvi, J-H. Hung, K. Okamura, Q. Dai, D. Bortolamiol-Becet, R. Martin, Y. Zhao, P.D. Zamore, G.J. Hannon, M.A. Marra, Z. Weng, N. Perrimon and E.C. Lai, Deep annotation of Drosophila melanogaster microRNAs yields insights into their processing, modification, and emergence. *Genome Res.*, 21:203-215, 2011.

[23] P. Landgraf *et al,*, A Mammalian microRNA Expression Atlas Based on Small RNA Library Sequencing. *Cell*, 129:1401-1414, 2007.

[24] A. Sewer, N. Paul, P. Landgraf, A. Aravin, S. Pfeffer, M.J. Brownstein, T. Tuschl, E. van Nimwegen and M. Zavolan, Identification of clustered microRNAs using an ab initio prediction method. *BMC Bioinformatics*, 6:267, 2005.

[25] J. Jurka, P. Klonowski, V. Dagman and P. Pelton, CENSOR - a program for identification and elimination of repetitive elements from DNA sequences. *Comput Chem.*, 20:119-21, 1996.

[26] T. Flutre, E. Duprat, C. Feuillet and H. Quesneville, Considering Transposable Element Diversification in De Novo Annotation Approaches. *PLoS ONE* 6:e16526, 2011.

[27] W.J. Kent, BLAT – The BLAST-Like Alignment Tool. *Genome Research*, 4:656-664, 2002.

[28] P.A. Fujita *et al,*, The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.*, 39:D876-82, 2011.

[29] MA Larkin and G Blackshields and NP Brown and R. Chenna and PA McGettigan and H McWilliam and F Valentin and IM Wallace and A Wilm and R Lopez and JD Thompson and TJ Gibson and DG Higgins., Clustal W and Clustal X version 2.0. *Bioinformatics*, 23:2947-8, 2007.

[30] J. Jurka, V.V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany and J. Walichiewicz, Repbase Update, a database of eukaryotic repetitive elements. *Cytogentic and Genome Research*, 110:462-467, 2005.

[31] BC. Meyers *et al,* Criteria for Annotation of Plant MicroRNAs. *The Plant Cell*, 20:3186-3190, 2008.

[32] Y. Yan, Y. Zhang, K. Yang, Z. Sun, Y. Fu, X. Chen, and R. Fang, Small RNAs from MITE-derived stem-loop precursors regulate abscisic acid signaling and abiotic stress responses in rice. *The Plant Journal*, 65:820-828, 2011.

Conserved Disorder

Its Role in Human Disease and Common Variations

Magali MICHAUT[1*], TaeHyung KIM[1*], Sangjo HAN[1], Jeremy BELLAY[2] and Philip M. KIM[1]

1 The Donnelly Centre, University of Toronto, 160 College Street, M5S 3E1, Toronto, Ontario, Canada
{magali.michaut, taehyung.kim, sangjo.han, pm.kim}@utoronto.ca
2 Department of Computer Science and Engineering, University of Minnesota, MN 55455, Minneapolis, USA
bellay@cs.umn.edu
*These authors contributed equally to this work.

Keywords Disorder, Intrinsically Disordered Protein, Phosphorylation, Cancer mutation.

1 Introduction

Intrinsically disordered regions are widespread, especially in proteomes of higher eukaryotes. Recently, protein disorder has been associated with a wide variety of cellular processes and has been implicated in several human diseases [1]. Despite its apparent functional importance, the sheer range of different roles played by protein disorder often makes its exact contribution difficult to interpret.

Recently we introduced a novel comparative genomics analysis to uncover that protein disorder can be split into biologically and biophysically distinct phenomena in the yeast *S. cerevisiae*. Even if the amino acid sequence evolves, the fact to be disordered for a residue can sometimes be conserved across species. We defined disorder conservation using disorder predictions (*DISOPRED2*) in ~23 species of the yeast clade. This conserved disorder is further split into constrained disorder where the amino acid sequence itself is conserved and flexible disorder where the amino acid sequence is not conserved [2].

In this work, we explore the characteristics of conserved disorder in human (flexible and constrained) and in particular its relationship to cancer and disease mutations. We find that conserved disorder plays a specific role in biological processes and that phosphorylation sites (or phosphosites) are enriched in constrained disorder but surrounded by regions of flexible disorder. In addition mutations show conserved disorder specificity leading to possible improvement in disease and cancer mutation prediction.

2 Results

2.1 Conserved Disorder Plays a Specific Role

We previously found in yeast that flexible disorder bears many of the characteristics commonly attributed to disorder and is associated with signaling pathways and multi-functionality whereas constrained disorder has markedly different functional attributes and is involved in RNA binding and protein chaperones [2]. Investigating protein disorder in higher eukaryotes, we find that constrained disorder is specifically associated to development and related biological processes. Conversely, flexible disorder is enriched in repair mechanisms and immune system. In addition, proteins enriched in conserved disorder are more multi-functional than the average on the whole proteome and expressed in fewer tissues. Together those results suggest that conserved disorder plays specific role in biological processes.

2.2 Phosphosites are Conserved but Positioned in Flexible Regions

Phosphosites often appear in disordered regions of proteins [3]. In a detailed analysis at the residue level, we find that disorder conservation is strongly correlated with the placement of phosphosites (Fig. 1A). In particular, we find that the relative density of phosphosites increases dramatically for residues with higher disorder conservation. In fact phosphosites are specifically enriched in constrained disorder, which is coherent with the sequence conservation of those specific residues. Nevertheless, when considering continuous regions of conserved disorder (a sequence of at least 10 continuous positions of conserved disorder with maximum 3 gaps), we find that regions harboring at least a phosphosite have a higher percentage of flexible over constrained disorder. This suggests that the residues being phosphorylated are

rather conserved and thus found in constrained disorder positions but surrounded by flexible disorder regions. These results highlight the functional importance of conserved disorder in signaling.

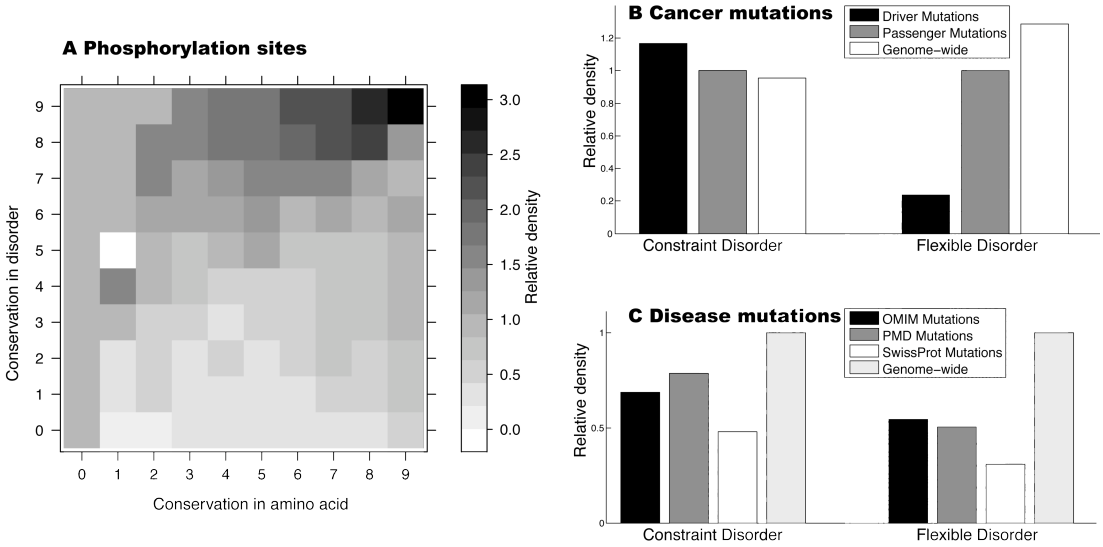


Figure 1. A) The heatmap illustrates the relative density of phosphosites in positions with various levels of disorder and amino acid conservation. The relative density of conserved disorder is represented for B) two classes of cancer mutations. C) three classes of disease mutations and compared to genome wide. All relative density values (passenger, driver, OMIM, PMD, SwissProt) are significantly different from genome-wide for both constrained and flexible disorder (Chi-square p-value < 0.05).

2.3 Mutations Show Conserved Disorder Specificity

Since various diseases are associated to dysfunction in signaling, we investigate next the relation of conserved disorder to mutations. One of the challenges in cancer genomics is to distinguish between mutations that confer a selective growth advantage (driver) and mutations happening coincidentally (passenger) [4]. We find that driver mutations are enriched in constrained disorder and depleted in flexible disorder (Fig. 1B), which could help distinguish driver/passenger mutations. In addition, disease mutations are depleted in conserved disorder (flexible and constrained) as compared to non disease mutations (Fig. 1C).

3 Conclusions

We show here that conserved disorder has a specific and important role in biological processes in human and that phosphosites are enriched in constrained disorder but surrounded by regions of flexible disorder. In addition, disease/cancer mutations show conserved disorder specificity. We are now investigating to which extent this information could improve the prediction of those mutations and help distinguish driver/passenger and disease/non disease mutations.

References

[1] H.J. Dyson and P.E. Wright, Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol*, 6(3):197-208, 2005.

[2] J. Bellay, S. Han, M. Michaut, T. Kim, M. Costanzo, B.J. Andrews, C. Boone, G.D. Bader *et al*, Bringing order to protein disorder through comparative genomics and genetic interactions. *Genome Biol*, 12(2):R14, 2011.

[3] L.M. Iakoucheva, P. Radivojac, C.J. Brown, T.R. O'Connor, J.G. Sikes, Z. Obradovic and A.K. Dunker, The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res*, 32(3):1037-1049, 2004.

[4] I. Bozic, T. Antal, H. Ohtsuki, H. Carter, D. Kim, S. Chen, R. Karchin, K.W. Kinzler *et al*, Accumulation of driver and passenger mutations during tumor progression. *Proc Natl Acad Sci U S A*, 107(43):18545-18550, 2010.

Sets of Symmetries by Base Substitutions in the Genetic Code

Jean-Luc JESTIN[1] and Christophe SOULE[2]

[1] UNITE DE VIROLOGIE STRUCTURALE, URA 3015 CNRS, INSTITUT PASTEUR

25 rue du Dr. Roux 75724 Paris, Cedex 15, France

jjestin@pasteur.fr

[2] INSTITUT DES HAUTES ETUDES SCIENTIFIQUES

35 Route de Chartres, 91440 Bures-sur-Yvette, France

soule@ihes.fr

Keywords coding, discrete symmetries, molecular evolution, mutations, tRNA-aminoacylation.

1 Introduction

The genetic code is quasi-universal among living organisms.

The question "Why is the genetic code the way it is?" remains open. In particular, models considering the genetic code as a product of biochemical evolution describe successfully a few properties of the genetic code [1].

2 Results

Here, we describe the set of symmetries by base substitutions for two essential properties of the genetic code: codon degeneracy [2,3] and 2' or 3' tRNA-aminoacylation [3].

The genetic code's sets of 64 codons can be dissected into two groups of 32 codons depending on whether or not the third codon base is required to define unambiguously the amino acid. Rumer identified the symmetry applied to the three codon bases and substituting G and T as well as A and C, which exchanges both groups [4]. The symmetry applied to the first codon base and substituting G and C as well as A and T leaves each group unchanged [2]. The proof that no further symmetries of degeneracy in the genetic code exist, apart from the combination of these symmetries, will be given [3].

Concerning tRNA-aminoacylation by aminoacyl-tRNA-synthetases, two classes can be defined depending on whether aminoacylation occurs on the 2' or on the 3' hydroxyl group of the last tRNA nucleotide. The two symmetries substituting G and T as well as A and C for the first codon base, substituting A and G as well as C and T for the second codon base and exchanging purines and pyrimidines for the third codon base (i.e. G and T as well as A and C or G and C as well as A and T) exchange the two classes [3]. No further symmetry exchanges both classes.

The observations are discussed within the context of the known minor changes found among different genetic codes and using available biochemical data on aminoacyl-tRNA synthetases.

References

[1] J. L. Jestin and A. Kempf, Chain termination codons and polymerase-induced frameshift mutations. *FEBS Lett.*, 419:153-156, 1997 (and references therein).

[2] J.L. Jestin, Degeneracy in the genetic code and its symmetries by base substitutions. *Comp. Rend. Biol.*, 329:168-171, 2006.

[3] J.L. Jestin and C. Soulé, Symmetries by base substitutions in the genetic code predict 2' or 3' aminoacylation of tRNAs. *J. Theor. Biol.*, 247:391-394, 2007.

[4] Y.B. Rumer, About the codon's systematization in the genetic code. *Proc. Acad. Sci. USSR*, 167, 1393-1394, 1966.

A pipeLine Dedicated to Oligonucleotides design (ALDO)

A Workflow for Molecular Diagnostics Assay Design

Iandry RABEARIVELO[1,2] and François PAILLIER[2]

[1] Université de Rouen, Master 2.1 Bioinformatique, 76130, Mont Saint-Aignan, France
iandry.rabearivelo@etu.univ-rouen.fr
[2] BioMérieux, 5 rue des Berges, 38000, Grenoble, France
{iandry.rabearivelo, francois.paillier}@biomerieux.com

Abstract *Primers and probes design is an important step of molecular diagnostics (MDx) reagents conception. Bioinformatics is very helpful for the design of these oligonucleotides, but also to validate candidates in silico before in vitro tests, an expensive and time-consuming task. ALDO is a bioinformatics pipeline dedicated to the conception of MDx qPCR primers and probes. ALDO was applied to Influenza primers and probes design and validation.*

Keywords primers and probes, assay design, *in silico* validation, qPCR simulation.

1 Introduction

Primers and probes are used in MDx reagents, with different technologies such as qPCR, to detect microorganisms by targeting taxon-specific genomic sequences. Some existing bioinformatics programs were developed for identifying taxon-specific regions (Insignia [1]) and designing candidate primers and probes. Others were defined to test these oligonucleotides with *in silico* amplification reaction simulations (isPCR [2]), but without any assay scoring function quantifying their *in silico* performance.

2 Objective

ALDO is a workflow dedicated to the conception of primers and probes, integrating both the steps of assay design and validation. Moreover, validated assays shall be ranked according to a performance score.

3 Principle

Three main parts compose ALDO: *(i)* integration of public and internal data, *(ii)* design of primers and probes and *(iii)* their *in silico* validation. ALDO relies both on internally-developed programs as well as public programs such as Blast [3], Primer3 [4], EMBOSS utilities [5] and NCBI eUtils programs [6].

3.1 Data Integration

The first step of the pipeline consists in collecting all targeted sequences of the studied organism or taxon. To this aim, a project sequence databank is created by the PePI builder program. Entries are selected from internal sequence collections (obtained from targeted sequencing campaigns) and from public sequence databases (obtained via Blackcell). Blackcell is a surveillance tool watching out for new sequences on public databases. It is based on NCBI eUtils programs and contributes to update ALDO's sequence collections. Using this sequence collection, a multiple sequence alignment (MSA) is constructed.

Polymorphism Diversity Estimator (PoDE), a program implemented in ALDO, was developed with the aim of estimating the diversity of a target sequence. It checks if the total number of sequences in the MSA is representative of the natural diversity, and if not, estimates the sequencing effort still needed. The method uses the chao2 non-parametric estimator of asymptotic SNPs richness [7].

3.2 Design of Oligonucleotides

The second part of the pipeline actually designs qPCR assays (combination of PCR primers and TaqMan® probe) using the reference MSA. The SLv8 program encapsulates Primer3 for this design step. Candidates are selected in accordance with more than 20 qPCR-specific design rules. Each qPCR assay is then a candidate assay whose *in silico* performance is computed.

3.3 Candidate Assay Validation

Third and final part is assay validation. Candidate assays are validated by simulating a qPCR reaction against each individual sequence of a particular databank (such as Genbank bacterial division). To do so, eNv3 model simulates the amplification/detection qPCR reaction by computing thermodynamic affinities of each individual oligonucleotide both for hybridization and primer extension step. Finally, all assays are ranked according to a global performance score integrating both a positive term (completeness of targeted species detection) and a negative term (unwanted cross-detection).

Mismatches between oligonucleotides and target sites impact the qPCR reaction efficiency in a nature and oligonucleotide position-dependent way. The Assay Mismatch Table (AMT) permits to locate positions of mismatches or deletions between oligonucleotides and target sequences. This table is performed to double-check both oligonucleotides' sensitivity and cross-validate eNv3 results. Moreover, the AMT allow to rank hybridization sites according to their frequency in sequence databases and shows the most suitable oligonucleotide sequences to target in order to insure the highest performance.

4 Results and Conclusion

ALDO was applied to design and validation of primers and probes targeting *Influenza A* segment 8. Briefly, we extracted 6442 sequences for *Influenza* segment 8 (average sequence size : 844nt, 43% GC) from IPDR Flu database. MSA was built using MUSCLE v3.6, and provided an alignment of 977nt and average pairwise sequence identity of 88%. PoDE estimated an overall of 284 polymorphic positions on this MSA but only 240 were observed. PoDE estimated we need 4831 additional sequences to be able to observe the unseen polymorphism positions (at the sampling risk level 5%). Assuming this non-completeness risk, assay design step can be performed but with significant risks to have oligonucleotides targeting a polymorphic site. Thus, a regular update of the MSA with new sequences is required to check validity of designed assays. Based on this MSA, SLv8 designed 196 candidates qPCR assays for an average amplicon length of 143nt.

To test AMT and eNv3 programs, an existing qPCR assay was tested against a collection of 2409 sequences of *Influenza* segment 8. Overall detection rate for this triplet was estimated to 87% according to AMT. Cross-reactivity risk was assessed using eNv3: only 14 entries out of 10'572'835 sequences (custom databank composed of bacteria, fungi, viruses and homo sapiens sequences) were reported as potentially cross-reactive at least for one of the oligonucleotides. Manual analysis revealed no major cross-reactivity risk because either these entries are in fact *Influenza* entries badly annotated or non-*Influenza* entries but without both primers hybridization sites being presents on the same sequence.

Each part of the workflow was validated individually but the validation of the whole process is on-going.

Acknowledgements

This work was supported both by bioMérieux and the University of Rouen. We thank Dr. Fritz Schwarzmann and Dr. Andrew Derome for reviewing this paper.

References

[1] A. M. Phillippy, K. Ayanbule, N. J. Edwards, and S. L. Salzberg, Insignia: a DNA signature search web server for diagnostic assay development, *Nucleic Acids Res.*, 37: W229-W234, 2009 July.

[2] J. Kent, UCSC In silico PCR, http://genome.ucsc.edu/cgi-bin/hgPcr?command=start

[3] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *Jmol. Biol.* 215(3): 403–410, 1990.

[4] S. Rozen, H. J. Skaletsky, Primer3 on the WWW for general users and for biologist programmers. In: S. Krawetz, S. Misener, *Bioinformatics Methods and Protocols: Methods* in *Molecular Biology*. Humana Press, Totowa, NJ, pp. 365-386, 2000.

[5] P. Rice, I. Longden, and A. Bleasby, EMBOSS: The European Molecular Biology Open Software Suite, *Trends in Genetics* 16(6): 276-277, 2000.

[6] E. Sayers, D. Wheeler, Building Customized Data Pipelines Using the Entrez Programming Utilities (eUtils), NCBI eUtils http://eutils.ncbi.nlm.nih.gov/

[7] A. Chao, R. K. Colwell, C-W. Lin, N. J. Gotelli, Sufficient sampling for asymptotic minimum species richness estimators, *Ecology*, 90(4): 1125-1133, 2009.

Bioinformatics Tools to Decrypt Pyoverdine Biosynthesis in *Pseudomonas sp.*

Aurélien VANVLASSENBROECK[1], Valérie LECLERE[1], Maude PUPIN[2], Bernard WATHELET[3] and Philippe JACQUES[1]

[1] ProBioGEM, UPRES EA 1026, Polytech'Lille/IUT A, Av P Langevin, Univ Lille Nord de France, Sciences et Technologies, 59655 Villeneuve d'Ascq cedex,
vanvlas008@hotmail.fr, {valerie.leclere, philippe.jacques}@univ-lille1.f

[2] LIFL, UMR8020 CNRS, INRIA, Bat M3, Univ Lille Nord de France, Sciences et Technologies, 59655 Villeneuve d'Ascq cedex, France
maude.pupin@lifl.fr

[3] Unity of Industrial Biological Chemistry, Pass des déportés, Gembloux Agricultural Univ, 5030 Gembloux, Belgium
bwathelet@ulg.ac.be

Keywords Protein annotation, sequence comparison, evolution, non-ribosomal peptide.

1 Introduction

Pyoverdines are chromopeptides produced by fluorescent *Pseudomonas sp.* during growth under iron-limiting conditions. More than fifty different structures of these siderophores have been elucidated so far. The biosynthesis follows a non ribosomal mechanism carried out by Non Ribosomal Peptide Synthetases (NRPS).These synthetases are multifunctional enzymes organized in sets of catalytic domains which constitute modules containing the information needed to complete an elongation step in peptide biosynthesis. The main catalytic functions are responsible for the activation of an amino acid residue (adenylation domain, called A-domain), the transfer of the corresponding adenylate to the enzyme-bound 4'-phosphopantetheinyl cofactor (peptidyl carrier protein domain) and the peptide bond formation (condensation domain). Additional domains can lead to modification of the substrates if required in the peptide synthesis. A thioesterase (Te) domain is usually present in final position to ensure the cleavage of the thioester bond between the nascent peptide and the last PCP-domain and, in several cases, to cyclise the peptide. The specificity of the selection of the amino acid residue is mostly conditioned by the A- domain. The aim of this work is to study the specificity/permissivity of the adenylation domains involved in the biosynthesis of pyoverdines and to study their capacity to recognize several (at least 2) amino acids. Those molecules are an appropriate model for this study because of their diversity.

2 Methods

Bioinformatics analysis was first performed on the seventeen *Pseudomonas sp.* genomes available following the process described in figure 1. The research of genes coding for NRPS in the genomes was performed through two different approaches, first by key words research in the MBGD databank [1] and by querying the complete genome sequences with a known NRPS given to tBLASTn.

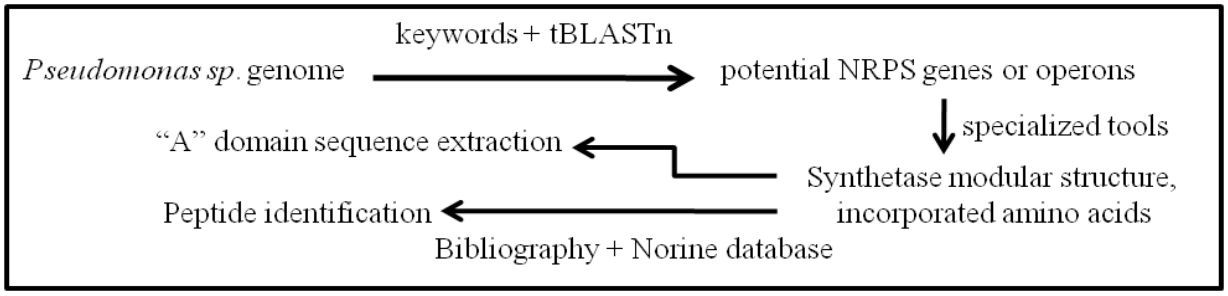


Figure 1. General strategy of genome analysis used in this work.

Potential NRPS extracted by this process are analyzed by prediction tools available on the web [2, 3, 4]. The first two tools predict the modular organization of the synthetases and the all three predict the amino acids incorporated by the A-domains. As pyoverdines are linear or partially cyclic peptides, the order of the A-domains coupled with the prediction of the selected amino acid allowed us to construct the potential product of the synthetases. To identify the predicted peptides, the structure search tools of Norine database [5, 6] and bibliography research were used. This analysis also allowed extracting the sequence of A-domains.

A high throughput technique was set up to detect less specific A-domains, using a feeding approach consisting in modifying the composition of the culture medium with different amino acid residues combined to whole cells MALDI-TOF mass spectrometry analysis. If the amino acid used in the feeding is incorporated by A-domain instead of the commonly accepted one, we detect the change of the peptide mass.

3 Results

Genes involved in the pyoverdine biosynthesis were for the first time identified for eight strains. For fifteen of the seventeen *Pseudomonas sp*. genomes studied a correlation between the NRPS sequences and the produced siderophores was established. Only the predicted sequence of the pyoverdine produced by one of these strains did not match with other known pyoverdine sequences. This sequence appeared to be a new form of pyoverdine. In the three strains of *Pseudomonas syringae* sequenced, the sequence predicted for the pyoverdine and the organization of the synthetase are strictly co-linear. Furthermore, seven cyclic lipopeptide synthetases recognizable by two Te-domains in tandem at the end of the synthetase and nine PKS/NRPS hybrid synthetases were identified in the genomes. In a second part, an alignment and a tree of the sequence of A-domains with known specificity was performed. We observed that few A-domain sequences are similar despite a different substrate activated. This observation can support the idea that some A-domains have a low specificity. The specificity/permissivity of A-domains was studied by amino acids feeding experiments. Amino acid substitutions were observed by this method, especially the substitution of a threonine by a serine occurred in the pyoverdines produced by several strains.

4 Conclusion

Using bioinformatics tools, we have predicted the production of 41 non-ribosomal peptides by seventeen *Pseudomonas sp*. species. Diversity of the peptides produced depended on species, pointing out a specific evolution. To further study one aspect of this evolution, we are now investigating the permissivity of the A-domains showing that they can select one or two amino acids, depending on their intracellular pool.

Acknowledgements

This work was supported by PPF Bioinformatique of Lille 1 University and FEDER.

References

[1] I. Uchiyama, T. Higuchi and M. Kawai, MBGD update 2010: toward a comprehensive resource for exploring microbial genome diversity. *Nucl. Acids Res.*, 38:D361-D365, 2010. (http://mbgd.genome.ad.jp/)

[2] M. Z. Ansari, G. Yadav, R.S. Gokhale and D. Mohanty, NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthetase. *Nucl. Acids Res.*, 32:405-413, 2004. (http://www.nii.res.in)

[3] B. O. Bachmann and J. Ravel, In Silico Prediction of Microbial Secondary Metabolic Pathways from DNA Sequence Data. *Meth. Enzymol.*, 458:181-217, 2009. (http://nrps.igs.umaryland.edu)

[4] C. Rausch, T. Weber, O. Kohlbacher, W. Wohlleben and D. H. Huson, Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using Transductive Support Vector Machines (TSVM). *Nucl. Acids Res.*, 33:5799-5808, 2005.(http://www-ab.informatik.uni-tuebingen.de)

[5] S. Caboche, M. Pupin, V. Leclère, A. Fontaine, P. Jacques and G. Kucherov, NORINE: a database of nonribosomal peptides. *Nucl. Acids Res.*, 36:D326-D331, 2008. (http://bioinfo.lifl.fr/norine/)

[6] S. Caboche, M. Pupin, V. Leclère, P. Jacques and G. Kucherov, Structural pattern matching of nonribosomal peptides. *BMC Structural Biology*, 9:15, 2009.

Predicting Protein Flexibility through the Prediction of Local Structures

Catherine ETCHEBEST[1], Aurélie BORNOT[1] and Alexandre G. de BREVERN[1]

[1] DSIMB, UMR-S 665 INSERM, Université Paris Diderot – Paris 7, Institut National de la Transfusion Sanguine, 6 rue Alexandre Cabanel, 75739, Paris, Cedex 15, France

{catherine.etchebest, aurelie.bornot, alexandre.debrevern}@univ-paris-diderot.fr

Abstract Protein *structures are valuable tools for understanding protein function. However, protein dynamics is also considered a key element in protein function. For fully understanding protein function at the molecular level now requires accounting for flexibility.*
Protein structure can be described by a limited set of recurring local structures. We established a library composed of 120 overlapping long structural prototypes (LSPs) representing fragments of 11 residues in length and covering all known local protein structures. A novel prediction method that proposes structural candidates in terms of LSPs along a given sequence was proposed. In this study, we utilise this methodology to predict protein flexibility. We first examine flexibility according two different descriptors, the B-factor and root mean square fluctuations from molecular dynamics simulations. We define three flexibility classes and propose a method based on the LSP prediction method for predicting flexibility along the sequence. This method competes rather efficiently with the most recent, cutting-edge methods based on true flexibility data learning with sophisticated algorithms.

Keywords Bioinformatics, protein structure, flexibility, protein dynamics, structural alphabet.

1 Introduction

Knowledge on protein 3D structures is essential for better understanding protein functions. In the case of enzymes, determination of 3D structures has helped elucidate why residues far apart in the sequence are involved in a given catalytic reaction. We have described global protein structures using a limited set of recurring local structures [1]. We have defined a library of 120 overlapping representative fragments of 11 amino acids in length named long structural prototypes (LSP). They encompass all known local protein structures and ensure good quality 3D local approximation [2]. The length of representative fragments makes it possible to account for long-range interactions and correlations. Using the sequence-structure relationships deduced from this library, prediction methods in terms of LSPs have been elaborated [3]. The prediction method is based on evolutionary information coupled with an efficient learning method called support vector machines (SVM) [3]. We have examined protein flexibility of fragments in representative datasets using two different approaches, X-ray experiments and *in silico* simulations [4].

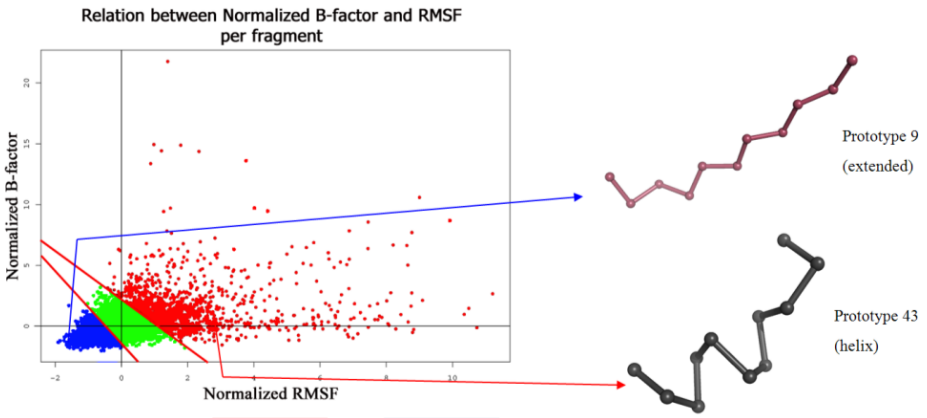
2 Material and Methods

A dataset of X-ray high-resolution (≤ 1.5 Å) globular protein structures was extracted from the Protein Data Bank (PDB). In this dataset, the proteins shared less than 10% sequence identity and differed by at least 10 Å Cα root mean square deviations (Cα RMSD). Selected protein structures were 70 to 200 residues long, composed of a single domain and were not involved in a protein complex, and did not have extensive number of contacts with ligands. A final dataset of 43 protein structures was obtained. We extracted normalized Cα B-factors from the PDB files of the protein structures dataset.

Predictions of flexibility were performed using the results of LSP prediction [3] treated with Support Vector Machines (SVMs). LSP prediction is based also on SVMs with the help of PSI-BLAST.

3 Results

We chose to define three flexibility classes from the most rigid to the most flexible using both two descriptors of protein dynamics, *i.e.*, normalized B-factor and normalized RMSF values. The 4 thresholds used to define the three classes were optimized to obtain the best prediction rate and equilibrated classes.



	Helix	Extended	Extended Edge	Connection	
Rigid	23.3 %	30.2 %	46.5 %	Ø	100 %
Intermediate	13.3 %	Ø	43.3 %	43.3 %	100 %
Flexible	4.3 %	Ø	14.9 %	80. 9 %	100 %

Figure 1. Distribution of the three classes of flexibility. Examples of flexible and rigid LSPs are shown.

Figure 1 shows the repartition of the three classes of flexibility with corresponding distributions. The rigid and intermediate flexibility classes were similarly populated with 40.4% and 36.7% of protein fragments, respectively, whereas only 22.9% were classified in the most flexible class. Interestingly all extended LSPs are in rigid class (*e.g.*, LSP 9) while no connection LSP is found rigid and helices are found in all classes (*e.g.*, LSP 43 in flexible class).

The prediction method led to an average, very well-balanced prediction rate of 49.4% for the three defined flexibility classes. Significant confusion with Class 2 (intermediate) is observed in the predicted states. Indeed, 86.5% of rigid protein fragments were predicted to be rigid or intermediate. Likewise, 94.2% of flexible fragments were predicted to belong to an intermediate or flexible class. In contrast, confusion between flexible and rigid classes was very low. Less than 13.5% of fragments observed in the rigid class were predicted to be flexible, whereas only 6.0% of fragments observed in the flexible class were predicted to be rigid. More importantly, this prediction rate was considerably higher than a random prediction rate. A random prediction rate would give 36.0%, with only 8.5 and 13.8% of rigid and flexible fragments correctly predicted.

Comparison with prediction methods based on two flexibility classes shows that our approach is similarly powerful when the three flexibilities are regrouped into two classes. A confidence index of the prediction has been also proposed.

Acknowledgements

This work was supported by the National Institute for Blood Transfusion (INTS), the French Institute for Health and Medical Research (INSERM), the University of Paris Diderot - Paris 7. AB benefited from a grant from the French Ministry of Research.

References

[1] B. Offmann, M. Tyagi and A.G. de Brevern, Local Protein Structures. *Current Bioinformatics*, 2:165-202, 2007.

[2] 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.

[3] A. Bornot, C. Etchebest and A.G. de Brevern, A new prediction strategy for long local protein structures using an original description. *Proteins*, 76:570-587, 2009.

[4] A. Bornot, C. Etchebest and A.G. de Brevern, Predicting Protein Flexibility through the Prediction of Local Structures. *Proteins*, 79:839-852, 2011.

Protein 3D Structure Comparison based on Sequence Alignment Approaches: Application of a Structural Alphabet

Agnel Praveen JOSEPH[1], Jean Christophe GELLY[1], N. SRINIVASAN[2] and Alexandre G. de BREVERN[1]

[1] INSERM, UMR-S 665, DSIMB, Université Paris Diderot - Paris 7, INTS, 6, rue Alexandre Cabanel, 75739 Paris cedex 15, France.

{agnel.praveen,jean-christophe.gelly,alexandre.debrevern}@univ-paris-diderot.fr

[2] MOLECULAR BIOPHYSICS UNIT, Indian Institute of Science, Bangalore 560012, India.

ns@mbu.iisc.ernet.in

Abstract *Protein Structure Comparison is a useful method for function characterization and evolutionary studies. We propose a method for three dimensional (3D) protein structure comparison based on similarities in local backbone conformation. A library of 16 frequently occurring penta-peptide backbone conformations, namely Protein Blocks, was used to transform 3D information as a sequence. This reduces the problem of structural comparison to a more classical sequence alignment. The use of an anchor based dynamic programming algorithm with specialized gap penalties resulted in a significant improvement over earlier studies based on simple global alignments. The alignment quality improved by about 82% and the efficiency in searching a structure databank for related folds was also enhanced by 6.2%.Comparison with other popular methods suggest that iPBA is one among the top two best approaches.*

Keywords. Protein Structure Comparison, Protein Blocks, Dynamic Programming.

1 Introduction

A large majority of structure alignment tools optimize regions of local structural similarities followed by global refinement. We had also developed an approach for structure comparison based on the use of a widely used library of local backbone conformations, a Structural Alphabet, *i.e.*, Protein Blocks (PBs) [1, 2]. PBs consists of a set of 16 penta-peptide backbone conformations described in terms of φ/ψ dihedral angles. A complete protein backbone can be approximated with an average (Root Mean Square Deviation) RMSD of 0.42 Å, using the prototypes from this library. As each of the PBs is represented by a letter (from *a-p*), then the 3D structure information can be converted to a one dimensional sequence. Hence two protein structures can be compared by the alignment of PB sequences [3]. In this study we try to improve the PB based structure comparison using an anchor-based alignment methodology and refined PB substitution matrices.

2 Methods

A database of structural alignments was used to generate a PB substitution matrix based on the preference for PB changes. The substitution data was normalized based on sequence and structural similarity to refine the scores of the matrix. The substitution scores guide the alignment of PB sequences based on dynamic programming. An anchor-based alignment methodology was designed, where the structurally similar stretches are first identified as local alignments and the intervening segments are then aligned with lower gap penalties (Figure 1A) [4]. The local alignments were obtained using a linear space dynamic programming algorithm [5]. The alignment quality was quantified using classical RMSD, GDT_TS [4] or similar scores.

3 Results

The use of anchor based dynamic programming algorithm with optimized gap penalties resulted in a significant improvement over the earlier approach (PBALIGN) (Figure 1A). The alignment was scored based on refined PB substitution matrices coupled with amino acid substitution weights.

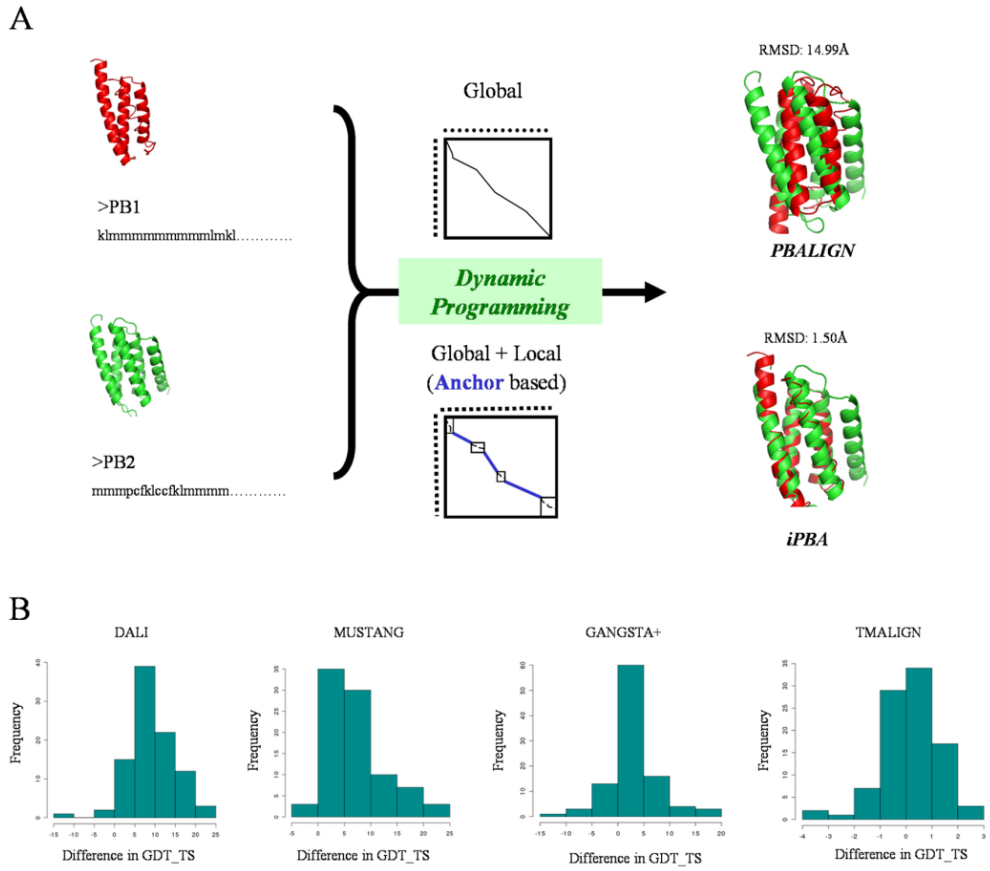


Figure 1. (A) Improvement in protein structure comparison using anchor based PB alignment (iPBA) (B) Comparison of iPBA with DALI, MUSTANG, GANGSTA+ and TMALIGN for alignment of 100 domain pairs.

With the new developments, about 82% of the alignments had better RMSD and the efficiency in finding homologues (from the same structural super-family), improved by 6.2%. The alignment quality (GDT_TS score) was better than DALI and MUSTANG in about 93.2% and 95.1% of the cases respectively [4]. Significant improvement was also achieved with respect to GANGSTA+ (81%), while comparable performance was obtained with TMALIGN (Figure 1B). A web server on this approach is also available: http://www.dsimb.inserm.fr/dsimb_tools/ipba [6].

Acknowledgements

This work was supported by grants from CEFIPRA (3903E), the French Ministry of Research, University of Paris Diderot – Paris 7, INTS, INSERM, France and Department of Biotechnology, India.

References

[1] 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.

[2] A.P. Joseph, *et al*., A short survey on protein blocks. *Biophys Rev*, 2:137-145, 2010.

[3] M. Tyagi, A.G. de Brevern, N. Srinivasan and B. Offmann, Protein structure mining using a structural alphabet. *Proteins*, 71:920-937, 2008.

[4] A.P. Joseph, N. Srinivasan and A.G. de Brevern, Improvement of protein structure comparison using a structural alphabet. *Biochimie (in press),* 2011.

[5] X. Huang. and W. Miller, A time-efficient linear-space local similarity algorithm. *Advances in Applied Mathematics*, 12: 337-357, 1991.

[6] J.-C.Gelly, A.P. Joseph, N. Srinivasan and A.G. de Brevern, iPBA : A tool for protein structure comparison using sequence alignment strategies. *Nucleic Acid Research (in press)*, 2011.

DOMIRE, a Web Server for Structural Domain Identification in Proteins

Franck SAMSON[1], Richard SHRAGER[2] , Chin-Hsien TAI[3], Vichetra SAM[2], Jean-Francois GIBRAT[1], Byungkook LEE[3] , Peter MUNSON[2] and Jean GARNIER[1,2]

[1] Mathématique, Informatique et Génome, INRA, Jouy-en-Josas, France
{franck.samson, jean-francois.gibrat, jean.garnier}@jouy.inra.fr
[2] Mathematical and Statistical Computing Laboratory, CIT, NIH, Bethesda, USA
{ shragerr, munson }@helix.nih.gov
[3] Laboratory of Molecular Biology, NCI, NIH, Bethesda, USA
{ vichetra.sam, taic, bklee }@mail.nih.gov

Abstract *DOMIRE (DOMain Identification from REcurrence) is a web server using VAST to define the domain boundaries in proteins from their 3 D structure recurrences with a list of structural neighbours in the Protein Data Bank.*

Keywords structural domains, protein structure evolution, protein modelling.

1 Introduction

Domains play an essential role in our understanding of protein evolution and function either because they appear as substructures of a protein or correspond to individual three-dimensional (3D) structures in their own right. The characteristic property of compactness of their structures have been extensively used to define the domains from their atomic coordinates, a comprehensive list of these methods is given in Tai et al[1]. Domains are currently forming the basis of the CATH (http://www.cathdb.info/) or SCOP (http://scop.mrc-lmb.cam.ac.uk) protein structures classifications.

Another definition of a domain is an exchangeable segment of amino acid sequence, that retains its 3D structure and its molecular function. The servers Prodom (http://prodom.prabi.fr) and Pfam (http://pfam.sanger.ac.uk/) apply this definition to identify the domains by comparing the protein amino acid sequences and their degree of conservation.

The two definitions, compactness and sequence conservation should converge notwithstanding some intrinsic limitations, there are considerably less determined protein structures than amino acid sequences and the sequence is less conserved than the structure and the function.

Using the VAST algorithm (Gibrat et al. [2], Madej et al. [3]) we observed that protein domains could be also assigned from the recurrence of small 3D common substructures found in proteins of the PDB. (Tai et al. [1]). Following this work we designed DOMIRE for DOMain Identification from REcurrence (http://genome.jouy.inra.fr/domire/).

2 Methods

In the VAST algorithm, proteins are represented by their secondary structure elements (SSEs), more specifically by the endpoints of vectors going through these SSEs. The basic task of VAST is to find the best 3D common substructures between a query and a target. A 3D common substructure is formally defined as a one-to-one correspondence between a subset of SSE vectors in the query and a subset of the SSE vectors in the target. This correspondence respects the type of SSE (i.e., helices are only paired with helices and strands with strands) and the topology. This ensemble is named a clique an example of which is given in fig 1. For further details about VAST see Methods and Appendix in Sam et al [4]. The secondary structures of the query and the target are determined with the program KAKSI (Martin et al [5]) from the atomic coordinates.

The server collects all the cliques having a Pcli > -10 and a rmsd < 5 Å that are found by comparing a query protein with a representative set of protein chains of the PDB. These cliques are listed in the file *.mathlab. Then the cliques are extended by including the residues between two secondary structural elements of the clique if they are less than 40 (fig.1b, query pLSSP (for padded Locally Similar Structural Piece) and target pLSSP) and all the query pLSSP are clustered as a binary matrix A along the query length. This matrix A is transformed into a co-occurrence N matrix presented as a heat-map/contour map (file *_Nmatrix.png/*_Nmatrix_contour.png) from which the domains are parsed by three different methods: PCM, SMF and SVD (Tai et al. [1]) in the file *_Domains.txt.

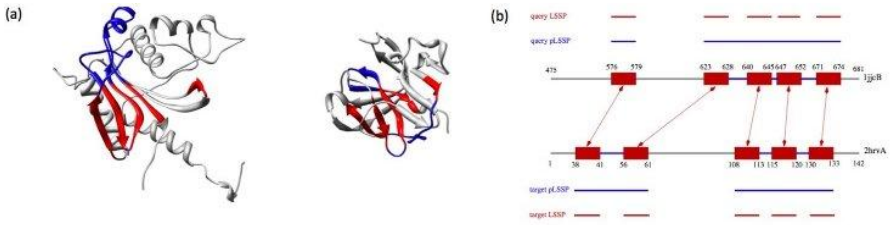


Figure 1. Example of one of the 12,282 cliques for the query protein 1jjcB:

- Panel (a) the query protein 1jjcB (residues 475-674) on the left and a target protein 2hrvA on the right. The structural similarity found by VAST for this clique corresponds to 4 anti-parallel strands of a sheet. Aligned segments of secondary structures are shown in red in both structures, and the gaps less than 40 residues in blue. The target protein 2hrvA is not a structural neighbour considering its low percent of aligned residues.
- Panel (b) The same clique, with (blue lines) and without (red lines) gaps included is "projected" along the query sequences (blue: pLSSP, red: LSSP) and the target. The red boxes correspond to the VAST clique segments of secondary structures with their numbering along the sequence. Arrows indicate the structural correspondence between the segments of secondary structures. Taken with permission from Tai et al [1].

The server introduces a concept of structural neighbours by selecting the targets according two criteria: the length of the target clique, gaps included, amounts to at least 80% of the target total number of residues and the quality of the alignments with 40% at least of the target CA (carbon alpha) aligned by VAST with the query. In other words, the domain in the query protein exists as an individual 3D structure in the PDB. The server provides a list, and a graphic representation of these neighbours. Inputs are a PDB accession code or a file of coordinates (PDB format) and the results are sent by email with an access to a web page.

3 Results and Discussion

From our studies in Tai et al. [1], the domain assignments by the algorithms SMF and SVD perform as well as the server PUU (87.5%, 87.3 and 86.7% respectively) but less than Domain Parser and PDP (92.9% and 93.1% respectively) in their agreement with CATH or SCOP classifications. Considering that the domain assignments by the algorithm PCM are closer to those of CATH or SCOP than SMF and SVD for chains having three or more domains. The web site presents a 3D model of the domain assignments.

The non redundant data base used for the VAST comparison are proteins of the PDB (http://www.pdb.org) having less or equal to 40% of identical residues. It is periodically updated taking advantage of the normal evolution of the PDB content.

4 Conclusion

The present server offers the possibility to determine the domain boundaries with a comparable accuracy with other servers of this type. It offers also a list of structural neighbours useful to detect remote homologues. The collection of small 3D common substructures (3 to 5 secondary structures, 4 in average) represents a global vision of the domain structure that we consider as reflecting its formation during evolution. The introduction of the notion of structural neighbours allows defining the structural domains as individual structures already existing in the Protein Data Base. They can differ in some cases from the assigned boundaries of the domain by the server, this suggests future works.

References

[1] CH. Tai, V. Sam, JF. Gibrat, J. Garnier, PJ. Munson and BK. Lee, Protein domain assignment from the recurrence of locally similar structures. *PROTEINS: Structure, Function, and Bioinformatics*., 79:853-866, 2011.

[2] JF. Gibrat, T. Madej and SH. Bryant, Surprising similarities in structure comparison. *Curr Opin Struct Biol.*, 6(3):377-385, 1996.

[3] T. Madej, JF. Gibrat and SH. Bryant, Threading a database of protein cores. *Proteins.*, 23(3):356-369, 1995.

[4] V. Sam, CH. Tai, J. Garnier, JF. Gibrat, B. Lee and PJ. Munson, ROC and confusion analysis of structure comparison methods identify the main causes of divergence from manual protein classification. *BMC bioinformatics.*, 7:206, 2006.

[5] J. Martin, G. Letellier, A. Marin, JF. Taly, A. de Brevern and JF. Gibrat, Protein secondary structure assignment revisited: a detailed analysis of different assignment methods. *BMC Struct Biol*., 5:17, 2005.

In Silico Insights into the Platelet Alloimmune Response to αIIbβ3 Polymorphisms

Pierre POULAIN[1], Vincent JALLU[2], Cécile KAPLAN[2] and Alexandre DE BREVERN[1]

[1] DSIMB, INSERM UMR-S 665, Université Paris Diderot - Paris 7, Institut National de la Transfusion Sanguine, 6 rue Alexandre Cabanel, 75015 Paris, France
{pierre.poulain, alexandre.debrevern}@univ-paris-diderot.fr

[2] Laboratoire d'Immunologie Plaquettaire, Institut National de la Transfusion Sanguine, 6 rue Alexandre Cabanel, 75015 Paris, France
{vincent.jallu, cecile.kaplan}@ints.fr

Abstract *Fetal / Neonatal alloimmune thrombocytopenia (FNAIT) is a severe bleeding syndrome in which fetal / neonatal platelet destruction is mediated by maternal antibodies directed to specific antigens (or alloantigens) inherited from the father. The integrin complex αIIbβ3 is highly immunogenic and is responsible for most FNAIT. We used state-of-the-art molecular modelling techniques to study the impact of Human Platelet Alloantigen (HPA) polymorphisms on the complex structure, and their role in antigenicity. We showed that accessibility is a key element in the immune response.*

Keywords αIIbβ3, HPA polymorphisms, molecular modelling, antigenicity.

Fetal / Neonatal alloimmune thrombocytopenia (FNAIT) is a severe bleeding syndrome in which fetal / neonatal platelet destruction is mediated by maternal antibodies directed to specific antigens (or alloantigens) inherited from the father. These antigens depend on polymorphisms of genes coding for several membrane glycoproteins (GPIb-IX-V, GPIIbIIIa, and GPIaIIa) or lipo-protein (CD109) receptors expressed at the platelet surface. These polymorphisms are classified in the Human Platelet Alloantige (HPA) nomenclature. αIIbβ3 (or GPIIbIIIa) carries the majority of the HPA systems described to date. This complex is highly immunogenic and is responsible for most FNAIT.

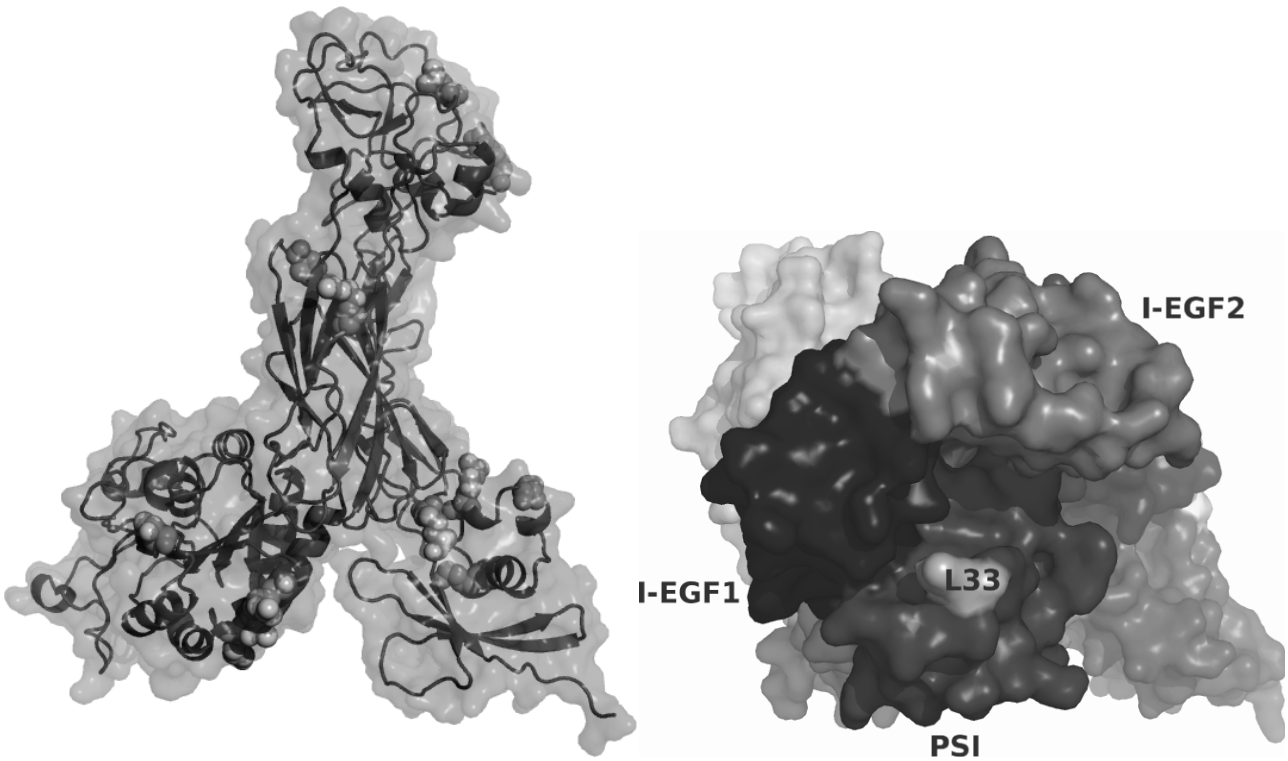
αIIbβ3 belongs to the large family of the integrins that is composed of heterodimeric membrane receptors involved in cell-cell or cell-matrix interactions. It mediates platelet aggregation as a receptor for fibrinogen, a major plasmatic adhesion molecule. Resting platelets express on their surface about 50 000 copies of αIIbβ3 and 30 000 additional copies when activated.

Protein 3D structures (or structural models) help to understand relationships between the protein dynamics and their biological functions. They provide new insights into atomic mechanisms of macromolecular recognition and conformational changes. We have rencently used a 3D structure of αIIbβ3 (PDB code 3FCS) to propose an explanation for the structure effect of the β3 Lys253Met substitution indentified in a Glanzmann patient, a mutation impairing αIIbβ3 expression [1].

Immune response relies on both immunogenicity and antigenicity. Antigenicity can depend on the 3D molecular structure surrounding the polymorphic site. We have used a 3D structure of αIIbβ3 and modelling experiments to study the impact of HPA polymorphisms on the complex structure, and their role in antigenicity. Different HPA allelic forms of αIIb and β3 were modeled and resulting structure characteristics of residue accessibility, mobility, and electrostatic change were analyzed [2].

β3 HPA polymorphisms 1a, 1b, 4a, 4b, 6a, 7a, 10a, 11a, 14a, 16a, 17a, 19a and 21a have been studied and reported in Fig. 1(a). The β3 backbone structure is represented as ribbon and HPA polymorphic amino acids are represented as spheres. Multiple glycosylation sites seem to not overlap with HPA polymorphism sites. Alloantibodies rely on the residue presence at the surface of the structure (accessibility) but do not tightly depend on its mobility or its electrostatic charge.

We focused our work on the the HPA-1 system of β3 that is the most frequent HPA system involved in FNAIT in Caucasian population in term of frequency and pathology severity. In the β3 structure, the HPA 1a



(a) (b)

Figure 1. (a, left panel) β3 HPA polymorphisms. The β3 backbone structure is represented as ribbon and HPA polymorphic amino acids as spheres. (b, right panel) HPA 1a polymorphism localisation on β3 with neighboring domains labeled.

polymorphism is characteristed by a leucine residue and a proline for HPA 1b. The residue of interest is located in a loop in the PSI domain of β3 (illustrated in Fig. 1(b)). This loop is also in close vicinity of the I-EGF1 and I-EGF2 domains that are strongly involved in the activation of αIIbβ3. Molecular dynamics studies were performed on HPA 1a and 1b structures showing a high flexibility of the I-EGF1 and I-EGF2 domains while retaining a similar local structure of the mutated loop. The antigenicity of HPA 1a and 1b seems preserved although a difference in the accessibility of both polymorphisms is observed.

Structural modelling of the different HPA forms of αIIbβ3 and comparative analyses of the structure characteristics suggest that, as expected, antigenicity mainly depends on residue accessibility. The other structure features such as residue mobility and electrostatic do not appear critical for the presence of an alloantibody although they can modulate its binding affinity. These analyses are performed on static structures obtained directly from the Protein Data Bank, or after modelling of a specific allele. However, αIIbβ3 protein chains are highly dynamic. We have performed dynamic analyses to go deeper in the understanding of polymorphisms structural properties.

Acknowledgements

This work was supported by the National Institute for Blood Transfusion (INTS), the French Institute for Health and Medical Research (INSERM), and the University of Paris Diderot - Paris 7.

References

[1] V. Jallu V, M. Dusseaux, S. Panzer, M.F. Torchet, N. Hezard, J. Goudemand, A. G. de Brevern and C. Kaplan, αIIb/β3 integrin: new allelic variants in Glanzmann Thrombasthenia, effects on ITGA2B and ITGB3 mRNA splicing, expression and structure-function. *Hum. Mutat.* 31: 237-246, 2010.

[2] V. Jallu, P. Poulain, C. Kaplan and A. G. de Brevern, 3D protein structure modeling: A tool to provide insight into the platelet alloimmune response. *Transfusion Today*, in press, 2011.

The Sequence-Structure Relationship in α-helical Transmembrane Proteins

Jérémy Esque[1], Aurélie Urbain[2], Catherine Etchebest[1] and Alexandre G. de Brevern[1]

[1] DSIMB, UMR-S 665 INSERM, Université Paris Diderot – Paris 7, Institut Nationale de la Transfusion Sanguine, 6 rue Alexandre Cabanel, 75739, Paris, Cedex 15, France
{jeremy.esque, catherine.etchebest, alexandre.debrevern}@univ-paris-diderot.fr

[2] Unité de recherche Mathématiques et Informatique Appliquées, INRA Jouy en Josas - Domaine de Vilvert, 78352, Jouy en Josas, France
aurelie.urbain@versailles.inra.fr

Abstract *Transmembrane proteins (TMP) are known to play essential roles in all living cells, like bacteria and eucaryota. They are involved in numerous major and essential biological processes, e.g., ion and small molecule transport and signal transduction. This kind of proteins is also the target of most of manufactured drugs. Although they account for about 20-30% of coding genome, they represent only less than 2% of structures available in the Protein Data Bank (PDB). This low number is due to the difficulty to obtain high-resolution structures of transmembrane proteins as they are embedded into lipid bilayers.*
To overcome the limitations in the number of available structures, comparative modeling is an interesting tool, but is also limited by the lack of structures usable as templates. We propose to study the sequence-structure relationship of TMP and to extract general features from various proteins. To perform our work, we use the Hybrid Protein Model (HPM), a learning approach able to compact protein 3D-structure and physico-chemical information. This methodology has already been successfully used for globular protein studies.

Keywords α-helical transmembrane protein, protein structures, protein blocks, sequence-structure relationship.

1 Introduction

Transmembrane proteins (TPMs) are involved in many essential functions. They are also linked to, many diseases and pathologies due to mutations, leading to misfolds or missassembly of TMPs, or the binding of unwanted partners, *e.g.*, Duffy Binding Protein of *Plasmodium vivax* [1]. So, understanding better the TPMs is an important research field especially for drug design strategies [2]. These last years, many researches have been led in the area of TPMs and their functions, particularly in analyzing the sequences, the topology and the mutation effects [3].

Even though the number of membrane protein structures is limited, the uniformity of their structures and interactions allow them to investigate computationally. Integral α-helical membrane proteins are composed of a bundle of helices crossing completely the membrane. So, α-helical membrane proteins are rich in reentrant regions, interfacial helices, structured extracellular or cytoplasmic loops. An exhaustive analysis of these structural features related to the sequence would be needed for enhancing the understanding of functions, modeling and drug design [4].

To study the sequence-structure relationship in transmembrane proteins, we use an adapted and original methodology, named Hybrid Protein Model (HPM). The power of HPM is to compact the information contained in a structural protein databank to analyze the relationship between sequence and structures. The advantage of this clustering method is that it can take into account the sequentiality of the protein without any *a priori*. The latter is an important point in proteins as they are polymer chains. It allows creating clusters which are learnt independently but having a sequentiality, *i.e.*, a cluster *i* overlaps with their neighbors *i*+1, *i*-1, … . It has been used to compact protein 3D-structures information and physicochemical properties of globular proteins [5]. As 3D local folds are clustered using HPM, the informativity on their sequence is an interesting analysis to study the sequence-structure relationship.

2 Material and Methods

From membrane protein databanks (OPM, TMPDB, PDB-TM ...), a non-redundant dataset of X-ray TMPα structures (<2.5 Å) was extracted. The pairwise identity percent was 40%, and no backbone atom was missing. The sequence-structure relationship was performed using HPM method [6] (see Figure 1).

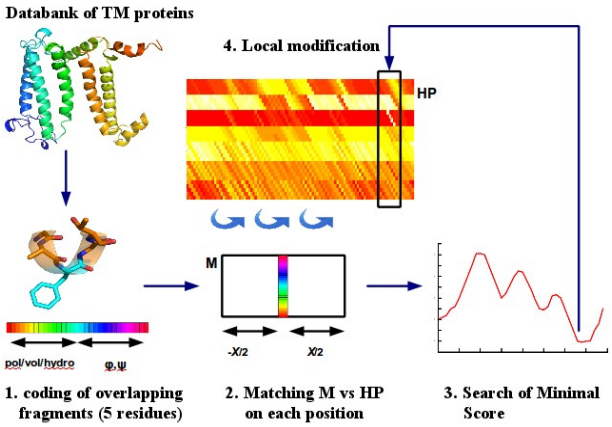


Figure 1. HPM method. 1) A databank of protein fragments is built. Each protein fragment is encoded into a vector encompassing physico-chemical properties and angles information. 2) Each fragment *F* with its environment (*X*=13) forms a sub-matrix. A compatibility score is computed at every positions of HP matrix. 3) The minimal score implies the best matching position. 4) A local modification at this position is performed to learn the fragment F. Process is reiterated through stabilization.

3 Results

After checking important parameters, as the HP length, the number of learning cycles, the most representative HP matrix is selected. Two major criteria to select the HP are: 1) a high continuity between consecutive hybrid positions, 2) a low redundancy within the HP. Analyses are performed, like the distribution of protein blocks (PBs) and their amino acid distribution, to underline interesting sequence – structure relationship. The main results will be presented, *e.g.*, the different kinds of helical regions defined by HPM. For example, after clustering analysis, two kinds of helices have been detected with different physico-chemical properties due to the composition in amino acids.

Acknowledgements

This work was supported by the National Institute for Blood Transfusion (INTS), the French Institute for Health and Medical Research (INSERM), the University of Paris Diderot – Paris 7.

References

[1] C. Sanders and J. K. Myers, Disease-related misassembly of membrane proteins. *Annu. Rev. Biophys. Biomol. Struct.*, 33:25-51, 2004.

[2] Y. Arinaminpathy, E. Khurana, D. M. Engelman and M. B. Gerstein, Computations analysis of membrane proteins: the largest class of drug targets. *Drug Discov. Today*, 14:1330-1335, 2009.

[3] A. Elofsson and G. von Heijne, Membrane Protein Structure: Prediction versus Reality. *Annu. Rev. Biochem.*, 76:125-140, 2007.

[4] A. Marsico, A. Henschel, C. Winter, A. Tuukkanen, B. Vassilev, K. Scheubert and M. Schroeder, Structural fragment clustering reveals novel structural and functional motifs in α-helical transmembrane proteins. *Bmc Bioinformatics*, 11:204, 2010.

[5] A. G. de Brevern and S. Hazout, Hybrid Protein Model (HPM): a method to compact protein 3D-structures information and physicochemical properties. *IEEE- Comp Soc*, S1:49-54, 2000.

[6] A. G. de Brevern and S. Hazout, 'Hybrid Protein Model' for optimally defining 3D protein structure fragments. *Bioinformatics*, 19:345-353, 2003.

Conformational Plasticity of the Adenylyl Cyclase CyaA from *Bordetella Pertussis*

Edithe Selwa[1], Elodie Laine[2] and Thérèse E Malliavin[1]

[1] Unité de Bioinformatique Structurale, Institut Pasteur and CNRS URA 2185, 25-28, rue du Dr. Roux, 75724, Paris, France

edithe.selwa@pasteur.fr

terez@pasteur.fr

[2] Laboratoire de Biologie et de Pharmacologie Appliquée Ecole Normale Supérieure de Cachan and CNRS UMR8113, 61, avenue du Président Wilson, 94235, Cachan cedex, France

elodie.laine@lbpa.ens-cachan.fr

Keywords Calmodulin, *Bordetella pertussis*, molecular dynamics, calcium, essential dynamics, X-Ray crystallography.

1 Introduction

The protein CyaA is an important virulence factor of *Bordetella pertussis*, the causative agent of whooping cough disease. The AC domain of CyaA is translocated into eukaryotic cells, and is activated as adenyl-cyclase to produce cAMP in an uncontrolled way leading to an alteration of the immune system, by interaction with the ubiquitous protein calmodulin (CaM).

CaM is a protein including two lobes, N-CaM and C-CaM, connected by a flexible linker. The structure of the AC bound to C-CaM was recently resolved by Guo et al, 2005 [1], by X-ray diffraction. In this crystallographic structure, the lobe N-CaM could not be determined. The AC domain includes three subdomains, named CA, CB, and Switch A (SA). The catalytic loop and the C terminal regions are included in the CA domain.

In a previous work on the homologous protein EF of *Bacillus anthracis* (Laine et al, 2008 [2]), MD simulations showed that the removal of Ca^{2+} from C-CaM induces a tension toward EF, and deformations in apo-EF which bring domains closer and induce a collapse catalytic site.

In order to get more information on the conformational behavior of AC domain in the absence of CaM and on the role of the Ca^{2+} ions in the interaction and to get an energy dependency map between subdomains, we performed molecular dynamics (MD) simulations of three systems : AC-(2Ca-C-CaM) corresponding to the PDB structure 1YRT [1], AC-(0Ca-C-CaM) where the calcium ions were removed from the structure, and the free AC domain where calmodulin was removed. This simulations showed that the removal of Ca^{2+} ions and C-CaM induced a large conformational variability of AC, possibly leading to a protein compaction. Energy dependency analyzes revealed a very simple energetic influence of C-CaM on CA, which disappears with the removal of Ca^{2+}.

2 Results

It's important to analyze the degree of convergence of a trajectory to check the stability of a simulation. This was evaluated by calculating the standard deviation (RMSD) of Cα positions with regard to the first structure of the trajectory. The global conformational drifts of the solute is the smallest in the presence of C-CaM with a plateau around 3 Å. The deletion of CaM make the Cα RMSD drop to 6 Å, after 12ns of trajectory. Thus, C-CaM seems to be essential for stabilizing AC. The conformational drift of C-CaM corresponds to a quite limited reorganization of the calmodulin lobe.

The global motions of the solute were analyzed by principal component analysis (PCA) of the covariance matrix of atomic coordinates. For AC, there is one dominant eigenvalue, while there are several ones for the complexes. 90% of the protein internal motion are explained by 10, 19, 30 eigenvectors in free AC, AC-

(0Ca-C-CaM) and AC-(2Ca-C-CaM) respectively. In all trajectories, the Switch A (SA) and the domain CB display the largest projections of PCA eigenvectors (Figure 1). A tendency of AC to become more compact by moving closer SA and CB can be noticed in free AC. The presence of C-CaM reduces the SA rotation.

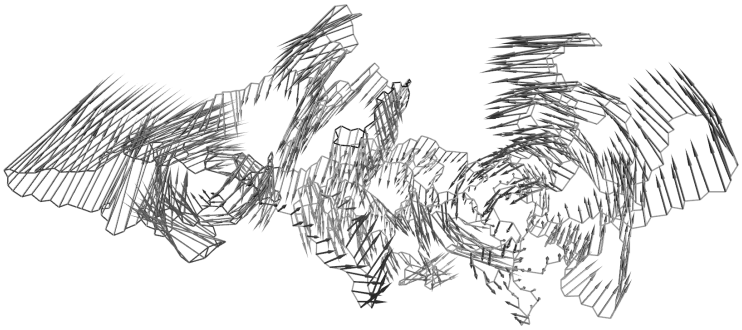


Figure 1. Projection of the first mode PCA on the backbone of free AC.

The energetic influences between the Switch A, CA, CB and C-CaM in the three molecular systems were determined by MMPBSA ΔG [3] binding energies calculation, which was proposed by Laine [4] on the homologous complex EF-CaM. No energetic influence was observed between domains for free AC and AC-(0Ca-C-CaM) trajectories. During the trajectory AC-(2Ca-C-CaM), only one influence was observed from C-CaM on CA. The interaction network is more simple than in the complex EF-CaM, in which CaM is captured between the domains Hel and CA.

3 Conclusion

The global motion observed for the AC domain during the trajectory AC-free shows a tendency of the protein to become more globular. This agrees with the hydrodynamics measurements performed by Karst and coworkers [5] on AC domain isolated in solution. The simplicity of the architecture of the complex revealed by energy dependency analyses allows us to imagine a synthetic biology approach, in which AC will be engineered so as to modify the control by calmodulin of the production of cAMP.

Acknowledgements

The authors thank the Institut Pasteur Paris, CNRS, UMPC and iViV.

References

[1] Q. Guo, Y. Shen, Y S. Lee, C S. Gibbs, M. Mrksich and W J. Tang, Structural basis for the interaction of Bordetella pertussis adenylyl cyclase toxin with calmodulin. *Embo J.*, 24:3190-3201, 2005.

[2] E. Laine, J.D. Yoneda, A. Blondel and T.E. Malliavin, The conformational plasticity of calmodulin upon calcium complexation gives a model of its interaction with the oedema factor of *Bacillus anthracis. Proteins*, 71:1813-1829, 2008.

[3] J. Srinivasan, T. Cheatham, J. Cieplak, P. Kollman and D. Case, Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate-DNA helices. *J.Am.Chem.Soc.*, 120:9401-9409, 1998.

[4] E. Laine, A. Blondel and T.E Malliavin, Dynamics and energetics : a consensus analysis of the impacf of calcium on EF-CaM protein complex. *Biophys J.*, 96(4):1249-63, 2009.

[5] JC. Karst, AC. Sotomayor Pérez, JI. Guijarro, B. Raynal, A. Chenal and D. Ladant, Calmodulin-induced conformational and hydrodynamic changes in the catalytic domain of *Bordetella pertussis* adenylate cyclase toxin. *Biochemistry*, 49(2):318-28, 2010.

The Dynamics Modes of the VanA D-alanyl:D-lactate Ligase are Similar to those of the D-alanyl:D-alanine Ligase

Nathalie Duclert-Savatier[1], Djalal Meziane-Cherif[2], Arnaud Blondel[1], Michael Nilges[1] and Thérèse E Malliavin[1]

[1] Unité de Bio-informatique Structurale, URA 2185 CNRS, 25-28, rue du Dr Roux, Institut Pasteur, 75724, Paris, cedex 15, France

nathalie.duclert-savatier@pasteur.fr

[2] Unité des Agents Antibactériens, 25-28, rue du Dr Roux, Institut Pasteur, 75724, Paris, cedex 15, France

Keywords D-Ala:D-Lac Ligase, antibiotic resistance, molecular dynamics simulations.

1 Introduction

The peptidoglycan layer is a key component of the bacteria cell wall to keep its shape and to prevent it from osmotic lysis. The peptidoglycan is a polymer resulting of the cross linking of identical chains of monomers, each of them made of two joined amino sugars with a pentapeptide tail. D-Ala:D-Ala ligase is involved in the first steps of the synthesis of the cell wall. It catalyzes the formation of the dipeptide D-Ala-D-Ala before its incorporation in peptidoglycan synthesis. Glycopeptide antibiotics, like vancomycin or teicoplanin, selectively bind to the D-Ala-D-Ala termini of the peptidoglycan precusors preventing them from crosslinking. In glycopeptide-resistant bacteria, an additional ligase, vanA, forms the depsipeptide D-Ala:D-Lac instead of the expected dipeptide D-Ala:D-Ala, preventing the binding of vancomycin. Such bacteria are still able to build an effective cell wall in the presence of the antibiotic.

Indeed, D-Ala:D-Lac ligase and D-Ala:D-Ala ligase share very close functions. Their amino acids involved in the binding of their ligands are mainly conserved, except for the ones involved in the selectivity of the second ligand and their structures show the same folding. D-Ala:D-Ala ligase has been crystallised in several configurations [1]. The empty form of the protein is open and exhibits an extended omega loop while the protein with its binding site filled with ADP, 2 Mg^{2+} and a ligand is much more compact. In the co-crystals, the omega loop closes the binding pocket through an H-bonds network constituted by a triad of amino acids connecting three loops. This lock keeps the binding cavity closed protecting the ligands bonding from hydrolysis. Up to now, the D-Ala:D-Lac ligase vanA was only co-crystallised with ADP, 2 Mg^{2+} and a phosphinate inhibitor in its binding cavity (PDB code : 1E4E).

We have investigated the molecular dynamics of the vanA D-Ala:D-Lac ligase [2], from resistant *Enterococcus faecium* in order to compare it to the published crystallographic conformations of the structures of D-Ala:D-Ala ligases.

2 Materials and Methods

Molecular dynamics (MD) trajectories were recorded over 30 ns using AMBER 10 in explicit solvent with TIP3P water parameters under periodic conditions at a constant pressure of 1 atmosphere regulated with isotropic position scaling and a relaxation time of 1 ps. The force field was FF99SB. A cutoff of 10 Å was used for Lennard-Jones interactions, and long-range electrostatic interactions were calculated with the Particule Mesh Ewald (PME) protocol. The system was neutralized using 4 Na^{+} counterions. The simulations were performed at 300 K, using a Berendsen thermostat to control it. The Shake algorithm kept rigid all covalent bonds involving hydrogens, with a time step of 2 fs.

The protein vanA (1E4E) was simulated independently 9 times with ATP, 2 Mg^{2+} and a phosphinate inhibitor in its binding pocket. The empty protein was obtained by removing the ligands from the same initial structure and 8 trajectories were run. The proteins were simulated in the presence (= vanA-ss) or in the absence (= vanA) of a disulphide bridge formed between C52 and C64 in the N-terminus, far away from the binding site. At least, 7 trajectories were run for VanA-ss empty and 9 for vanA-ss with its ligands.

3 Results

In the absence of ligands, vanA does not display conformational drift. On the contrary, vanA-ss shows an increase of its radius of gyration in 5 out of 9 cases. This expansion is due to the beginning of the opening of the omega loop involved in the closure of the binding site. Indeed, a principal component analysis (PCA) (Figure 1) of the dynamics covariance of vanA-ss underline the omega loop and the part of the central domain facing it in the structure. These two regions build up the ligands binding site.

An opening of the structures was also observed for the vanA and vanA-ss simulated with their ligands. In all cases, the most mobile parts of the protein were restricted to the omega loop and to its counterpart central domain. The binding mode of the ligands is an ordered Ter Ter mechanism with ATP binding first prior to the D-Ala followed by the D-Lac. The obtained motions can be compared to the structural variations [1] observed by superimposing different crystal structures of the D-Ala:D-Ala ligases.

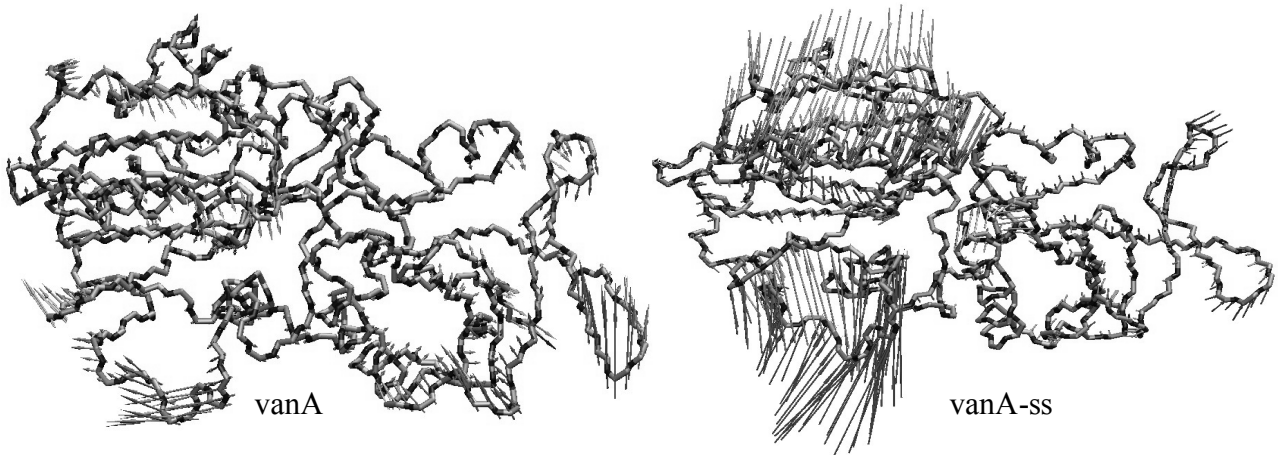


Figure 1. The first PCA mode projected on the trajectory structures of vanA and vanA-ss.

4 Conclusion

The MD simulations performed on the protein vanA allowed us to relate its internal dynamics to the ligand interactions. As the same relations were observed on several D-Ala:D-Ala ligase structures [1], D-Ala:D-Lac ligase internal dynamics should aim at the same characteristics. These observations allow to speculate that the two families D-Ala:D-Ala and D-Ala:D-Lac ligases have a similar opening pocket mechanism.

Acknowledgements

We thank Institut Pasteur and CNRS for the funding.

References

[1] Y. Kitamura, A. Ebihara, Y. Agari, A. Shinkai, K. Hirotsu. and A. Kuramitsu, Structure of D-alanine-D-alanine ligase from Thermus thermophilus HB8: cumulative conformational change and enzyme-ligand interactions. *Acta Cryst.*, 65:1098-1106, 2009.

[2] D. I. Roper, T. Huyton, A. Vagin and G. Dodson, The molecular basis of vancomycin resistance in clinically relevant *Enterococci*: Crystal structure of D-alanyl-D-lactate ligase (VanA). *Proc. Natl. Acad. Sci. USA.*, 97(16):8921-8925, 2000.

Analysis of the Full Orthosteric Cavity to Discriminate Agonist from Antagonist Ligands in AChBP

Julien BURATTI, Arnaud BLONDEL, Thérèse E. MALLIAVIN and Michael NILGES

UNITE DE BIOINFORMATIQUE STRUCTURALE, INSTITUT PASTEUR and URA CNRS 2185, 25-28 rue du Docteur Roux, 75015 Paris, France

{jburatti, ablondel, terez, nilges}@pasteur.fr

Abstract *The binding mode of agonist and antagonist on the AChBP structures is usually considered as being determined by the C-loop position. We are proposing a new geometric parameter to discriminate the ligand type: the full orthosteric cavity located at the interface between the two subunits in ECD. Indeed, the spherical harmonics decomposition of this cavity allows to discriminate between agonist and antagonist binding mode.*

Keywords acetylcholine binding protein, spherical harmonics, ligand, surface calculation.

1 Introduction

The acetylcholine nicotinic receptors (nAChR) are homo- or hetero- transmembrane pentamers, which belong to the Cys-loop receptors family. Each subunit contains two domains: the extra-cellular domain (ECD) forms the ligand orthosteric site at the interface between two subunits and the transmembrane domain forms the ionic channel. The agonist acetylcholine is the natural ligand of nAChR but other molecules can bind to the receptor, displaying an agonist (stabilizing the open channel) or an antagonist (stabilizing the closed channel) effect. Up to now, only one structure of nAChR is available at a resolution of 4Å [1]. Nevertheless, numerous high resolution crystallographic structures involving agonist and antagonist ligands are available for the soluble acetylcholine binding protein (AChBP), which is homologuous to the nAChR ECD.

The binding mode of agonist and antagonist was previously analyzed on the AChBP structures [2] and the position of the C-loop over the binding pocket is usually considered as a discriminated factor. Herein we are proposing a new geometric parameter to discriminate the ligand type: the full orthosteric cavity located at the interface between the two subunits in ECD.

2 Materials and Methods

32 PDB structures: 1I9B, 1UV6, 1UW6, 1UX2, 1YI5, 2ZJU, 2ZJV (*Lymnaea stagnalis*), 2BG9 (*Torpedo marmorata*), 2BJ0 (*Bulinus truncatus*) et 2BR7, 2BR8, 2BYN, 2BYP, 2BYQ, 2BYR, 2BYS, 2C9T, 2PGZ, 2PH9, 2UZ6, 2W8E, 2W8F, 2W8G, 2WN9, 2WNC, 2WNJ, 2WNL, 2WZY, 2X00, 3C79, 3C84, 3GUA (*Aplysia californica*) were included in the analysis. The 220 subunit dimers, extracted from these PDB structures, were superimposed to the dimer AB of 2BYP using ProFit [3].

The orthosteric cavities are detected by the software mkgrid (A. Blondel, unpublished result) by rolling a variable sized probe on the structures. The solvent accessible surface of the cavity is determined using a 10Å probe, whereas the surface accessible to the protein is determined using a 1.4Å probe. The atoms N, H, O, C, S, P of the protein are represented by spheres with van der Waals radii of 1.6, 0.6, 1.6, 2.3, 1.9 et 2Å respectively. We obtain on the overall structure cavities represented by 3D grids of 0.5Å resolution. The orthosteric grid is manually selected as the grid located at the interface of the two subunits and lying behind the C-loop.

The spherical harmonics development of the orthosteric cavity is obtained using SpharmonicKit/s2kit [4,5]. The distance of the surface of the orthosteric cavity to the origin of the axes, can be expressed as a scalar function $f(\theta, \varphi)$, θ and φ being the angular spherical coordinates. f is developed in spherical harmonics:

$$f(\theta,\varphi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{+l} C_l^m Y_l^m(\theta,\varphi) \quad (1)$$

where C_l^m are the development coefficients, and $Y_l^m(\theta, \varphi)$ are the spherical harmonics:

$$Y_l^m(\theta, \varphi) = \sqrt{\frac{2 \cdot (l-m)!}{(l+m)!}} P_l^m(\cos\theta) \mathrm{e}^{+i\,m\,\varphi} \tag{2}$$

where $P_l^m(\cos\theta)$ is the Legendre polynomial. Using the previously determined C_l^m coefficients, a principal component analysis (PCA) is realized with R [6] and the package FactoMineR [7]. The PCA is centered but not unit-scaled in order to take into account the large coefficients variability.

3 Results and Discussion

The projection of the dimers on the two first PCA modes (Fig. 1) seems to well discriminate between agonist and antagonist ligands. The apo dimers, as well as the dimers containing crystallographic co-factors (not shown in the figure) can have variable configuration. Hence, their scattering over the whole projection is not surprising, which agrees with the non-specific effect of these ligands.

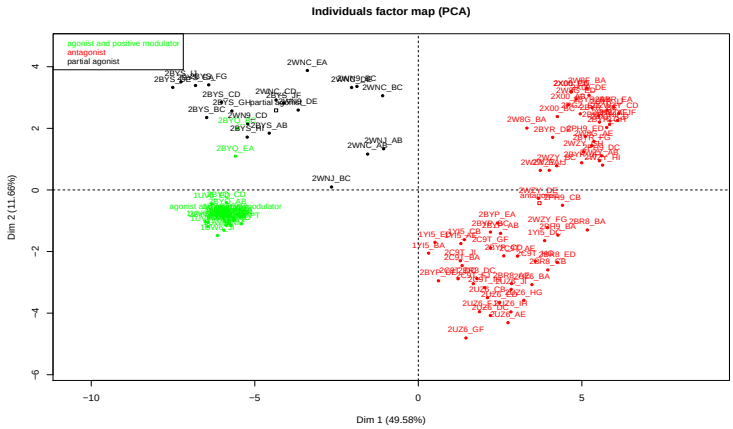


Figure 1. Projection of the subunits dimers on the two first modes of the PCA, based on spherical harmonics development.

4 Conclusion

The PCA projection based on spherical harmonics development of the full orthosteric cavity shows a good discriminative power, according to the agonist and antagonist ligand effect. This projection can be used as a reference to classify crystallographic structures of conformation produced by molecular dynamic simulation. The selection rules derived from the projection could be used during molecular docking approach in order to orient the search towards agonist or antagonist ligands.

Acknowledgements

This work was supported by Fondation pour la Recherche Médicale, Institut Pasteur and CNRS.

References

[1] N. Unwin, Refined structure of the nicotinic acetylcholine receptor at 4Å resolution, *J Mol Biol.*, 346(4):967-89, 2005.
[2] P. Taylor, T.T. Talley, Z. Radic, S.B. Hansen, R.E. Hibbs and J. Shi, Structure-guided drug design: conferring selectivity among neuronal nicotinic receptor and acetylcholine-binding protein subtypes, *Biochemical Pharmacology*, 74:1164-1171, 2007.
[3] http://www.bioinf.org.uk/software/profit
[4] A. Kahraman, R.J. Morris, R.A. Laskowski and J.M. Thornton, Shape variation in protein binding pockets and their ligands, *J Mol Biol.*, 368:283-301, 2007.
[5] http://www.cs.dartmouth.edu/ geelong/sphere
[6] http://www.r-project.org/
[7] http://factominer.free.fr/classical-methods/analyse-en-composantes-principales.html

Developments in NMR Structure Calculation Protocol in order to Improve the Structure Quality and Convergence

Fabien MAREUIL[1,2], Christophe BLANCHET[2], Thérèse E MALLIAVIN[1] and Michael NILGES[1]

[1] Unité de Bioinformatique Structurale, CNRS URA 2185, Institut Pasteur, 25-28 rue du Dr Roux, F-75724, Paris Cedex 15, France

fabien.mareuil@pasteur.fr

[2] Université de Lyon 1, Univ Lyon, France; CNRS; FR 3302; Institut de Biologie et Chimie des Proteines, IBCP, 7 passage du Vercors, F-69367, France

Keywords Nuclear Magnetic Resonance (NMR), ARIA, log-harmonic restraint potential, Bayesian statistics, restraints violation.

1 Introduction

Ambiguous Restraints for Iterative Assignment (ARIA) is a software for efficient NMR structure determination of proteins by automated assignment of the nuclear Overhauser effects (NOEs) measured on the NOESY spectra. ARIA follows an iterative protocol. In each iteration, the NOEs determined from the previous iteration conformations, and the data points most inconsistent with the conformations are removed with a simple statistical analysis [1,2]. Protein conformations are then calculated with distance restraints, which are based on the current set of assignments but usually retain a large level of ambiguity. Convergence of the structures improve from iteration to iteration, and the calculation is terminated when a converged set of conformations is obtained with good restraints fit.

The automated NOE assignment and the structures determinations are hampered by low spectral resolution or missing structural data. The NMR restraints are usually applied through an harmonic restraint potential, but a log-harmonic potential has been recently developed from the observation of the distribution of distances and NOE intensities [3]. The advantage of this log-harmonic potential is twofold: it has a single minimum and it is more tolerant for large violation [4]. Nevertheless, some difficulties in structure calculation have been encountered with log-harmonic potential. They may be detected as bad quality Molprobity scores [5], or as a lack of structure convergence.

2 ARIA Developments

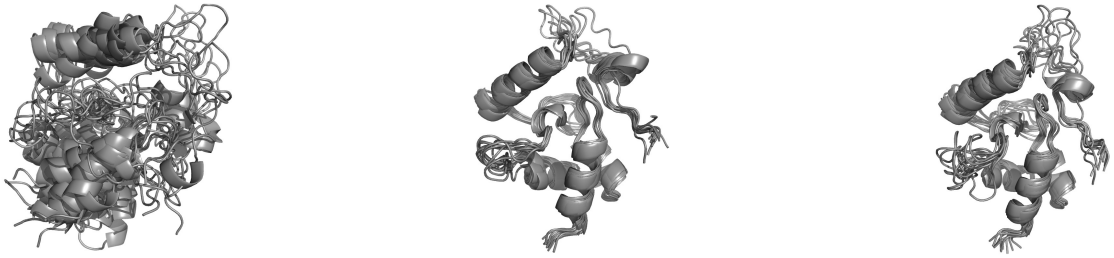
The developments realized in ARIA concern three aspects: (a) the implementation of ARIA to the grid computing, (b) the modification of the force field used during the conformers generation, (c) the determination of the violation tolerance from the current set of protein conformations.

The iterative ARIA cycle is based on three major steps: (i) input preparation for the generation of conformations, (ii) generation of conformations by a simulated annealing procedure using the software CNS [6], (iii) generation of conformations to generate a new set of NOE assignments and restraints. ARIA is straightforwardly parallelized by distributing the step (ii) over a few or many CPU units. To offer to the greatest number of people the opportunity to use ARIA with powerful CPU, a new version of ARIA have been developed that can run on a grid computing. This development has been realized in the frame GRISBI (Grid Support to Bioinformatics: www.grisbio.fr) [7].

Low Molprobity score have been observed when using the log-harmonic potential. The force field used during the conformer generation displays high force constants and small Van der Waal radii that are not compatible with the single minimum of the log-harmonic potential. A new force field have been developed, the energy constants for covalent angles and improper dihedrals have been decreased (by factor 10) and the Van der Waals radii for hydrogen have been increased, to soften the force field. The softer force field has

been tested on several proteins, which were proposed as targets in the **C**ritical **A**ssessment of **A**utomated **S**tructure **D**etermination of Proteins from **NMR** data (CASD-NMR: http://www.e-nmr.eu/CASD-NMR) [8]. These modifications significantly improved the Molprobity scores without affecting the coordinates RMSD difference to the target.

An additional feature of the calculations (with the log-harmonic potential) is the lack of convergence. The CASD target VpR247 displays a high level of restraints rejection producing an unconverged set of conformers (Figure 1a). To overcome this problem, the violation tolerance was calculated as root mean square deviation between the distances obtained from the restraints and from the best conformers, adapting automatically the tolerance value to the data quality. A good convergence of VpR was thus obtained (Figure 1c) without having to increase the number of conformers calculated by iteration and/or to manually adjust the violation tolerance (Figure 1b).



a) 50 conformers and no automatic monitoring of the violation tolerance. Average RMSD convergence: 6.7 Å. Average RMSD with the PDB 5.22 Å.	b) 200 conformers and manual monitoring of the violation tolerance. Average RMSD convergence: 1.54 Å Average RMSD with the PDB 1.57 Å.	c) 50 conformers and automatic monitoring of the violation tolerance. Average RMSD convergence: 1.6 Å, Average RMSD with the PDB 1.69 Å.

Figure 1. ARIA VpR247 calculations using log-harmonic potential.

Acknowledgements

The authors thank Dr. Anja Böckmann for fruitful discussions at the beginning of the ARIA grid implementation. Funding: ANR THALER ”Massively parallel simulation and analysis of protein structure and dynamics”, CNRS, Institut Pasteur, GIS IBiSA (www.ibisa.net)

References

[1] P. Guntert, Automated NMR structure calculation with CYANA, *Methods Mol Biol*, 278:353–378, 2004.

[2] W. Rieping, M. Habeck, B. Bardiaux, A. Bernard, T.E. Malliavin and M. Nilges, ARIA2: automated NOE assignment and data integration in NMR structure calculation, *Bioinformatics*, 23:381-382, 2006.

[3] A. Bernard, F.W. F. Vranken, B. Bardiaux, M. Nilges and T.E. Malliavin, Bayesian estimation of NMR restraint potential and weight: A validation on a representative set of protein structures. *Proteins,* 79(5):1525-1537, 2011.

[4] M. Nilges, A. Bernard, B. Bardiaux, T. Malliavin, M. Habeck, and W. Rieping, Accurate NMR structures through minimisation of an extended hybrid energy. *Structure* , 16(9):1305-12, 2008.

[5] I.W. Davis, A. Leaver-Fay, V.B. Chen, J.N. Block, G.J. Kapral, X. Wang, L.W. Murray, W.B. Arendall, J. Snoeyink, J.S. Richardson and D.C. Richardson, MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Research* 35: Web Server issue, W375-W383, 2007.

[6] A. Brunger, P.D. Adams, G.M. Clore, W.L. DeLano, P. Gros, R.W. Grosse-Kunstleve, J.S. Jiang, J. Kuszewski, M. Nilges, N.S. Pannu, R.J. Read, L.M. Rice, T. Simonson and G.L. Warren, Crystallography & NMR system: A new software suite for macromolecular structure determination, *Acta Crystallo D Biol Crystallo*, 54:905-921, 1998.

[7] C. Blanchet, R. Mollon, D. Thain and G. Deleage, Grid deployment of legacy bioinformatics applications with transparent data access, *in 7th IEEE/ACM International Conference on Grid Computing*, 120-127, 2006.

[8] A. Rosato, A. Bagaria, D. Baker, B. Bardiaux, A. Cavalli, J.F. Doreleijers, A. Giachetti, P. Günter, T. Herrmann, Y.J. Huang, H.R. Jonker, B. Mao, T.E. Malliavin, G.T. Montelione, M. Nilges, S. Raman, G. van der Schot, W.F. Vranken, G.W. Vuister and A.M Bonvin, CASD-NMR: critical assessment of automated structure determination by NMR. *Nature Methods*, 6:625-626, 2009.

The Superfamily of Beta- and Gamma-Crystallins

Evolution history and Sequence-Structure-Function relationships

Elodie DUPRAT, Windy LUSCAP, Feriel SKOURI-PANET and Stéphanie FINET

IMPMC, UMR7590 CNRS/UPMC/Paris Diderot/IPGP/IRD, 4 place Jussieu, 75252 Paris, Cedex 05, France
{elodie.duprat, windy.luscap, feriel.skouri, stephanie.finet}@impmc.upmc.fr

Keywords Protein superfamily, duplication events, evolution, supervised clustering, feature selection, sequence-structure-function relationships.

1 Background and Aims

The beta- and gamma-crystallins are the major components of the vertebrate eye lens; these long-life and unusually stable proteins account for the lens transparency [1].

These proteins share two conserved structural domains, generated by duplication. Each domain comprises two Greek key motifs, and belongs to the same structural family. However, these proteins differ by their structural assembly: the gamma-crystallins are monomers, whereas the beta-crystallins are oligomers.

Structural homologues of this family were identified in all life kingdoms. These non-lens proteins share a various number of beta/gamma-like domains (at least one), which probably arose from a common ancestor, by multiple duplication events [2].

These proteins together form the beta/gamma-crystallin superfamily. Only few information about their assembly state is available. Moreover, these proteins are involved in a wide variety of functions, which remain mainly unclear. Some members of this superfamily (bacteria and lower eukaryotes) are stress-related proteins, and main members share a calcium-binding site [3].

In this context, we aim to understand the evolution history of these proteins, and identify the determinants of their sequence-structure-function relationships. We present an explicative approach to detect sites and their physicochemical properties critical for beta- and gamma-crystallin assembly and domain evolution, based on aligned sequences.

This in silico approach provides prediction of interactions and deleterious mutations for new proteins of the superfamily, according to these critical features.

2 Methods

The beta/gamma-crystallin homologous protein sequences were retrieved from Uniprot (*http://www.uniprot.org*) by a similarity search approach based on the HMMER suite [4]. A profile Hidden Markov Model (HMM-profile) is built from an initial multiple sequence alignment (obtained by combination of sequence and 3D structure data), calibrated, and further used to detect similar sequences, split them into regions (N-ter, linker, C-ter, and homologous domains), and add their domain sequences to the initial multiple sequence alignment.

We detect 256 complete protein sequences which belong to the beta/gamma-crystallin superfamily (94 beta-crystallins, 112 gamma-crystallins, 50 non-lens proteins), corresponding to 527 domains. We then use PhyML [5] to reconstruct the phylogenetic tree of these domains, and the one of the two-domain superfamily members.

We apply a supervised clustering method (based on [6]) on these aligned sequences; this method includes two steps: (1) selection of discriminant binary features according to the mutual information criteria [7], each feature associating an alignment position with an amino acid group, and (2) learning of the classifier by estimating the frequencies of selected features, conditionally to the assembly state (monomer or oligomer) or

the domain orthologous group. The most relevant features are predicted to be the structural determinants of assembly or function.

We analyze the available 3D structures of beta- and gamma-crystallins (8 beta, 25 gamma) in order to determine the atomic contacts between the two homologous domains (D1 and D2), and the solvent accessible surface (SAS) of each domain residue. The package CCP4 [8] was used to detect the atomic contacts; the distance cut-off was set to 3.65 Å for all contacts, and the hydrogen bonds were detected according to angle and atom type. The residue SAS was computed by NACCESS [9].

3 Results

We identify 32 sites (15 beta-specific, 17 gamma-specific), whose amino acid composition determines the protein assembly (monomer or oligomer). We also identify 6 specific sites and physicochemical properties determining the domain orthologous group.

These sites are mainly located out of the domain interface. For most of these amino acid sites, their SAS significantly differs between beta- and gamma-crystallins, or between D1 and D2 domains. These features are useful to understand and predict the sequence-structure-function relationships of the beta/gamma-crystallin superfamily members.

References

[1] H. Bloemendal, W. de Jong, R. Jaenicke, N.H. Lubsen, C. Slingsby and A. Tardieu, Ageing and vision: structure, stability and function of lens crystallins. *Prog. Biophys. Mol. Biol.*, 86:407-485, 2004.

[2] G. Kappé, A.G. Purkiss, S.T. van Genesen, C. Slingsby and N.H. Lubsen, Explosive expansion of $\beta\gamma$-crystallin genes in the ancestral vertebrate. *J. Mol. Evol.*, 71:219-230, 2010.

[3] P. Aravind, A. Mishra, S.K. Suman, M.K. Jobby, R. Sankaranarayanan and Y. Sharma, The $\beta\gamma$-crystallin superfamily contains a universal motif for binding calcium. *Biochemistry*, 48:12180-12190, 2009.

[4] S.R. Eddy, Profile hidden Markov models. *Bioinformatics*, 14:755-763, 1998.

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

[6] E. Duprat, M.-P. Lefranc and O. Gascuel, A simple method to predict protein-binding from aligned sequences – application to MHC superfamily and beta2-microglobulin. *Bioinformatics*, 22:453-459, 2006.

[7] T.M. Cover and J.A. Thomas, Elements of information theory. John Wiley and Sons, New York, 1991.

[8] Collaborative Computational Project, The CCP4 Suite: Programs for Protein Crystallography. *Acta Cryst.* D50:760-763, 1994.

[9] S.J. Hubbard and J.M. Thornton, NACCESS, Computer Program, Department of Biochemistry and Molecular Biology, University College London, 1993.

Enzyme Classification Using 3D Signatures of Protein Binding Sites

Ahmed EL HAMADI [1], Joël MOUTOUSSAMY[2], Edwin CARLINET[3] AND Jean-Yves TROSSET[1]

[1] LABORATOIRE, BIRL Sup'Biotech, 66, rue Guy Môquet, 94800, Villejuif, France
el.hamadi.ahmed@gmail.com, jean-yves.trosset@supbiotech.fr

[2] RIABILITY, Chatenay Malabry, France
info@riability.fr

[3] LABORATOIRE, LRDE EPITA, 14-16, rue Voltaire, 94276 Le Kremlin-Bicêtre, France
edwin.carlinet@gmail.com

Abstract *This article presents a structural approach to reveal similarities between evolutionary distant enzymes. As 3D shape of enzyme's binding site is more closely related to function than sequences, we describe the shape of the binding site using chemical structural features (SCF) surrounding the ligand. Pairwise structural alignment of these SCF using graph matching algorithm can be used to compare proteins. Classification and structural phylogeny reconstruction using the enzyme distance matrix can reveal similarity that could not be seen from sequence alignment. A proof of concept of the approach is presented in this paper using 16 kinases selected from different branches of the human kinome phylogenic tree plus 2 plasmodium kinases. The scope of this study is to identify chemical class of human kinases inhibitors to discover new anti-malarial kinases inhibitors.*

Keywords Enzyme classification, structural chemical features, kinase, plasmodium, MedSUMO.

1 Introduction

In this paper, we present a proof of concept to classify enzymes of a given family based on the 3D shape of their binding site. The 3D shape signature is captured by Structural Chemical Features (SCF) of the protein atoms around the ligand [1]. Common SCF between two proteins are used to make a 3D superimposition of enzyme binding sites. Quality of alignment is measured by the percentage of common SCF between the two proteins, similarly to sequence alignment approach. Such structural comparison reveals similarities between evolutionary distant enzymes that remain unseen from protein sequence alignment approach.

This classification will be useful in the scope of studying cross-reactivity of a given class of inhibitors within the members of a protein target family or to cluster enzymes according to their "similarity" of functions. As a case study, we use a small set of 16 human kinases cherry picked from each of the main branches of the human kinome phylogenic tree. Two additional kinases from *Plasmodium falciparum* were included to investigate relationship with human kinome. We used the Med-SuMo protein surface comparison software to make a pairwise structure alignment of this set of kinase. The resulting distance matrices from both the ClustalW multi- sequence alignment and the SCF Med-SuMo SCF alignment are processed for cluster distance analysis [2,3].

We show that the enzyme classification based on structural shape of the binding site reveal different neighborhood profile as the one deduced from sequence similarity.

2 Materiel and Methods

Med-SuMo Structural alignment of the 18 kinases was carried out using the MED-SMA server structures. This was compared with the ClustalW sequence alignment from Pipeline Pilot (Accelrys). Distance matrices from both approaches, were analyzed using the statistical R package.

3 Results

The comparison between the protein sequence and SCF Med-SuMo alignment is presented in Figures 1 and 2.

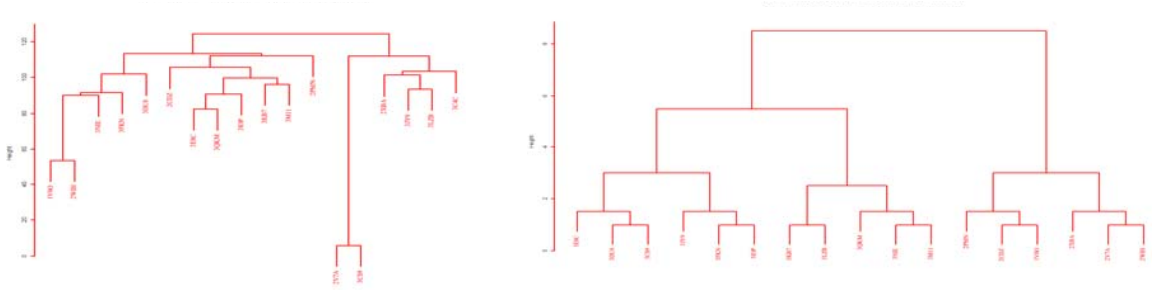


Figure 1. Dendogram of 18 kinases. ClustalW sequence alignment (left). Med-SuMo structural chemical features alignment (right).

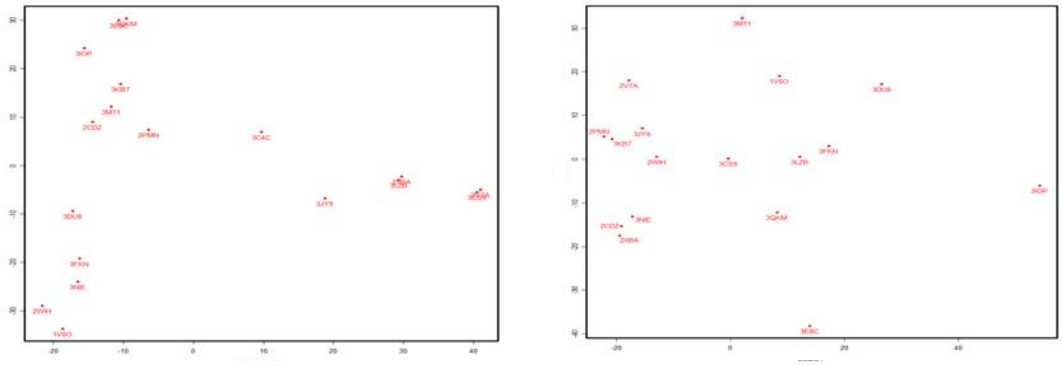


Figure 2. Distance scaling clusters of 18 kinases. ClustalW sequence alignment (left). Med-SuMo structural chemical features alignment (right).

4 Conclusion

This paper shows a proof of concept to use the SCF Med-SuMo features to classify enzyme based on the 3D shape of the binding site. This gives a similarity landscape different from the one derived from sequences only. This approach is currently used to select chemical classes from human anti-kinase inhibitors that are potential candidates for related plasmodium kinase as structural 3D shape binding site is concerned.

Acknowledgements

The authors are indebted to Olivia Doppelt-Azeroual and François Delfaud from MEDIT SA, for their help in the MEDP-SiteClassifier server configuration and for their useful comments on this study.

References

[1] M. Jambon, A. Imberty, G. Deléage and C. Geourjon. A new bioinformatics approach to detect Common 3D Sites in Protein Structures. *Protein Structure Function and Genetics*, 52: 52: 137-145, 2003.

[2] O. Doppelt-Azeroual, F. Delfaud, F. Moriaud and A.G. de Brevern, Classification of binding sites with MED-SuMo Multi approach: an application on Purinome, *Protein Science* 19(4):847-67, 2010.

[3] O. Doppelt-Azeroual, F. Moriaud, F. Delfaud and A.G. de Brevern, Analysis of HSP90 related folds with MED-SuMo classification approach. *Drug Design, Development and Therapy* 3:59–72, 2009.

Protein Structure Prediction with a Half Coarse Grained Model and Empirical Functions

Tristan BITARD-FEILDEL[1,2], Antoine VIGNERON[3] and Jean-François GIBRAT[1]

[1] Mathématique, Informatique et Génome, INRA, Domaine de Vilvert, 78352 Jouy en Josas, Cedex France
{jean-francois.gibrat, tristan.bitardfeildel}@jouy.inra.fr
[2] Mathématiques Informatique Appliquées, INRA, Domaine de Vilvert, 78352 Jouy en Josas, Cedex France
[3] Geometric Modeling and Scientific Visualization Center, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia
antoine.vigneron@kaust.edu.sa

Keywords Coarse Grain, Protein Prediction, Empirical Energy Function.

1 Introduction

Ab initio protein structure prediction is one of the main approach to protein prediction. Ab initio methods consist of a protein model, an energy function, and an algorithm for searching the conformational space. In particular, these methods do not use sequence comparison information.

The energy function applied to all the possible conformations given by a protein model defines the potential energy surface (PES). The goal of ab initio structure prediction is to find, with an appropriate algorithm, the most statistically probable conformation of a query molecule in the PES.

In this paper, we present an ab initio methodology based on a new coarse grained model, an empirical energy function, and a simulated annealing algorithm for exploring the conformational space.

2 Protein Model

The choice of a protein model has a direct incidence upon the number of variables, the energy function and, thus, the running time of the prediction algorithm. In the literature, many coarse grained models have been proposed, more or less representative of the physico-chemical properties of the residues. We developed a new coarse grained model with an all atomic description of the protein backbone and a coarse grained simplification of the amino acid side chain for non aromatic residues. For the latter, an all atomic description is used to enforce planarity.

Our model uses internal variables, i.e. valence angles, dihedral angles and bond lengths to describe the protein conformation. Bond lengths and valence angles are kept constant, and therefore, the only variables are the dihedral angles. A protein conformation C_i is thus defined by a sequence of dihedral angles :

$$C_i = \left((\phi_1 \dots \phi_n)_i, (\psi_1 \dots \psi_n)_i, (\chi_1 \dots \chi_m)_i\right).$$

A difficulty with internal variables is to perform appropriate moves to explore the PES: a small variation in a dihedral angle may change the whole conformation dramatically.

3 Empirical Energy Function

The energy function is a critical step to define the PES and the sets of solutions corresponding to local or/and global minima. We defined a simple and realistic energy function using five terms: a solvent accessible surface area (SASA) term for hydrophobic and hydrophilic atoms, a contact term to avoid steric clashes and to favor contact between atoms, a torsion term to mimic the Ramachandran distribution, and a term for secondary structure elements (SSE) so as to favor their construction:

$$E(C_i) = E_{sasa}(C_i) + E_{sasa_{H_2O^+}}(C_i) + E_{contact}(C_i) + E_{torsion}(C_i) + E_{SSE}(C_i) \quad (1)$$

SASA Potential
We approximate spherical atoms by geodesic domes. Each geodesic dome has a fixed number of vertices. Estimating the accessible surface of an atom reduces to counting the number of its vertices of that are not covered by any other atom. The SASA score is the sum of the surface areas of all the atoms. We defined two scores, one for hydrophobic atoms and another one for hydrophilic atoms. The SASA of hydrophobic atoms has to be minimized since hydrophobic atoms are supposed to be buried inside the proteins. By contrast, the SASA of hydrophilic atoms has to be maximized, so as to increase the surface of interaction with water molecules.

Contact Potential
We implemented a pseudo van der Waals potential based on distances observed in real three-dimensional structures.

Torsional Potential
It is well known that dihedral angles ϕ and ψ are not uniformly distributed between $-\pi$ and π: the residue angle pairs (ϕ, ψ) have preferred locations in the Ramachandran plot. To take this into account, we compute for each residue a torsional potential that depends on its location in the Ramachandran plot, based on two-dimensional Gaussian distributions.

SSE Potential
In order to favor the formation of SSEs, we used two terms in our function, one for α-helixes and the other for β-sheets formation. Each term consists of a sum of Gaussian scores based on atomics distances [1] and/or angles. For the α-helixes, we use local C_α distances and for β-sheets, we use the H-bond distances, and angles involving the CO groups and the N atoms on opposite strands.

4 Conformational Space Search

The exploration of the PES is the third crucial element for protein structure prediction. Our method uses a simulated annealing algorithm to locate the most probable structure corresponding to a sequence query. Two types of moves are allowed : global and local moves.

Global Move. This procedure updates a (ϕ, ψ) pair from a randomly chosen amino acid. The update can be an addition or subtraction of $\Delta_{dihedral}$, or it can be a (ϕ, ψ) pair drawn at random from a residue-dependent representative distribution. As a result, all the positions in the chain after the selected amino acid will be modified.

Local Move. We use a concerted rotation algorithm [2] to locally update the dihedral angles of four successive residues. Atomics positions before and after these residues do not change.

5 Discussion

Our first results regarding the prediction of small α-protein (1LP1, 1GYZ) 3D-structure are encouraging. We obtain topologically similar structures with an RMSD around 5 Å. We must now carry out simulation of proteins containing β strands.

Acknowledgements

This work was supported by the MIA department of INRA.

References

[1] J. Martin, G. Letellier, A. Marin, J. F. Taly, A. G. de Brevern and J. F. Gibrat, Protein secondary structure assignment revisited: a detailed analysis of different assignment methods. *BMC Struct. Biol*, 5:17, 2005.
[2] L. R. Dodd, T. D. Boone and D. N. Theodorou, A concerted rotation algorithm for atomistic Monte Carlo simulation of polymer melts and glasses. *Molecular Physics*, 4:961-996 ,1993.

Can Aspecific Docking Predict Protein-Protein Binding Sites ?

Juliette MARTIN[1]

Institut de Biologie et Chimie des Protéines, Bases Moléculaires et Structurales des Systèmes infectieux UMR5086 CNRS, Université de Lyon 7 passage du Vercors, 69367 Lyon, Cedex 07, France
juliette.martin@ibcp.fr

Keywords Protein-protein interaction, binding site prediction, docking, GPU.

1 Introduction

We are interested in the prediction of binding sites from the structures. A previous study of cross-docking using 6 protein-protein complexes revealed a clear tendency of (presumably) not interacting proteins to use their native binding sites to form the complexes [1]. In this contribution, we further investigate the capability of aspecific docking, i.e., docking of a candidate protein with random partners, to predict the residues involved in the binding site.

2 Material and methods

2.1 Data Sets

We use the version 4.0 of the docking benchmark data set assembled by Hwang et al [2]. This set consists of 176 protein-protein complexes, for which structures are available in both bound and unbound forms. After removal of antibody-antigen complexes, our benchmark data set encompass 299 proteins. Only the unbound forms of the proteins are used in this study.

Random partners are taken from Nh3D [3], a data set of representative structures of each topology of the CATH structural classification database. We removed structures with gaps, high radius of gyration, and structures classified in the same CATH topology as the benchmark proteins.

2.2 Docking

We used the Hex software [4], version 6.3, accelerated on GPU, with following parameters: grid size=0.6Å 13 and 25 correlations used in the *scan* and *search* steps respectively. Each docking experiment takes between 20s to 1 minute.

2.3 Prediction Assessment

A residue belongs to the binding site if it is less than 5Å away from the interacting chain. Prediction is assessed on a per-residue basis using :

- sensibility, i.e., the fraction of the real binding that is recovered by the prediction,
- specificity, i.e., the fraction of the predicted binding site that is correctly predicted
- the F1 measure, given by

$$\frac{2 \times sensibility \times specificity}{sensibility + specificity}$$

3 Results

We present here preliminary results obtained by docking each benchmark protein to 10 random partners with length between 25 and 75 residues. For each docking experiment, we extract a predicted binding site. For a given benchmark protein, we then compute a "naive" consensus by majority voting. The size is chosen according to the mean predicted size in the 10 experiments. We compare our results to the prediction given by Meta-PPISP, a meta-server making a consensus prediction from three other predictors. It is currently one of the top-performing methods for binding site prediction [5]. Results are presented in Figure 1, for the 56 proteins for which Meta-PPISP returned a prediction at the time of the submission. Overall, the consensus prediction by aspecific docking yields higher or F1 measure for 25 out of 56 proteins when compared to Meta-PPISP. It is worth to note that the apparent difficulty of the prediction differs between the two methods: some binding sites are well predicted by aspecific docking and not by meta-PPISP, and inversely.

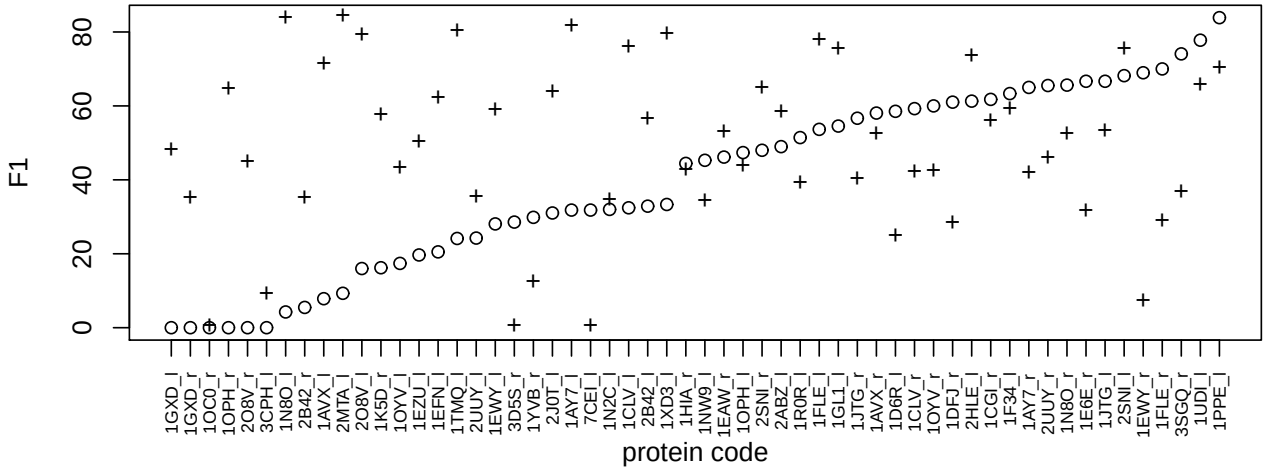


Figure 1. F1 measures on 56 proteins from the benchmark data set. Circles denote the prediction using aspecific docking (consensus from 10 docking experiments using random partners in the length range 25-75) , and crosses, results obtained using Meta-PPISP.

4 Conclusion

These preliminary results show that aspecific docking can be used as a predictive tool for protein binding site from unbound structures We plan to improve the prediction using two directions: (i) the choice of random partners and (ii) a more sophisticated way to build consensus prediction from the set of aspecific docking experiments.

References

[1] S. Sacquin-Mora, A. Carbone and R. Lavery, Identification of protein interaction partners and protein-protein interaction sites. *J Mol Biol*, 382(5):1276-89, 2008.

[2] H. Hwang, T. Vreven, J. Janin and Z. Weng, Protein-protein docking benchmark version 4.0. *Proteins*, 78(15):3111-4, 2010.

[3] B. Thiruv, G. Quon, S.A. Saldanha and B. Steipe, Nh3D: a reference dataset of non-homologous protein structures. *BMC Struct Biol*, 5:12, 2005.

[4] D.W. Ritchie and V. Venkatraman, Ultra-Fast FFT Protein Docking On Graphics Processors. *Bioinformatics*, 26, 2398-2405, 2010.

[5] S.B. Qin and H.X. Zhou, Meta-PPISP: a meta web server for protein-protein interaction site prediction. *Bioinformatics*, 23, 3386-3387, 2007.

Analysis of Protein-protein Interactions at the Subdomain Level

Dirk STRATMANN[1,2], Nicolas PRUDHOMME[1], Maya CHLIOUI[1], Jananan Sylvestre PATHMANATHAN[1], Mathilde CARPENTIER[1,2] and Jacques CHOMILIER[1,2]

[1] IMPMC, UMR7590 CNRS, UPMC, Case courrier 115, 4 place Jussieu, 75252 Paris, Cedex 05, France
dirk.stratmann@impmc.upmc.fr
jacques.chomilier@impmc.jussieu.fr

[2] RPBS, Bâtiment Lamarck, 35 rue Hélène Brion, 75205 Paris, Cedex 13, France

Abstract *Globular proteins can be decomposed into subdomain structural fragments with tight ends. These "tightened end fragments (TEF)" play an important role for the fold of a protein. We propose here an alternative approach for the TEF decomposition of a protein structure aiming at a better cover of the chain. We further investigate the role of the TEFs in protein-protein interactions and their use for binding site prediction.*

Keywords Structural bioinformatics, protein-protein complexes, closed loops, tightened end fragments, TEF.

The compactness of common globular protein folds requires that the amino acid polymer chain returns back to itself, forming "closed loops" [1], also named "tightened end fragments (TEF)" [2]. The two ends of a TEF are structurally nearby, typically less than 10 Å between Cα-atoms. These structural fragments have two remarkable properties: their length distribution shows a peak at about 25 amino acids [1] and their ends are mainly in the hydrophobic core [3] at highly conserved hydrophobic [4] sequence positions [2]. The TEFs that are "locked" by tight van der Waals interactions are also identified as "loop-n-lock" structures [5] which have been shown to be a universal basic unit of protein folds [6]. Globular proteins can therefore be considered as an assembly of these subdomain structural fragments.

Our current research explores the putative role of the "tightened end fragments (TEF)" in protein-protein interactions. Interactions between two globular proteins can often be reduced to a few "hot spots" or "core" residues [7]. The spatial arrangements of the hot spot residues, and their surrounding binding surface formed by the "rim" residues, are maintained by the global fold of the globular domain. Drugs, in form of small molecules, that can inhibit protein-protein interactions, are difficult to find, as they have to target the hot spot residues in the absence of a well-defined binding pocket [8]. In the search for new drugs, cyclized stable peptides are quite promising [9-11], as they can mimic the spatial arrangement of the important part of the binding site [12]. The decomposition of a protein into TEFs may help to identify a candidate peptide that can be stabilized by tight van der Waals interactions at its close ends. Beside this pharmaceutical application of TEF-decomposition, TEFs may also help to identify the binding sites in protein-protein interactions. To identify the role of TEFs in protein-protein interactions, we assembled a database of interacting TEF-pairs starting from dimer structures of the PDB [13]. A non-redundant set of homo-dimers was taken from the PiQSi database, where biological and crystallographic contacts have been differentiated by manual curation [14]. For hetero-dimers, we used the « Protein-protein docking benchmark 4.0 » [15] and assembled our own database starting from a list of hetero-dimers obtained by PISA [16]. We chose a non-redundant subset of the PISA-list and removed the special case [17] of antibody/antigen complexes. PDB structures with holes in the backbone coordinates were excluded from all databases used.

The decomposition of a protein structure into TEFs can be done in different ways. Our in-house decomposition program, available at the RPBS server [18], selects the TEFs with the tightest ends in terms of distance between Cα-atoms. This approach is also used by the DHcL server from Berezovsky et al. [19]. Some redundancy exists since one can find a TEF fully included in a longer one, and the best coverage is not taken into account in the algorithms. In order to improve the splitting of a domain into its constituting TEFs, we present here an alternative approach. It selects a distribution of TEFs that minimizes the number of residues unassigned to any TEF. For this task, we developed a complete search algorithm testing all combinations of TEFs yielding a decomposition of the protein structure, within limits in the amino acid length (between 15 and 50 by default).

Once the TEF decomposition is obtained on each chain of the dimer, we further classify the TEF pairs according to various structural and functional characteristics. The structural classification cannot be done on simple RMSD based calculations, as the lengths of the TEFs are in general different. Therefore, we developed topological descriptors of TEF conformations. A coarse-grained description of a TEF conformation is the height and the radius of its enveloping cylinder. It allows the comparison of TEF of various lengths and has proven to be successful in the structural classification of loops [20]. We are currently testing these and other descriptors. As this work is still in progress, the latest results from the classification of the TEF pairs will be shown at the conference as well as its use for binding site prediction.

References

[1] I. N. Berezovsky, A. Y. Grosberg, and E. N. Trifonov, Closed loops of nearly standard size: common basic element of protein structure. *FEBS Letters*, 466:283-286, 2000.

[2] M. Lamarine, J. P. Mornon, N. Berezovsky and J. Chomilier, Distribution of tightened end fragments of globular proteins statistically matches that of topohydrophobic positions: towards an efficient punctuation of protein folding? *Cellular and Molecular Life Sciences: CMLS*, 58:492-498, 2001.

[3] I. N. Berezovsky, V. M. Kirzhner, A. Kirzhner and E. N. Trifonov, Protein folding: looping from hydrophobic nuclei. *Proteins*, 45:346-350, 2001.

[4] A. Poupon and J. P. Mornon, Populations of hydrophobic amino acids within protein globular domains: identification of conserved "topohydrophobic" positions. *Proteins*, 33:329-342, 1998.

[5] I. N. Berezovsky and E. N. Trifonov, Van der Waals locks: loop-n-lock structure of globular proteins. *Journal of Molecular Biology*, 307:1419-1426, 2001.

[6] I. N. Berezovsky, Discrete structure of van der Waals domains in globular proteins. *Protein Engineering*, 16:161-167, 2003.

[7] J. Janin, R. P. Bahadur and P. Chakrabarti, Protein-protein interaction and quaternary structure. *Quarterly Reviews of Biophysics*, 41:133-180, 2008.

[8] J. A. Wells and C. L. McClendon, Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature*, 450:1001-1009, 2007.

[9] C. J. Capini, S. M. Bertin-Maghit, N. Bessis, P. M. Haumont, E.M. Bernier, E.G. Muel, M.A. Laborie, L. Autin, S. Paturance, J. Chomilier, M. C. Boissier, J. P. Briand, S. Muller, J. M. Cavaillon, A. Therwath and J. F. Zagury, Active immunization against murine TNFalpha peptides in mice: generation of endogenous antibodies cross-reacting with the native cytokine and in vivo protection. *Vaccine*, 22:3144-3153, 2004.

[10] S. M. Bertin-Maghit, C. J. Capini, N. Bessis, J. Chomilier, S. Muller, A. Abbas, L. Autin, J. L. Spadoni, J. Rappaport, A. Therwath, M. C. Boissier and J. F. Zagury, Improvement of collagen-induced arthritis by active immunization against murine IL-1beta peptides designed by molecular modeling. *Vaccine*, 23:4228-4235, 2005.

[11] P. Vlieghe, V. Lisowski, J. Martinez and M. Khrestchatisky, Synthetic therapeutic peptides: science and market. *Drug Discovery Today*, 15:40-56, 2010.

[12] N. London, B. Raveh, D. Movshovitz-Attias and O. Schueler-Furman, Can self-inhibitory peptides be derived from the interfaces of globular protein-protein interactions?. *Proteins*, 78:3140-3149, 2010.

[13] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28:235 -242, 2000.

[14] E. D. Levy, PiQSi: protein quaternary structure investigation. *Structure (London, England: 1993)*, 15:1364-1367, 2007.

[15] H. Hwang, T. Vreven, J. Janin and Z. Weng, Protein-protein docking benchmark version 4.0. *Proteins*, 78:3111-3114, 2010.

[16] E. Krissinel and K. Henrick, Inference of macromolecular assemblies from crystalline state. *Journal of Molecular Biology*, 372:774-797, 2007.

[17] S. J. de Vries and A. M. J. J. Bonvin, How proteins get in touch: interface prediction in the study of biomolecular complexes. *Current Protein & Peptide Science*, 9:394-406, 2008.

[18] C. Alland, F. Moreews, D. Boens, M. Carpentier, S. Chiusa, M. Lonquety, N. Renault, Y. Wong, H. Cantalloube, J. Chomilier, J. Hochez, J. Pothier, B. O. Villoutreix, J. F. Zagury and P. Tufféry, RPBS: a web resource for structural bioinformatics. *Nucleic Acids Research*, 33:W44-49, 2005.

[19] G. Koczyk and I. N. Berezovsky, Domain Hierarchy and closed Loops (DHcL): a server for exploring hierarchy of protein domain structure. *Nucleic Acids Research*, 36:W239-245, 2008.

[20] J. Wojcik, J. P. Mornon and J. Chomilier, New efficient statistical sequence-dependent structure prediction of short to medium-sized protein loops based on an exhaustive loop classification. *Journal of Molecular Biology*, 289:1469-1490, 1999.

Graph Clustering Analysis of a Boolean Model

Case of Iron Homeostasis in *Saccharomyces cerevisiae*

Mickael NERI[1], Jean Michel CAMADRO[1] and Denis MESTIVIER[1]

Institut Jacques Monod, UMR7592 CNRS, 15 rue Hélène Brion, Université Paris Diderot , 75205 Paris, Cedex 13, France

neri@ijm.univ-paris-diderot.fr

Keywords Graphs clustering, stochastic boolean networks, iron, homeostasis.

Analyse d'un Modèle Booléen par Classification de Graphe

Cas de l'Homéostasie du Fer chez *Saccharomyces cerevisiae*

Mots-clés Classification de graphes, réseaux booléens stochastiques, fer, homéostasie.

Le fer est un élément essentiel à l'Homme. C'est un co-facteur de certaines enzymes métaboliques et il est impliqué dans de nombreuses voies essentielles, telles que la respiration cellulaire, la synthèse d'ADN, le cycle de Krebs, le transport du dioxygène ou la protection contre le stress oxidatif [1,2,3,4,5]. Cependant, bien qu'étant nécessaire, le fer est également toxique pour l'organisme. En effet, le fer ferreux (Fe^{2+}) est capable de réagir avec l'eau oxygénée (également produite par le fer en présence de dioxygène) pour produire un radical hydroxyle, une espèce réactive de l'oxygène très toxique. De plus, le fer ferrique (Fe^{3+}) peut avoir tendance à s'accumuler. Pour ces raisons et son implication dans un grand nombre de voies métaboliques, le fer est impliqué dans de nombreuses maladies : anémies [6], hémochromatoses [1,5] ou des maladies neurodégénératives telles qu'Alzheimer, Parkinson et l'ataxie de Friedreich [7].

Cette dernière maladie est causée par le dysfonctionnement d'une petite protéine de la matrice mitochondriale, la frataxine, qui intervient de manière directe ou indirecte dans l'assemblage des centres Fer-Soufre, la résistance des cellules au stress oxydant, la synthèse d'hème et l'homéostasie mitochondriale du fer. Le rôle précis de cette protéine est encore controversé, et, dans ce contexte, un modèle de l'homéostasie du fer permet alors d'étudier les conséquences sur la physiologie cellulaire d'une dérégulation de cette protéine.

L'équipe ≪ Modelisation en Biologie Intégrative ≫ de l'Institut Jacques Monod, a développé un modèle Booléen stochastique de l'homéostasie du Fer chez la levure *Saccharomyces cerevisiae* [8]. Ce modèle, composé de 643 éléments et 1029 règles, a été validé par simulations et confrontation avec 147 mutants phénotypiques (91% de consistence) et 11 expériences de flux métaboliques (10/11 des résultats d'expériences sont reproduits avec le modèle).

Dans ce travail, nous étudions le réseau d'intéractions du modèle sous l'angle d'un graphe, où les noeuds correspondent à des espèces ou à des règles (graphe bi-parti) et les arêtes représentent les liens entre ces espèces et règles. Le graphe comporte 1672 noeuds et 3369 arêtes. Cependant, sa grande taille rend son analyse difficile.

Nous proposons d'analyser le graphe en utilisant une approche récente de classification de graphe (MCL, Markov Cluster Algorithm [9]), qui permet de découper un graphe en *modules*, qui sont des sous-parties plus facilement analysables. Cet algorithme a été utilisé dans plusieurs situations d'analyse de graphes [10,11], d'identification de modules fonctionnels dans des réseaux d'interactions de protéines [12] ou encore pour la classification de documents dans PubMed [13].

Nous avons utilisé MCL dans une nouvelle configuration peu exploitée dans les travaux précédemment mentionnés. En effet, MCL peut utiliser un poids sur chacune des arêtes du graphe lorsqu'il découpe le graphe. Nous avons déterminé les poids des arêtes par simulation du modèle booléen stochastique en estimant la fréquence d'occurence des règles durant la simulation. En modifiant un paramètre du modèle, nous pouvons mimer une dépletion jusqu'à l'excès en frataxine, et obtenir différents jeux de poids des arêtes. Il s'agit donc d'une

situation où l'on dispose d'un graphe d'interaction, *statique*, sur lequel on rajoute une information caractérisant la *dynamique* du réseau correspondant.

L'ajout de poids obtenus par simulation fait passer le nombre de modules obtenus après classification de 388 à 221, dont 106 d'au moins 5 composants. Ces modules peuvent être représentés comme des sous-graphes de celui du modèle, et apparentés à des motifs, reconnaissables.

Nous avons ainsi modélisé des déficiences graduelles en frataxine, et montré qu'il existe un seuil à partir duquel la production de cette protéine perturbe l'organisation du graphe. Les modules ainsi affectés nous permettent d'identifier des mécanimes de régulation cellulaire qui seraient impliqués dans la réponse à une déficience en frataxine.

Références

[1] B.R. Bacon and R. Britton, Clinical penetrance of hereditary hemochromatosis. *The New England Journal of Medicine*, 358 :1533-4406, 2008.

[2] G. Anderson and C.D. Vulpe, Mammalian iron transport. *Cellular and Molecular Life Sciences : CMLS*, 66 :1420-9071, 2009.

[3] C. Camaschella and P. Strati, Recent advances in iron metabolism and related disorders. *Internal and Emergency Medicine*, 1970-9366, 2010.

[4] N.C. Andrews, Forging a field : the golden age of iron biology. *Blood*, 112 :1528-0020, 2008.

[5] R.R. Crichton, *Iron Metabolism : From Molecular Mechanisms to Clinical Consequences*, John Wiley and Sons, 2009.

[6] S.F. Clark, Iron deficiency anemia : diagnosis and management. *Current Opinion in Gastroenterology*, 25 :1531-7056, 2009.

[7] S. Altamura and M.U. Muckenthaler, Iron toxicity in diseases of aging : Alzheimer's disease, Parkinson's disease and atherosclerosis. *Journal of Alzheimer's Disease : JAD*, 16, 2009.

[8] F.Achcar, J.M. Camadro and D. Mestivier, A Boolean probabilistic model of metabolic adaptation to oxygen in relation to iron homeostasis and oxidative stress. *BMC Systems Biology*, 5 :51, 2011.

[9] S. van Dongen, *A New Cluster Algorithm for Graphs*, National Research Institute for Mathematics and Computer Science, 1998.

[10] L. Zinger, E. Coissac, P. Choler and R.A. Geremia, Assessment of microbial communities by graph partitioning in a study of soil fungi in two Alpine meadows. *Applied and Environmental Microbiology*, 75 :1098-5336, 2009.

[11] B.S. Lattimore, S. van Dongen, M. James and C .Crabbe, GeneMCL in microarray analysis. *Computational Biology and Chemistry*, 29 :1476-9271, 2005.

[12] N. Sohaee and C.V. Forst, Identification of functional modules in a PPI network by bounded diameter clustering. *Journal of Bioinformatics and Computational Biology*, 8 :0219-7200, 2010.

[13] T. Theodosiou, N. Darzentas, L. Angelis and C.A. Ouzounis, PuReD-MCL : a graph-based PubMed document clustering methodology. *Bioinformatics (Oxford, England)*, 24 :1367-4811, 2008.

Study of Statistical Parameters to Perform a Convenient Prediction of Different Endocrine Phenotypes in Sportsmen Based on Metabonomic Data

Alain PARIS[1], Boris LABRADOR[2], François-Xavier LEJEUNE[1], Aziz ZOUBAÏ[3], Cécile CANLET[3], Jérôme MOLINA[3], Michel GUINOT[4], Armand MEGRET[5], Jean-Christophe THALABARD[6], Michel RIEU[7] and Yves LE BOUC[2]

[1] Unité Mét@risk, INRA, 16, rue Claude Bernard, 75231 Paris, Cedex 05, France
aparis@paris.inra.fr

[2] INSERM U 938, Centre de Recherche Saint-Antoine, Hôpital Saint-Antoine, Bâtiment Kourilsky, 184, rue du Faubourg Saint-Antoine, 75012 Paris, France

[3] UMR1089 – Xénobiotiques, INRA, 180, chemin de Tournefeuille, BP 93173, 31027 Toulouse, Cedex 3, France

[4] Médecine du Sport, Antenne Médicale de Prévention du dopage Rhône Alpes, Pôle Rééducation et Physiologie, Hôpital Sud, Grenoble, France

[5] FFC, 5, rue de Rome, 93561 Rosny sous Bois, France

[6] Unité Gynécologie-Endocrinologie, Hôtel-Dieu, Assistance Publique-Hôpitaux de Paris, 1, place du Parvis de Notre-Dame, 75004 Paris, France

[7] Agence française de lutte contre le dopage, 229, boulevard Saint-Germain, 75007 Paris, France

Keywords Metabonomics, discrimination, multivariate statistics, prediction.

1 Background

The longitudinal endocrine follow-up of sportsmen achieved by conventional methods allows detection of clinical abnormalities that may be related to some prohibited doping practices. Indeed, some disturbed physical performance can be explained by atypical physiological deviations. However, recent events revealing doping cases have shown some limitations of the principles of anti-doping control currently prevailing in the establishment of doping practices.

Screening of pharmacological substances, which use is prohibited (and some of them are often designed to improve athletic performance), basically consists in detecting presence in urine or blood of these compounds or their metabolites. To do so, the direct detection of doping involves very sophisticated physicochemical methods. Nevertheless, these methods are expensive. In addition, they specifically target known molecules.

Yet, an alternative to this direct strategy is to measure in the serum concentration of circulating endogenous hormones or their metabolites in order to get a hormonal fingerprint of subjects that may give indirect proofs of doping practices.

In the case of endogenous hormones is raised the problem of definition of what are normal concentrations and how to define clinical thresholds. Indeed, when a hormonal doping practice is used, homeostatic regulation may have some repercussion on the hormonal fingerprint. Observation of this hormonal anomaly is the first step in indirect detection of doping practice. In France, it is currently done in the frame of the medical longitudinal follow-up. Besides, these hormonal variations may induce some metabolic adjustments which can be detected in a global metabolic assessment in biofluids. This metabolic fingerprinting is called metabonomics.

In this context, statistical and computational approaches used by metabonomics may be helpful to solve such a problem designed as a numerical analysis of the multidimensional metabolic response when metabolic fingerprints, which correspond to the large quantification of the general metabolism of an organism, are used in complement of hormonal fingerprints. Since the last decade, metabonomics has been efficiently applied and developed in various biological domains including plant genotype discrimination, toxicological mechanisms, disease aetiology, and drug discovery among others. In our context, we aim at improving the ability to predict the endocrine phenotype of any individual based on metabolic fingerprints with the constraint of avoiding the risk of false-negative. This requires fine modelling of the relationship between

metabolic profiles and the endocrine status given by the determination of the hormonal concentration class for three hormones, *i.e.* cortisol, IGF-1 and testosterone.

2 Results

We have studied a cohort of 655 individuals described with 419 metabolites (variables) obtained by ^{1}H NMR (Nuclear Magnetic Resonance) spectrometry from a fingerprinting of serum. For the whole cohort, we also get in parallel three endocrine phenotypes (cortisol, IGF-1 and testosterone), for which 3 classes have been defined *a priori* for "low", "normal" and "high" concentrations. In the procedure presented here, we consider a classification method of any of these classes, which relies to the metabolic fingerprints.

The core method combines a data regularization step based on orthogonal signal correction to a shrinkage discriminant analysis (SDA) [1], which is well suited to deal with the multicollinearity carried out by the metabolites. Thus, in our situation, SDA outperforms other usual discriminant methods such as LDA, QDA and PLS-DA. However, for all these methods, it is noteworthy that classification is substantially improved when data are pre-processed using orthogonal signal correction based on partial least squares regression [2]. To improve the level of confidence on the prediction, assignment to a given class is then obtained using bootstrap techniques. Using bootstrap, we have also studied how the prediction rates vary depending on cohort size, choice of metabolites (variables) and phenotypes.

With the same protocol, we have displayed in abacus the prediction rates obtained with different sizes for the cohort together with increasing the number of selected metabolites. Thus, it can be observed that, for each phenotype, a well suited choice in parameter modelling is achievable to expect the highest rates of prediction for each given hormonal phenotype.

Extensive calculus in the resampling step also revealed special classes of individuals for which the classification systematically failed. Identification of such classes may indicate new further and distinct investigations to improve their hormonal characterization.

3 Conclusion

The procedure combining the orthogonal signal correction based on PLS regression and the shrinkage discriminant analysis is very promising to better detect endocrine disruptions from a metabolomic fingerprint. Abacus appear instructive to decide of how many statistical units we need to consider and therefore to control the cost of such experiments. It is also important to build decision rules from a cohort which is large enough with a selected set of metabolites. A quantitative analysis of changes in general metabolism performed in this physiological context may indirectly provide some tangible biochemical suspicion of doping. In other words, use of indirect methods to routinely phenotype the endocrine status of sportsmen from their metabolomic fingerprints is very promising to better detect endocrine disruptions.

Acknowledgements

This work was supported by AFLD (Agence Française de Lutte contre le Dopage) and WADA (World Anti-Doping Agency).

References

[1] M. Ahdesmäki and K. Strimmer, Feature selection in omics prediction problems using cat scores and false non-discovery rate control. *Ann. Appl. Stat.* 4:503-519, 2010.

[2] S. Wold, H. Antti, F. Lindgren and J. Ohman. Orthogonal signal correction of near-infrared spectra. *Chemometrics Intell. Lab. Syst.*, 44: 175-185, 1998.

A Web-Oriented Platform for Gene Regulatory Network Inference

Application to Seed Storage Proteins in Wheat

Jonathan VINCENT[1,2], Pierre MARTRE[1], Catherine RAVEL[1], Alexandre BAILLIF[1] and Marie AGIER[2]

[1] INRA - Université Blaise Pascal, UMR1095 GDEC, 234 Avenue du Brézet, Clermont-Ferrand, F-63 100, France

pierre.martre@clermont.inra.fr

[2] Université Blaise Pascal - CNRS, UMR 6158 LIMOS, BP 10 448, F-63 173 Aubière

agier@isima.fr (author for correspondence)

Abstract *The development of –omics methods has been followed by an explosion of the amount of data produced in functional genomics research projects. Retrieving biologically relevant information and knowledge on molecular regulations from such massive amount of data is therefore an important challenge. The work presented herein aims at elucidating gene regulatory networks from quantitative –omic data using rules discovery algorithms. This approach grants users the possibility to specify genes of interest and the rule discovery algorithm discovers interactions implying these genes. The user can choose between different semantics for the rules. Each semantics is relevant and has its own biological meaning, thus providing the user with global information. This work has led to the development of an efficient web-oriented platform. Eventually, the application was evaluated with the reconstruction of transcriptional networks involved in the regulation of seed storage protein genes for wheat.*

Keywords Gene regulatory networks, network inference, rules discovery algorithm, wheat seed storage proteins, transcription factors, transcriptional network.

1 Introduction

Several tools are now available to infer regulatory networks from transcriptomic data. Most of them are based on Boolean, relevance, or Bayesian networks or association rules [1]. Knowledge database such as String (string-db.org/) allowing the manipulation of gene networks have also been developed. However, most of these approaches do not take into account the quantitative nature of –omic data. Moreover, considering the complexity of eukaryotic organisms, it is essential to take into account more than one type of interaction, which is not the case in most of the available tools.

Wheat grain end-use value is mainly determined by its storage protein (SSP) composition [2]. SSP accumulate in grains from *ca.* 14 to 42 days after flowering. SSP genes are essentially regulated at the transcriptional level [3] and several transcription factors involved in this regulation have been identified [4].

In this work a web-based platform for regulatory network inference based on semantic rules [5] has been developed. The platform was evaluated with the reconstruction of transcriptional networks for the regulation of SSPs expression using temporal transcriptomic data.

2 Gene Regulatory Networks Inference Method

The approach consists in interacting with biologists in order to establish rule semantics that meets their research objectives and are adapted to their data. Currently, three semantics for gene expression data are available [5]. The first one generates rules between genes according to their expression levels. The second one generates rules between genes according to the variation of their expression levels. The third semantics generates rules between genes according to the evolution of their expression levels over time. Through this approach, it becomes possible to have a global view of interactions inferred from a dataset.

In this work, the method was implemented as a web-oriented platform. A simple process allows the user to load his own dataset onto the server and then to choose among the different semantics, according to the type of data. The dataset is then processed on the server's side and the results are loaded into a visualization platform (Fig. 1).

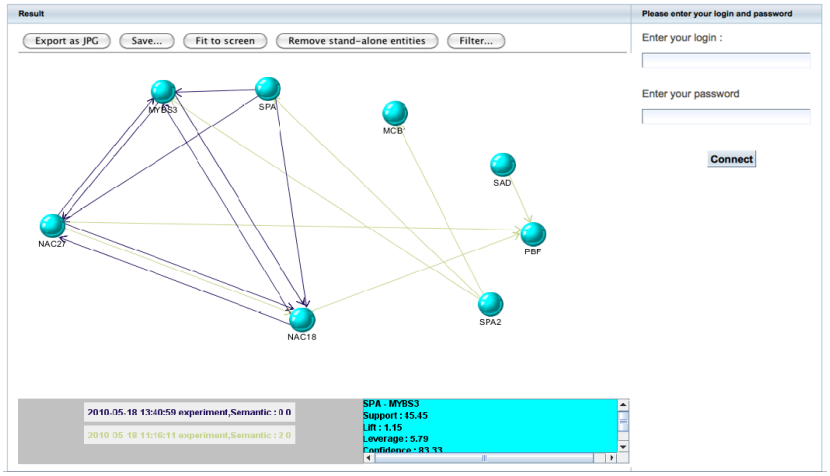


Figure 1. Screen capture of the application showing interactions between transcription factors putatively involved in the regulation of SSP expression for wheat. Different interaction colors correspond to different semantics.

3 Application to Seed Storage Proteins Transcriptional Network in Wheat

We used a temporal dataset comprising 11 data points covering the period of SSPs accumulation [6]. This allowed partially reconstructing a transcriptional network previously established for barley based on direct experimental evidences [4]. Based on this work several new putative interactions can be proposed.

4 Discussion

This work led to the integration of an existing method into a web oriented platform. Despite the low quantity of gene expression data used in this study, the method allowed reconstructing a well-characterized model of interactions implying nine transcription factors [4]. It also allowed us to propose new putative interactions between these genes. The results also highlighted two NAC transcription factors, not previously shown to participate in the regulation of SSP in cereals. Transcriptomic analysis has shown that the level of expression of one of these transcription factors is strongly influenced by sulfur supply, which is a major determinant of SSPs composition [7].

Acknowledgements

The authors thank J-B Perez for his contribution to the visualization platform. The authors also thank A. Besson and I. Romeuf for sharing some of the transcriptomic data used in this work.

References

[1] H. Hache, H. Lehrach, and R. Herwig, Reverse Engineering of Gene Regulatory Networks: A Comparative Study, EURASIP *Journal on Bioinformatics and Systems Biology*, vol.2009. doi:10.1155/2009/617281, 2009

[2] CW Wrigley and F. Békés, Glutenin-protein formation during the continuum from anthesis to processing. *Cereal Foods World* 44:562-565, 1999.

[3] J. Verdier and RD. Thompson, Transcriptional regulation of storage protein synthesis during dicotyledon seed filling. *Plant Cell Physiol.* 49:1263-1271, 2008.

[4] I. Rubio-Somoza, M. Martinez, Z. Abraham, I. Diaz and P. Carbonero, Ternary complex formation between HvMYBS3 and other factors involved in transcriptional control in barley seeds. *Plant J.* 47:269-281, 2006.

[5] M. Agier, JM. Petit and E. Suzuki, Unifying framework for rule semantics: Application to gene expression data. *Fundamenta Informaticae* 78:543-559, 2007.

[6] I. Romeuf, Identification in silico des facteurs de transcription du blé tendre (Triticum aestivum) et mise en évidence des facteurs de transcription impliqués dans la synthèse des protéines de réserve. *Ph.D. thesis,* Université Blaise Pascal, Clermont-Ferrand, France, 2010.

[7] H. Wieser, R. Guster and S. von Tucher, Influence of sulphur fertilisation on quantities and proportions of gluten protein types in wheat flour. *J. Cereal Sci.* 40:239-244, 2004.

A Software Architecture for *de Novo* Induction of Regulatory Networks from Expression Data

Frank RÜGHEIMER[12], Ashish ANAND[12] and Benno SCHWIKOWSKI[12]

[1] Institut Pasteur, Laboratoire de Biologie Systémique, Dept Génomes et Génétique, F-75015 Paris, France
frueghei@pasteur.fr, anand.ashish@gmail.com, benno@pasteur.fr
[2] CNRS, URA2171, F-75015 Paris, France

Keywords systems biology, regulatory networks, structure learning, bioinformatics tools

1 Introduction

Understanding the regulatory networks in cells to explain or even predict phenotypical effects has been one of the key goals of systems biology. Although extensive information is available for specific, well understood systems, that knowledge still covers only a fraction of the expected global regulatory interactions. Unfortunately, the huge number of potential regulatory structures limits purely computational approaches to the network learning task, as even significant numbers of large-scale transcriptomics experiments do not supply sufficient data to discriminate between all alternative solutions.

Here we present a modular architecture for the induction of biologically meaningful regulatory networks from quantitative data and their iterative refinement by new experiments (Section 2). By integrating the experiment selection with the computional framework we aim to maximize the benefits of conducted experiments with respect to their impact on searching the space of regulatory hypotheses. The approach is supported by tools that identify hypotheses consistent with observed data and propose experiments to further assess them.

An application of the proposed approach is presented in Section 3. In that application we use a set genes with potential regulatory function as identified by existing annotations. Similar annotations, such as those using the Gene Ontology standard [1], are now available for an increasing number of model organisms.

2 Global Search Strategy

The proposed strategy (Fig. 1) iterates between three phases: the induction of connectivity scores for individual edges from data (I), a search for likely regulatory hypotheses and the selection of suitable experiments (II), and the assessment of these hypotheses in direct experiments (III). During the first phase pairwise interaction measures are used to compute a connectivity score that assesses the plausibility of pertubations being transmitted along edges between pairs of regulators and targets. The second phase integrates these local evaluations into a global hypotheses of regulatory paths form the origin of a pertubation to the effectors. From superpositions of high scoring pathway hypotheses it is then possible to identify critical experiments that allow to distinguish between large subsets of hypotheses and test important putative links. Both phase (I) and phase (II) allow for the integration of external data via the selected connectivity measures and scoring functions. Results of phase (III) are fed back to the next interation, e.g. by penalizing links that failed validation. The same method can used to draw on prior knowledge from public databases if so desired.

3 Application Example and Software

We applied our approach to transcriptome data collected from *bacillus subtilis* grown in liquid culture. In these experiments a pertubation was caused by the addition of malate to a glucose-based medium. Expression change was subsequently detected in pathways previously thought to be unconnected to carbon metabolism.

First we identified genes involved in carbon metabolism, genetic regulation and affected pathways using a recently released functional annotations available via http://www.subtiwiki.uni-goettingen.de

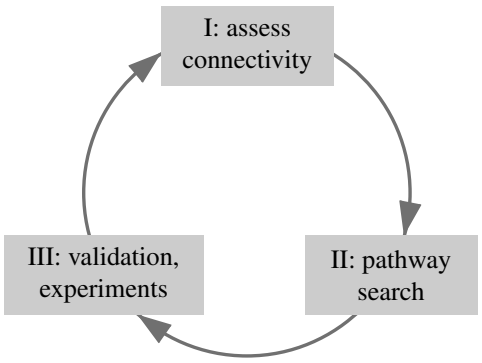


Figure 1. Three phase architecture for regulatory network induction

gene	score	gene	score	gene	score
glcR	0.006311	hxlR	0.006262	rapH	0.006081
comGA	0.006311	manP	0.006229	ybdK	0.005989
cggR	0.006311	yydG	0.006147	phrH	0.005803
kinD	0.006294	yydF	0.006147	yydI	0.005739
ccpA	0.006294	phoR	0.006147	comK	0.005711

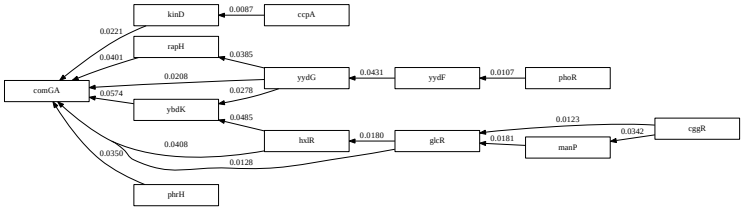


Figure 2. 15 highest scoring nodes and superposition of top 10 pathways for nutrient shift problem (extracted from 409 genes)

For the assessment of link connectivity (phase I) we use the GENIE3 algorithm [2] to induce edge weights from transcriptome data. For the second phase we implemented a path search strategy in the scoreKO command line tool. ScoreKO reads a list of weighted edges in a table format and can be configured to either directly report node assessments or to produce pathway reconstructions. The latter mode reports superpositions of all regulatory pathways up to a specified quality rank allowing to visualize critical players in the selection of regulatory hypotheses. Supplementary scripts that convert the program output for visualization and further processing in Cytoscape [3] are included in the source code archive.

The edge weights induced by GENIE3 were aggregated into pathway scores using the Hamacher product. Our implementation draws on the monotonicity property of that operator for efficient search. This monotonicity property is a natural requirement for any conjunctive aggregation function. Depending on the interpretation of edge weights the operator can be replaced, e.g. by other t-norms. As of version 1.2 our tool also supports the minimum and product as pathway scoring operators.

Results of the second phase (Figure 2) including proposed experiments are assessed in an ongoing collaboration with the Medical Microbiology group at the University of Groningen.

The command line-based pathway search and network induction program we developed has been made available via the the website http://www.ruegheimer.org/scoreKO. A complementary tool named findGenes (available from our software website http://proteomics.fr/Sysbio/Software) calculates interaction measures, which serve as input for scoreKO, from expression data. We are planning to provide plug-in versions of both tools for the upcoming version 3.0 of the Cytoscape software.

Acknowledgements

This work was supported by a grant of the European Union (FP6, BaSysBio, grant LSHG-CT-2006-037469)

References

[1] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock, Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

[2] V. A. Huynh-Thu, A. Irrthum, L. Wehenkel and P. Geurts, Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE*, 5(9), 2010.

[3] M. S. Cline, M. Smoot, E. Cerami, A. Kuchinsky, N. Landys, C. Workman, R. Christmas, I. Avila-Campilo, M. Creech, B. Gross, K. Hanspers, R. Isserlin, R. Kelley, S. Killcoyne, S. Lotia, S. Maere, J. Morris, K. Ono, V. Pavlovic, A. R. Pico, A. Vailaya, P.-L. Wang, A. Adler, B. R. Conklin, L. Hood, M. Kuiper, C. Sander, I. Schmulevich, B. Schwikowski, G. J. Warner, T. Ideker and G. D. Bader, Integration of biological networks and gene expression data using cytoscape. *Nat Protoc*, 2(10):2366–82, 2007, URL http://www.ncbi.nlm.nih.gov/pubmed/17947979.

Modeling Stochastic Switched Systems with BioRica

Rodrigo ASSAR, Alice GARCIA and David J. SHERMAN

INRIA[1] Bordeaux Sud-Ouest.

{rodrigo.assar, alice.garcia, david.sherman}@inria.fr

Abstract *Modeling physycal and biological dynamic systems needs to combine different types of models in a non-ambiguous way. We present an approach to integrate continuous, discrete, stochastic, deterministic and non-deterministic elements by using Transition Systems theory, reuse, composition of models, and the framework BioRica. The systems are described by interacting continuous and discrete models, and in addition continuous models are decomposed into two components: controlled and controller model. We define Stochastic Switched Systems whose continuous dynamics is modeled by differential equations and its discrete dynamics by transition systems, allowing stochastic and non-deterministic behaviours. We illustrated the use of our approach with examples of intrinsically and approximated hybrid systems. Our approach allows us to give a first step to integrate and to extend models of complex systems, such as cell differentiation.*

Keywords Hybrid Systems, Transition Systems, Switched Systems, Cell differentiation.

1 Introduction

[1] Physical phenomena often are described by combinations of different types of equations. These systems are called *Hybrid systems* ([1], [2]) due to the use of continuous and discrete features or *Switched systems* because they switch its equations over time ([3], [4], [5]). Switched systems are a way to introduce discrete behavior into continuous models. They are Hybrid systems with discipline.

In *dynamic models*, one considers two groups of variables: dependent and factors. Models give the dynamics of the dependent variables, considering factors affecting it. One talks of *continuous model* if the variables change continuously over time and it is relevant to know the behaviour at any time. Continuous models use functional relations between the variables, being a common type the differential equations. On the other hand, if the variables make discrete changes at instantaneous points in time, the model is *discrete*. *Hybrid models* join both types of models: some variables have continuous dynamics while other ones have discrete dynamics. These models allow the interaction of diverse components of a system to contribute to complete descriptions of the behaviour. Switched systems are one kind of Hybrid system that restricts way discreteness is added. Discrete variables modify the behaviour of the continuous model by controlling its coefficients.

Models try to accurately represent the reality, by using empirical observations and knowledge. With limited observations one wants to build models that are valid to explain the system in general conditions without testing it on all the conditions. As result, models are strongly dependent of the studied conditions. In order to get most valued out of existing models and to refine models to include more complex behaviours, it is necessary to define how to compose models. A system is built in a hierarchical way, composed of subsystems, where behaviours emerge from the association of components and its diversity ([6]). Complex biological processes can, in this way, be defined by interactions between basic functional entities called *modules* ([7]) and, to explain its behaviour, to each module is associated a model that represents it.

The existence of different types of models to explain connected processes makes it necessary to define theory and tools to integrate them. Such theory must be able to unify processes with different timescales and whose models have different stiffness levels. Sometimes, to see the changes in the behaviour it is necessary to compare nearby times, but other times the changes happen in distant times. Equations with high stiffness require

[1] Project-team (EPC) MAGNOME common to INRIA, CNRS, and U. Bordeaux 1, an EMR of UMR 5800 LaBRI

a superior level of discretization to obtain good approximations. Both characteristics, different timescales and stiffness affect computation times and accuracy. To compose models one must consider these characteristics, and be capable of giving an unique and non-ambiguous semantics to the composition. Modeling biological dynamic systems by composition requires a framework in which existing models of different types can be combined without need of rewriting them.

Many dynamic biological systems, such as physiological processes ([8]), are represented by ordinary differential equations ([9]). Changes in the environmental conditions modify the development of the processes, switching coefficients in its continuous dynamics. In Gene Regulatory Networks, some biological facts such as a gene is activated by a transcription factor or regulator give power to the idea of using switched models.

We can group the different approaches of Hybrid Systems in two kinds: *function* and *implementation* oriented. Function oriented approaches favor human comprehension of models, while implementation orientation focuses in descriptions easy to interpret by machines. In the first orientation, the dynamics of systems are defined by functions. Continuous dynamics is commonly represented by differential equations, and the discrete dynamics affects it by switching its equations ([3], [4], [5]). Systems are called *Switched System*. The models are easy to understand, but too restrictive with respect to the dynamics. Implementations oriented approaches present more general descriptions of Hybrid systems, by using an abstract representation to implement the model ([1], [2]). Models describe the rules of the dynamics allowing many types of continuous dynamics.

Here, we relax the concept of Switched Systems to allow possible stochastic or non-deterministic changes in the continuous dynamics. Such systems, called *Stochastic Switched Systems*, are described from function and implementation orientations using Transition Systems ([10], [11]). We analize Gene Regulatory Networks ([12]), by approximating them by Switched Systems. Our representation of the recent osteo-chondro differentiation model ([13]), as Stochastic Switched System composed by two interacting components, allows us to improve the differentiation stimuli models separately and so improve the complete model. We suggest some experiments to model the effect of the Wnt pathway on the bone formation (osteogenic lineage) and include it as stimulus.

To simulate Hybrid models we use *BioRica*, available in BioRica. It is a high-level modeling framework that integrates discrete, continuous, stochastic, non-deterministic and timed behaviors in a non-ambiguous way allowing multi-scale dynamics, composition of models and hierarchical relations. The modeling language is an extension of the AltaRica Dataflow language ([14]), allowing hybrid systems and stochastic behaviors.

2 Approach

2.1 Modeling

The dependent variables are called *state variables* (in analogy with Transition Systems), while continuous and discrete factors are considered *controllers*. These systems are described using a mixture of continuous, discrete dynamics and logical relations to allow multiple interacting components.

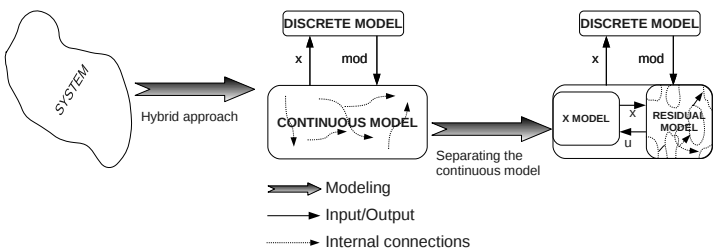


Figure 1. Modeling schema of Complex Biological Systems by Hybrid models. First it is identified the discrete and continuous interacting dynamics, then the continuous dynamics is separated into two interacting models: the *X MODEL* that describes the dynamics of X, and the residual model.

Let $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ be the *state variables* of the model. The variables $u = (u_1, \ldots, u_k) \in \mathbb{R}^k$ are the continuous control variables, and mod are the discrete controllers (Figure Fig. 1). We consider that the

dynamics of state variables is modeled by ordinary differential equations (explicit representation, equation 1) including continuous and discrete variables.

$$\dot{x}(t) = F(x(t), u(t), mod(t)) \tag{1}$$

The continuous dynamics is given by the changes in x over time, with F a function from $\mathbb{R}^n \times \mathbb{R}^k \times M$ to $\mathbb{R}^n$. The discrete dynamics is given by the evolution of the *mode* variable denoted $mod \in M$, where M is a finite or enumerable set and we denote $M = \{1, \ldots, M\}$ by identification of its elements.

The next step is to define how the discrete variables mod evolve over time. By considering hybrid systems as switches between continuous systems, represented by sets of differential equations, one talks of *Switched systems* and *Switched models* ([3], [4], [5]). Independent of x, the value of the right hand side function F changes as a function of the value of discrete variable mod (equation 1) that affects the form of the equations. We will say that changes in discrete variables values carry *switches* between model configurations.

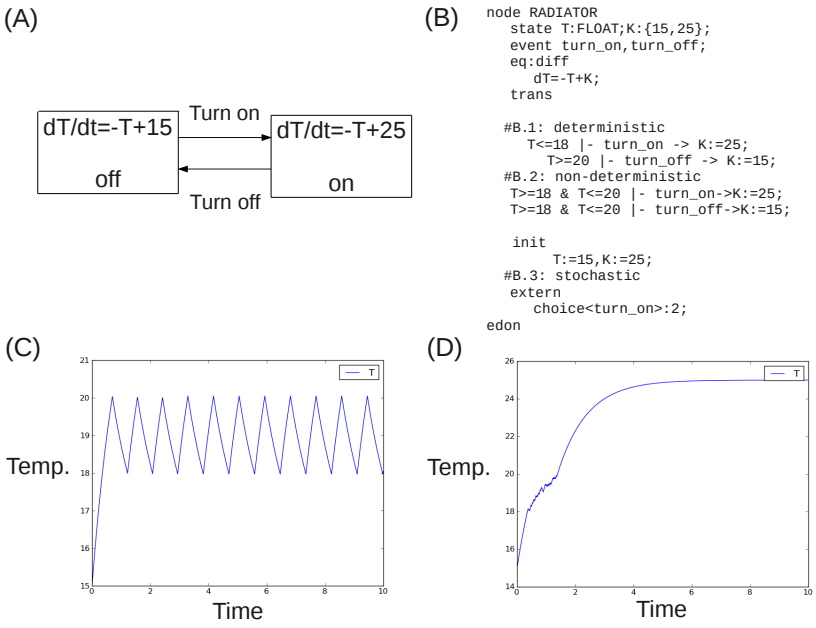


Figure 2. The radiator. (A)Schema of the Switched system. (B)Three models. (B.1) deterministic: it is turned on if $T \leq 18°C$ and turned off if $T \geq 20°C$, (B.2) non-deterministic: both events can happen if $18 \leq T \leq 20$, and (B.3) stochastic: turn on with probability $\frac{2}{3}$. (C)Temperature dynamics for (B.1), (D) For (B.3).

An example of Switched System is the behaviour of a radiator that controls the temperature (T) of a room (Fig. 2). A thermostat is activated when the temperature is detected to be low and it is regulated, if the temperature is high the system is turned off. This behaviour can be modeled in different ways: deterministic, non-deterministic or stochastic.

2.2 Composing Models

The act of building a model that is made of two or more modules is called *Composition*. The composition of two models is the model that explains the behaviour of both interacting modules. This allows us to the model to learn and to integrate knowledge of diverse types. The model can be extended and improved by introducing new modules that relate different functions. To be capable of describing the behaviour of a biological system over time, one needs to combine different types of models in a non-ambiguous way. A good implementation of this concept is essential to take advantage of the modularity of biological systems to build accurate and complete models. It must be sufficiently flexible to be capable of joining modules defined with different types of models, and reuse modules that have been *a priori* defined.

Our approach is based on the use of a global semantics to compose interacting modules. Flow conections and synchronizations allow to connect modules. Local clocks and solvers allow us to consider diverse types of dynamics.

The BioRica semantics is based on the automata semantics of AltaRica ([10], [14]), and Stochastic Transition Systems ([11]) that allow the inclusion of randomness and non-determinism. Given a BioRica node, one computes the probability of the state dynamics and considers non-deterministic decisions solved by random schedulers. The resulting semantics it is preserved with respect to flow relations and event syncronizations.

An important fact we approach here is that different processes have different timescales and stiffness levels. The use of modules solves in part this problem: each module has an specific timescale and discretization level. With this strategy, we improve the precision and the cumputation time. The processes with small timescales are observed at small time steps, while in case of long timescales we use longer time steps to reduce the number of simulations. So, if the equations solved by a module are stiff one uses small time intervals only at that module.

We implement models with these considerations in BioRica. A common specification of Systems Biology models is SBML ([15]), maybe the most popular abstraction for biochemical reactions models governed by temporal differential equations. Our framework includes a SBML parser that translates SBML models into BioRica models. So, it is possible to reuse and compose models previously specified in SBML to obtain more general models.

2.3 Stochastic Switched Systems

Our approach mixes both point of view of Hybrid Systems: function and implementation oriented. It is adapted to Hybrid models whose continuous dynamics is represented by differential equation systems, but we give more flexibility to the discrete dynamics allowing not only deterministic behaviours. We give function and implementation oriented descriptions of such systems.

The system dynamics is represented by transitions between different states. The interpretation of a model as dynamic entity, with possible changes on its form, turns it into a *Transition Systems* ([10]) with two types of transitions: state transitions and mode transitions. The state transitions are internal and controlled by the continuous dynamics of the model. The mode transitions are transitions in the sense of Transition Systems theory and they can be deterministic, non-deterministic or stochastic. The transitions can be modeled with stochastic components or including non-determinism to allow different behaviours (Fig. 2).

In the general theory of Stochastic Transition Systems ([11]) the transitions have possible stochastic behaviours. Given an action producing a transition, in this case a mode change, the next mode is randomly chosen according to transition probabilities if they exist. We will call *Stochastic Switched Systems* the extension of Switched systems that allows stochastic or non-deterministic transitions between modes. We consider two randomness sources: the moment (time) at which happens the action of changing mode, and the new mode that is chosen. The conditions that provoke the mode transitions are called *guards*. They are boolean formulas defined over state variables, external controllers and modes values. For each mode i, the mode transition arriving to i corresponds to an event (action in Transition Systems theory, [10] and [11]). In the radiator example (Fig. 2(B.1)), $T \leq 18$ is a guard condition provoking that event $turn\_on$ assigns the value 25 to the mode variable K.

We formally define a Stochastic Switched System as a hybrid system whose model is given by the equations 2-5. The first one defines the continuous dynamics and 3,4,5 the discrete dynamics at any time t. We denote $P(ev|(x(t), u(t), mod(t))$ the probability of choosing the event ev when the values of the state variables x are $x(t)$, the values of the continuous control variables u are $u(t)$ and the value of the mode variable mod is $mod(t)$, $time((x(t), u(t), mod(t)), ev)$ denotes the delay time of the event ev that is modeled to have distribution

$Dist_{ev}\{p_{event,1}, \ldots, p_{event,m}\}$.

$$\dot{x}(t) = F_{mod(t)}(x(t), u(t)), \tag{2}$$

$$P(ev|(x(t), u(t), mod(t)) = \frac{w_{ev}}{\sum_{e \in A(x(t),u(t),mod(t))} w_e}, \tag{3}$$

$$time((x(t), u(t), mod(t)), ev) \sim Dist_{ev}\{p_{event,1}, \ldots, p_{event,m}\}, \tag{4}$$

where w_e is the probability weight assigned to the event e, and $A(x(t), u(t), mod(t))$ is the set of available actions when x takes the value $x(t)$, u the values $u(t)$ and the mode variable m the value $mod(t))$ given by the equation 5.

$$A(x(t), u(t), mod(t)) = \{ev \in EVENTS : G_{ev}(x(t), u(t), mod(t)) = TRUE\}\,, \tag{5}$$

and $EVENTS$ is the set of events of the system (in Fig. 2 $EVENTS = \{turn\_on, turn\_off\}$).

Non-determinism appears with the presence of two or more available actions given a tuple $\langle(x(t), u(t), mod($ We simulate non-deterministic systems by using *random schedulers* with weights given by the external directives *choice*, so the event $turn\_on$ has probability $\frac{2}{3}$ for the non-deterministic radiator (Fig. 2(B.3)). The time delays can be stochastic, with $law < ev >$: $Dist\_ev\{pev\_1, \ldots, pev\_m\}$ denoting that the time delay of the event ev has distribution $Dist_{ev}\{pev_1, \ldots, pev_m\}$ with parameters $\{pev_1, \ldots, pev_m\}$. Between the possible distributions we include T (a deterministic time T), the Gaussian distribution $Normal\{\mu, \sigma\}$, the exponential distribution $Exponential\{r\}$ with r the rate and the uniform with parameters a, b $(a < b)$ $Uniform\{a, b\}$.

In the radiator example (Fig. 2), the simpler model is to consider a deterministic behaviour of the thermostat, where at any temperature at most one mode transition is observed: the radiator can be or active or not. For this model, it is obtained by considering the temperature of activation lower than the temperature where the radiator is turned off (18 and 20 respectively, Fig. 2(B.1)). With more ambiguous guard rules, given a temperature the radiator is accepted to be activated and turned off. This non-deterministic behaviour is obtained in case of activation if temperature is between 18 and 20 and deactivation if it is in the same interval (Fig. 2(B.2)).

2.4 BioRica Description of Stochastic Switched Systems

We represent and simulate Stochastic Switched Systems with BioRica. BioRica uses automata theory to represent *Stochastic Switched Systems* as a particular type of Hybrid Automaton ([1], [2]) where the continuous dynamics is given by differential equations. The continuous dynamics (equation 3) is described in *eqdiff* while the discrete one (equations 4 and 5) is described by transitions that change the form of the model.

In Fig. 2 is defined the BioRica node *RADIATOR*, with state variable T (temperature) and mode K. The possible events are defined with the keyword *event*, $turn\_on$ and $turn\_off$, and its effect is described in *trans*: $turn\_on$ provokes the assignment $K := 25$ when $T \leq 18$, $turn\_off$ assigns $K := 15$ if $T \geq 20$. With the keyword *eqdiff* one codes a set of differential equations, where $dxi = fi(x1, \ldots, xn, \_u1, ..., \_uk, mod)$ means that the rate of change of xi with respect to the time is equal to $fi(x1, \ldots, xn, \_u1, ..., \_uk, mod)$. In *init*, one defines the initial values of the variables. It is possible to define constant, $const$, and formula expressions, $formula$, to use in the equations.

In cases of non deterministic behaviours, BioRica decides what event select considering an aleatory decision between the possible events with weights given in the *choice* option of *extern*. In Fig. 2(B.3) the event *turn on* has weight 2, then at arriving to temperatures between 18 and 20 it selects the event *turn on* (activate the radiator) with probability $\frac{2}{3}$ and *turn off* with probability $\frac{1}{3}$. The keyword *extern* allows the inclusion of external directives about distributions of event delays and priority between events.

To decompose the dynamics we use the ideas of Fig. 1. With the keyword *flow*, one includes inputs (outputs) from (to) other BioRica nodes. Continuous and discrete dynamics can be modeled separately, the node *MAIN* represents the complete system and the keyword *sub* is used to define instances of other node (Fig. 3).

3 Application: Approximating Regulatory Systems by Stochastic Switched Systems

The dynamics of the radiator (Fig. 2) is hybrid in its nature because the switches are directly associated to the value of the mode. So, the mode K of the radiator is a piecewise-constant function. In less restrictive cases, where one identifies different underlying behaviours, we can use Switched Systems too. One can choose a set of factors to consider as *piecewise-order k* and approximate them to obtain a Switched model. A particular case is given by the regulation models with reduced Hill functions.

3.1 Reducing Hill Functions

Hill functions [16] are sigmoidal curves used to measure the continuous influence of an element on a target, depending on the concentration of the affecting element x, an exponent m to control the curve steepness, and on the mean point of influence θ. We denoted $h^+(x,\theta,m) = \frac{x^m}{x^m+\theta^m}$ the positive influence and $h^-(x,\theta,m) = 1 - h^+(x,\theta,m)$ the negative influence.

Dynamics systems with interacting elements often generate differential equation systems with Hill functions. The solution of such differential equation systems, equations 7-9 in case of the osteo-chondro cell differentiation model, can be complicate and use high computation times. More influence relations more difficult to solve the system. To simplify them, we reduce them into switched systems choosing some influence functions to be represented with piecewise-dependent behaviours. Thus, the system dynamics is obtained from the interaction of continuous and discrete dynamics.

Here, we show two reductions of the Hill functions: Piecewise-constant and Piecewise-linear approximation. The first idea is considering Hill function as step functions. It is to say, piecewise-constant functions to be 0 when x is lower or equal than the threshold θ and 1 after this threshold. With this simplification, the model moves between different modes in function of how high or low are the concentrations x with respect to the thresholds θ. The thresholds divide the state variables space into cuboids, each one with an associated system of equations. Despite one obtains information of the system behaviour by looking the form of the equations in each cuboid of the state spaces, the observations are only qualitative.

The second reduction of Hill functions is the Piecewise-linear approximation, in which the transition between 0 and 1 is smoothed by a linear function. It is to say, we use the approximation of equation 6 below:

$$h^+(x,\theta,m) \approx l^+(x,\theta_1,\theta_2) = \begin{cases} 0 \text{ if } x \leq \theta_1 \\ \frac{x-\theta_1}{\theta_2-\theta_1} \text{ if } \theta_1 < x < \theta_2 \\ 1 \text{ if } x \geq \theta_2 \end{cases} \quad (6)$$

With this second alternative, the switched system transits between a big set of modes according to how high are the mRNA concentrations compared with the thresholds θ_1 and θ_2 of each influence function.

3.2 An Osteo-chondro Differentiation Model

An application of Gene Regulatory Networks is cell differentiation modeling. Each possible differentiation of a cell is associated to the mRNA concentration of an specific gene. Here we use the model of Schittler *et al.* ([13]) to differentiate progenitor cells into osteoblasts (bone cells) or chondrocytes (cartilage cells). They are considered two mutually inhibiting genes, so called the *osteo-chondro switch*, one associated to the osteogenic differentiation (*Runx2*) and another (*Sox9*) to the chondrogenic option. A third gene (*Tweak*) is associated with the progenitor maintenance role that inhibits both genes of the osteo-chondro switch. The mRNA concentration associated to the progenitor state is denoted x_P, the mRNA concentration of the osteogenic state is denoted x_O and the associated to the chondrogenic differentiation is denoted x_C. To incorporate the external pro-differentiation, pro-osteogenic and pro-chondrogenic stimulus are included three inputs: z_D, z_O and z_C with positive value. The increase of any differentiation stimulus provokes an increase of the expression of the

```
const n=2;ap=0.2;bp=0.5;mp=10;cpp=0.1;kp=0.1;ao=0.1;ac=0.1;bo=1;
const n=2;ap=0.2;bp=0.5;mp=10;cpp=0.1;kp=0.1;ao=0.1;ac=0.1;bo=1;
const bc=1;mo=1;mc=1;coo=0.1;ccc=0.1;coc=0.1;cco=0.1;cop=0.5;
const ccp=0.5;ko=0.1;kc=0.1;

node DIFF
      state xp,xo,xc:FLOAT;
      flow zd,zo,zc:[0,1]:in;
      eqdiff
            dxp=(ap*xp^n+bp)/(mp+zd+cpp*xp^n)-kp*xp;
            dxo=(ao*xo^n+bo+zo)/(mo+coo*xo^n+coc*xc^n+cop*xp^n)
                  -ko*xo;
            dxc=(ac*xc^n+bc+zc)/(mc+ccc*xc^n+cco*xo^n+ccp*xp^n)
                  -kc*xc;
      init
            xp:=12,xo:=0,xc:=0;
edon

node STIMULUS
      state _zd,_zo,_zc:FLOAT;on_d,on_o,on_c:BOOL;
      flow zd,zo,zc:FLOAT:out;
      event to_diff,to_osteo,to_chondro;
      trans
            on_d=False |-to_diff -> _zd=1,on_d=True;
            on_o=False |-to_osteo -> _zo=0.8,on_o=True;
            on_c=False |-to_chondro -> _zc=0.8,on_c=True;
      init
       _zd:=0,_zo=0,_zc=0,on_d:=False,ond_o=False,on_c=False;
      extern
             law<to_diff>:Exponential{0.01};
            law<to_osteo>:Exponential{0.002};
            law<to_chondro>:Exponential{0.001};
      assert
            zd=_zd;
            zo=_zo;
            zc=_zc;
edon
node MAIN
      sub
            S:STIMULUS;
            D:DIFF;

      assert
            D.zd=S.zd;
            D.zo=S.zo;
            D.zc=S.zc;
edon
```

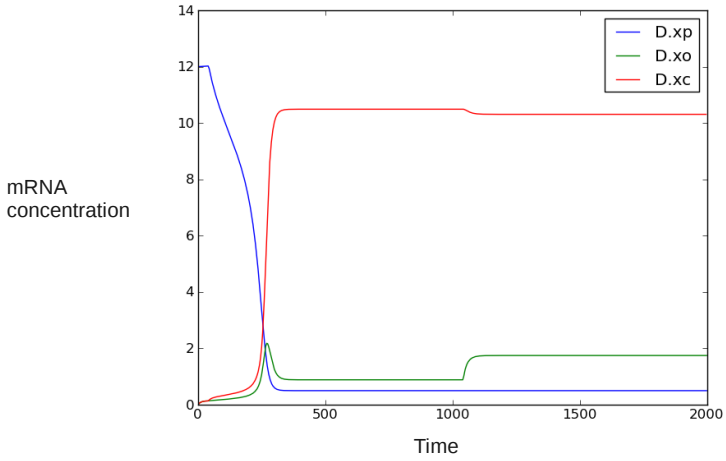


Figure 3. BioRica code and simulation of an osteo-chondro differentiation model ([13]). The pro-differentiation stimulus happens at time exponential with rate 0.01 (expected value $E(t) = 100$), the pro-osteogenic stimulus happens with rate 0.002 ($E(t) = 500$) and the pro-chondrogenic with $E(t) = 1000$.

associated gene. The model is given by the equations 7-9 above.

$$\dot{x}_P(t) = \frac{a_P \cdot x_P^n + b_P}{m_P + z_D + c_{PP} \cdot x_P^n} - k_P \cdot x_P, \tag{7}$$

$$\dot{x}_O(t) = \frac{a_O \cdot x_O^n + b_O + z_O}{m_O + c_{OO} \cdot x_O^n + c_{OC} \cdot x_C^n + c_{OP} \cdot x_P^n} - k_O \cdot x_O, \tag{8}$$

$$\dot{x}_C(t) = \frac{a_C \cdot x_C^n + b_C + z_C}{m_C + c_{CC} \cdot x_C^n + c_{CO} \cdot x_O^n + c_{CP} \cdot x_P^n} - k_C \cdot x_C, \tag{9}$$

with $n = 2$, $a_P = 0.2$, $b_P = 0.5$, $m_P = 10$, $c_{PP} = 0.1$, $k_P = 0.1$, $a_O = a_C = 0.1$, $b_O = b_C = 1$, $m_O = m_C = 1$, $c_{OO} = c_{CC} = c_{OC} = c_{CO} = 0.1$, $c_{OP} = c_{CP} = 0.5$, $k_O = k_C = 0.1$ known parameters.

We obtain the same results for the scenarios analyzed in [13], but the BioRica representation gives flexibility to the model. In Fig. 3 we considered another scenario, where pro-differentiation, pro-osteogenic and pro-chondrogenic stimulus happen with exponential probabilities over time (a Poisson process). The system corresponds to a strict Stochastic Switched System, in which delay times have random behaviours.

Since we separate between stimulus (node *STIMULUS*) and differentiation dynamics (*DIFF*), to specify each differentiation stimulus one needs only to modify such a node. The dynamics of *STIMULUS* controls the lineage decision by switching the values of the z coefficients, and depends on external factors. A factor that affects the lineage decision is the activation of the Wnt/$\beta-$catenin pathway. Since *Runx2* is a Wnt target gene, the accumulation of $\beta-$catenin in the nucleus stimulates the expression of *Runx2*, and consequently favors the bone formation ([17]). One can include this effect, on the synamics of z_0, by measuring the concentration of nuclear $\beta-$catenin over time and using *LiCl* to activate the pathway.

4 Conclusions and Discussion

We used the theory of Stochastic Transition Systems ([10] and [11]) to define an special type of Hybrid System: Stochastic Switched System (section 2.3). The model is formed by the continuous dynamics of the state variables, given by differential equations, and the discrete dynamics of the modes that change over time to transform the differential equations. We allow stochastic and non deterministic behaviours, and implemented such systems using BioRica.

We defined the osteo-chondro cell differentiation model in [13] as a Stochastic Switched System by composing *STIMULUS* and *DIFF* (differentiation) components giving more flexibility to extensions. As example, stimulis are considered with aleatory behaviour. By considering the activation of the Wnt/$\beta-$catenin pathway as factor of bone formation and measuring its effect, it is possible to improve the model.

We defined a non-ambiguous way to describe a complex system by decomposing it into different types of interacting models. Behaviour laws change over time, which is modeled by discrete changes of mode variables that transform the continuous dynamics, and complex processes are modeled by composing diverse models with flow connections and synchronization of events. Our approach allows us to reuse SBML specified models and exploits modular properties of systems, which can be separated into modules in function of the type of process and its timescale, and the complexity or type of model.

Acknowledgements

RA supported by an INRIA CORDI-S scolarship.

References

[1] T.A. Henzinger, The theory of hybrid automata. Technical Report UCB/ERL M96/28, EECS Department, University of California, Berkeley, 1996.
[2] K.H. Johansson, Hybrid control systems. Technical Report EOLSS,6.43.28, Dept. of Signals, Sensors & Systems, Royal Institute of Technology, 1044, 2003.
[3] M.S. Branicky, Stability of switched and hybrid systems. *Decision and Control, 1994., Proceedings of the 33rd IEEE Conference on Decision and Control*, 4:3498–3503, 1994.
[4] R. Shorten, F. Wirth, O. Mason, K. Wulff and Ch. King, Stability criteria for switched and hybrid systems. *SIAM REVIEW*, 49:545–592, 2005.
[5] L. Tavernini, Differential automata and their discrete simulators. *Nonlinear Analysis*, 11(6):665–683, 1987.
[6] H. Kitano, Computational systems biology. *Nature*, 420(6912):206–210, 2002.
[7] W. Callebaut and D. Rasskin-Gutman, Modularity: Understanding the Development and Evolution of Natural Complex Systems. *The MIT Press, illustrated edition edition*, 2005.
[8] J.J. Tyson, Modeling the cell division cycle: cdc2 and cyclin interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 88(16):7328 –7332, 1991.
[9] Ch. Li, M. Donizelli, N. Rodriguez, H. Dharuri, L. Endler, V. Chelliah, L. Li, E. He, A. Henry, M. Stefan, J. Snoep, M. Hucka, N. Le Novere and C. Laibe, BioModels database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Systems Biology*, 4(1):92, 2010.
[10] A. Arnold, D. Bégay and P. Crubillé, Construction and analysis of transition systems with MEC. *World Scientific Publishing Co., Inc.*, 1994.
[11] L. De Alfaro, Stochastic transition systems. *Proceedings CONCUR 98*, 1466:423—438, 1998.
[12] J. Gebert, N. Radde, and G.W. Weber, Modeling gene regulatory networks with piecewise linear differential equations. *European Journal of Operational Research*, 181(3):1148–1165, 2007.
[13] D. Schittler, J. Hasenauer, F. Allgöwer and S. Waldherr, Cell differentiation modeled via a coupled two-switch regulatory networks. *Chaos*, 20(4), 2010.
[14] A. Arnold, G. Point, A. Griffault and A. Rauzy, The AltaRica formalism for describing concurrent systems. *Fundam. Inf.*, 40(2-3):109–124, 1999.
[15] M. Hucka, F. Bergmann, S. Hoops, S. Keating, S. Sahle and D. Wilkinson, The Systems Biology Markup Language SBML: Language Specification for Level 3 Version 1 Core Release 1 Candidate. *Nature Precedings*, 2010.
[16] A.V. Hill, The possible effects of the aggregation of the molecules of hæmoglobin on its dissociation curves. *J.Physiol*, 40:iv–vii, 1910.
[17] V. Krishnan, H.U. Bryant, and O.A. Macdougald, Regulation of bone mass by wnt signaling. *The Journal of Clinical Investigation*, 116(5):1202–1209, 2006.

MoonGO: Predicting Moonlighting Proteins from PPi Networks Based on GO Annotations

Charles E. CHAPPLE, Benoît ROBISSON, Carl HERRMANN and Christine BRUN

TAGC Inserm U928, Université de la Méditerranée, Parc Technologique de Luminy,
Case 928, 13288, Marseille, Cedex 9, France
{cchapple},{robisson},{herrmann},{brun},@tagc.univ-mrs.fr

Keywords PPI network, moonlighting, GeneOntology, protein function, graph partitioning.

1 Introduction

Moonlighting proteins are a specific subset of multifunctional proteins. To be considered moonlighting, a protein must have multiple, unrelated and independent functions (disabling one does not affect the other). For example, human aconitase, in addition to its primary function of catalyzing the isomerization of citrate to isocitrate is also involved in iron homeostasis [1]. Identification of moonlighting proteins is important for many reasons. First of all, correct identification of moonlighting proteins is essential for the complete functional annotation of a proteome. In addition, since moonlighting proteins can link unrelated processes, they are likely to play a regulatory role. Some proteins perform their moonlighting function only in tumor cells and are important for disease progression [2]. Finally, an unknown moonlighting function of a drug target may result in unexpected side effects. The moonlighting proteins that have so far been identified were serendipitous discoveries and are probably but a fraction of the whole. The large scale prediction of moonlighting proteins, however, requires the development of specific tools.

Protein-protein interaction networks (PPI networks) can be represented as simple graphs where each vertex represents a protein, and each edge a direct interaction. By definition, moonlighting proteins will have multiple, unrelated interacting partners. Large scale PPI networks are built using data from high throughput techniques and are therefore context-free. A PPI network, therefore, represents the set of all known interactions between each of its constituent proteins. When combined with functional annotations of its constituent proteins, PPI networks are particularly well-suited for the identification of moonlighting proteins. Here, we present `MoonGO`, the first method for the high-throughput prediction of moonlighting proteins from PPI networks.

2 Methods

2.1 Class Annotation

To ensure the quality of our data, we only use 'high confidence' interactions, i.e., those that have been identified by experimental methods that find direct, binary, interactions such as yeast two hybrid assays. We start with a network that has been partitioned into a system of overlapping classes (for the work presented here, we have used the OCG algorithm ([3]). Each class is then annotated according to the annotations of the proteins it contains. We have tried different annotation methods and found that the best results are obtained when using a simple majority rule. We therefore annotated a given class to a specific GO term *iff* $\geq 50\%$ of annotated proteins in that class share that GO term. While very extensive, the GO annotations are far from complete. For example, of the 9006 proteins in our human interactome, 1540 (17.2%) lack annotations. A class is, therefore, considered for annotation *iff* at least two thirds of its constituent proteins have GO annotations. If not already annotated to them, member proteins inherit the annotation(s) of the class.

2.2 `MoonGO`

Many GO terms are either semantically similar, such as "cell death" and "cytolysis" or refer to closely linked biological processes, such as "cytokinesis" and "mitosis". Therefore, a method of evaluating the similarity, or "closeness" of GO annotations is needed to distinguish between 'normal' multifunctional and *bona*

fide moonlighting proteins. To that end we have calculated the hypergeometric probability of association for all combinations of GO term pairs that annotate the human proteome. GO term pairs whose probability of association lies below the 99th percentile of the calculated probabilities are considered 'dissimilar'. For the work presented here, we have used the 'biological process' ontology but any other may also be used.

MoonGO uses the annotated classes and GO term association probabilities to search the network for proteins found at the intersection of two classes annotated to dissimilar GOs (Figure 1A). In addition, it can also search for proteins whose interactors bridge two dissimilar classes (Figure 1B) and proteins whose direct annotations are dissimilar to those of their class (Figure 1C). Finally, it can also search for interactions between proteins annotated to dissimilar GOs and that belong to different classes (Figure 1D).

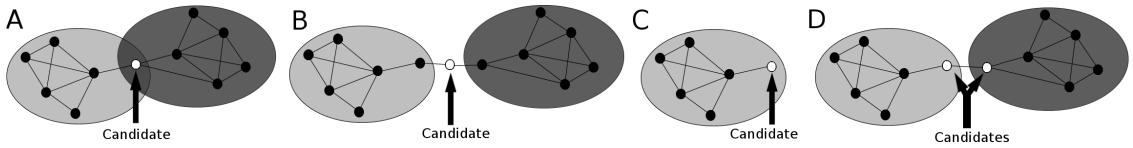


Figure 1. Moonlighting candidates. The situations recognized by MoonGO as indicative of moonlighting function. Cases A,B and C identify a moonlighting protein, while case D a 'moonlighting interaction' as either, or indeed both, interactors can be considered as candidates. Moonlighting candidates are shown as white circles. The large grey circles represent classes annotated to dissimilar GO terms.

3 Results

When using the methods illustrated in Figure 1, which incorporate information from the annotated classes, mode A finds 7 candidate proteins, mode B 52 and mode C 1565. Mode D finds 2396 interactions involving 2763 proteins. These sets of proteins are preliminary results and will be further filteredikel as described below.

4 Discussion/Perspectives

Our next step will be to integrate other data such as expression, subcellular localisation and protein structural disorder to refine our predictions. We will apply our method to the interactomes of five species: human, mouse, fly, worm and yeast. As well as predicting possible moonlighting proteins in these species, we hope to identify special cases where a protein is not only moonlighting in multiple species but also has different secondary functions in different organisms. This is the case for aconitase which in addition to its primary function is involved in mitochondrial DNA maintenance in yeast but in iron homeostasis in mammals [4]. Cases like this suggest that some proteins may be intrinsically amenable to the acquisition of novel functions.

Acknowledgements

We would like to thank Alain Guenoche and Anaïs Baudot for helpful discussions. This work is supported by an ANR PIRIBIO grant to CB (RO9127AA, Moonlight project).

References

[1] K. Volz, The functional duality of iron regulatory protein 1, *Curr Opin Struct Biol*, 18:106-111, 2008.

[2] C.A. Maxwell, J. McCarthy and E. Turley , Cell-surface and mitotic-spindle RHAMM: moonlighting or dual oncogenic functions? , *J Cell Sci*, 121:925-932, 2008.

[3] E. Becker, A. Guénoche and C. Brun, Systèmes de classes chevauchantes pour la recherche de protéines multifonctionnelles, *Actes des Journées Ouvertes Biologie Informatique Mathématiques (JOBIM)*, p. 49-54,2009.

[4] The moonlighting function of pyruvate carboxylase resides in the non-catalytic end of the TIM barrel, *Biochim Biophys Acta*,1803:1038-1042, 2010.

Analysis of the Functionalisation Process for Duplicated Genes of *Arabidopsis thaliana* in Protein-Protein Interactions Networks and Transcriptomic Data

Justin WHALLEY, Etienne BIRMELÉ, Claudine DEVAUCHELLE and Carène RIZZON

LABORATOIRE STATISTIQUE ET GÉNOME, UMR8071 CNRS, Tour Évry II - 2ème étage, 523, places des Terrasses F-91000 Évry , France

{justin.whalley, etienne.birmele, claudine.devauchelle, carene.rizzon}@genopole.cnrs.fr

Abstract *Using a clustering method to classify genes into families and taking advantage of available knock-out data, transcriptomic data and protein-protein interactions data, our purpose is to characterise the functionalisation process underlying the maintenance of duplicated genes in* Arabidopsis thaliana*. We aim to answer the following questions: which characteristics of* Arabidopsis thaliana *families are prone to functionalisation? Which topological measures can we use in biological networks to classify duplicate genes according to the functionalisation processes of maintenance?*

Keywords Gene duplication, Evolutionary genomics, Clustering methods.

Gene duplication is readily accepted as a primary mechanism for generating organism complexity. However, the mechanisms responsible for the maintenance of duplicated genes, at the genome scale, are still poorly understood. Analysis of biological networks can help us to understand better which evolutionary forces are acting on duplicated genes, as their interacting context is taken into account. Taking advantage of available knock-out data, microarrays transcriptomic data and protein-protein interactions data from literature, we look to characterise the functionalisation processes acting on duplicated genes in biological networks of *Arabidopsis thaliana*. We take into account the underlying mechanism of duplication (whole genome duplication, tandem duplication and segmental duplication) as it can influence the maintenance of duplicated genes [1,2].

The mechanisms mainly proposed to explain the maintenance of duplicated genes in genomes imply (see for review [3]):

- Neofunctionalisation, where duplicated genes can acquire new functions, is modelled as a mutation happening to a duplicate gene after duplication and being fixed due to directional selection or genetic drift.
- Subfunctionalisation, which can describe two different models: the duplication, degeneration, complementation model (DDC), where duplicate genes are maintained, as the function of the parent gene is divided between them, or the escape from adaptive conflict model (EAC), where the duplicate genes undergo adaptive mutations that cause specialisations of subfunctions of the parent gene. We can distinguish subfunctionalisation where both duplicated genes are needed to complete ubiquitly the ancestral function and subfunctionalisation where duplicated genes maintain each the ancestral function but in different tissues.
- Redundancy of function, where the duplicate genes are maintained as copies of their parent gene, due to increased gene dosage.

Hanada et al. [4] examined *Arabidopsis thaliana* paralogous gene pairs associated with morphological diversification and classified them into high, low, and no morphological diversification groups, based on knock-out data. They found that the divergence rate of both gene expression and protein sequences were significantly higher in either high or low morphological diversification groups compared with those in the no morphological diversification group. Considering that high morphological diversification duplicated gene pairs can correspond to a neofunctionalisation underlying process and that no morphological diversification duplicated gene pairs can correspond to a gene maintenance by redundancy of function, we took advantage of these data: we showed that the evolutionary process of functionalisation can be in part characterised using the PPI network [5].

We clustered the *Arabidopsis thaliana* genes into families using a walktrap clustering algorithm developed at the laboratory with which we can control the inter connections level inside families. Using available knock-out data, microarrays transcriptomic data and protein-protein interactions data from literature, we aim to answer the following questions:

1. Which gene families are homogeneous in function? Which are the families showing the more divergence in function (showing neofunctionalisation)? Can we retrieve this in the PPI network or in the expression data?
2. Using knock-out data, can we distinguish the high, low and no morphological diversification groups of duplicated pairs as defined by Hanada et al. [4] in the expression data? We expect a correlation of expression for genes maintained by redundancy, a divergence of expression for genes maintained by neofunctionalisation and either a correlation of expression across all tissues or a divergence of expression according to tissues for genes maintained by subfunctionalisation.
3. Is there specificity of correlation of expression according to the underlying mechanism of duplication?

References

[1] I. Wapinski, A. Pfeffer, N. Friedman and A. Regev, Natural history and evolutionary principles of gene duplication in fungi. *Nature*, 6;449(7158):54-61, 2007.

[2] K. Hanada, C. Zou, MD Lehti-Shiu, K. Shinozaki and SH Shiu, Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. *Plant Physiol*, 148(2):993-1003, 2008.

[3] GC Conant and KH Wolfe. Turning a hobby into a job: how duplicated genes find new functions. *Nat Rev Genet*, 9(12):938-50 2008.

[4] K. Hanada, T. Kuromori, F. Myouga, T. Toyoda and K. Shinozaki, Increased Expression and Protein Divergence in Duplicate Genes Is Associated with Morphological Diversification. *PLoS Genet*, 5(12):e1000781, 2009.

[5] R. Zaag, E. Birmelé and C. Rizzon, Topological characteristics of the functionalisation process for duplicated genes in PPI networks of *Arabidopsis thaliana*. *JOBIM 2010*, Montpellier 7-9 September 2010.

[6] P. Pons and M. Latapy, Computing communities in large networks using random walks. *Most*, 10(2):1-20, 2005.

SORGOdb: Superoxide Reductase Gene Ontology Curated DataBase

Céline LUCCHETTI-MIGANEH[1], David GOUDENEGE[1], David THYBERT[2], Gilles SALBERT[1] and Frédérique BARLOY-HUBLER[1]

[1]CNRS UMR 6026, ICM, Equipe Sp@rte, Université de Rennes 1, Campus de Beaulieu, 35042 Rennes, France
{celine.lucchetti, david.goudenege, gilles.salbert, fhubler}@univ-rennes1.fr
[2] EMBL-EBI Wellcome Trust Genome Campus; Hinxton, Cambridgeshire, CB10 1SD,U.K
dthybert@ebi.ac.uk

Keywords Superoxide Reductase, Database, Ontology, Prokaryote.

1 Introduction

The superoxide anion is one of the deleterious reactive oxygen species. To survive and protect themselves from the toxicity of superoxide anion, many species have developed defence mechanisms [1]. Dismutation of O2- into molecular oxygen and hydrogen peroxide by Superoxide dismutase was the only biological mechanism identified for scavenging superoxide anion radicals until the early 1990's [2,3,4]. Two novel iron-sulphur-containing proteins that detoxify superoxide molecules were then discovered in sulphate-reducing anaerobes: desulfoferrodoxin (Dfx) [3], and neelaredoxin (Nlr) [4]. Superoxide reductase proteins (SOR) catalyzes the reduction of superoxide, rather than it dismutation [5]. SOR proteins have different numbers of iron sites: both contain a similar C-terminal single iron-containing site (centre II) but Dfx also has a second N-terminal site (centre I) [6]. SOR were first thought to be restricted to anaerobic prokaryotes but were subsequently discovered in some micro-aerophilic and micro-aerotolerant Bacteria and Archaea [4, 7]. More recently, a SOR encoding gene was also discovered in an eukaryote [8]. A first classification of these enzymes was proposed according to the number of metal centres: neelaredoxin or 1Fe-SOR and desulfoferrodoxin or 2Fe-SOR [9,10]. An additional class was proposed after the isolation of a *Treponema pallidum* SOR that contains an extended non-iron N-terminal domain of unknown function [11,12]. Add to the problem of ambigous SOR classification, there are several mistakes with the annotation of superoxide reductase genes, partly a consequence of heterogeneous transfer of annotations from previously characterized neelaredoxin, desulfoferrodoxin, superoxide réductase.

For all these reasons, we developed SORGOdb, the first resource specifically dedicated to superoxide reductase genes in entirely sequenced and in-draft genomes. SOR sequences were curated manually, analysed and stored using a new ontology in a publically available resource (http://sorgo.genouest.org/).

2 Construction and Content

For collection of SOR, we have extensively searched the Pubmed database and identified all relevant literature concerning any protein with "superoxide reductase" activity (13 SOR published in 12 organisms). We therefore enriched the database using manually curated sequences described as desulfoferrodoxin, superoxide reductase or neelaredoxin in EntrezGene and/or GenBank entries. As the "centre II" is the active site for the SOR activity, we also included all proteins with a domain of this type as described in databases (PRODOM, PFAM…). All sequences collected were cleaned up to remove redundancy and unrelated proteins. This non-redundant and curated dataset was used to investigate the complete and in-draft genomes available in the NCBI database through a series of successive BlastP and tBlanstN searches. At the end of this integrative research, we had a collection of 325 non-redundant and curated predicted SOR in 274 organisms, covering all the three kingdoms: Bacteria (270 genes), Archaea (52 genes) and Eukaryota (3 genes).

3 New Classification and Ontology

We propose a new unambiguous SOR classification based on their domain architectures (sequential order of domains from the N- to the C-terminus). Considering both domain compositions and arrangements,

this classification contains seven functionally relevant classes. Briefly, the 144 proteins that contain only the active site II (SOR) have been classified as Class II-related SOR. Class III-related SOR correspond to proteins which have the active site II and enclose an additional N-terminal region of unknown function. Class-IV related SOR correspond to very recently new class of methanoferrodoxin which have the active site II and an additional iron sulfur domain. The TAT-SOR have the active site II and include an extra twin-arginine N-terminal signal peptide. The 152 proteins composed of a desulforedoxin (Dx) domain preceding the SOR unit were clustered in a class named Dx-SOR. The 19 proteins that combined a N-terminal helix-turn-helix domain (HTH) before the Dx-SOR module were gathered in a class called HTH-Dx-SOR. Finally, 10 SOR proteins that correspond to exceptional domains fusion or that encompass a mutated ncDx domain were classified in a disparate class labelled "Atypical-SOR".

4 Conclusion

The SORGOdb server is the first web server that centralizes and provides an interface for information concerning superoxide reductase proteins. SORGOdb provides integrated features: (1) Multiple options for data browsing and searching (2) Complete descriptions of SOR and a new domain-based classification (3) Synthetic and downloadable synopsis for each locus tag (4) A SOR-homology analysis tool using BlastP similarity searches with the SORGOdb-positive dataset. SORGOdb is a unique mining tool that can assist researchers with diverse interests to retrieve, visualize and analyse superoxide reductase genes and proteins.

Acknowledgements

CLM is supported by Agence Nationale de la Recherche and DG by the Ministère de la Recherche. We wish to thank the bioinformatics platform of Biogenouest of Rennes for providing the hosting infrastructure.

References

[1] J.A. Imlay, Cellular defenses against superoxide and hydrogen peroxide. *Annu. Rev. Biochem.,* 77:755-776, 2008.
[2] J.M. McCord and I. Fridovich, Superoxide dismutase. An enzymic function for erythrocuprein (hemocuprein). *J. Biol. Chem.* 244(22):6049-6055, 1969.
[3] I. Moura, P. Tavares, J.J. Moura, N. Ravi, B.H. Huynh, M.Y. Liu and J. LeGall, Purification and characterization of desulfoferrodoxin. A novel protein from *Desulfovibrio desulfuricans* (ATCC 27774) and from *Desulfovibrio vulgaris* (strain Hildenborough) that contains a distorted rubredoxin center and a mononuclear ferrous center. *J. Biol. Chem.* 265(35):21596-21602, 1990.
[4] V. Niviere and M. Fontecave, Discovery of superoxide reductase: an historical perspective. *J. Biol. Inorg. Chem.*, 9(2):119-123, 2004.
[5] S.I. Liochev and I. Fridovich, A mechanism for complementation of the sodA sodB defect in *Escherichia coli* by overproduction of the rbo gene product (desulfoferrodoxin) from *Desulfoarculus baarsii*. *J. Biol. Chem.*, 272(41):25573-25575, 1997.
[6] L. Chen, P. Sharma, J. Le Gall, A.M. Mariano, M. Teixeira and A.V. Xavier, A blue non-heme iron protein from *Desulfovibrio gigas*. *Eur. J. Biochem.,* 226(2):613-618, 1994.
[7] F. Rusnak, C. Ascenso, I. Moura and J.J. Moura, Superoxide reductase activities of neelaredoxin and desulfoferrodoxin metalloproteins. *Methods Enzymol.,* 349:243-258, 2002.
[8] A.F. Pinto, J.V. Rodrigues and M. Teixeira, Reductive elimination of superoxide: Structure and mechanism of superoxide reductases. *Biochim. Biophys. Acta.,* 1804(2):285-297, 2010.
[9] D.M. Jr. Kurtz, E.D. Coulter, The mechanism(s) of superoxide reduction by superoxide reductases in vitro and in vivo. *J. Biol. Inorg. Chem.***,** 7**(**6):653-658, 2002.
[10] S.A. Pereira, P. Tavares, F. Folgosa, R.M. Almeida, I. Moura and J.J.G. Moura, European Journal of Inorganic Chemistry. *European Journal of Inorganic Chemistry,* 2007(18):2569-2581, 2007.
[11] M. Lombard, D. Touati, M. Fontecave and V. Niviere, Superoxide reductase as a unique defense system against superoxide stress in the microaerophile *Treponema pallidum*. *J. Biol. Chem.,* 275(35):27021-27026, 2000.
[12] T. Jovanovic, C. Ascenso, K.R. Hazlett, R. Sikkink, C. Krebs, R. Litwiller, L.M. Benson, I. Moura, J.J. Moura, J.D. Radolf *et al*: Neelaredoxin, an iron-binding protein from the syphilis spirochete, *Treponema pallidum*, is a superoxide reductase. *J. Biol. Chem.*, 275(37):28439-28448, 2000.
[13] D. Goudenege, S. Avner, C. Lucchetti-Miganeh and F. Barloy-Hubler, CoBaltDB: Complete bacterial and archaeal orfeomes subcellular localization database and associated resources. *BMC Microbiol.,* 10:88, 2010.
[14] A. Dolla, M. Fournier and Z. Dermoun, Oxygen defense in sulfate-reducing bacteria. *J. Biotechnol.*, 126(1):87-100, 2006.

The Biogemix Knowledge Base Project: Cross-species and Network-based Data Integration for Huntington's Disease Research

François-Xavier LEJEUNE[1], Lilia MESROB[1], Frédéric PARMENTIER[1], Cédric BICEP[1], Jean-Philippe VERT[2], Christian NÉRI[1] and the Working Group 'Biological Modifiers' of the European Huntington's Disease Network*

[1] Laboratory of Neuronal Cell Biology and Pathology, INSERM Unit 894, 2 ter rue d'Alésia, 75014 Paris, France
christian.neri@imserm.fr

[2] Institut Curie, INSERM Unit 900, 26 rue d'Ulm, 75005 Paris, France

* http://www.euro-hd.net/html/network/groups/biomodifiers

Keywords Disease, omics data integration, unbiased data integration, target prioritisation, cross-species analysis, network-based analysis, knowledge base.

1 Background

The identification and validation of neuroprotective targets is of primary importance in research on neurodegenerative diseases such as Huntington's disease (HD). The development of genetically tractable models of disease and their use in genome-wide screens has generated a large amount of data in several species. A current challenge is the unbiased integration of these data sets in order to prioritize candidate target genes. The Biogemix knowledge base project has been developed with the European HD network (Euro-HD) to integrate 'omics' data from models of HD pathogenesis as available in several species (invertebrates, mammalian cells, mice, human samples). This project relies on the combination of network-based and cross-species procedures to unlock the biological information buried into disease data sets. Ultimately, the aim is to make the Biogemix knowledge base v 1.0 publically available on-line.

2 Results

The Biogemix procedure is a method that relies on the use of molecular networks for the unbiased integration of 'omics' data across different species. This method is suited to the analysis of data sets for which the number of genes analyzed clearly exceeds the number of conditions tested. Single data sets are firstly analyzed with respect to a reference molecular network (for instance, use of WormNet to analyze worm data). To this end, the core method is the Fourier Transform of the data using prior knowledge of gene connectivity to gradually remove unreliable information [1]. Clusters of highly interconnected genes (modules) are then extracted, and they are annotated for their biological content using information from databases such as Gene Ontology, KEGG and Panther. In a second step, cross-species clusters (meta-modules) are calculated using pairwise cluster alignment driven by gene/protein connectivity and protein similarities [2] and the resulting graphs annotated for their biological content. In a third step, all of the Biogemix products are ranked according to topological features, which is part of a larger prioritisation system that uses biological and drug discovery criteria to classify modules and genes of high interest. Preliminary results suggest that the Biogemix procedure is able to identify pathways previously associated to HD pathogenesis and to emphasize genes and pathways of novel interest in HD research.

3 Conclusion

Current developments aim at fine-tuning data analysis. Another aim is to develop a user-friendly query system that will allow the users to easily localize and visualize the information of interest. Results will be shown to illustrate the performances of the Biogemix procedure, its value for research on HD and potential for research on other diseases.

References

[1] F. Rapaport, A. Zinovyev, M. Dutreix, E. Barillot and J.-P. Vert, Classification of microarray data using gene networks. *Bioinformatics*, 8 (35), 1-15, 2007.

[2] R. Singh, J. Xu and B. Berger, Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proc Natl Acad Sci U S A*., 105 (35), 12763-8, 2008.

SIDR, a Public Data Repository for Multi-assay Experiments

Issues on Metadata Biocuration

Alain ZASADZINSKI[1], Marie-Christine JACQUEMOT[1], Florian MAZUR[1], Damien FLEURY[1], Yahia BERCHI[1], Morad MECHREF[1], Claude NIEDERLENDER[1] and Magali ROUX[1,2]

[1]INSTITUT DE L'INFORMATION SCIENTIFIQUE ET TECHNIQUE, UPS76 CNRS, 2 allée du Parc de Brabois, 54519, Vandoeuvre-lès-Nancy, Cedex, France

{alain.zasadzinski, marie-christine.jacquemot, florian.mazur, yahia.berchi, morad.mechref, claude.niederlender, magali.roux}@inist.fr

[2]LABORATOIRE D'INFORMATIQUE DE PARIS VI, UMR 7606 CNRS, 4 Place Jussieu, 75005, Paris, France

magali.roux@lip6.fr

Keywords Data repository, multi-omics, metadata, biocuration, standard, ontology.

1 Introduction

The Standards-based Infrastructure with Distributed Resources (SIDR) is a data and metadata repository for multi-assay experiments. To overcome the "silo" organization ground on a discipline and/or technology basis, the purpose of the SIDR initiative is to provide the community with a data resource that collide all omics experiments in a standard format and to contribute to a worldwide network of interoperable data repositories. To achieve this, the CNRS Institute for Biological Science (INSB) launched the SIDR project in November 2008 to help collecting and sharing standardized data and metadata in biology. Developments are currently performed at the Institute for Scientific and Technical Information (INIST), a CNRS facility, which has expertise in the field of biocuration of digital resources and bibliographic databases.

The objectives of SIDR being the integration, the preservation and the sharing of data resulting from multi-omics and other biological techniques, the main constraints are the interoperability of metadata. To achieve this goal, SIDR collects and structures metadata by combining different approaches. Metadata are collected: (i) in a standard umbrella technical format (ISA-Tab) [1] which overlays technology-specific omics data models in a metamodeling-like architecture [2]; as such, SIDR figures among ISA case studies (see: http://isatab.sourceforge.net/case_studies.html), (ii) according to the appropriate community/technological standards and guidelines (available at http://mibbi.org/), and (iii) annotated using ontologies (available at http://bioportal.bioontology.org/). Finally, metadata are centralized in a repository, linked to data files and identified by Digital Object Identifiers (DOI). A prototype was released in December 2009 and the actual (V1) version was put online at http://sidr-dr.inist.fr in November 2010.

2 The SIDR V1 Release

Metadata and data architecture. The SIDR metadata object model is intended to be generic and applicable to any omics or other biological techniques. To that aim, SIDR has mapped the ISA-Tab specifications to the FuGE (Functional Genomics Experiment) object-oriented model [3]. This mapping uses a two-stage process: (i) a metamodel architecture was delineated to elicit the syntax of the mapping and resulted in producing an ISA-OM model; (ii) the ISA-OM model was utilized as a "helper" model to define the semantic of the mapping. The objects of the SIDR model are then persisted into a PostgreSQL relational database. Data files and metadata are distinct objects with their own identifiers, and metadata reference their related data files; data may be either stored in the SIDR repository or on a third-party site, including the data production site.

Metadata and data collection. In the development phase, metadata were collected according to the ISA-Tab format, by using ISA-Tab spreadsheet templates or the ISA-Creator® tool [4]. Metadata are transformed in XML dialect (ISA-ML) before being stored in a database. For now, SIDR repository contains 520 data files related to 24 studies grouped in 13 investigations. Main features concerning the current metadata content of the repository are listed in Table 1.

Feature type	Content
Organism	Homo sapiens, Mus musculus, Caenorhabditis elegans, Escherichia coli, Cucumis melo
Measurement type	Molecular structure, Molecular interaction, Cell counting, Transcription profiling, Transcription binding site identification, Protein-protein interaction, Metabolite profiling
Technology type	Mass and NMR spectroscopy, DNA microarray, Flow cytometry, Nucleotide sequencing
Vocabularies/Ontologies used for annotations	BTO, CHEBI, CL, EFO, FIX, GO , IMR, MI, MS, MSH, NCBITaxon, NCIt, OBI, PO, SO, SwissProt, SWO, UO, WBbt, WBls, WBPhenotype.

Table 1. Main features of the metadata in SIDR V1 release.

Metadata searching. To anticipate performance issues in querying large datasets, a search engine allows eliciting and retrieving complex data. Actually, SIDR V1 allows simple queries and the search engine retrieves datasets by browsing straightforward indexed terms. Metadata files can be downloaded in both ISA-Tab and FuGE formats. In the next (V2) version (December 2011), complex queries will be allowed; for example: "give all *Metabolite profiling* (MeasurementType) and *Transcription profiling* (MeasurementType) performed with *any organism* (Organism) treated for *12 hours* (Factor) with *Salicylic acid* (Factor)".

3 Issues in Metadata Biocuration

The submission of metadata by researchers themselves, owning most of the times insufficient knowledge about standards, annotation and ontologies, may result in discrepancies and inaccuracies in metadata description and hamper the comparison and the reuse of the datasets. This highlights the importance of biocurators recognized for their expertise in organizing knowledge, retrieving valuable information and designing databases, in producing consistently annotated experimental metadata. Looking forward scalability issues in data curation, we are currently developing tools to alleviate the biocuration effort.

4 Conclusion

Rapid advances in genome-wide technologies coupled with extensive access to large amounts of highly detailed scientific data have dramatically increased the difficulties of researchers in data handling, storage and retrieval. We think that academic data repositories might provide public, well-recognized places to store and share data. SIDR further developments will focus on methods and tools to improve data interoperability.

Acknowledgements

We acknowledge the kind contributions and cooperation of all data providers to the SIDR repository.

References

[1] S-A. Sansone, P. Rocca-Serra, M. Brandizi, A. Brazma, D. Field, J. Fostel, A. G. Garrow, J. Gilbert, F. Goodsaid, N. Hardy, P. Jones, A. Lister, M. Miller, N. Morrison, T. Rayner, N. Sklyar, C. Taylor, W. Tong, G. Warner and S. Wiemann, The first RSBI (ISA-TAB) workshop: "can a simple format work for complex studies?" *Omics*, 12:143-149, 2008.

[2] M-N. Terrasse and M. Roux, Metamodelling architectures for complex data integration in systems biology. *In.t J. Biomed. Eng. Technol.*, 3:22-42, 2010.

[3] A. R. Jones, M. Miller, R. Aebersold, R. Apweiler, C. A. Ball, A. Brazma, J. DeGreef, N. Hardy, H. Hermjakob, S. J. Hubbard, P. Hussey, M. Igra, H. Jenkins, R. K. Jr. Julian, K. Laursen, S. G. Oliver, N. W. Paton, S-A. Sansone, U. Sarkans, C. J. Jr. Stoeckert, C. F. Taylor, P. L. Whetzel, J. A. White, P. Spellman and A. Pizarro, The Functional Genomics Experiment model (FuGE): an extensible framework for standards in functional genomics, *Nature Biotech.*, 25,1127-1133, 2007.

[4] P. Rocca-Serra, M. Brandizi, E. Maguire, N. Sklyar, C. Taylor, K. Begley, D. Field, S. Harris, W. Hide, O. Hofmann, S. Neumann, P. Sterk, W. Tong and S-A Sansone, ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics*, 26:2354-2356, 2010.

eDystrophin: a Database from *DMD* in-frame Mutations to Dystrophin Structure

Aurélie NICOLAS[1], Céline LUCCHETTI-MIGANEH[2], Frédérique BARLOY-HUBLER[2] and Elisabeth LE RUMEUR[1]

[1] Team RMN-ILP, UMR CNRS 6026, Avenue du Professeur Léon Bernard, CS 34317, 35043 Rennes Cedex
{aurelie.nicolas, elisabeth.lerumeur}@univ-rennes1.fr

[2] Team SP@RTE, UMR CNRS 6026, 2 Avenue du Général Leclerc, 35042 Rennes Cedex
{celine.lucchetti, frederique.hubler}@univ-rennes1.fr

Keywords Dystrophin, Becker Muscular Dystrophy, database, protein structure.

1 Introduction

Dystrophin, encoded by the largest human gene *DMD*, is a 427 kDa sarcolemmal protein found predominantly in skeletal and cardiac muscles [1,2]. Dystrophin is composed of four structural domains with the major central rod domain composed by 24 repeats of about 100-110 residues. Each repeat has an identifiable structure in alpha-helical triple coiled-coil homologous to spectrin repeats. Dystrophin is also characterized by various binding domains (actin, lipids, nNOS, ß-dystroglycan…) [3].

Hundreds of mutations that conserve the open reading frame or not have been observed in human patients and lead to respectively, Becker (BMD) and Duchenne (DMD) Muscular Dystrophies. DMD is a severe disease that considerably reduces patient's life span whereas BMD is less severe with a large spectrum of severity [4]. Generally, DMD patients have a reduced level or an absence of dystrophin whereas BMD patients show a reduced level of mutated dystrophin. The mutations are for a large part deletions or duplications of one or several exons. When the reading frame is conserved, the mutations lead to internally truncated or lengthened dystrophin molecules. These mutated dystrophins are largely used as patterns for the design of dystrophin to be expressed in gene therapy strategies. Therefore, it is highly relevant to gain great knowledge of their effects in muscle cells.

In order to have a more detailed view of the consequences of the in-frame mutations, we developed a new database called eDystrophin specifically dedicated to the proteins produced by these mutations. For this purpose, eDystrophin gathers and builds wild type and mutation information going from the *DMD* gene to the dystrophin molecules. Concerning the mutated proteins, we have been able to model the three dimensional structure of the mutated parts of the dystrophin by homology modeling, giving some new insights above the consequences of the mutation on the structure and function of the dystrophin. This database is the first that focuses on the dystrophin protein and that correlates information between protein isoforms and structures with pathology phenotypes.

2 The Database: eDystrophin

eDystrophin is a database with only in-frame mutations of the *DMD* gene. The website has three distinct parts: Knowledge (dystrophin state of art), Wild type dystrophin data and Mutated dystrophin data.

Gene data: The cDNA sequences and exons of 18 *DMD* variants were obtained from GenBank. eDystrophin contains 209 mutations that conserve the open reading frame gathered from literature and from J. Chelly group (Institut Cochin – Paris). The mutation nomenclature is unified according to HGVS (http://www.hgvs.org/mutnomen/).

Protein data: Eighteen isoforms of dystrophin were gathered from GenBank such as cDNA sequences. The structural and functional domains are defined as from data of the literature. Thirty-five structural domains and 14 binding domains are available in eDystrophin. Only two X-Ray structures of dystrophin domains (ABD1 and Cys rich domain) were available from PDB (http://www.pdb.org/). 3D model of 24 rod domain repeats obtained in our group were added to eDystrophin. In a second part, mutated protein sequences were calculated from the mutated cDNA, and the presence of structural and binding domains were

deduced while lacking domains are highlighted. A structural homology model of each mutated protein domain was generated with I-TASSER when the mutation occurs in the central rod domain. All models were thermodynamically and stereochemistry checked using various softwares.

Clinical data: At the present time, eDystrophin contains 884 patients. When available, information about disease state (DMD, BMD or IMD) and the severity grading, the dystrophin expression in muscles, the presence of an additional cardiomyopathy and/or a mental retardation is done. The presentation of all these data is standardized in order to allow a statistical analysis of the database. For each mutation type, the repartition of this information is available through histograms.

3 Cases Studies

We can study some mutations with eDystrophin such as exon deletion 45-55. In the left menu, the user can choose Explore database/Mutated dystrophin/Search exon deletion to find this mutation. Here, he can select “deletion” and “exon 45” to display a list of mutations that involve exon 45. In this list we can select del 45-55 and the result page appears. The summary array provides general information like the protein size and sequence as well as the phenotype associated with the mutation. More information about patients is available by the detail “link”. As a result, this deletion appears to be associated with a majority of BMD patients (more than 91%) with a mild grading of severity. Then three boxes are provided to study protein in detail. In a first box, a protein scheme is available to easily visualize the mutation localization. In the deletion of exon 45-55, repeats 18 to 21 and hinge 3 are lacking and repeats 17 and 22 are truncated. In a second box, 3D structural model of repeats 16 to 23 is displayed. The model appears like a crystallographic structure. In the deletion studied here, the two truncated repeats are able to reconstitute fold that organizes in a triple coiled-coil as in wild type repeats. Therefore, it appears that this particular mutation may not have deleterious consequence on the central rod domain structure. In a third box, binding domains affected by the mutation are notified. In our case, ABD2, LBD2 and nNOS are partially lacking. ABD2 and LBD2 are large domains with electrostatic interactions so we can assume the binding property is always possible. This mutation may affect strongly nNOS binding. All this protein information can be correlated with clinical data. In this case, the absence of a priori deleterious effect of the mutation may explain the observed mild phenotype. This exon deletion can be a candidate for exon therapy.

4 Conclusion

eDystrophin database allows, for the first time, integration of data from the mutation on the *DMD* gene to the consequences of mutation on the protein structure the functional binding domains. In that sense,eDystrophin database can help to choose the best targets for gene therapies.

Acknowledgements

We thank Jamel Chelly, France Leturcq and Rahah Ben Yaou from Institut Cochin (Paris) to provide clinical data. We also thank plateforme Genouest which host database and Swiss Prot Institute of Bioinformatics to adapted their tool MyDomains to our needs. AN is supported by the Centre national de la recherche scientifique and CLM by Agence Nationale de la Recherche.

References

[1] E.P. Hoffman, R.H. Brown, Jr. and L.M. Kunkel, Dystrophin: the protein product of the Duchenne muscular dystrophy locus. *Cell*, 51:919-928, 1987.

[2] M. Koenig, E.P. Hoffman, C.J. Bertelson, A.P. Monaco, C. Feener and L. M. Kunkel, Complete cloning of the Duchenne muscular dystrophy (DMD) cDNA and preliminary genomic organization of the DMD gene in normal and affected individuals. *Cell*, 50:509-517, 1987.

[3] E. Le Rumeur, S.J. Winder and J.F. Hubert, Dystrophin: More than just the sum of its parts. *Biochim Biophys Acta*, 1804:1713-1722, 2010.

[4] A.P. Monaco, C.J. Bertelson, S. Liechti-Gallati, H. Moser and L.M. Kunkel, An explanation for the phenotypic differences between patients bearing partial deletions of the DMD locus. *Genomics*, 2:90-95, 1988.

`mixOmics`: an `R` Package for the Integration of 'omics' Data

Ignacio GONZÁLEZ[1], Kim-Anh LÊ CAO[2] and Sébastien DÉJEAN[1]

[1] Institut de Mathématiques, UMR5219 CNRS, Université Paul Sabatier, 31062 Toulouse Cedex 9, France
{ignacio.gonzalez,sebastien.dejean}@math.univ-toulouse.fr

[2] Queensland Facility for Advanced Bioinformatics, University of Queensland, 4072 St Lucia, QLD, Australia
k.lecao@uq.edu.au

Abstract *In the current 'omics' era, the integrative or joint analysis of large amount of data, such as genomics, proteomics, metabolomics, interactomics, is becoming crucial to unravel the relationships between different biological functional levels and to better understand biological systems as a whole. Integrating multiple highly dimensional omics data represent both computational and analytical challenges. New methodologies need to be developed to extract and visualise meaningful information. We introduce `mixOmics`, an `R` package dedicated to the integrative analysis of biological data that implements several recently developed statistical methodologies to enlighten correlation between two matching data sets and perform simultaneous variable selection in both sets. We also propose useful graphical outputs to aid the interpretation of these promising and flexible analysis tools. `mixOmics` has been successfully applied in various biological integrative studies and is undoubtedly useful to give more insight into biological systems.*

Keywords integrative analysis, regularized Canonical Correlation Analysis, sparse PLS.

1 Background

'Omics' data now form a core part of systems biology by enabling researchers to understand the integrated functions of a living organism. However, the available abundance of such data (genomics, transcriptomics, proteomics ...) is not a guarantee of obtaining useful information in the investigated system if the data are not properly processed and analyzed to highlight this useful information. A major challenge with the integration of omics data is therefore the extraction of discernable biological meaning from multiple omics data. The statistical integration of two highly dimensional data sets, combined with variable selection from both sets has attracted considerable attention these last few years. Regularized and sparse variants of Canonical Correlation Analysis (CCA) [1,2,3,4,5] and Partial Least Squares regression (PLS) [6,7] were subsequently proposed. However, most of these articles are limited to numerical results, and little attention has been paid to either the interpretation of the results or the graphical outputs. It is therefore crucial to propose a software that combines both computationally efficient statistical methodologies and graphical outputs that can aid the interpretation.

2 Methods

To address this issue, we have developed and implemented several exploratory tools. To describe briefly *regularized* CCA and *sparse* PLS, let first denote the two data matrices $X_{n\times p}$ and $Y_{n\times q}$ with standardized columns, where the p and q variables are measured on the same n samples. Both approaches seek for p- and q-dimensional weight vectors (*loading vectors*), and n-dimensional vectors (*scores* or *latent vectors*).

Regularized Canonical Correlation Analysis. CCA maximizes the correlation between linear combinations of the variables from each data set. However, CCA requires the computation of the inverses of the covariance matrices XX' and YY' that are singular if $p >> n$ and $q >> n$. The introduction of l_2 penalties [8,9] make them invertible in a regularized CCA (rCCA).

Sparse Partial Least Squares Regression. On the contrary to CCA, PLS circumvents the issue of ill-conditioned matrices by performing local regressions. In order to give interpretable results and remove noisy variables, [6,4] proposed to add l_1 penalizations to each PLS loading vector, in which the magnitude of the coefficients

indicate the importance of the variables in the integrative model. As a result, many coefficients in these vectors are set to zero. This naturally allows for a simultaneous variable selection in the two data sets. Recently, a discriminant analysis framework was introduced (sPLS-DA) for a supervised setting [10].

3 Graphical Outputs

In addition, a strong focus is given in `mixOmics` for graphical outputs to make the interpretation of the obtained results easier to understand [13]. Typical plots in 2D and 3D such as samples representation or correlation circles to represent variables are available. New developments were also recently proposed to generate Clustered Image Map and infer Relevance Networks. A further assessment of the biological relevance of such graphical tools showed that the inferred networks were relevant to the system under study [14].

4 Conclusion

`mixOmics` is a computationally efficient `R` package that enables the integrative analysis of large data sets using exploratory techniques. The package provides relevant graphical outputs to explore the relationships between two data sets. The relevancy of the implemented approaches and the graphical outputs has been previously demonstrated [1,6,11,12]. `mixOmics` is easily applicable to systems biology studies and will undoubtedly help in addressing fundamental biological questions and in understanding systems as a whole. `mixOmics` is freely available from `http://cran.R-project.org` or from the website companion `http://math.univ-toulouse.fr/biostat/mixOmics` that provides full documentation, tutorials and case studies.

References

[1] I. González, S. Déjean, P. Martin, O. Gonçalves, P. Besse and A. Baccini, Highlighting relationships between heteregeneous biological data through graphical displays based on regularized Canonical Correlation Analysis. *Journal of Biological Systems*, 17:173-199, 2009.

[2] S. Waaijenborg, P.C. Verselewel de Witt Hamer and A. Zwinderman, Quantifying the association between gene expressions and dna-markers by penalized canonical correlation analysis. *Statistical Applications in Genetics and Molecular Biology*, 7:1, 2008.

[3] E. Parkhomenko, D. Tritchler and J. Beyene, Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology*, 7, 2009.

[4] K-A. Lê Cao, P. Martin, C. Robert-Granié and P. Besse, Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics*, 10:34, 2009.

[5] D.M. Witten, R. Tibshirani and T. Hastie, A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10:515-534, 2009.

[6] K-A. Lê Cao, D. Rossouw, C. Robert-Granié and P. Besse, A sparse PLS for variable selection when integrating omics data. *Statistical Applications in Genetics and Molecular Biology*. 7:29, 2008.

[7] H. Chun and S. Keles, Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 72:3-25, 2010.

[8] H.D. Vinod, Canonical ridge and econometrics of joint production. *Journal of Econometrics*. 6:129-137, 1976.

[9] I. González, S. Déjean, P. Martin and A. Baccini, CCA: An R Package to extend Canonical Correlation Analysis. *Journal of Statistical Software*, 23:12, 2008.

[10] K-A. Lê Cao, S. Boitard and P. Besse, Sparse PLS Discriminant Analysis: biologically relevant feature selection and graphical displays for multiclass problems. *Technical report*, 2011.

[11] E. Yergeau, S. A. Schoondermark-Stolk, E.L. Brodie, S. Déjean, T.Z. DeSantis, O. Gonçalves, Y.M. Piceno, G.L. Andersen and G.A. Kowalchuk, Environmental microarray analyses of Antarctic soil microbial communities. *The International Society for Microbial Ecology Journal*, 3(3):340-351, 2009.

[12] S. Combes, I. González, S. Déjean, A. Baccini, N. Jehl, H. Juin, L. Cauquil, B. Gabinaud, F. Lebas and C. Larzul, Relationships between sensorial and physicochemical measurements in meat of rabbit from three different breeding systems using canonical correlation analysis. *Meat Science*, 3:835-841, 2008.

[13] K-A. Lê Cao, I. González and S. Déjean, integrOmics: an R package to unravel relationships between two omics data sets. *Bioinformatics*, 25(21):2855-2856, 2009.

[14] I. González, K-A. Lê Cao, M.J. Davis, S. Déjean, Insightful graphical outputs to explore relationships between two 'omics' data sets. *Technical report*, 2011.

Building CoryneCyc Database

A “Pathway/Genome” Database For *Corynebacterium glutamicum*

Thomas Duigou[1,2] and Bruno Bost[3]

[1] Institut de Génétique et Microbiologie, UMR8621 CNRS, Université Paris-Sud 11, 91405, Orsay Cedex, France

[2] Master Bioinformatique, Université de Rouen, France

t.duigou@gmail.com

[3] Institut de Génétique et Microbiologie, UMR8621 CNRS, Université Paris-Sud 11, 91405, Orsay Cedex France

bruno.bost@u-psud.fr

Keywords *Corynebacterium glutamicum*, “Pathway/Genome” Database Generation, Pathway Tools, Additional Developments.

Mise en place de la Base de Données CoryneCyc

Une Base de Données de type « Pathway/Genome » pour *Corynebacterium glutamicum*

Mots-clés *Corynebacterium glutamicum*, Génération d'une Base de Données de type « Pathway/Genome », Pathway Tools, Développements Additionnels.

1 Introduction

Corynebacterium glutamicum est une bactérie largement utilisée dans l'industrie agroalimentaire, en particulier pour la production d'acides aminés (glutamate, lysine).

L'équipe *Physiologie et Métabolisme des Corynébactéries* de l'IGM étudie le fonctionnement et la régulation du métabolisme de *C. glutamicum* en condition de stress depuis de nombreuses années. Les recherches de l'équipe sont menées en utilisant une approche pluridisciplinaire associant des méthodes expérimentales et bioinformatiques. Au cours des années, l'équipe a accumulé un ensemble de données hétérogènes (prédictions bioinformatiques, niveaux d'expression, phénomènes de régulation, flux de métabolites) sur *C. glutamicum*.

À l'heure actuelle, les informations disponibles sur *C. glutamicum* sont « dispersées », aussi bien à l'échelle de l'équipe, qu'à l'échelle des données consultables dans les bases de données généralistes (NCBI, PubMed, EBI, KEGG, Brenda) et spécialisées (CoryneRegNet). L'absence d'un outil dédié à *C. glutamicum*, qui fournirait un point d'entrée unique pour accéder aux données portant sur le sujet, complique la recherche d'informations, et les analyses bioinformatiques « inter-bases » sont limitées par le schéma propre à chaque base de données.

2 Objectif et Contexte

Au regard de ces constats, l'équipe *Physiologie et Métabolisme des Corynébactéries* a proposé de regrouper un maximum d'informations dans une base de données unique, suivant un schéma qui reflèterait les « briques » fonctionnelles élémentaires et les interactions existantes chez un organisme tel que *C. glutamicum*. Cette base de données doit intégrer les données génomiques, les données disponibles au laboratoire et dans la littérature, sur les réseaux métaboliques et de régulations chez cet organisme.

Cette proposition s'est concrétisée au cours de la mission que je réalise dans le cadre de ma formation en apprentissage et en alternance (master 2 de bioinformatique de l'université de Rouen). Au cours de cette mission de deux ans (2009 – 2011), l'objectif est de mettre en place un outil permettant d'une part d'accéder

facilement aux données relatives à *C. glutamicum* et d'autre part de générer ensuite d'autres bases de données pour des organismes proches à des fins d'études comparatives.

Pour mettre en place un tel outil, nous utilisons une base de données de type « Pathway/Genome » (Pathway/Genome Database, PGDB), qui regroupe des informations à la fois sur le génome et sur les voies métaboliques, ainsi que sur les entités intermédiaires comme les régulateurs transcriptionnels, les complexes protéiques ou les modulateurs de réactions enzymatiques. Parmi les modèles de PGDB existants, nous avons choisi d'utiliser une PGDB de type « BioCyc » [1].

3 Résultats

J'ai consacré une première partie de ma mission à la mise en place de l'infrastructure technique nécessaire au fonctionnement de l'application Pathway Tools [2] qui permet de générer des PGDB de type BioCyc. Cette infrastructure autorise l'utilisation de toutes les fonctionnalités offertes par l'application : génération, édition et consultation d'une base de données. J'ai ensuite développé des programmes afin d'améliorer le processus de génération de bases de données par Pathway Tools. Enfin, j'ai défini et appliqué une procédure intégrant les outils de génération de Pathway Tools et mes propres développements pour générer une première version de CoryneCyc.

Dans une deuxième partie, j'ai développé une fonctionnalité qui permet aux scientifiques de contribuer à l'amélioration du contenu de la base de données CoryneCyc directement depuis une interface web. J'ai conçu un site « privé » en vue de proposer divers outils d'analyses de données, et d'entreposer des résultats issus d'expériences sur *C. glutamicum* utilisant des technologies à haut débit.

Enfin, je suis actuellement en train d'exploiter la base de données CoryneCyc en réalisant une étude sur l'usage des codons dans le génome de *C. glutamicum*. L'objectif de cette étude est de regrouper les séquences codantes du génome sur le critère de leur usage des codons, puis de rechercher à travers l'ensemble des informations disponibles dans CoryneCyc si des caractéristiques « biologiques » (classe fonctionnelle, processus biologique, domaine protéique, etc.) sont partagées par des séquences contenues dans un même regroupement, ou discriminent des séquences contenues dans des regroupements différents.

4 Perspectives

CoryneCyc est déjà utilisé par l'équipe *Physiologie et Métabolisme des Corynébactéries* pour interpréter des données issues de technologies à haut débit (transcriptomique, protéomique). Par la suite, d'autres PGDB pour des organismes proches de *C. glutamicum* vont être générées, en utilisant comme base pour la génération les annotations de la base de données MetaCyc et de la PGDB de *C. glutamicum*. D'ici l'été 2011, l'accès à CoryneCyc sera ouvert à la communauté scientifique.

Les données qui font partie de CoryneCyc pourront servir de point de départ pour des analyses bioinformatiques fonctionnelles, comme la modélisation automatisée de réseaux métaboliques et de régulations. Ces analyses peuvent être grandement facilitées par l'utilisation d'API dédiées aux bases de données de type « BioCyc ».

Outre l'utilisation des API « PerlCyc » et « JavaCyc » déjà existantes, la mise au point d'une API « RCyc » permettrait de coupler la puissance du logiciel de statistique R avec l'ensemble des informations contenues dans les PGDB de type « BioCyc ». Dans le cadre de mon étude de l'usage des codons, j'ai commencé à développer cette API « RCyc ».

Références

[1] R. Caspi, H. Foerster, C.A. Fulcher, P. Kaipa, M. Krummenacker, M. Latendresse, S. Paley, S.Y. Rhee, A. Shearer, C. Tissier, T.C. Walk, P. Zhang and P.D. Karp, The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.*, 36(suppl 1):D623-631, 2007.

[2] P.D. Karp, S.M. Paley, M. Krummenacker, M. Latendresse, J.M. Dale, T.J. Lee, P. Kaipa, F. Gilham, A. Spaulding, L. Popescu, T. Altman, I. Paulsen, I.M. Keseler and R. Caspi. Pathway Tools version 13.0: Integrated Software for Pathway/Genome Informatics and Systems Biology. *Brief. Bioinform.*, 11(1): 40-79, 2010.

Overview of The Universal Protein Resource (UniProt)

What's in and How to Use UniProt Databases ?

Benoît Bely[1], Eleanor Stanley[2] and Maria Martin[1]

[1] The EMBL outstation - The European Bioinformatics Institue,
Wellcome Trust Genome Campus, Hixton, Cambridge CB11 1SD, UK
{benoit.bely, martin}@ebi.ac.uk

[2] Swiss Institute of Bioinformatics - Swiss-Prot Group
CMU - 1, rue Michel Servet, CH-1211 Geneva 4, Switzerland
eleanor.stanley@isb-sib.ch

Keywords Protein database, UniProtKB, UniParc, UniRef, UniMES, UniSave, UniMart, programmatic access, REST, UniProtJAPI, complete proteome dataset, evidence code.

1 Introduction

UniProt is the central resource for storing and interconnecting information from large and disparate sources, and the most comprehensive catalog of protein sequence and functional annotation. UniProt is built upon the extensive bioinformatics infrastructure and scientific expertise at European Bioinformatics Institute (EBI), Protein Information Resource (PIR) and Swiss Institute of Bioinformatics (SIB).

2 Different components optimized for different uses

The UniProt Knowledgebase (UniProtKB) [1] is a central access point for integrated protein information with cross-references to multiple sources. The UniProtKB contains two sections. UniProtKB/Swiss-Prot contains records with full manual annotation (or computer-assisted, manually-verified annotation) performed by biologists and based on published literature and sequence analysis. UniProtKB/TrEMBL contains computationally generated records enriched with automatic classification and annotation.

The UniProt Archive (UniParc) [2] is a comprehensive and non-redundant database of protein sequences extracted from public databases UniProtKB/Swiss-Prot, UniProtKB/TrEMBL, PIR-PSD, EMBL, EMBL WGS, Ensembl, IPI, PDB, PIR-PSD, RefSeq, FlyBase, WormBase, H-Invitational Database, TROME database, European Patent Office proteins, United States Patent and Trademark Office proteins (USPTO) and Japan Patent Office proteins.

UniProt Reference Clusters (UniRef) [3] consist of three databases of clustered sets of protein sequences from UniProtKB and selected UniParc records. In the UniRef90 and UniRef50 databases no pair of sequences in the representative set has >90% or >50% mutual sequence identity. The UniRef100 database presents identical sequences and sub-fragments as a single entry with protein IDs, sequences, bibliography, and links to protein databases.

The UniProt Metagenomic and Environmental Sequences (UniMES) database is a repository specifically developed for the expanding area of metagenomic and environmental data.

UniProt include other developments such as the ID mapping service (http://www.uniprot.org/mapping) which allows users to map between UniProtKB and more than 85 other data sources. Also The UniProtKB Sequence/Annotation Version Archive (UniSave) is a repository of UniProtKB/Swiss-Prot and UniProtKB/TrEMBL entry versions. UniMart (http://www.ebi.ac.uk/uniprot/biomart/martview) is a BioMart database for querying UniProtKB data, with a cross-querying facility to join to/from other BioMart databases. UniProtJAPI [6] is a Java library which facilitates the integration of UniProt data into Java-based software applications.

UniProt is updated and distributed every four weeks and can be accessed online for searches or download at http://www.uniprot.org.

3 Complete Proteome Data Sets

The species coverage of the UniProtKB Complete proteome data sets will be extended to include all the species within the International Protein Index (IPI) (Arabidopsis thaliana, Bos taurus, Danio rerio, Gallus gallus, Homo sapiens, Mus musculus, Rattus norvegicus) andan additional five requested by users (Canis familiaris, Sus scrofa, Caenorhabditis elegans, Drosophila melanogaster and Saccharomyces cerevisiae). Due to strong collaborations between UniProt, Ensembl, NCBI and Vega, the quality of gene predictions from a reference genome has increased greatly. A consequence being that the requirement for IPI has diminished and is scheduled for closure summer 2011. All IPI users are being encouraged to use UniProtKB complete proteome sets as the approved replacement. The complete UniProtKB proteome sets will be based on existing UniProtKB sequences supplemented by high quality predictions imported from Ensembl. Incorporation of Ensembl sequences into UniProtKB has been achieved for all the IPI species and will expand to other species of interest in the near future. The proteome sets are available for download from the UniProt FTP and web site. We expect these will be very useful for our users as they will eliminate the need to combine data from different databases.

4 Evidence Code in UniProt XML

Each UniProtKB entry combines information from a wide range of sources and becomes a central hub for the collection of functional information on proteins. An entry will be generated from a DDBJ/ENA/GenBank nucleotide record or other sequence database (for example Ensembl and PDB), and be enriched with cross references to a large number of other databases, output from automatic annotation predictions, sequence analysis programs and, if selected for manual curation, experimental characterisation data from scientific literature and curator-evaluated computational analysis will be added. Because of this huge variety and number of data sources, it is vital that users are provided with a way of tracing the origin of each piece of information in an entry and given the opportunity to evaluate it. The UniProt Consortium has begun to approach this challenge by the use of an evidence attribution system. Evidence is attached to most data items in a UniProtKB entry thereby identifying the source(s) and/or method(s) used to generate the data. The evidences will be standardised using the widely known Evidence Code Ontology (ECO). In future, any database from which we import data that also has evidence attributions, this data will also be incorporated.

5 Demonstration

The demonstration will cover:

1. A brief description of the UniProt databases.
2. Accessing UniProt using simple query syntax. The user will be presented with helpful suggestions and hints.
3. Exploration of sequence similarity searches, alignments and ID mapping tools provided.
4. Accessing UniProt data programmatically.

This demonstration will also encourage user interaction and feedback.

References

[1] The UniProt Consortium, Ongoing and future developments at the Universal Protein Resource.. *Nucleic Acids Res.,* 38:D142-D148, 2011

[2] R. Leinonen, FG. Diez, D. Binns, W. Fleischmann, R. Lopez and R. Apweiler, UniProt archive. *Bioinformatics,* 20:3236–3237,2009.

[3] BE. Suzek, H. Huang, P. McGarvey, R. Mazumder and CH. Wu, UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics.* 23:1282–1288, 2007.

[4] S. Patient, D. Wieser, M. Kleen, E. Kretschmann, MJ. Martin and R. Apweiler, UniProtJAPI: a remote API for accessing UniProt data. *Bioinformatics.* 24:1321–1322, 2008.

Djeen: A High Throughput Multi-Technological Research Information Management System for the Joomla! CMS

Olivier Stahl[1,2,3], Arnaud Guille[1,2,3], Fanny Blondin[1,2,3], Pascal Finetti[1,2,3], Samuel Granjeaud[3,4] and Ghislain Bidaut[1,2,3]

[1] Centre de Recherche en Cancérologie de Marseille, U891 Inserm, 13009, Marseille, France.
[2] Institut Paoli-Calmettes, Marseille, 13009, Marseille, France
[3] Université de la Méditerranée, Marseille, 13009, Marseille, France.
[4] Techniques Avancées pour le génome et la Clinique, Inserm U928 Inserm, Marseille, France
{olivier.stahl; arnaud.guille; fanny.blondin; ghislain.bidaut}@inserm.fr,
finettip@marseille.fnclcc.fr, granjeaud@tagc.univ-mrs.fr

Keywords RIMS, database, minimum information, Joomla! CMS.

1 Introduction

The current growth of high-throughput experimentation in biological research implies a huge challenge to store and share generated data and their annotations [1]. Manipulating and integrating heterogeneous data types in large scale collaborative projects involving several geographically separated laboratories remains an even higher hurdle for which proper management systems must be deployed. Data annotations need to be recorded and homogenized for proper integration, by using controlled vocabularies [2], while respecting Minimum Information standards, such as MIAME [3] or MIFlowCyt or others.

In contrast with previously proposed databases and Laboratory Information Management Systems (LIMS), the Database for Joomla's Extensible Engine (Djeen) is conceptually simple (simple database scheme and organization of data within a file system), usable for any data types (no technology-dependent semantics was used in the core code) and adapted to the manipulation of large numbers of data files and their annotations, generated in a high throughput way.

Djeen is designed as a Research Information Management System [4] which means that its structure is not modeled on data generated by a particular device, but around laboratory experimentation work flows (figure 1). Longitudinal data integration concepts has also been implemented to address four previously identified fundamental issues in high throughput biomedical data management: data organization, data sharing, collaboration and publication [4].

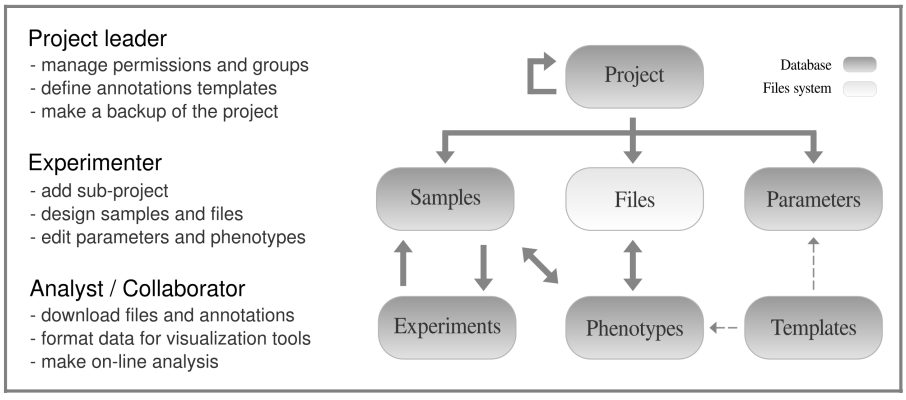


Figure 1. Data architecture and workflow.

2 Architecture and Workflow

Figure 1 shows the internal Djeen structure. It is based on a hierarchy including laboratory experimental workflow composed of 'Project' element at its top, as well as 'Samples', and 'Experiments'. Projects can contain one or several sub-projects. Each one is a coherent set of data - samples and files - described by annotations. Projects are owned by a 'Project Administrator', who controls annotations and users/groups

permissions. Samples and Experiments are connected structures defined as follows. Samples describe biological objects, and Experiments describe processes to transform these objects into other Samples.

Sample and file annotations are managed using 'Templates' to enforce the type of annotations to be provided by experimenters, permitting further data integration and minimum information gathering across collaborative work. Templates are specified by Project Administrators who defined their structures and elements that must be populated. Two types of annotations are stored, global experimental annotations that describe experimental design (stored as 'Parameters'), and phenotypes that varies between samples (for example, patients clinical annotations, stored as 'Phenotypes'). Moreover, Project Administrators can backup a whole project through the 'History' function.

Djeen has the capability to manage instruments configurations (for instance a flow cytometer). To do so, a user has to define a number of samples to be processed and directly export the experimental design to the device. Instrument export feature is currently implemented for flow cytometry data (FCS format) to the BD Diva® software XML format and other instruments interfaces are currently under development. After samples have been processed and measurements generated, data is loaded in Djeen through a network share.

Designed to be user friendly and simple to administrate, Djeen user interface allows quick access to major elements and presents information into clear and simple views. By using templates, it allows streamlining data annotation for high-throughput projects. User management allows to share data only with designated collaborators and release them publicly after publication.

3 Software and Availability

Djeen web interface has been developed as a Joomla! component. Joomla! is an open-source Content Management System (www.joomla.org) featuring a documented API to create advanced extensions that can re-use basic features, such as authentication, back-end administration, database access and web interface. Embedding Djeen within a CMS helps saving costs on in-house development while focusing onto scientific development.

Djeen is available for download at http://bioinformatique.marseille.inserm.fr/djeen under CeCILL licence, and a test instance is available at http://bioinformatique.marseille.inserm.fr/djeentest. Installation is greatly facilitated by a step by step process.

Future developments consists on the creation of exportable instruments configuration templates for high throughout repetitive projects and experiments. Export function to other instruments (Affymetrix GeneChips® system) and visualization software exports are planned for next release.

Acknowledgements

This project has been funded by an Institut National du Cancer grant. Servers running Djeen were funded by a Fondation pour la Recherche Médicale Grant. Additional support was obtained from the Université de la Méditerranée.

References

[1] N. R. Anderson et al., Issues in biomedical research data management and analysis: needs and barriers. *J Am Med Inform Assoc*, 14:478-488, 2007.

[2] G. Bidaut and C. J. Jr. Stoeckert, Large scale transcriptome data integration across multiple tissues to decipher stem cell signatures. *Methods Enzymol*, 467:229-245, 2009.

[3] A. Brazma, Minimum Information About a Microarray Experiment (MIAME)--successes, failures, challenges. *ScientificWorldJournal*, 9:420-423, 2009.

[4] S. Myneni and V. L. Patel, Organization of Biomedical Data for Collaborative Scientific Research: A Research Information Management System. *Int J Inf Manage*, 30:256-264, 2010.

PARYS, a Web Server for Managing Reverse-Phase Protein Array Platform Data

Stéphane Liva[1] , Patrick Poullet[1] , Leanne De Konig[2] , Bérangère Marty[2] , Sylvie Troncale[1] , Carine Danelski[2] , Philippe Hupé[3] , Thierry Dubois[2] and Emmanuel Barillot[1]

[1] U900 INSERM – Mines ParisTech – Institut Curie, Bioinformatics and Computational System Biology of Cancer, Institut Curie, 26 rue d'Ulm, 75248 Paris Cedex 05, France.
{stephane.liva, patrick.poullet}@curie.fr

[2] Translational Research Department – Institut Curie, Hôpital Saint, Louis, 75475 Paris Cedex 10, France.

[3] UMR144, CNRS, Institut Curie, 26 rue d'Ulm, 75248 Paris Cedex 05, France.

Keywords Protein array, protein quantification, antibody.

1 Introduction

The Reverse Phase Protein Array (RPPA) is an emerging high-throughput technology relying on highly specific antibodies to quantify proteins and post-translational activation levels. Because it can be used with nanogram amounts of extracts, thus technology has become a promising approach for proteome analysis of cancer patients. A RPPA platform was set up at Curie Institute (Paris, France) as a result of a partnership with Servier pharmaceutical company. Among the research projects utilizing the platform, a pilot study focuses on finding new therapeutic strategies to treat sub-types of breast cancers by examining the phosphokinome of patient biopsies.

2 Results

To fully exploit the potential of RPPA technology, we have developed PARYS (Protein ARraY Server), a comprehensive bioinformatics environment for the platform. PARYS is composed of :

A LIMS-like section to track major laboratory reagents (e.g. antibodies, tumour samples, cell lines, proteins extracts, arrays) and key processes such as extracts preparation, spotting, antibody labelling and quantification.

A statistical module for optimal antibody dilution calculation, for array normalization (with SuperCurve [1]), for exploratory analysis (Principal Component Analysis and Clustering) [2].

A project section to coherently organize, visualize the data generated. Data-mining tools is implemented to help users retrieve meaningful biological information from the data.

Future improvements include methods for confronting RPPA-based proteomic quantifications with genomic and transcriptomic array data, and implementation of differential analysis.

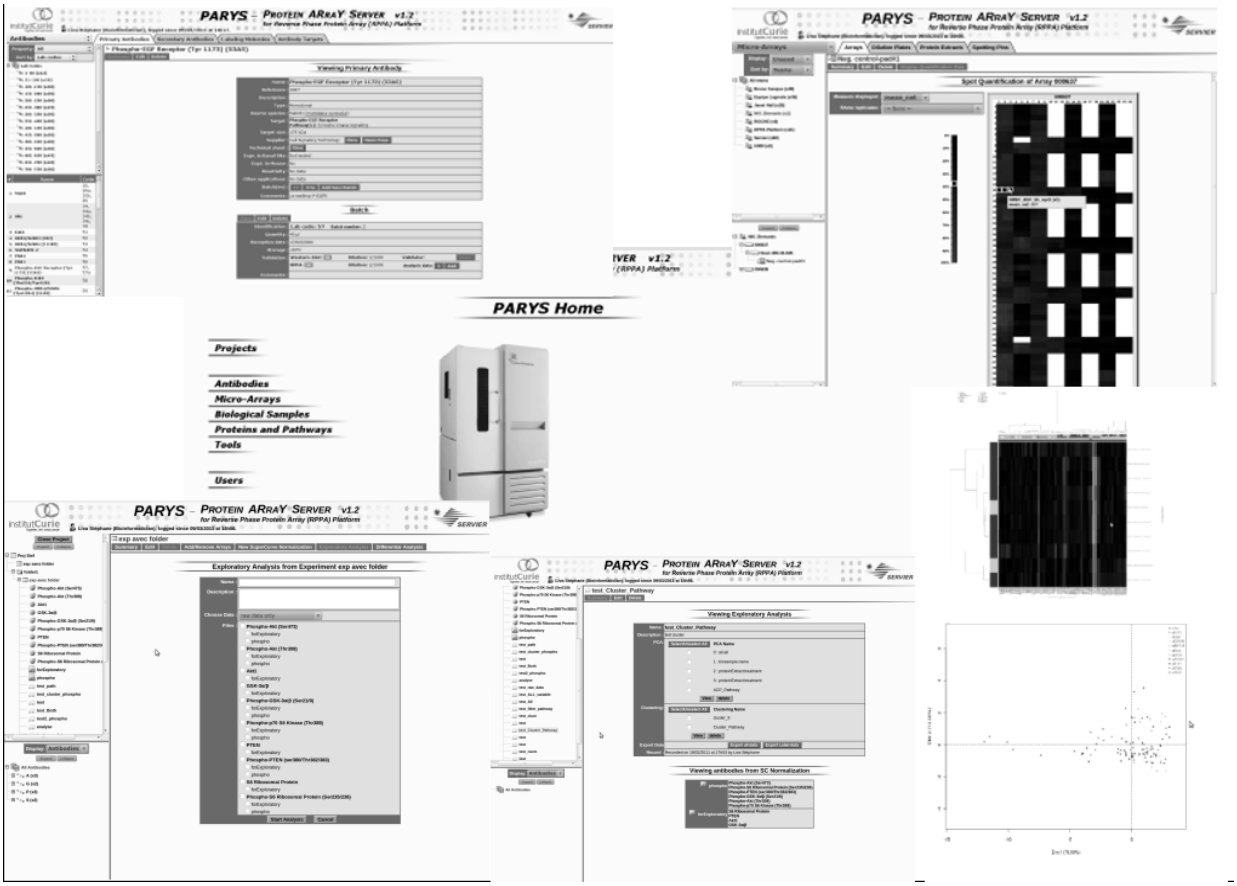


Figure 1.PARYS Interface.

References

[1] J. Hue, X. He, K. Baggerly, K. Coombes, B. Hennessy and G. Mills, Non-parametric quantification of protein lysate arrays. *Bioinformatics*, 23:1986-1994, 2007.

[2] N. Servant, E. Gravier, P. Gestraud, C. Laurent, C. Paccard, A. Biton, I. Brito, J. Mandel and B. Asselain, EMA: A R package to Easy Microarray data Analysis, *BMC research note*, 3:277, 2010.

dbWFA: A Web-Based Database for Functional Annotation of Wheat Transcripts

Jonathan VINCENT[1,2], Marie AGIER[2], Catherine RAVEL[1], M. Fouad BOUZIDI[1], Saïd MOUZEYAR[1] and Pierre MARTRE[1]

[1] INRA - Université Blaise Pascal, UMR1095 GDEC, 234 Avenue du Brézet, Clermont-Ferrand, F-63 100, France
pierre.martre@clermont.inra.fr (author for correspondence)
[2] Université Blaise Pascal - CNRS, UMR6158 LIMOS, BP 10 448, F-63 173 Aubière
agier@isima.fr

Abstract *The search for homologous sequences and orthologs is very useful for analyzing biological sequences, especially when studying a species whose genome is still to be entirely sequenced such as wheat. Steady growth of the size of genomic database makes homologous sequences analysis very time-consuming. Moreover, the functional annotation of the genes based on their sequence homology with model species genomes requires querying unrelated databases. Expressed sequence tags and transcripts represent one of the most important sources of information in gene expression study; however, raw sequences must be annotated before they are of value to the research community. The aim of the present work was to develop a functional annotation database dedicated to bread wheat (*Triticum aestivum*), gathering information about model plant species and linking them to wheat through BLAST results.*

Keywords BLAST, database, homology, orthologs, sequences, transcripts, *Triticum aestivum*.

1 Introduction

In order to annotate expressed sequence tags (ESTs) or transcripts one has to navigate through unrelated databases. Efforts are in progress to gather this kind of information for the model species *Arabidopsis thaliana* and rice (*Oryza sativa*) [1]. To the best of our knowledge, no database is available to relate wheat transcripts information to functional annotation using model species information.

The aim of the present work was to develop a functional annotation database dedicated to wheat, gathering information about model plant species and linking them to wheat through BLAST [2] results. We developed an open-access database relating the transcripts of the wheat set UniGene and the wheat Transcription Factor DataBase (wDBTF) [3] that have been used to design a Custom Wheat Gene Expression 12x135k NimbleGen cDNA microarray [4] to *A. thaliana* and rice databases.

2 Database Structure and Website

The database contains information about Gene Ontologies (geneontology.org), functional annotations (mips.helmholtz-muenchen.de/plant/athal/), MapMan information (mapman.gabipd.org), [5] gene families (arabidopsis.org) and metabolic pathways (plantcyc.org) for model species. More than four million BLAST results were stored, allowing an efficient use of the database. Following the recommendation of the International Wheat Genome Sequencing Consortium (IWGSC) for homology research, the percentages of coverage and identity are used to assign functional annotations to wheat transcripts [6].

The database was built so that one can use two different query's approaches (Fig. 1). From wheat transcripts IDs or NimbleGen gene expression microarray probes to putative family, ontology, function, and/or metabolic pathway through BLAST results, and vice versa. BLAST results allow one to characterize the homology and assign putative functional annotation between two sequences based on coverage and identity threshold values specified by the user.

A website (http://urgi.versailles.inra.fr/Species/Wheat/Tools/dbWFA) was developed in order to grant users with the most common queries that can be applied to the database. It includes searching for genes

involved in specified metabolic pathways, functional annotations, gene ontologies or belonging to specified families. Or inversely, searching for putative annotations related to a wheat transcript or a list of wheat transcripts. Results are available as detailed web/html pages or text files.

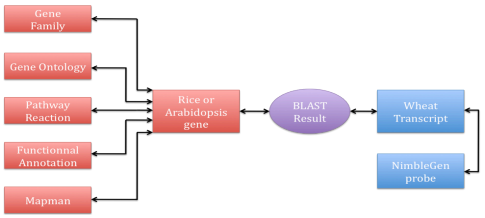


Figure 1. Simplified model of the dbWFA database.

3 Annotation of the Wheat Gene Expression NimbleGen Array

A total of 40 349 transcripts are spotted on the Custom Wheat Gene Expression 12x135k NimbleGen array [4]. Out of these, 14 233 have a BLAST result with an identity percentage greater than 45% and a coverage percentage greater than 50%, which are the cutoff values recommended by the IWGSC [6]. Thus, 35.3% of the transcripts of the NimbleGen array can be linked to A. *thaliana* and rice genes. Previous studies have shown that 18 140 genes are expressed during grain development [8], 8 160 (45%) of these genes have a BLAST matching the IWGSC recommended cutoff values.

This database is part of a larger project leaded by INRA Clermont-Ferrand and the LIMOS aiming at elucidating Gene Regulatory Networks (GRNs) involved in the transcriptional regulation of seed storage protein genes in wheat using data mining methods. Network inference and mining tools are being developed as a web-oriented platform [7]. The aim is to develop integrated biological resources for wheat containing functional annotation data and providing network inference, visualization and mining tools.

Acknowledgements

The authors thank Etienne Paux (INRA, Clermont-Ferrand) for useful discussions and advices during this work. The tables of the database and its structure can be downloaded from the website.

References

[1] RA. Gutiérrez, DE. Shasha, and GM. Coruzzi, Systems Biology for the Virtual Plant. *Plant Physiol.,* 138:550–554, 2005.

[2] SF. Altschul, W. Gish, W Miller, EW. Myers and DJ. Lipman, Basic local alignment search tool. *J. Mol. Biol.,* 215:403–410, 1990.

[3] I. Romeuf, D. Tessier, M. Dardevet, G. Branlard, G. Charmet and C. Ravel, wDBTF: an integrated database resource for studying wheat transcription factor families. *BMC Genomics,* 11:185, 2010.

[4] C. Rustenholz, F. Choulet, C. Laugier, J. Safar, H. Simkova, J. Dolezel, F. Magni, S. Scalabrin, F. Cattonaro, S. Vautrin, A. Bellec, H. Bergès, C. Feuillet and E. Paux, A 3000-loci transcription map of chromosome 3B unravels the structural and functional features of gene islands in hexaploid wheat. *Submitted for publication* (2011).

[5] HW. Mewes, A. Ruepp, F. Theis, T. Rattei, M. Walter, D. Frishman, K. Suhre, M. Spannagl, KF. Mayer, V. Stümpflen and A. Antonov, MIPS: curated databases and comprehensive secondary data resources in 2010. *Nucleic Acids Res.,* 39:D220-4, 2011.

[6] International Wheat Genome Sequencing Consortium. Guidelines for annotating wheat genomic sequences: release 1. Available at <http://www.wheatgenome.org/content/download/794/8948/file/wheat_gene_annotation_Release1-1.pdf>, 2006.

[7] J. Vincent, P. Martre, C. Ravel, A. Baillif and M. Agier, A web-oriented platform for gene regulatory network inference. Application to seed storage proteins in wheat. *This volume*.

[8] I. Romeuf, Identification in silico des facteurs de transcription du blé tendre (Triticum aestivum) et mise en évidence des facteurs de transcription impliqués dans la synthèse des protéines de réserve. *Ph.D. thesis*, Université Blaise Pascal, Clermont-Ferrand, France, 2010.

High Precision Approximations for the Significance Score of a Motif

Grégory NUEL[1]

MAP5, CNRS 8145, Dept. of Applied Mathematics, Paris Descartes University, F-75006, France
gregory.nuel@parisdescartes.fr

Keywords Markov chains, Markov chain embedding, Edgeworth's expansion, Large deviations.

We present here two new approximations for the computation of the significance score of a motif in a biological sequence. The first one, called Near Gaussian (NG) approximation, dramatically extends the reliability of the classical Gaussian approximations. The second one is a precise large deviation type result which supersedes previous ones by providing explicit formulas for the derivatives of the cumulant generative function Λ. Both approximations are compared to classical ones (Gaussian, Compound Poisson, see [1,2,3]) on a toy-example.

Significance score. Given $\mathcal{R}$ a motif of interest (from simple strings to complex regular expressions), a recurrent question is: "how surprising is it to observe n occurrences of $\mathcal{R}$ in my dataset ?". In statistical terms, this is equivalent to compute the p-value of observation n in respect with a relevant reference model. More precisely, if $X_{1:\ell}$ is length ℓ random sequence generated by our reference model, and if N_ℓ denotes the random number of occurrences of $\mathcal{R}$ in $X_{1:\ell}$, for any $n \geqslant 0$ or objective is to compute the significance score of observation n:

$$S(n) = \begin{cases} +\log_{10} \mathbb{P}(N_\ell \leqslant n) & \text{if } n \leqslant \mathbb{E}[N_\ell] \\ -\log_{10} \mathbb{P}(N_\ell \geqslant n) & \text{if } n > \mathbb{E}[N_\ell] \end{cases}$$

this score representing the p-value in a decimal log-scale, negative (resp. positive) values being associated to under- (resp. over-) representation events.

Exact computation. Using Deterministic Finite Automata and Markov chain embedding techniques [4,5] one can compute exactly $S(n)$ with a complexity $O(L^3 \times n^2 \times \log(\ell))$ (power) or $O(L \times n \times \ell)$ (recursion) where L is a parameter related to the complexity of the motif $\mathcal{R}$ (see [6,7] for details). Unfortunately, these complexities prevent to use exact methods for long sequences (large ℓ) when considering either high complexity motifs (large L) or motif with large number of occurrences (large n). For such cases where exact computations are not tractable, the need of efficient approximations remains critical.

Near-Gaussian Approximation. Thanks to recent advances in the field [8] it is now possible to compute exactly the first k-th moments of N_ℓ complexity $O(L^3 \times k^2 \times \log(\ell))$ (power), or $O(L \times k \times \ell)$ (recursion) which hence allows to obtain efficiently the moments of N_ℓ for any pattern problem. A tempting idea is then to take advantage of these moments to derive Near Gaussian (NG) approximations for N_ℓ using Edgeworth's expansions. We compare these approximations to the exact distribution on Fig. 1: NG approximations are well suited for center distribution events (ex: $|S(n)| < 5.0$) but are completely irrelevant for tail distribution events.

Bahadur-Rao Bound. Since we are usually more interested by tail distribution events, it might be interesting to turn to the large deviation theory where the behavior of N_ℓ is commanded by $\Lambda(t) \stackrel{\text{def}}{=} \log \mathbb{E}[e^{tN_\ell}]$ from which we can derive the following precise Bahadur-Rao (BR) type bound:

$$\text{BR}(n) \stackrel{\text{def}}{=} \frac{e^{-\Lambda^*(n)}}{\left(1 - e^{-|\tau|}\right) \sqrt{\Lambda''(\tau) 2\pi}}$$

with $\Lambda^*(n) \stackrel{\text{def}}{=} \max_t\{tn - \Lambda(t)\}$ and $\tau \stackrel{\text{def}}{=} \arg\max_t\{tn - \Lambda(t)\}$, and we have: $|S(n)| \geqslant \log_{10} \text{BR}(n)$ for all n. By providing explicit formulas for the two derivatives $\Lambda'(t)$ and $\Lambda''(t)$, we derive an efficient algorithm to compute $\text{BR}(n)$ with complexity $O(L^3 \times \log(\ell))$ (power), or $O(L \times \ell)$ (recursion). On Fig. 1, we observe that BR unsurprisingly performs well for tail distribution events but is irrelevant for the center of the distribution.

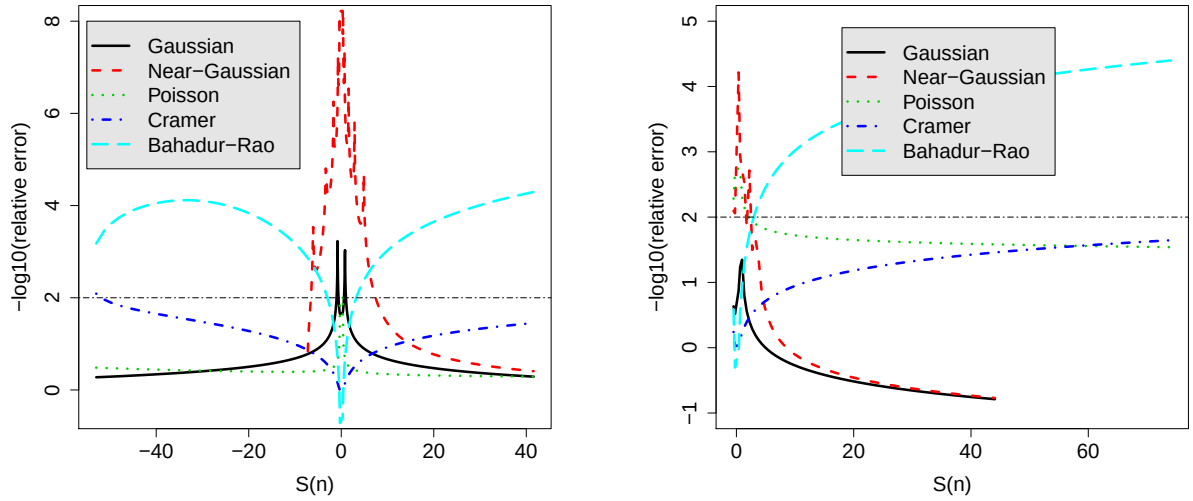


Figure 1. Relative error in log-scale for various approximations on a random sequence $X_{1:\ell}$ generated by a M0 model with parameters $\pi(\mathtt{A}) = \pi(\mathtt{T}) = 0.10$ and $\pi(\mathtt{C}) = \pi(\mathtt{G}) = 0.40$, and with $\ell = 1200$ for: (a) the frequent motif $\mathtt{G}(\mathtt{G}|\mathtt{C})\mathtt{G}$; (b) the rare motif $\mathtt{A}(\mathtt{A}|\mathtt{T})\mathtt{A}$.

Discussion. We can see on Fig. 1 that the classical approximations (Gaussian, compound Poisson) both lack of precision and induce a bias which is not even conservative: significance is too large for over-represented motifs with Gaussian approximations or for under-represented motif with compound Poisson approximations. On the other hand, by combining NG approximation (for small $S(n)$) and BR (for large $S(n)$), one can obtain a very reliable the approximation of $S(n)$ for any value of n, thus providing a reasonable alternative to exact computations with a dramatic reduction of the computational cost.

References

[1] M. Reignier, A unified approach to word occurrences probabilities. *Discrete Applied Mathematics*, 104(1):259-280, 2000.

[2] M. Lothaire, Statistics on Words with application to Biological Sequences. In Applied Combinatorics on Words, *Cambridge University Press*, 2005.

[3] G. Nuel, Numerical solutions for Patterns Statistics on Markov chains. *Stat. App. in Genet. and Mol. Biol.*, 5(1):26, 2006.

[4] P. Nicodème, B. Salvy and P. Flajolet, Motif statistics. *Theoretical Com. Sci.*, 287(2):593–617, 2002.

[5] M. Crochemore and V. Stefanov, Waiting time and complexity for matching patterns with automata. INFO. PROC. LETTERS, 87(3):119–125, 2003.

[6] M. E. Lladser, Mininal Markov chain embeddings of pattern problems, *Information Theory and Applications Workshop*, 251-255, 2007.

[7] G. Nuel, Pattern Markov chains: optimal Markov chain embedding through deterministic finite automata. *J. of Applied Prob.*, 45(1):226-243, 2008.

[8] G. Nuel, On the first k moments of the random count of a pattern in a multi-states sequence generated by a Markov source. *J. of Applied Prob.*, 47:1-19, 2010.

3D Axis Clustering for Mapping Cell Polarity in the Complex Geometry of the Heart

Jean-François LE GARREC [1,2], Chiara RAGNI[1,2] and Sigolène MEILHAC[1,2]
[1] INSTITUT PASTEUR, Unité de Génétique Moléculaire du Développement, Département de Biologie du Développement, F-75015 Paris, France
[2] CNRS, URA2578 , F-75015 Paris, France
jean-francois.le-garrec, sigolene.meilhac@pasteur.fr

Keywords Polarity, clustering, 3D mapping, heart morphogenesis.

1 Introduction

Evidence from several systems indicates that oriented tissue growth is a major factor contributing to the shape of organs during embryonic development. Pioneering work for mapping cell polarity dealt with the simple geometry of the fly wing [1]. Most studies in mammals focused on organs where one or two principal axes allow an easier 2D analysis of individual cell polarities. For example in the kidney tubule the orientation of cell division in 2D is fully defined by the angle relative to the axis of the tubule. Thus in 2D, the tissue may be represented as a scalar field, and the analysis of tissue polarity mainly consists in showing that the angles deviate significantly from a random distribution. However, in an organ such as the embryonic heart, the tube axis is looped and therefore polarity cannot be analyzed in 2D. We have shown previously that growth of the developing mouse myocardium is oriented [2]. We now aim to map cell polarities in relation to the geometry of the heart. We analyze the direction of the centrosome-nucleus axis as a classical indicator of cell polarity. We propose here a 3D fluorescent image processing framework to analyze tissue polarity.

Polarity analysis in 3D presents specific challenges: 1) The data resolution must be of sufficient quality in each of the three dimensions, to allow proper object segmentation and precise axis measurement. 2) The aim of the analysis is to map regions where cell polarities are coordinated. Since such regions may not coincide with those determined by morphological or molecular criteria, the search must be conducted objectively as a clustering problem in 3D. 3) The complex geometry of the heart prevents a straightforward application of standard spherical or circular statistics to test the significance of polarization in a given region.

2 Acquisition of Centrosome-Nucleus Axis Data

3-color 3D images of isolated E (embryonic day) 8.5 mouse hearts are acquired as previously described [3]. To optimize resolution we use confocal microscopy on samples with fluorescently labeled nuclei and membranes, and immunostained centrosomes. The size of the data is 1024 x1024 x 40 voxels, with a resolution of 0.379 x 0.379 x 1µm. Appropriate 3D segmentation tools are used to compute automatically the center of gravity of the centrosomes (wavelet-based) and nuclei [3]. Centrosomes and nuclei are paired by a sorting procedure based on distances. The data to be analyzed hence consist of a set of 6 coordinates per cell: 3 for the centrosome position and 3 for the associated centrosome-nucleus axis. Typically, data from about 200 cells is extracted from each confocal stack, and images from different embryos are combined by Procrustes alignment and scaling.

3 K-means Clustering of the Centrosome-Nucleus Axes

To map regions where cell polarities are coordinated, we investigate how neighboring cells can be grouped together into regions where their centrosome-nucleus axes are similar. We have adapted a K-means clustering algorithm initially designed for computer graphics [4]. A measure of the quality of polarization in a given region is first defined. Since we are dealing with axial (and not vector) data, the mean direction of a region is the third eigenvector of the scatter matrix computed from the axes of all the cells in the region: the higher the associated third eigenvalue, the more concentrated the axes around this mean direction and the higher the quality of polarization. After choosing K random seeds among the data points, the algorithm

progressively grows regions around these seeds by allocating their closest neighbors. For this purpose, spatial neighbors of a given seed are sorted in ascending order of the angle made by the neighbor axis and the seed axis. This process leads to a partition of all the data points in K regions. For the next iteration, the previous seeds are replaced in each region by the data point associated with the axis making the smallest angle with the mean direction of the region. A new partition is built from these seeds and leads to a further iteration. In order to escape local minima, the basic algorithm is improved by inserting and deleting regions at regular intervals during iterations. For typical data sets, the algorithm does not fully converge but leads to very similar partitions, depending on the initial seeds.

4 Optimization of Regions where Axes are Statistically Correlated

Standard spherical statistical tests, applied to the axial dataset, show that there is a planar bias in the distribution of the centrosome-nucleus axes, which are preferentially aligned parallel to the envelope of the heart. We therefore divide the axes into planar and transmural components, in reference to this envelope, and focus the polarity analysis on the planar component, which is more relevant to cardiac chamber expansion. No meaningful statistical test is available to assess whether such planar components along a complex surface are uniformly distributed. We therefore use a bootstrap method in order to rate regions according to the quality of their polarization, i.e. the concentration of the axes of their cells around a mean direction. Random partitions are generated and for each region-size a threshold (0.05) third eigenvalue is computed; regions are rated by the ratio of their third eigenvalue to this threshold. Since the clustering algorithm may be run with varying K values, we end up with a list of partly overlapping regions. The final map, which is independent of K, is determined by an optimization criterion according to the biological question asked. Only the most significant regions may be selected and thus the map is built by adding non-overlapping regions in decreasing order of significance. Alternatively, the map providing the best coverage of the whole organ may be selected, with the highest number of data points lying within significant regions. The statistical p-value of each selected region may finally be approximated in 2D by standard tests of uniformity in a plane tangential to the point of the heart envelope nearest to the center of the region.

5 Conclusion

Our 3D polarity analysis in the developing heart of the mouse, by optimized K-means clustering, permits the objective identification of regions where neighboring cells coordinate their polarity, as evidenced by the centrosome-nucleus axis. This method may be used more generally for analyzing any polarity axis within a complex 3D organ geometry.

Acknowledgements

This work was performed in the laboratory of M. Buckingham and we are grateful for her support. The work was funded by the Institut Pasteur (PTR 335). SM is an INSERM research scientist, CR has benefited from a fellowship of the French Ministry of Research, and JFLG is financed with a grant from the E.U. Project ‘CardioCell’ (LT2009-223372) to M. Buckingham.

References

[1] L. Baena-Lopez, A. Baonza and A. Garcia-Bellido, The orientation of cell divisions determines the shape of *Drosophila* organs. *Curr. Biol.*, 15:1640-1644, 2005.

[2] S. Meilhac, M. Esner, M. Kerszberg, J. Moss and M. Buckingham, Oriented clonal cell growth in the developing mouse myocardium underlies cardiac morphogenesis. *J. Cell Biol.* 164: 97-109, 2004.

[3] S. Pop, A. Dufour, J.F. Le Garrec, C. Ragni, M. Buckingham, S. Meilhac and J.C. Olivo-Marin, A fast and automated framework for extraction of nuclei from cluttered 3D images in fluorescence microscopy. *Proc. of IEEE Internat. Symposium on Biomedical Imaging: from Nano to Macro* - ISBI2011:2113-2116, Chicago, 2011.

[4] D. Cohen-Steiner, P. Alliez and M. Desbrun, Variational shape approximation. *ACM Transact. On Graphics* 23 : doi 10.1145/1015706.1015817, 2004.

biomanycores.org: Open-Source Parallel Bioinformatics

Jean-Frédéric BERTHELOT[1], Charles DELTEL[2], Mathieu GIRAUD[1,3], Stéphane JANOT[1,3], Laetita JOURDAN[1,3], Dominique LAVENIER[2,4], Hélène TOUZET[1,3] and Jean-Stéphane VARRÉ[1,3]

[1] INRIA Lille, 40 avenue Halley, 59650 Villeneuve d'Ascq, France
[2] INRIA Rennes, Campus de Beaulieu, 35042 Rennes Cedex, France
[3] LIFL, UMR 8022 CNRS, Cité Scientifique, 59655 Villeneuve d'Ascq Cedex, France
[4] IRISA, UMR 6074 CNRS, Campus de Beaulieu, 35042 Rennes Cedex, France
contact@biomanycores.org

Abstract *biomanycores is a repository of open-source parallel bioinformatics code in OpenCL (and, temporarily, in CUDA). We aim to bridge the gap between research in high-performance computing and the everyday work of bioinformaticians and biologists through the BioJava, BioPerl and Biopython frameworks.*

Keywords bioinformatics software, high-performance computing, GPU.

biomanycores.org : un Portail de Codes Libres pour la Bio-informatique Haute Performance

Résumé *biomanycores.org est un portail pour la diffusion de codes libres en bio-informatique pour processeurs massivement multi-cœurs. Biomanycores propose des interfaces de ces programmes aux frameworks BioJava, BioPerl et Biopython.*

Mots-clés pipelines bioinformatiques, calcul haute-performance, cartes graphiques, GPU.

1 Contexte

La bio-informatique est une discipline traditionnellement gourmande en ressources de calculs. La situation s'est encore amplifiée depuis l'arrivée des nouvelles technologies de séquençage. Dans certains cas, les grilles de calcul sont une solution. Aujourd'hui, les architectures massivement multi-cœurs, notamment les cartes graphiques (GPU), offrent une alternative interessante pour du calcul haute-performance à un coût bien plus faible. Ces processeurs proposent un parallélisme à grain fort (*work-groups* ou *blocks* de calculs indépendants), tout comme un parallélisme à grain fin, similaire à du SIMD (*work-items* ou *threads*).

Depuis 2007, avec l'apparition du langage CUDA (www.nvidia.com/cuda), de nombreuses applications ont été développées. On en compte aujourd'hui plus d'une quinzaine provenant d'une dizaine d'équipes, que cela soit en France ou dans le monde [1,2,3,4,5,6,7] (revue dans [8]). Défini en 2009, le standard OpenCL [1] améliore la portabilité des applications multi-cœurs, en permettant de développer à la fois pour des CPU multi-coeurs et des GPU de différents constructeurs.

Les programmes CUDA ou OpenCL offrent des accélérations de 5× à 50× par rapport à un processeur mono-cœur. Cependant, la valorisation de ces résultats de recherche reste faible. Les outils développés restent à l'état de prototype et ont peu de visibilité du fait de leur nouveauté et du changement de culture que cela suppose. De plus, ils ne proposent pas d'intégration aisée dans les frameworks d'analyse bio-informatique couramment utilisés.

2 Biomanycores

Biomanycores (www.biomanycores.org) est une collection d'applications bioinformatiques pour architectures massivement multi-cœurs, conçue pour faire le lien entre la recherche en calcul haute-performance

1. www.khronos.org/opencl

et le quotidien des biologistes et des bio-informaticiens. Les services suivants sont proposés : portail pour les codes sources CUDA et OpenCL, développement d'interfaces aux frameworks BioJava [9], BioPerl [10], et Biopython [11], et définition et mise à disposition de benchmarks construits autour de données biologiques.

3 État du Projet et Développements Futurs

Biomanycores intègre actuellement cinq applications : alignement local de séquences (Smith-Waterman) [4], recherche de motifs protéiques modélisés par des modèles de Markov cachés (HMMER) [12], recherche de sites de fixation de facteurs de transcription sur une séquence d'ADN avec des matrices position-poids [1], prédiction de structures secondaires d'ARN (RNAfold) [13], et détection de pseudo-nœuds (pKnotsRG) [14]. D'ici fin 2011, au moins cinq nouvelles applications seront ajoutées, ainsi que des benchmarks et des tutoriaux d'installation et d'utilisation.

Depuis fin 2010, Biomanycores bénéficie d'un ingénieur à temps plein grâce au soutien d'une ADT (action de développement technologique) INRIA. Biomanycores est un projet collaboratif, ouvert à la communauté : nous invitons les équipes souhaitant l'utiliser ou développant des applications CUDA/OpenCL à nous contacter (`contact@biomanycores.org`).

Références

[1] M. Giraud and J.-S. Varré, Parallel position weight matrices algorithms. *Parallel Computing*, 2010.

[2] G. Rizk and D. Lavenier, GPU accelerated rna folding algorithm, in G. Allen, J. Nabrzyski, E. Seidel, G. Dick van Albada, J. Dongarra, and P. M. A. Sloot (eds), *Proceedings of the 9th International Conference on Computational Science*, pages 1004–1013. Springer, 2009.

[3] M. C Schatz, C. Trapnell, A. L. Delcher, and A. Varshney, High-throughput sequence alignment using graphics processing units. *BMC Bioinformatics*, 8 :474, 2007.

[4] S. A Manavski and G. Valle, CUDA compatible GPU cards as efficient hardware accelerators for Smith-Waterman sequence alignment. *BMC Bioinformatics*, 9 Suppl 2 :S10, 2008.

[5] H. Shi, B. Schmidt, W. Liu, and W. Mueller-Wittig, A parallel algorithm for error correction in high-throughput short-read date on cuda-enabled graphics hardware. *Journal of Computational Biology*, 17(4) :603–615, 2010.

[6] Y. Liu, B. Schmidt, and D. L. Maskell, Parallel reconstruction of neighbor-joining trees for large multiple sequence alignments using cuda, in M. Taufer, S. Aluru, and D. A. Bader (eds), *Proceedings of the 8th IEEE International Workshop on High Performance Computational Biology*, pages 1–8. IEEE, 2009.

[7] P. D. Vouzis and N. V. Sahinidis, GPU-BLAST : using graphics processors to accelerate protein sequence alignment. *Bioinformatics*, 27(2) :182–188, 2011.

[8] J.-S. Varré, B. Schmidt, S. Janot, and M. Giraud, *Advances in Genomic Sequence Analysis and Pattern Discovery*, chapter Manycore high-performance computing in bioinformatics. World Scientific Publishing Company, 2011.

[9] R. C. G. Holland, T. A. Down, M. Pocock, and al, BioJava : an open-source framework for bioinformatics. *Bioinformatics*, 24(18) :2096–2097, 2008.

[10] J. E. Stajich, D. Block, K. Boulez, and al, The Bioperl toolkit : Perl modules for the life sciences. *Genome Research*, 12(10) :1611–1618, 2002.

[11] P. J. A. Cock, T. Antao, J. T. Chang, and al, Biopython : freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, page btp163, 2009.

[12] E. Roberts, J. Stone, L. Sepulveda, W.-M. Hwu, and Z. Luthey-Schulten, Long time-scale simulations of in vivo diffusion using GPU hardware, in M. Taufer, S. Aluru, and D. A. Bader (eds), *Proceedings of the 8th IEEE International Workshop on High Performance Computational Biology*, pages 1–8. IEEE, 2009.

[13] P. Steffen, R. Giegerich, and M. Giraud, GPU parallelization of algebraic dynamic programming, in R. Wyrzykowski, J. Dongarra, K. Karczewski, and J. Wasniewski, editors, *Proceedings of the 8th International Conference on Parallel Processing and Applied Mathematics*, pages 290–299. Springer, 2009.

[14] J. Reeder, P. Steffen, and R. Giegerich, pknotsRG : RNA pseudoknot folding including near-optimal structures and sliding windows. *Nucleic Acids Research*, 35(S2) :W320–324, 2007.

MGX – Montpellier GenomiX

A Next Generation Sequencing and Microarray Facility integrating Data Production and Analysis Tools

Christelle Dantec, Jean-Pierre Desvignes, Emeric Dubois, Grégory Baronian, Clémence Genthon, Hugues Parrinello, Dany Severac and Laurent Journot

MGX – Montpellier GenomiX; c/o Institut de Génomique Fonctionnelle; 141, rue de la cardonille; 34094 Montpellier Cedex 05, France

{Christelle.Dantec, Jean-Pierre.Desvignes, Emeric.Dubois, Gregory.Baronian, Clemence.Genthon, Hugues.Parrinello, Dany.Severac, Laurent.Journot}@igf.cnrs.fr

Abstract *BioCampus Montpellier hosts several technical platforms including MGX - Montpellier GenomiX, a microarray and next-gen sequencing facility (Illumina GAIIx et HiSeq2000). This platform is aimed at seamless integration of data production with 1st-3rd level data analysis tools. MGX team comprises 3 molecular biologists, 3 bioinformaticians and 1 manager. The facility is accessible to both academic and industry/biotech scientists.*

Keywords *Genomics, next generation sequencing, microarrays, bioinformatic analysis.*

MGX – Montpellier GenomiX

Plateforme de Service en Génomique

1 Introduction

La plateforme de service en génomique de Montpellier (MGX – Montpellier GenomiX) est une plateforme labellisée par le réseau IBiSA et la Cancéropôle Grand Sud-Ouest. Elle propose des services en microarrays, séquençage nouvelle génération et bioinformatique pour les communautés scientifique et industrielle, aussi bien dans le domaine végétal qu'animal. Depuis octobre 2010, en reconnaissance de sa conformité au référentiel ISO 9001 : 2008, le plateau technique IGF/IGH est certifié pour ses activités de développement et de réalisation de prestations en génomique (microarray, séquençage et bioinformatique).

La plateforme apporte son expertise pour :

- conseiller les utilisateurs dans la réalisation de leurs expériences et leur proposer les plans expérimentaux les mieux adaptés

- générer les données microarrays et séquençage nouvelle génération

- analyser les données générées sur la plateforme

- former les utilisateurs aux outils mis en place et les accompagner dans l'interprétation des données.

2 Procédure pour la Réalisation d'un Projet

2.1 Définir Ensemble la Prestation

Afin de comprendre la problématique de chaque équipe et de cibler au mieux le service à mettre en œuvre, la plateforme organise une réunion au démarrage de chaque projet. Suite à cette réunion, un compte-rendu et un devis sont rédigés. Nous avons mis en place un gestionnaire de projet accessible via une interface web pour que les utilisateurs puissent visualiser à tout moment la progression de leur projet.

2.2 Traiter et Analyser pour vous les Echantillons

La politique de la plateforme pour le traitement des projets est la méthode FIFO (First In, First Out) : un utilisateur entre dans la file d'attente à partir du moment où la plateforme reçoit la totalité des échantillons à traiter. MGX réalise ensuite le traitement biologique des échantillons et l'analyse bioinformatique des données générées (détaillée plus bas) dans l'ordre d'arrivée.

2.3 Accompagner et Former

Une fois l'analyse réalisée, un rapport est rédigé, décrivant étape par étape le traitement des échantillons et des données. En complément du rapport, il est proposé à l'utilisateur de venir nous rencontrer afin de discuter des résultats obtenus et du contenu du rapport. Une formation aux outils mis à disposition des utilisateurs est alors réalisée. Une aide à la publication des données est proposée notamment en écrivant les "matériels et méthodes" et en déposant les données dans les dépôts publics, *e.g.* Gene Expression Omnibus.

3 Analyses de Données

3.1 Microarray

L'équipe de bioinformatique MGX analyse les données générées sur la plateforme

- analyses primaires : en réalisant les contrôles qualité des données, la normalisation, les analyses statistiques

- analyses secondaires : analyses Gene Ontology, classification hiérarchique...

3.2 Séquençage Haut Débit

La plateforme a depuis 2008 un Genome Analyzer (GAIIx) et depuis juin 2010 un HiSeq2000. La plateforme réalise essentiellement des applications de comptages d'étiquettes telles que la ChIP-Seq, Digital Gene Expression, small RNA-Seq, RNA-Seq mais peut réaliser également des séquençages de novo. Les prestations comprennent : alignement – contrôle qualité : les bioinformaticiens réalisent l'épuration des tags si nécessaire et l'identification des tags sur des banques de données spécialisées (banque de transcrits, génome de l'espèce étudiée) ; analyse statistique ; visualisation ; analyses secondaires : analyses Gene Ontology, classification hiérarchique...

3.3 Développement – Collaboration

La plateforme travaille en collaboration avec différentes équipes de recherche :

En biologie pour développer de nouvelles applications: 1ère puce café, plateforme microarrays infrarouge, identification à l'échelle génomique des régions d'interaction avec la matrice nucléaire (MAR-Seq).

En bioinformatique avec les bioinformaticiens et biostatisticiens des équipes de recherches de l'Institut de Génomique Fonctionnelle et de l'Institut de Génétique Humaine, ainsi qu'avec les chercheurs statisticiens de l'université de Montpellier pour proposer de nouveaux outils adaptés aux besoins des utilisateurs.

La politique de la plateforme est d'utiliser au mieux des logiciels libres (soutient au projet 'Plume') dont notamment les outils du BioConductor (www.bioconductor.org).

4 Projets réalisés sur la Plateforme

MGX a réalisé plus de 100 projets microarrays, séquençage et bioinformatique dans de nombreuses espèces (mammifères, insectes, poissons, invertébrés, protozoaire, plantes, levures, bactéries...)

Combining Combinatorial Optimization and Statistics to Mine High-throughput Genotyping Data

Julie HAMON[1,3], Clarisse DHAENENS[1], Julien JACQUES[2] and Gaël EVEN[3]

[1] LIFL, Université Lille 1 / INRIA Lille-Nord Europe
clarisse.dhaenens@lifl.fr, julie.hamon@inria.fr
[2] Laboratoire Painlevé, UMR CNRS 8524 & Université Lille 1 / INRIA Lille-Nord Europe
Julien.Jacques@inria.fr
[3] GENES DIFFUSION, 3595 Route de Tournai, BP 70023, 59501 DOUAI Cedex
g.even@genesdiffusion.com

Keywords genomic selection, optimization, regression.

Coopération entre Optimisation Combinatoire et Statistiques pour l'Analyse de données de Génotypage haut-débit

Depuis quelques années, la génomique a grandement évolué avec le développement de nouvelles technologies telles que le séquençage et le génotypage haut-débit. En ce qui concerne le domaine animal, nous sommes aujourd'hui capables de lire les informations génomiques sur près de 800 000 marqueurs sur des ensembles d'individus de plus en plus larges (de 3 000 à 10 000). Ces données peuvent donner lieu à des études d'association entre les marqueurs (Genome-Wide Association Studies). Outre les contraintes biologiques (stockage des échantillons, manipulations longues et coûteuses…), la partie analyse de données (extraction de connaissances) doit aussi être adaptée en terme de méthodologie et d'architecture matérielle et logicielle. L'objectif est d'élaborer des modèles prédictifs permettant, à partir des données génomiques, de déterminer les individus les plus performants selon certains critères quantitatifs de sélection animale. Pour cela, l'objectif théorique est à terme de définir de nouvelles méthodes permettant la coopération entre statistique et optimisation combinatoire spécifiquement dédiées aux données issues de génotypage haut débit en vue d'une implémentation.

1 La Sélection Génomique dans le Domaine Animal

En génétique, on admet que plusieurs zones chromosomiques, portant un ou plusieurs génes, sont impliquées dans le contrôle de caractères quantitatifs (production de lait, fertilité...), et que de nombreux allèles (identifiés sous forme de marqueurs) sont responsables de la variabilité. On appelle ces zones QTL : Quantitative Trait Loci. La Sélection Génomique est une méthode d'évaluation génétique des animaux grâce à leur ADN (suite à un prélèvement biologique de type sang, poils ou biopsie), qui utilise un nombre très élevé de marqueurs couvrant l'intégralité du génome. Le principe de base a été établie par Meuwissen, Hayes et Goddard en 2001 [1]. Elle ne prend pas en compte les régions chromosomiques (QTL) mais exploite une densité de marqueurs suffisante si bien que chaque QTL se trouve en déséquilibre de liaison avec au moins un marqueur. Les effets des SNP (Single Nucleotide Polymorphism) sont estimés en associant les génotypes aux valeurs phénotypiques d'animaux déjà indexés. Grâce à leur détection, on peut calculer l'index propre d'un animal. Dans ce contexte, une problématique importante de la sélection génomique consiste à rechercher des marqueurs explicatifs (ou combinaisons de marqueurs) pour un phénotype sous étude. Il est à noter que l'augmentation actuelle du nombre de marqueurs (777 000 marqueurs en bovins) rend l'application de méthodologies séquentielles (analyse des marqueurs un par un par régression linéaire) non adaptée et extrêmement coûteuse en temps de calcul, et ne prend en compte aucune interaction éventuelle entre marqueurs. Nous proposons d'aborder ce problème en combinant deux approches de la littérature.

2 Approches Statistiques Existantes

Deux types de modèles statistiques sont généralement utilisés pour prédire un trait quantitatif à partir d'un grand nombre de marqueurs génétiques [2] :

- les méthodes de régression pénalisées :
  - la régression Ridge qui consiste à imposer une pénalité L^2 sur les coefficients de la régression.
  - la régression LASSO (Least Absolute Shrinkage and Selection Operator), qui en imposant une pénalité L^1, réduit des coefficients à 0, et donc sélectionne des variables (marqueurs génétiques) [3].
  - la régression Oscar dont la pénalité conduit à annuler certains coefficients de régression et à en regrouper d'autres en mettant leurs coefficients égaux [4].
- Les méthodes de régression sur combinaison des variables d'entrées :
  - PCA (Principal Components Analysis) - MCA (Multiple Correspondance Analysis)
  - SPCA (Sparse Components Analysis) : intégration d'une pénalité de type LASSO dans la détermination des composantes principales.
  - PLS (Partial Least Square)

On notera que les méthodes de régression sur combinaison des variables d'entrées ne permettent pas de sélectionner un nombre réduit de SNP et sont difficilement interprétables.

3 Optimisation Combinatoire

Les problématiques d'analyse liées aux données génomiques peuvent également être vues, dans la plupart des cas, comme des problèmes d'optimisation combinatoire. L'utilisation de méthodes d'optimisation combinatoire pour l'extraction de connaissances permet d'accélérer l'analyse en présélectionnant des sous-ensembles d'attributs intéressants, et de proposer de nouvelles méthodes permettant d'identifier des zones d'intérêts. Etant donnée la taille de l'espace de recherche (constitué de toutes les combinaisons de marqueurs possibles), seules des méthodes de types méta-heuristiques peuvent être mises en œuvre de façon efficace. Parmi ces méthodes, nous pouvons citer les algorithmes de recherche locale (descente, recherche tabou...) et les algorithmes à base de population (algorithme génétique) qui ont déjà fait leur preuve dans ce domaine [5].

4 Approche proposée : Optimisation Combinatoire et Statistique

L'objectif de ce travail est de définir un modèle de prédiction des traits des animaux à partir des marqueurs génétiques, utilisant à la fois la puissance exploratoire des algorithmes d'optimisation combinatoire et la spécificité des modèles statistiques de régression [1]. Nous choisissons comme première approche d'effectuer une sélection d'attributs en combinant une méthode d'optimisation de type recherche locale avec une régression RIDGE. A chaque étape de la recherche locale, nous évaluons la sélection d'attributs à l'aide d'un critère de type CVE (Cross Validation Error) calculé sur les prédictions par un modèle de régression, pour au final converger vers une sélection d'un nombre réduit de SNP, et vers un modèle de régression sur ces SNP.

Afin d'évaluer la qualité de la méthode nous utiliserons les données de XII QTLMAS 2008 et comparerons nos résultats et performances avec ceux des méthodes de sélection génomique présentées lors de ce Workshop.

Pour ce faire, nous utiliserons la plateforme ParadisEo développée par des membres de l'équipe DOLPHIN-INRIA, en C++. La solution proposée pourra également être parallélisable et déployable sur des architectures de calcul haute-performance (Cluster ou Grille de Calcul).

Références

[1] T.H.E. Meuwissen, B. J. Hayes and M. E. Goddard, Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps, *Genetics Society of America*, 2001.

[2] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning*, 2009.

[3] T.T. Wu, Y.F. Chen, T. Hastie, E. Sobel and K. Lange, Genome-wide association analysis by lasso penalized logistic regression, *bioinformatics*, vol. 25, no. 6, 2009.

[4] H.D. Bondell and B.J. Reich, Simultaneous regression shrinkage, variable selection and clustering of predictors with OSCAR, *Biometrics*, vol. 64, p. 115-123, 2008.

[5] C. Dhaenens, *Optimisation Combinatoire Multi-Objectif : Apport des Méthodes Coopératives et Contribution à l'Extraction de Connaissances*, HDR, Université Lille 1, 2005.

New Types of Services in Mobyle 1.0

Hervé MÉNAGER[1], Vivek GOPALAN[2], Bertrand NÉRON[1], Sandrine LARROUDÉ[1], Julien MAUPETIT[3], Adrien SALADIN[3], Pierre TUFFÉRY[3], Yentram HUYEN[2] and Bernard CAUDRON[1]

[1] Groupe Projets et Développements en Bioinformatique, Institut Pasteur,
28, rue du Dr Roux, 75724 PARIS Cedex, France
{hmenager, bneron, slarroud, caudron}@pasteur.fr
[2] Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD 20892, USA, Bioinformatics and Computational Biosciences Branch (BCBB))
{gopalanv, huyeny}@niaid.nih.gov
[3] MTi, INSERM UMR-S 973, Université Paris Diderot (Paris 7), Paris, France
{julien.maupetit, adrien.saladin, pierre.tuffery}@univ-paris-diderot.fr

Keywords web, portal, integration, web-services, workflows, visualisation.

Les Nouveaux Types de Services de Mobyle 1.0

Mots-clés web, portail, intégration, services web, workflows, visualisation.

1 Introduction

Performing bioinformatics analyses requires the selection and combination of tools and data to answer a given scientific question. Many bioinformatics applications are command-line only and researchers are often hesitant to use them based on installation issues and complex command requirements. Mobyle [1] is a framework and web portal specifically aimed at the integration of bioinformatics software and databanks. It allows to run bioanalyses through a web interface without installing anything locally. In addition to a web interface to command-line tools, the latest release of Mobyle, version 1.0, offers the possibility to execute predefined workflows, and enhances visualization possibilities with browser-embedded client components, the viewers. We focus here on these major improvements.

2 Chaining Automation with Workflows

To publish command-line applications as a set of homogeneous web-interfaces, Mobyle uses an XML-based data model. The description of service parameters and user data includes a description of their type, both at the semantic and syntactic levels. These elements describe the nature and format of the information conveyed by the data or processed by the parameters and determine the compatibility between them. In the interface, this allows to (1) suggest the relevant options to interactively chain successive programs using an intelligent piping suggestion system, and (2) facilitate the reuse of data over successive analyses by storing data bookmarks that can be directly loaded into a form.

Based on the need to automate these chainings, the data model has been extended to incorporate **Workflows**, which define a dataflow-based coordination of programs that run successive and/or parallel tasks to perform an analysis (see Fig. 1). Similarly to programs, workflows are viewed as services, sharing most of their description with programs, with the exception of the execution, which consists of a coordination of subtasks rather than the generation and execution of a command line.

3 Data Visualization with Viewers

When running an analysis in Mobyle, job result files can be directly pre-visualized in the portal. However, the understandability of the result is still often hindered by the necessity to browse potentially large and

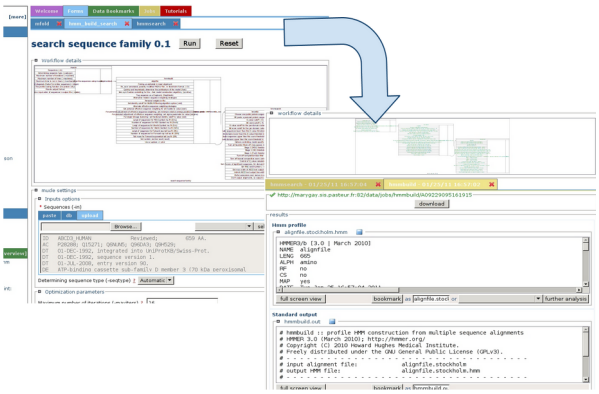


Figure 1. Use HMM to search a sequence family. The first step aligns a set of sequences with muscle, the second builds an HMM profile from the alignment produced by muscle, the last uses the profile to search for significantly similar sequences in a databank from the hmmprofile.

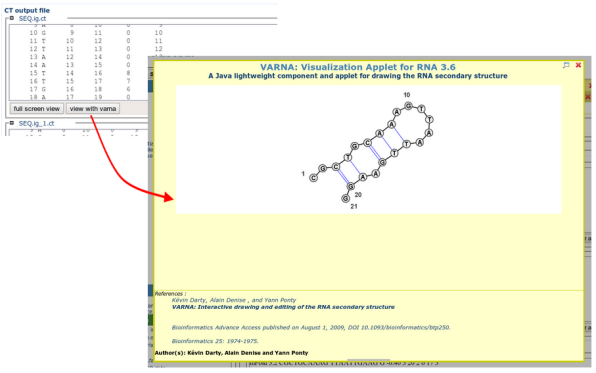


Figure 2. VARNA viewer example. Any compatible RNA secondary structure displayed in the portal can be accessed using this applet.

complex text-based files. To overcome this limitation, we created a specific type of service, **Viewers**. Viewers are a way to embed type-dependent visualization components for the data displayed in the Mobyle Portal. As opposed to programs and workflows, viewers are not executed on the server side, but rather rely entirely on browser-embedded code. The XML description files provide a way to incorporate custom interface code that will display data of a given type in the browser, incorporating HTML-embeddable components such as Java or Flash applets, Javascript code, etc. For instance, using viewers, we automate the inclusion of the VARNA [2] applet to visualize RNA secondary structures wherever it is relevant in the portal, such as in the results of RNA secondary structure prediction tools like MFOLD [3] (see Fig. 1).

4 Conclusion

The new version of Mobyle, v1.0, extends the spectrum of services available to include workflows and viewers. Current and future works include (1) the development of an interface that allows the "de novo" creation of workflows directly by users, and the automation of interactive chainings into workflows, and (2) the extension of the integration capabilities for client-side components beyond simple visualization, to the edition of user data. Mobyle is an open-source project available at https://projets.pasteur.fr/wiki/mobyle.

Acknowledgements

The MobyleNet project is funded by the IBISA (http://www.ibisa.net) initiative. The Mobyle/BCBI collaboration is funded by the NIH-Pasteur partnership http://nihpasteurpartnership.niaid.nih.gov.

References

[1] B. Néron, H. Ménager, C. Maufrais, N. Joly, J. Maupetit, S. Letort, S. Carrere, P. Tuffery, and C. Letondal, Mobyle: a new full web bioinformatics framework. *Bioinformatics*, 25(22):3005–3011, 2009.

[2] K. Darty, A. Denise, and Y. Ponty, VARNA: interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, 25(15):1974, 2009.

[3] M. Zuker, D.H. Mathews, and D.H. Turner, Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide. *RNA biochemistry and biotechnology*, 70:11–43, 1999.

An Organizational Environment for "In Silico" Experiments in Molecular Biology

Yuan LIN[1], Marie-Angélique LAPORTE[1,3], Lucile SOLER[4], Isabelle MOUGENOT[1,2] and Thérèse LIBOUREL[1,2]

[1] LIRMM, UMR5506 CNRS-UM2, 161, rue Ada, 34095 Montpellier, Cedex 5, France
firstname.lastname@lirmm.fr
[2] UMR ESPACE DEV IRD-UM2, 500 rue J.F. Breton, 34093 Montpellier, Cedex 5, France
firstname.lastname@univ-montp2.fr
[3] Centre d'Ecologie Fonctionnelle et Evolutive, UMR5175 CNRS, 1919, route de Mende, 34293 Montpellier, Cedex 5, France
firstname.lastname@cefe.cnrs.fr
[4] CIRAD-PERSYST, Campus International de Baillarguet, 34398 Montpellier cedex 5, France
firstname.lastname@cirad.fr

Abstract *Molecular biologists, just like geneticists, make use of various experimental mechanisms and devices to conduct research and to validate – or invalidate – their theories or initial hypotheses. Mechanisms powered by information technology, called "in silico", put data and analysis tools at the centre of the experiments, and are thus different from in vivo, ex vivo and in vitro mechanisms.*

Multiple resources (data sources as well as analysis tools) are widely available and, very often, allow various modes of operation, requiring certain expertise for their optimal use. This is especially true when drawing up complex analysis scenarios based on the sequential use of appropriate processing tools. To facilitate the construction of these experimentation mechanisms, we propose a scientific workflow infrastructure which uses an organizational environment to allow abstract planning of the experimentation, followed by its concretization. The concretization phase includes a verification of the conformity of the planned process chain's composition to avoid any error during execution.

Keywords Scientific workflow, analysis pipeline, specification language, validation aspects of service composition.

1 Introduction

Life sciences domains require the drawing up of experimentation process chains using various resources (data and processes). These resources, while available in ever-increasing quantities, remain, for the most part, expensive – and thus their reuse becomes almost a necessity.

To design these complex experiments, scientists often need to locate suitable resources and then to organize or reorganize them. In addition, each experiment deserves to be saved so that it can be re-executed several times, either in various different configurations or with diverse test data. In such a context, the use of a scientific workflow proves to be an invaluable help. Several dedicated software applications for this purpose now exist, most notably in the financial sector, and research in the field is relatively advanced. A first study [1] presented our approach based on the concept of the scientific workflow environment. Its objective is to help the user to:

– design experimentation process chains (in as abstract a manner as possible),
– better organize resources (data and processes) which will be elements in the concretization of these process chains,
– capitalize on the existing by constructing new processes from previously devised experimentation plans.

This article develops our research advances in terms of resource organization and semi-automatic verification of validity of workflows designed within a prototype.

This article is structured as follows: section 2 presents a brief state of the art, section 3 proposes an architecture for implementing a scientific workflow and section 4 provides a glimpse of the organization brought about. Section 5 covers the proposed verification of conformity, section 6 illustrates with an example the validation of conformity of a concrete process chain, and section 7 presents perspectives in progress.

2 State of the Art

A study was conducted based on characteristics we deemed relevant [2]:

- The existence of a meta level for describing and creating process chains. In fact, the generic aspect conferred by meta-modelling appears to be fundamental for all of us.
- Taking the experimental aspect into account. The unique characteristics of scientific data and processes should show through at the formalism level.

We present here only two representative projects, Kepler[3] and Taverna[4], which gain a certain amount of popularity among workflow scientists.

2.1 KEPLER

KEPLER [1] is a complete scientific workflow environment based on the Ptolemy II platform of the University of Berkeley. As far as process chains are concerned, KEPLER adopts a "human organization" metaphor. It is Actor-Based and considers all components of a process chain as actors. Actors (services) are accessed via a structure corresponding to the business ontology of the concerned domain.

The workflow is represented using a graphical language in the form of a graph linking *ports* (input/output parameters) of *actors* via *channels*. One or more actors in charge, *Directors*, plan tasks for other actors of the organization; they do so based on the available ontology. The execution plan of a process chain (or a portion of a process chain) is therefore created by a *Director* of the system. Any necessary adaptations are achieved by intermediary *sender* and *receiver* programs, which ensure the compatibility of data transferred over a channel. The process chain is saved in the form of MoML (Modelling Markup Language) files. (MoML is an XML-based language.) At the environment-interface level, a specific zoom feature is associated with the concept of an *opaque actor* (cf. Fig. 1). An *opaque actor* appearing in a process chain can be opened, thus revealing its constituent details.

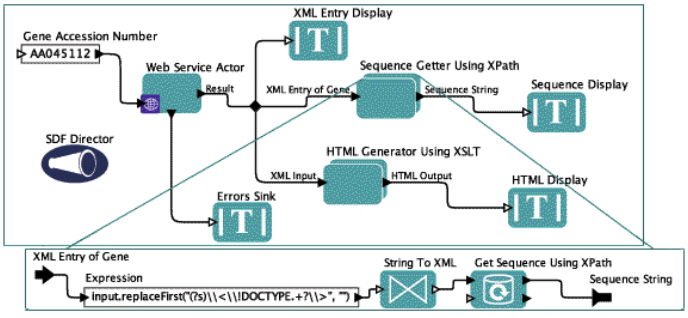


Figure 1. Overview of a process chain in the KEPLER environment.

2.2 Taverna

Taverna is a workflow project created by the ^{my}Grid team in England and used mainly in the life sciences. A workflow in Taverna is considered as a process graph in which processes are connected by data links or control links. Processes used are essentially web services (which can be supplemented by local libraries, manuscript scripts, etc.). During process composition, the user manually couples input/output parameters of web services or invokes *shim services*, specific adaptors existing from couplings constructed and tested for experiments. In

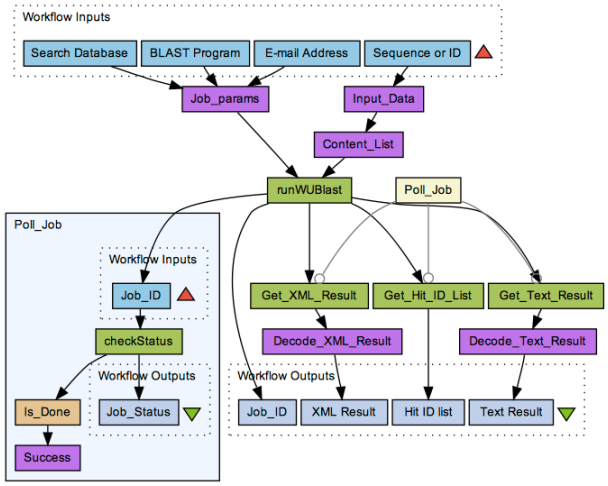


Figure 2. A concrete workflow in Taverna (taken from the *myExperiment* Taverna sharing site).

addition, the process chain is saved in the form of a SCUFL (Simple Conceptual Unified Flow Language) file. (SCUFL is an XML-based language.)

Taverna and Kepler projects both provide generic models for instantiation and composition of services. Our proposal introduces an additional abstraction level whose purpose is to describe the business domain before creating the process chains. This additional modelling level should facilitate the construction of process chains by allowing biologists to use their expertise of their domain but without requiring them to have expert and often exacting knowledge of the underlying resources and their locations. It also plays the role of a prescription model to which instantiation and service composition models have to conform.

3 Workflow architecture

Our efforts have been guided by the "business" point of view, that of the experimenters. Designing an experimental protocol corresponds to general model with three stages: 1) *Definition*: abstract definition of a process chain corresponding to an experimentation sequence (planning the experiments), 2) *Instantiation*: a more specific definition after identifying the various elements of the chain (data/processes), 3) *Execution*: customized execution (according to strategies corresponding to the requirements).

Based on this experimental life cycle, and inspired by the architectural styles proposed by OMG [5], we propose the following 3-level architectural vision (cf. Fig. 3):

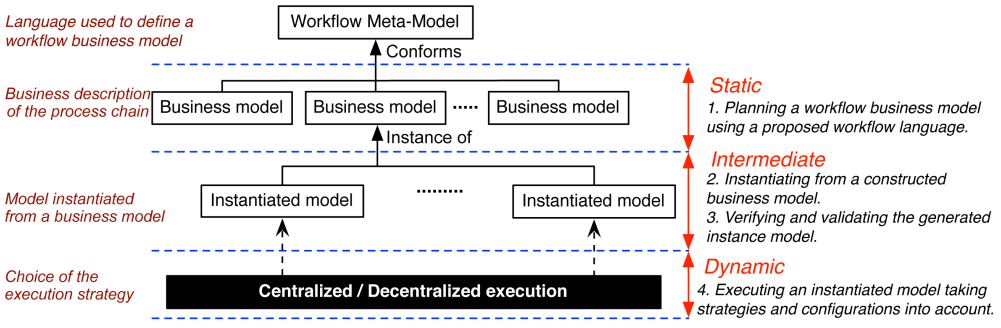


Figure 3. 3-level architecture of a workflow component.

The *static* level concerns the design phase. It is a matter of constructing (abstract) business-process models using a simple language. The *intermediate* level represents an instantiation and pre-verification phase. Using

1. http://kepler-project.org/

the business process model, the user constructs the real process chain by selecting and locating the processes and data most appropriate to the planned experimentation. The pre-verification is semi-automatized (cf. section 4). The *dynamic* level concerns the actual execution phase. It takes place based on the various strategies defined by both the user and the operational configurations.

The *static* level has been studied in some detail in our [2,1]. We have analyzed various language standards such as UML (activity diagram) [6] and SPEM [7], as also various existing projects such as BioSide [8], Meta-model WDO-It! [9] and CIMFlow [10]. Following this study, we proposed a simple but complete language. It is based on a language defined by a meta-model whose abstract elements, *tasks* or *processes*, are connected by unidirectional links and by the intermediary of *ports*. To facilitate the manipulation of abstract process chains, a corresponding graphical language was created within a prototype (cf. Fig. 4) [2].

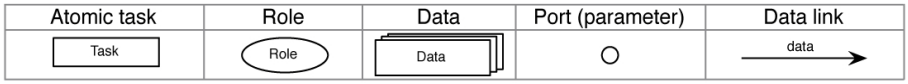


Figure 4. Some essential elements of our graphical language.

We currently focus on the *intermediate* level, which consists of two essential stages:
– instantiation of the abstract model with existing resources (data/processes);
– validation of the concrete model instantiated from the organizational environment.

4 Organizational Environment

To carry out the experimental protocols, the abstract model *instantiation* stage consists of finding and reusing existing resources. To facilitate this search, we base ourselves on the concept of *organizational environment*. This environment relies on the description of resources (data and processes) in the form of metadata (expressed in XML schema format). The resource descriptions are hierarchized in resource categories and in concrete resources. As shown in figure Fig. 5, it consists of:

– an organization relating to processes. It manages the hierarchy of descriptions of process categories and of concrete processes. The concept of *Converter* corresponds to the concept of a specific process responsible for adapting data between different formats of the same data category.
– an organization relating to data. It manages a hierarchy of descriptions of data categories, of concrete data and of the various associated data formats [3].

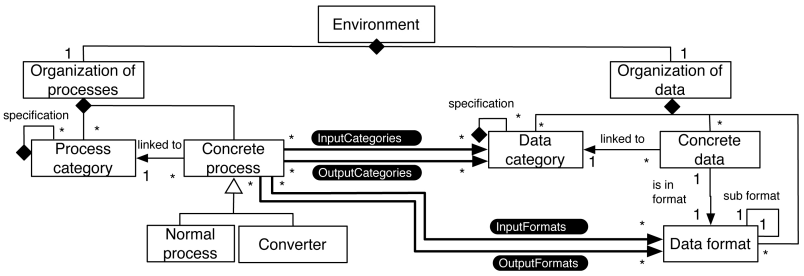


Figure 5. Organizational environment.

To illustrate this concept of the environment, we take an example from the world of molecular biology (cf. Fig. 6). The upper part of each hierarchy (processes and data) represent a set of categories (shown as ovals) sorted according to the generalization/specialization relationship. The descriptions of concrete resources (data or processes) are then associated to their category. The description of a concrete data describes its format, whereas that of a concrete process corresponds to its *signature*, which we formalize thus:

2. It will be used in the examples to follow.
3. Remark: It should be noted that several data categories can share the same format.

Name (Input parameter list, Output parameter list), where each parameter is described by the doublet (Data category : data format).

A set of data formats (Fasta, xml, MultiFasta, Clustal, Newick, Jpeg) is also presented. Figure (Fig. 6) is therefore complemented by the description of signatures of some example concrete processes:

Blastp(ProteinSeq:Fasta) : (SeqPairs:xml)
ClustalW(ProteinDataBank:MultiFasta) : (MultipleAlignment:Clustal)
InteractiveSelection(SeqPairs:xml) : (ProteinDataBank:MultiFasta)
Logo(MultipleAlignment:Clustal) : (Image:jpeg)
PhyML(MultipleAlignment:Clustal) : (PhylogeneticTree:Newick)

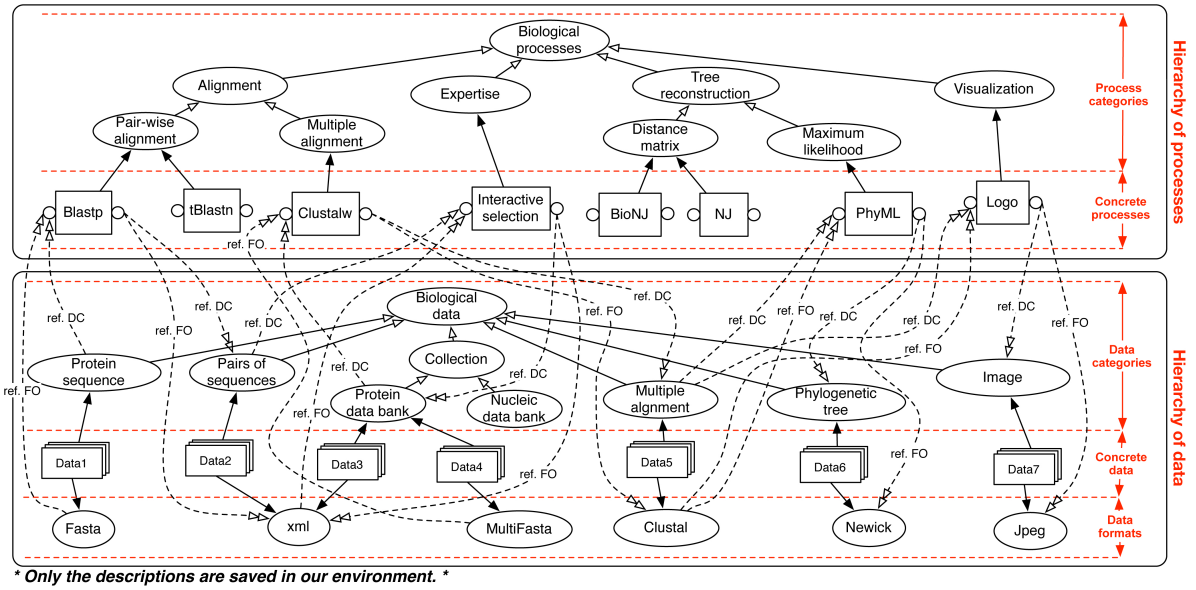


Figure 6. Illustration of an organizational environment in a biological context.

5 Conformities

5.1 The Problem

As already mentioned, the second important stage of the *intermediate* level consists of validating the concrete model instantiated from the abstract model.

Let us take an example described by using the workflow language, corresponding to an abstract process chain model that a biologist designs with the intention of characterizing a protein sequence which interests him in the context of his putative functional domains.

At the concrete level, the idea is to begin by using the *Blast* similarity-search tool to compare the protein sequence under consideration with a data bank of protein sequences and to thus identify segments with high similarity shared both by the protein sequence under consideration and by various sequences in the sequence data bank. These similar segments indicate the possible presence of functional domains. The biologist then continues his study by reusing the results output from the *Blast* tool [11], either to construct a phylogenetic tree and retrace the evolutionary history of the sequence via the *PhyML* tool [12] or to display the preserved positions common to all the similar segments via the *Logo* tool [13]. This simplified example of a process chain in molecular biology allows us to highlight the difficulties encountered by the biologist in using the results output by one tool as input to another tool. The difficulties relate, at the same time, to the nature of the data (here characterized as data category), to the format of this data, and, finally, to the biologist's expertise. In the example, we make willing use of the discrepancy which arises between the Blast tool, which outputs a collection of simple alignments, and the PhyML and Logo tools, which require multiple alignments to run. In fact, Blast leads to multiple discrepancies two-by-two, involving the sequence under consideration and one of

the sequences from the sequence data bank which is similar to it; whereas PhyML and Logo use the shared similarity by a set of sequences which includes the sequence under consideration. This example highlights what we will subsequently term *semantic incompatibility*.

In its upper part, Fig. 7 shows the abstract process chain and in the lower the concrete chain obtained after locating data descriptions *S1* and adapted processes *Blastp* and *PhyML*. The problem which we designate as one of *validation of the instantiated (concrete) model* consists of verifying the *compatibility* of each *composition*. A *composition* corresponds to the link between an output parameter p1 of a process T and an input parameter p2 of the process following T; we denote it *(p1 → p2)*.

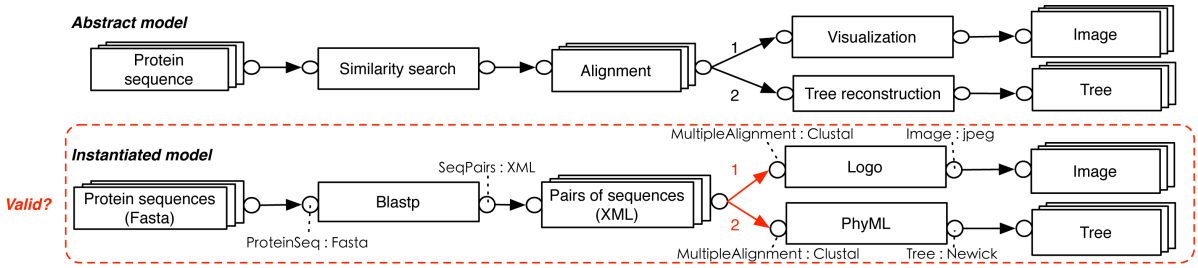


Figure 7. Problem at hand.

5.2 Identifying Situations of Compatibility

Verification is undertaken by analyzing the signatures of linked processes. To do so, we have to take two important aspects into account:

- the *syntactic* aspect, relating to the data formats used by the parameters.
- the *semantic* aspect, relating to the process's functionality. It not only depends on the process's name but also on the signification of the input/output parameters.

For two processes *T1(dc1:fo1) : (dc2:fo2, dc3:fo3)* and *T2(dc4:fo4) : (dc5:fo5)*, let us suppose that there exists a composition, denoted *p1→p2*, between the *p1 (dc3:fo3)* output parameter of process *T1* and the *p2 (dc4:fo4)* input parameter of process *T2*.

Syntactic and semantic compatibilities are defined as follows:

- *Syntactic compatibility: p1 → p2* is syntactically compatible if *(fo3 = fo4) ∨ (fo3 is a sub-format of fo4)*, denoted $p1 \overset{Syn}{\rightarrow} p2$. Two parameters are syntactically compatible if they use the same data format or if they use an output format which is a sub-format of the input format. Else $p1 \overset{Syn}{\not\rightarrow} p2$.
- *Semantic compatibility: p1 → p2* is semantically compatible if *(dc3 = dc4) ∨ (dc3 is a sub-category of dc4)*, denoted $p1 \overset{Sem}{\rightarrow} p2$. Two parameters are semantically compatible if they use the same category, or if they use an output category which is a sub-category of the input category. Else $p1 \overset{Sem}{\not\rightarrow} p2$.

The verification of a composition's compatibility is thus done at two levels: syntactic and semantic. Three types of situations can arise:

- Situation 1 $(p1 \overset{Sem}{\rightarrow} p2) \wedge (p1 \overset{Syn}{\rightarrow} p2)$: *p1* and *p2* are compatible at the semantic and syntactic levels. This is the ideal situation in our context; we designate it as valid.
- Situation 2 $(p1 \overset{Sem}{\rightarrow} p2) \wedge (p1 \overset{Syn}{\not\rightarrow} p2)$: *p1* and *p2* are compatible at the semantic level but not at the syntactic level. The composition is syntactically adaptable. An adaptation between the two data formats will be necessary (cf. converters).
- Situation 3 $p1 \overset{Sem}{\not\rightarrow} p2$: The two parameters are not semantically compatible. In such a case, it is pointless to proceed to verify their syntactic compatibility (in fact, for us, two parameters with different significations cannot be paired). The composition is semantically adaptable.

From these definitions, we develop our proposed approach for resolving the incompatibilities.

6 Validation of the Experimental Chain

Of the three compatibility situations identified, the latter two require an adaptation stage before going on to the execution phase. It is a matter of finding one or more intermediate processes which can overcome the composition's incompatibility. For situations 2 and 3, two types of adaptations are proposed:

- *semantic* adaptation (for situation 3). The incompatibility of situation 3 represents the case where the two parameters of a composition use incompatible data categories. The adaptation here consists of finding a possible intermediate process chain between these two categories.
- *syntactic* adaptation (for situation 2). In situation 2, where the composition is already semantically compatible, the problem can be expressed as a divergence between the data formats used by the two connected parameters. All that is required is to find *converters* to convert one data format into the other.

These adaptations are based on the organizational environment. The search for intermediate processes can be equated to a search for itineraries between two incompatible data categories or formats. We will illustrate this using the example and the organizational environment constructed earlier (cf. Fig. 6).

Let us consider again the previous example. The verification conducted on the instantiation of the abstract model detects a semantic incompatibility in the composition between Blastp and Logo or between Blastp and PhyML due to difference in categories *Pairs of sequences* and *Multiple Alignment* (*Incompatibility situation 3*). The (semantic) adaptation will be applied; it consists of finding in what we call the (semantic) resource graph the path allowing the conversion of categories.

The construction of the (semantic) resource graph consists of extracting, from the organizational environment, the descriptions of processes and of data categories referenced by their parameters. Such a (semantic) resource graph generated from the environment described in Fig. 6 is shown in Fig. 8.

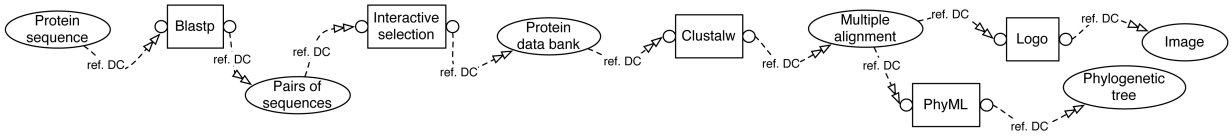


Figure 8. (Semantic) resource graph generated from the organizational environment of Fig. 6.

A graph traversal algorithm is used to find all the possible "paths" between the two concerned data categories (*Pairs of sequences* and *Multiple Alignment*). A single path is found in the graph: *Pairs of sequences* → *InteractiveSelection* → *ProteinDataBank* → *ClustalW* → *Multiple Alignment*. The two processes, *InteractiveSelection* and *ClustalW*, will therefore be added to the incompatible chain (cf. figure Fig. 9).

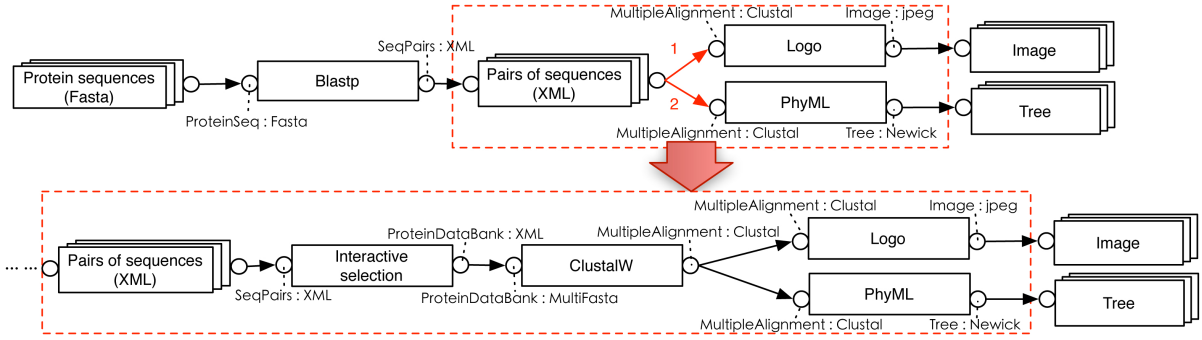


Figure 9. Semantic adaptation.

Once this adaptation is done, there still remains the existing syntactic incompatibility of the composition between the *InteractiveSelection* and *ClustalW* processes because even though *InteractiveSelection* outputs the same data category that is accepted for input by *ClustalW*, their data formats are different (*xml* and *MultiFasta*). Syntactic adaptation consists of finding specific *converters*, or compositions of *converters*, necessary for these conversions. We will not cover this stage in detail; it is simply enough to understand that converters (or their composition) can be added to obtain the required validity.

7 Conclusion and Perspectives

A prototype (http://www.lirmm.fr/~lin/project/) illustrating the key aspects of our approach for designing and validating scientific process chains is currently being developed. This prototype serves as a basis for an inductive experimental approach using data of BAC and EST nucleic sequences as well as physical and genetic maps for identifying and characterizing genetic markers relating to sex of the Nile tilapia (Oreochromis niloticus). Over a longer term, we intend to integrate the current prototype into a platform with a search engine based on resource descriptions to be able to undertake the execution using real resources, after requisite validation of experimentation chain. It will eventually also use open-source controlled vocabularies such as PFO (Protein Feature Ontology)[14], SO (Sequence Ontology)[15], and GO (Gene Ontology)[16] to enrich data categories by additional representations and thus extend the descriptive capacities of the organizational environment.

References

[1] T. Libourel, Y. Lin, I. Mougenot, C. Pierkot, JC. Desconnets, A Platform Dedicated to Share and Mutualize Environmental Applications, *Proceedings of 12th International Conference on Enterprise Information Systems*, Madere, 2010.

[2] Y. Lin, T. Libourel, I. Mougenot, A Workflow Language for the Experimental Sciences, *Proceedings of 11th International Conference on Enterprise Information Systems*, Milan, 2009.

[3] I. Altintas, B. Ludäscher, S. Klasky, M. A. Vouk, S04 - introduction to scientific workflow management and the kepler system, *In SC*, pp. 205, 2006.

[4] D. Hull, K. Wolstencroft, R. Stevens, C. A. Goble, M. R. Pocock, P. Li, T. Oinn, Taverna: a tool for building and running workflows of services. *Nucleic Acids Research*, 34(Web-Server-Issue):729–732, 2006.

[5] Object Management Group (OMG), Meta Object Facility (MOF) Core Specification OMG Available Specification Version 2.0, *OMG Document Number: formal/06-01-01*.

[6] Object Management Group (OMG), OMG Unified Modeling LanguageTM (OMG UML), Infrastructure Version 2.3, *OMG Document Number: formal/2010-05-03*.

[7] Object Management Group (OMG), SPEM - Software & Systems Process Engineering Meta-Model Specification, Version 2.0, *OMG Document Number: formal/2008-04-01*.

[8] M. Hallard & al, Bioside : faciliter l'accès des biologistes aux ressources bio-informatiques, *JOBIM*, Montréal , pp. 64, 2004.

[9] P. Pinheiro da Silva, L. Salayandia, A.Q. Gates, WDO-It! A Tool for Building Scientific Workflows from Ontologies *Departmental Technical Reports (CS). Paper 201*, 2007.

[10] L. Haibin, F. Yushun, CIMFlow: A Workflow Management System Based on Integration Platform Environment, *Proceedings of 7th IEEE International Conference on Emerging Technologies and Factory Automation*, Barcelona : ETFA, pp.187-193, 1999.

[11] S. Altschul, W. Gish, W. Miller, E. Myers, D. Lipman, Basic local alignment search tool, *Journal of Molecular Biology*, 215:403-410, 1990.

[12] S. Guindon and O. Gascuel, A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood, *Systematic Biology*, 52:696-704, 2003.

[13] T. D. Schneider and R. M. Stephens, Sequence Logos: A New Way to Display Consensus Sequences. *Nucleic Acids Res.*, 18:6097-6100, 1990.

[14] G.A. Reeves, K.Eilbeck, M.Magrane, C.O'Donovan, L.Montecchi-Palazzi, M.A. Harris, S.E. Orchard, R.C. Jimenez, A.Prlic, T. J. P. Hubbard, H.Hermjakob, J.M. Thornton, The Protein Feature Ontology: a tool for the unification of protein feature annotations, *Bioinformatics*, 24:2767-2772, 2008.

[15] K.Eilbeck, S.E Lewis, C.J Mungall, M.Yandell, L.Stein, R.Durbin, M.Ashburner, The Sequence Ontology: a tool for the unification of genome annotations, *Genome Biology*, 6:R44, 2005.

[16] M.Ashburner, C.A. Ball, J.A. Blake, D.Botstein, H.Butler, J. Michael Cherry, A.P. Davis, K.Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L.Issel-Tarver, A.Kasarskis, S.Lewis, J.C. Matese, J. E. Richardson, M.Ringwald, G.M. Rubin, G.Sherlock, *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium*. Nature Genetics, 25:25-29, 2000.

Communications affichées tardives

Les contributions contenues dans cette section correspondent aux communications affichées reçues après le second appel.

Phylogenomic Evidence Supports the Phagocytosis Model of the Evolutionary Origin of Eukaryotes

Nicolas ROCHETTE[1] and Manolo GOUY[1]

Laboratoire de Biométrie et Biologie Évolutive (UMR CNRS 5558)
Bât. Mendel, 43 bd du 11 novembre 1918, 69622 Villeurbanne cedex, France
{nicolas.rochette, manolo.gouy}@univ-lyon1.fr

Keywords Origin of eukaryotes, phylogeny, genomics.

Although it is widely regarded as a critical step in the evolution of life on earth, early eukaryotic evolution remains essentially unknown. The last eukaryotic common ancestor (LECA) [1] was fully compartmentalized, bore a mitochondrion derived from alphaproteobacteria, and its genome was made up of both archaeal-like and bacterial-like genes. Two views coexist concerning its origin. According to the symbiosis-fusion hypotheses, including the hydrogen hypothesis [2], a symbiosis between two prokaryotes evolved to the point that one was engulfed by the other and most of its genes transfered to it, what allowed, in combination with metabolic superiority, for a rapid complexification. According to the Archaezoa hypothesis [3], an archaea-related organism gradually evolved compartmentalization and phagocytosis, possibly collecting genes from its environment, and acquired mitochondrion lately.

We show that the bacterial signal in the eukaryotic genome is spread over many phyla, not just alphaproteobacteria. 289 homologous protein families that were traceable to LECA and widely represented in either archaea or bacteria were identified. The analysis was performed with two independent methods, both taking into account the statistical uncertainty associated with single-gene tree inferences. The two methods unequivocally recovered the archaeal and bacterial subsets, which were of even importance. Surprisingly, few families supported the three-domains view [4], and our data rather support archaeal paraphyly, like in the eocyte hypothesis [5,6]. At the phylum level, the methods agreed on bacterial families but not on archaeal families, indicating that phylogenetic resolution might be lower in that part of the tree. Among bacterial families, 32 linked to alphaproteobacteria, 67 were unequaly distributed over various phyla, and 30 could not be attributed to any individual phylum.

Our results indicate that the complexification of the eukaryotic genome, and thus that of the cell, is best explained by a phagocytosis-driven "you are what you eat" [7] process.

References

[1] T. M. Embley and W. Martin, Eukaryotic evolution, changes and challenges. *Nature*, 440:623-30, 2006. Review.

[2] W. Martin and M. Müller, The hydrogen hypothesis for the first eukaryote. *Nature.* 392:37-41, 1998.

[3] T. Cavalier-Smith, The origin of eukaryotic and archaebacterial cells. *Ann N Y Acad Sci.*, 503:17-54. 1987. Review.

[4] C. R. Woese, O. Kandler and M. L. Wheelis, Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *PNAS*, 87:4576-9, 1990.

[5] J. A.Lake, E. Hendersen, M. Oakes and M. W. Clark, Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes. *PNAS*, 81:3786-3790, 1984.

[6] P. G. Foster, C. J. Cox and T. M. Embley, The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 364:2197-207, 2009.

[7] W. F. Doolittle, You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends Genet.*, 14:307-11, 1998. Review.

The COaLA Model: a Time Non-Homogeneous Model of Evolution Based on a Correspondance Analysis

Mathieu GROUSSIN[1], Bastien BOUSSAU[2] and Manolo GOUY[1]

[1] LBBE, UMR5558 CNRS, UCB Lyon 1 – Bât Grégor Mendel, 43 bd du 11 Novembre 1918, 69622, Villeurbanne, Cedex, France

mathieu.groussin@etu.univ-lyon1.fr

bastien.boussau@univ-lyon1.fr

manolo.gouy@univ-lyon1.fr

Keywords Phylogeny, Non-homogeneous models, correspondance analysis.

Two probabilistic methods are now commonly employed to either reconstruct phylogenies or estimate evolutionary parameters on a fixed topology: the Maximum Likelihood (ML) and the Bayesian approaches. This study focuses only on the ML context. The probabilistic evolutionary models used in ML are Markovian substitution processes. The original models were much simplified, essentially for practical reasons. In this way, strong assumptions were made, as for example (i) the constancy of evolutionary rates over time for all lineages and among sites and (ii) the constancy of the evolutionary process over time and among sites. The last hypothesis leads to homogeneous and stationary models of evolution. Thus, not only the process of evolution remains constant over time and among sites, but all species share the same equilibrium frequencies in terms of base or amino-acid. In reality, variations of compositions are frequently observed between species. The too simplified evolutionary models can thus estimate wrong phylogenies by grouping unrelated species just because they share similar compositions. It has been shown as well that homogeneous models may also be poor to estimate ancestral compositions and sequences along a phylogenetic tree [1].

In this work, we are more concerned on the development of a new time non-homogeneous evolutionary model for protein sequences. To relax the homogeneity hypothesis, one can allow the different branches of a phylogenetic tree to have their own equilibrium frequencies to better fit possible shifts of composition over time in particular lineages. However, such an approach is too parameter-rich since the model would have n x 19 parameters to optimize, n being the number of branches. This major issue prevented to develop non-homogeneous evolutionary models in the ML framework so far. We propose here a way to reduce considerably the number of parameters in the model, although each branch still conserves its own equilibrium frequencies.

Thus, from the original alignment, a matrix of amino-acid frequencies for each species is built. From this matrix of frequencies, a Correspondance Analysis is performed to decompose the total variance in orthogonal factors, which specifies a 19-dimensional space. The first factor is the axis that represents most of the variance present in terms of amino-acid compositions. The second factor is the axis that represents most of the variance given the first orthogonal factor and so on. In this way, to one point in the new space characterized by the factors corresponds another point in the 20-dimensional space of amino-acid frequencies. The COaLA model, for COrrespondance and Likelihood Analysis, proposes to optimize positions along the factors instead of directly optimize the 19 frequencies. For a particular branch, by optimizing only a few positions along the first axes, the COaLA model allows to indirectly optimize a vector of equilibrium frequencies. The number of parameters is thus drastically reduced from 19 x n to P x n, P being the number of positions to optimize along the P first axes ($P \in [1:19]$). Consequently, it becomes feasible to determine the maximum likelihood values of the evolutionary parameters along a phylogenetic tree with a non-homogeneous protein model.

Simulation experiments were realised and showed that the model performs really well in estimating the ancestral amino-acid frequencies. The COaLA model was also used on different biological data sets and interesting results will be presented.

[1] B. Boussau, S. Blanquart, A. Necsulea, N. Lartillot and M. Gouy, Parallel adaptation to high temperature in the archaean eon. *Nature.*, 456:942-945, 2008.

A Web-Based Collaborative Platform for Comparing Phylogenies

Nicolas FIORINI[1], Antoine BISCH[1], Florent DUMOND[1], Mawusse AGBESSI[1], François CHEVENET[2], Vincent LEFORT[2], Anne-Muriel CHIFOLLEAU[2] and Vincent BERRY[2]

[1] Université Montpellier 2, Place E. Bataillon, 34095, Montpellier, France, Cedex 5

[2] LIRMM, UMR5506 CNRS, 161 rue Ada, 34095, Montpellier, France

{Francois.Chevenet, Vincent.Lefort, Anne-Muriel.Chifolleau, vberry}@lirmm.fr

Keywords Real-time collaboration, Online resource, Topological tree handling.

Comparing phylogenies is a routine process in bioinformatics, performed to analyze alternative trees for a sequence alignment, to detect an outlier tree in a collection of trees or to view differences between gene and species trees. This in turn allows to identify recombinant evolution such as horizontal gene transfers or to evaluate the ability of a supertree/supermatrix tree to summarize a collection of individual gene trees.

To locate (di)similar topologic features of analyzed trees, it is often useful to color particular leaves. When comparing trees with differing taxa sets, it helps to temporarily restrict these trees to their common taxa. Highlighting their common topologic structure is also important when they disagree on the position of few taxa [1]. Moreover, during the course of the analysis, the content of the tree collection or the names of the trees usually need to be modified to reflect the analysis progress.

The online *CompPhy* platform allows distant coworkers to upload and manage a collection of phylogenetic trees into a project environment. *CompPhy* offers a visual comparison of two trees chosen in a collection of source trees and supertrees. It implements tools able to perform taxa colorization, tree restriction to common taxa and supertree computation for selected source trees. As *CompPhy* relies on the *Scriptree* system [2] to display trees, complex highlighting and annotation operations can be performed manually. The platform is still under development and collaborative operations will be added such as discussion threads and the possibility to invite guests, including anonymous reviewers for trees appearing in a paper submitted to a journal. Advanced highlighting operations will also be proposed such as indicating the largest common substructure of selected trees (MAST) or automatic coloring of subtrees based on taxonomic information. The *CompPhy* website can be tested by exploring the sample tree collection made available at http://compphy.creatox.com/ .

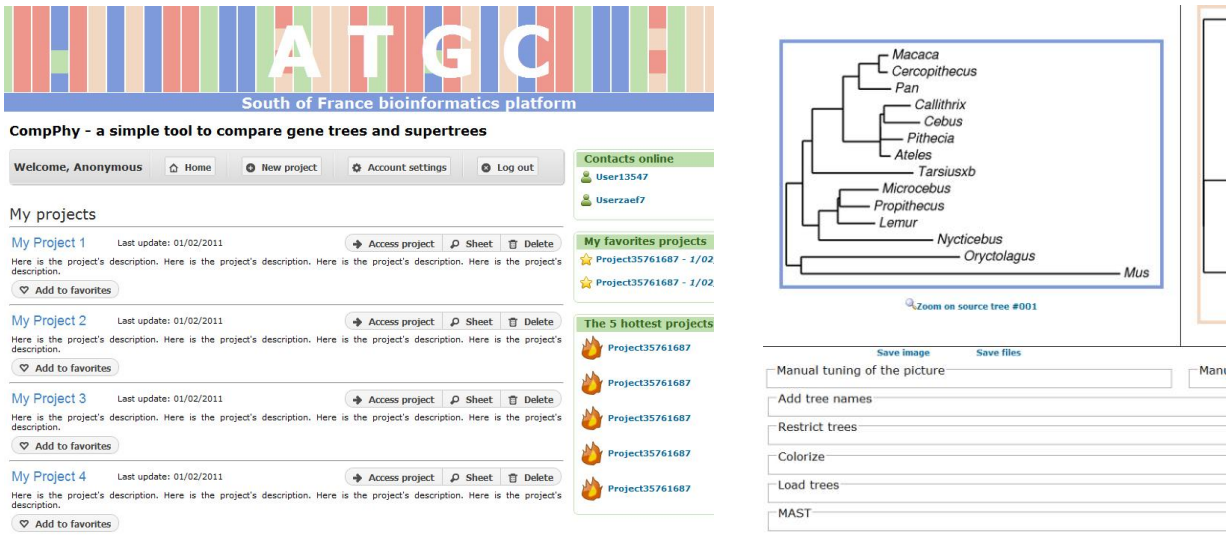


Figure 1. Extract of two *CompPhy* pages: project managing (left); tree comparison and tools (right).

[1] T. Nye, P. Lio and W. Gilks, A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. *Bioinformatics*, 22:117–119, 2006 (website http://www.mas.ncl.ac.uk/cgi-bin/ntmwn/pairwise.cgi).

[2] F. Chevenet, O. Croce, M. Hebrard, R. Christen and V. Berry, ScripTree: scripting phylogenetic graphics. *Bioinformatics*, 26, 8:1125-6, 2011.

PELICAN: Orthologous Groups and Gene Lateral Transfers for Comparative Genomic Analysis of Marine Cyanobacteria

Christophe HABIB[1,2], Gildas LE CORGUILLE[2], Erwan CORRE [2], Loraine BRILLET[2], Mark HOEBEKE[2], Wilfrid CARRE[2], Laurence GARCZAREK[1], Frédéric PARTENSKY[1] and Christophe CARON[2]

[1] UMR7144 CNRS, Station Biologique, Place G. Teissier, 29680, Roscoff, France
{christophe.habib, laurence.garczarek, frederic.partensky}@sb-roscoff.fr
[2] ABiMS, FR2424 CNRS-UPMC, Station Biologique, Place G. Teissier, 29680, Roscoff, France
{gildas.le-corguille, erwan.corre, loraine.brillet, mark.hoebeke, wilfrid.carre, christophe.caron}@sb-roscoff.fr

Keywords cyanobacteria, orthologous groups, annotation.

Competition of photosynthetic organisms for light has triggered the development of an amazing variety of pigments and chromophorylated proteins during evolution. This diversity is stunning in the ubiquitous marine cyanobacterium *Synechococcus*, the second most abundant oxygenic phototroph on Earth after the closely related genus *Prochlorococcus*. In the ANR PELICAN project (2010-2014), we intend to study the ecology, diversity and evolution of cyanobacterial pigment types in the marine environment using an integrative genomic approach.

The first step of this approach consisted in evaluating the sensitivity and performances of several clustering tools to build protein families. The dataset used is composed of 29 genomes of *Synechococcus* and *Prochlorococcus*. The result of this evaluation result will help selecting software applications that will be integrated in a fully automated pipeline for building CyOGs (Cyanobacterial Orthologous Groups). We also tested software based on the Markov Cluster algorithm such as *OrthoMCL* [1], and other approaches such as *Uclust* [2] or *Domclust* [3].

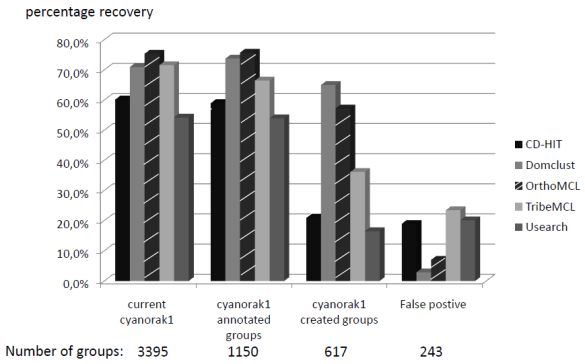


Figure 1. Comparative results of orthologous clustering tools.

The second step was the study of lateral transfers within and between the *Synechococcus* and *Prochlorococcus* genera. Several tools based on phylogenetic analyses (DarkHorse), nucleotides composition, genome organization and genome comparison (Alien Hunter, IslandPick, SIGI-HMM, IslandPath) were tested. Comparative analyses of the results were compared to detect putative genomic islands in the 29 genomes and phylogenetic origins of these islands.

These preliminary results will allow building a pipeline to integrate annotated sequences data into a relational database, usable through a collaborative platform with a user-friendly interface.

[1] Li Li, Christian J. Stoeckert, Jr., and David S. Roos, OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes, *Genome Res.*, 3 13: 2178-2189, 2003.

[2] Robert C. Edgar, Search and clustering orders of magnitude faster than BLAST, *Bioinformatics*, 26(19):2460-1, 2010.

[3] Ikuo Uchiyama, Hierarchical clustering algorithm for comprehensive orthologous-domain classification in multiple genomes, *Nucl. Acids Res.*, 34(2): 647-658, 2003.

AnnotQTL: a New Tool to Gather Functional and Comparative Information on a Genomic Region

Frédéric LECERF[1,2], Anthony BRETAUDEAU[3], Olivier SALLOU[3], Colette DESERT[1,2], Yuna BLUM[1,2], Sandrine LAGARRIGUE[1,2] and Olivier DEMEURE[1,2]

[1] INRA, UMR598 Génétique Animale, F-35000 Rennes, France
{blum, demeure}@rennes.inra.fr

[2] Agrocampus OUEST, UMR598 Génétique Animale, F-35000 Rennes, France
{lecerf, desert, lagarrigue }@agrocampus-ouest.fr

[3] GenOuest Platform, INRIA/Irisa – Campus de Beaulieu, F-35042 Rennes Cedex, France
{anthony.bretaudeau, olivier.sallou}@irisa.fr

Keywords Genomic regions, Annotation, Gene ontology, QTL.

The final steps of genetic mapping research programs require close analysis of several QTL regions to select candidate genes for further studies. Despite several websites (NCBI genome browser, Ensembl Browser, UCSC Genome Browser) or web tools (Biomart, Galaxy) developed to achieve this task, the selection of candidate genes remains a laborious process. The information made available on the more prominent websites differs slightly in terms of gene prediction and functional annotation, while other websites provide extra information that researchers may want to use (HGNC approved gene symbols, Gene Ontology Annotation or functional data, conservation of synteny with other species, etc.). It is possible to manually merge and compare this information for one QTL containing few genes, but not for many different QTL regions containing dozens of genes.

Here, we propose a web tool that, for a given region of interest, merges the list of genes available in NCBI and Ensembl, removes redundancy, adds functional annotations from different prominent web sites, and highlights the genes for which functional annotation fits the biological function or diseases of interest. The tool is dedicated to sequenced species of livestock including cattle, pig, chicken, and horse as well as dog, i.e. species that have been extensively studied (with over 8000 QTLs detected; see http://www.animalgenome.org/cgi-bin/QTLdb/index). Nevertheless, the family designs and the low number of animals used in these species, most of the studies use linkage analysis, and the QTL regions identified remain large (containing dozens of genes). Conversely, in human and model species, most analyses now draw heavily on association studies involving large cohorts, thus providing more power and accuracy, and the web tools already available focus on these species through functional annotation of SNPs in association with the trait [1-3]. As most of these tools focus on the SNP annotation itself, describing whether the SNP is located in a gene, then a coding sequence could have a functional effect, etc. While these web tools are highly efficient in providing a good annotation for specific SNPs, they clearly cannot be used to collect information on the large regions obtained in livestock species.

AnnotQTL is a web tool designed to gather the functional annotation of different prominent websites while minimizing redundant information. Using all known information substantially accelerates the gene analysis of QTL regions for livestock species traits and improves the selection of candidate genes. The AnnotQTL web tool is available at http://annotqtl.genouest.org.

[1] S. Goodswen, C. Gondro, N. Watson-Haigh, and H. Kadarmideen, FunctSNP: an R package to link SNPs to functional knowledge and dbAutoMaker: a suite of Perl scripts to build SNP databases. *BMC Bioinformatics*, 11(1): 311-311, 2010.

[2] J. Reumers, S. Maurer-Stroh, J. Schymkowitz, and F. Rousseau, SNPeffect v2.0: a new step in investigating the molecular phenotypic effects of human non-synonymous SNPs. *Bioinformatics (Oxford, England)*, 22(17): 2183-2185, 2006.

[3] M. Ryan, M. Diekhans, S. Lien, Y. Liu, and R. Karchin, LS-SNP/PDB: annotated non-synonymous SNPs mapped to Protein Data Bank structures. *Bioinformatics (Oxford, England)*, 25(11): 1431-1432, 2009.

Génolevures : Policy for Automated Annotation of Genome Sequences

Tiphaine MARTIN[1] and Pascal DURRENS[1]

[1] LaBRI, UMR5800 CNRS, 351 cours de la Libération, 33405, Talence, Cedex, France
{tiphaine.martin, pascal.durrens}@labri.fr

Keywords automatic annotation, yeast.

Recent DNA sequencing technologies (NextGen Sequencing) lead to an inflow of sequence for which adaptation of bioinformatics tools and processing policy are necessary. Nowadays, a typical project in genomics includes the sequencing of several genomes of species evolutionary related to a reference species. Thus, annotation of genome sequences by the Génolevures Consortium [1] becomes automatic, in the framework of its new projects.

Genome sequence annotation comprises 2 phases: syntactical annotation which consists in the prediction of various chromosomal elements (protein coding genes, tRNA genes, rRNA, transposons, centromeres, …), followed by functional annotation of each element often based on comparison with known sequences.

The automatic annotation pipeline assembled for the projects of the Génolevures Consortium gathers predictions of several types of objects. (1) The protein coding genes are predicted by 7 different algorithms using the same training set of sequences, which contains sequences with and without introns. The intron definition in this training sample can result either from experimental data (ESTs) or from comparisons with sequences of related genomes. The predictions are filtered according to the intron motifs and the values assessed by GeneMark [2]. (2) The contigs are compared to (a) non-coding elements of reference species with BLASTn [3], (b) proteomes of reference species and/or Uniprot with tBLASTn, and (c) PSSM representative of protein families (Génolevures protein families [4] for yeasts) with PSI-tBLASTn. (3) The other chromosomal elements are either predicted by Consortium experts or are the outcome of specific bioinformatics tools. (4) The overlap conflicts are solved by taking into account predicted models, other chromosomal elements, and similarity regions. (5) The functional annotation is then applied on the set of resulting elements, based on a decision tree inspired by previous semi-automated annotation projects. The functional annotation of a predicted model is composed by the level of similarity with the most meaningful hit, the ID and name of the hit and its functional annotation if any. This pipeline puts together bioinformatics tools widely used in the domain as well as specific scripts; it uses and produces data files in standard formats for Genomics and Bioinformatics (EMBL, GenBank, Fasta, GFF3).

The policy for automated annotation, implemented by the pipeline, allows the treatment of new genome sequences from the basis of reference related genomes, and under the annotation standards which result from the acquired experience of the Génolevures Consortium.

Moreover, all the data can feed into the MAGUS [5] genome annotation system and thus be visualized on a web navigator. The curator can see annotation of these genomes by contig (Gbrowse [6]), by element, or by homolog group and add/modify the chromosomal elements.

[1] www.genolevures.org.

[2] M. Borodovsky and J. McIninch, Recognition of genes in DNA sequence with ambiguities. *Biosystems,* 30:161-171, 1993.

[3] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 25:3387-3402, 1997.

[4] D.J. Sherman, T. Martin, M. Nikolski, C. Cayla, J.L. Souciet and P. Durrens,Génolevures: protein families and synteny among complete hemiascomycetous yeast proteomes and genomes. *Nucleic Acids Res*. 36:D550-D554, 2009.

[5] http://magus.gforge.inria.fr.

[6] L. D. Stein, The Generic Genome Browser: A building block for a model organism system database.*Genome Res.*, 12:599-1610, 2002.

Vector Genome Annotation at VectorBase

Karyn MEGY[1], Daniel LAWSON[1], Daniel HUGHES[1], Gautier KOSCIELNY[1], Derek WILSON[1] and Paul KERSEY[1]

[1] EMBL - EBI, Wellcome Trust Genome Campus, CB10 1SD, Hinxton Cambridge, United-Kingdom,
{kmegy, lawson, dsth, koscieln, dwilson, pkersey}@ebi.ac.uk

Keywords Genome annotation, vector human pathogens, VectorBase.

Tropical diseases such as malaria, filariasis or trypanosomiasis are transmitted via vector species directly through blood meals. Previously largely neglected, these vectors now have their share in the amount of genomic data generated. VectorBase [1, http://www.vectorbase.org] is a Bioinformatics Resource Centre responsible for the storage, organisation and updating of these data – currently hosting data mainly for the mosquitoes, the tick and the body louse.

Although it provides access to these data, it also generates some of them. It is involved in the annotation of all five species it represents [2, 3, 4, 5], and is currently annotating two more vector genomes: the kissing bug and the tsetse fly.

The annotation process, largely based on similarities, is derived from the Ensembl annotation pipeline [6]. After an initial repeat masking step, several independent gene sets are built, based on data of various origins and confidence degrees: (i) manual annotation (high confidence), (ii) species-specific ESTs/cDNA (medium confidence), (iii) validated genes from closely related species (medium confidence), (iv) Uniprot sequences (low confidence), (v) ab initio (low confidence). These sets are concatenated in a single set following a gap filling method: highest confidence genes are placed first on the genome, then lower confidence genes fill the gaps. The final gene set is then polished (UTR addition, functional annotation, link to external data) before being submitted to GenBank and released in VectorBase. Gene set updates occur when enough new data are available for the species, or if its assembly changes.

The pipeline is adapted for each species, based on its taxonomic position and data abundance. A new genome can benefit from the existing annotations of closely related species (e.g.: mosquitoes) or, this data source is limited (e.g. tick or kissing bug), it relies more on transcriptomic data and *ab initio* predictions. Regardless of its taxonomic position, if many species-specific data are available for a given genome (e.g. tsetse fly), those take precedence over closely related species data.

VectorBase has strong links with its users and encourages them to provide gene models for their species of interest, via a custom-made submission pipeline. Such annotations improve greatly the gene sets, and in some cases compose more than half of the models (e.g. *Anopheles gambiae*). This community involvement is fundamental and allows us to deliver a high standard of annotation.

The pipeline is deliberately very conservative, with a low false-positive rate compared to *ab initio* methods (K.M., D.L., pers. comm.). However, it remains a several month process incompatible, in its present state, with the increase of genomes to analyze. We are currently adjusting it in order to accelerate the procedure and take better advantage of the proliferation of transcriptomic data, while keeping a good standard of annotation.

[1] D.Lawson, P.Arensburger, P.Atkinson, *et al*. VectorBase: a data resource for invertebrate vector genomics. *Nucleic Acids Res*., 37:D583-7, 2009.

[2] M.Sharakhova, M.Hammond, N.Lobo, *et al*. Update of the Anopheles gambiae PEST genome assembly. *Genome Biol*.,8(1):R5, 2008.

[3] V.Nene, J.Wortman, D.Lawson, *et al*. Genome sequence of Aedes aegypti, a major arbovirus vector. *Science*, 316(5832):1718-23, 2007.

[4] P.Arensburger, K.Megy, R.Waterhouse, *et al*. Sequencing of Culex quinquefasciatus establishes a platform for mosquito comparative genomics. *Science*, 330(6000):86-8, 2010.

[5] E.Kirkness, B.Haas, W.Sun, *et al*. Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. *Proc Natl Acad Sci*,107(27):12168-73, 2010.

[6] V.Curwen, E.Eyras, T.Andrews, *et al*. The Ensembl automatic gene annotation system. *Genome Res.,*14(5):942-50, 2004.

RNA-seq Data Analysis: Lost in Normalization ?

Marie-Agnès DILLIES[1], on behalf of the Statomique Consortium[2]

[1] Plate-forme Transcriptome et Epigénome, Institut Pasteur, 28 rue du Dr Roux, 75724, Paris, Cedex 15, France
marie-agnes.dillies@pasteur.fr

[2] http://vim-iip.jouy.inra.fr:8080/statomique

Keywords RNA sequencing, normalization, differential analysis.

The continuing technical improvements and decreasing cost of next-generation sequencing technologies have made RNA sequencing (RNA-seq) a popular choice for gene expression studies in recent years. Because the data collected from such studies differ considerably from those measured using microarray technology, the statistical tools used for analysis must be adapted accordingly. In particular, several methods for the normalization and differential analysis of RNA-seq data have been proposed in recent years. However there are no clear indications on the best solution to be chosen.

In this work, we focus on a comparison of seven different proposed normalization methods for RNA-seq data in the context of analysis of differential expression: total count (TC), upper quartile (UQ, [3]), median, full quantile (FQ, [2]), Trimmed Mean of M values (TMM, [6]), Reads Per Kilobase per Million (RPKM, [4]), and the normalization method implemented in the DESeq package in R [1]. Using graphical analyses and the results of differential analysis (e.g., Table 1) we compare the normalization methods to one another using RNA-seq data from a human melanoma cell line [7]. As suggested in Table 1, a change in the normalization method can lead to a very different list of differentially expressed genes. Groups of methods appear to provide close results, while some others behave in a radically different manner. In particular, the RPKM method may not be relevant in the context of differential analysis, as correcting for gene length introduces a bias in the variance estimation [5]. In addition, a classification of normalized samples suggests that the median method may be more sensitive to differences in read count distribution between samples. From these results, we make preliminary recommendations to biologists about their use in practice.

	RPKM	TC	UQ	Median	DESeq	TMM	FQ
Total	403	1604	1462	389	1569	1581	1830

Table 1. Total number of differentially expressed genes for each normalization method (the DESeq package is used for the statistical test of differential expression).

[1] S. Anders and W. Huber, Differential expression analysis for sequence count data. *Genome Biology*, 11:R106, 2010.

[2] B. Bolstad, R. Irizarry, M. Astrand, and T. Speed, A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. *Bioinformatics*, 19:185-193, 2003.

[3] J. Bullard, E. Purdom, K. Hansen and S. Dudoit, Evauation of statistical methods for normalization and differential expression in mRNA-seq experiments. *BMC Bioinformatics*, 11:94, 2010.

[4] A. Mortazavi, B. Williams, K. McCue, L. Schaeffer, and B. Wold, Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nature Methods*, 5: 621-628, 2008.

[5] A. Oshlack and M. J. Wakefield. Transcript length bias in RNA-seq data confounds systems biology. *Biology Direct*, 4:14, 2009.

[6] M. Robinson and A. Oshlack, A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11:R25, 2010.

[7] T. Strub, S. Giuliano, T. Ye, C. Bonet, C. Keime, D. Kobi, S. Le Gras, M. Cormont, R. Ballotti, C. Bertolotto, and I. Davidson, Essential role of microphtalmia transcription factor for DNA replication, mitosis and genomic stability in melanoma. *Oncogene*, Epub ahead of print, 2011.

On the Use of the Negative Binomial Regression Model for Comparing Differential Expression or Abundance with NGS Data

Julie AUBERT[1] and Jean-Jacques DAUDIN[1]

[1] LABORATOIRE, UMR518 INRA AgroParisTech, 16 rue Claude Bernard, 75231, Paris, Cedex 05, France
{julie.aubert, jean-jacques.daudin}@agroparistech.fr

Keywords Biological variability, Comparative NGS Experiments, Generalized mixed linear model, Negative Binomial, Next Generation Sequencing, Robustness, Technical variability.

Next generation sequencing (NGS) are these days one of the key technologies in biology. Comparative RNA-Seq experiments allow to detect differentially expressed genes in two conditions, and comparative NGS-Metagenomics experiments allow to identify species or genes which are more frequent in one condition than in an another one. The whole pipeline from the raw data to the scientific conclusions includes three steps: alignment of the reads and counting, normalization and statistical analysis to identify differentially expressed genes. The last step is not yet stabilized and several competitors are proposed. Recent work is mainly focalized on the distribution of the technical error and the studies taking into account the biological variability are based on less than 4 biological replicates. The biological variability is generally considered as the most relevant source of variability in comparative studies.

We used a real metagenomic data set with many biological replicates [3] to compare the performances of different statistical methods for comparative studies: Negative Binomial model, methods implemented in DESeq [1] and edgeR [2] packages in R, Wilcoxon test, linear model on the log (GLM-log). The aim of the study is to detect differentially abundant species among 155 species present in the human gut between 41 obese danish patients and 44 non-obese danish patients. Data were normalized by the total number of reads and we used the following criteria to compare the methods : agreement between lists of declared differentially abundant species, false discovery rate evaluated by a simulation study based on the data from an homogeneous population (the 44 non-obese patients), robustness of the findings when suppressing one replicate.

The statistical method has a great impact on the results. We can separate the methods in two groups (i) methods based on the Negative Binomial (ii) Wilcoxon and GLM-log. The methods in the first group failed in controlling the type I error rate (see Table 1.). Moreover, these methods are not robust. We have made the comparative analysis of the “obese” versus “non-obese” groups. When we suppress one replicate among the 85 biological replicates, the lists of differentially abundant species are strongly modified when using methods of the first group. Therefore we recommend to be careful when analyzing comparative experiments using any method based on the Negative Binomial law.

	R=0	R>=1	mean(R)
Negative Binomial	0	50	6,32
DESeq	0	50	2,82
edgeR	0	50	8,5
Wilcoxon	48	2	0,04
GLM-log	48	2	0,04

Table 1. Number of simulations with no species (R=0) and at least one species (R>=1) declared differentially abundant between 2 groups taken at random in the non-obese population. R is the number of species (wrongly) declared significantly differently abundant between the two groups. The P-values have been adjusted using the Bonferroni procedure at level 0,05. The expected numbers for (R=0) and (R>0) are respectively 47,5 and 2,5.

[1] S. Anders and W. Huber, Differential expression analysis for sequence count data. *Genome Biology*, 11:R106, 2010.

[2] MD. Robinson, DJ. McCarthy, GK. Smyth and S. Dudoit, edgeR : a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26:136-140, 2009.

[3] J. Qin et al., A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464:59-70, 2010.

SMETHILLIUM: Spatial normalisation Method for ILLumina infinIUM HumanMethylation BeadChip

Camille SABBAH[1,2,3], Gildas MAZO[1,2,3], Caroline PACCARD[1,2,3], Fabien REYAL[1,4,5] and Philippe HUPÉ[1,2,3,4]

[1] INSTITUT CURIE, 26 rue d'Ulm, 75248, Paris, France
{camille.sabbah, gildas.mazo, caroline.paccard, philippe.hupe}@curie.fr
[2] INSERM, U900, 26 rue d'Ulm, 75248, Paris, France
[3] MINES PARISTECH, 77300, Fontainebleau, Paris, France
[4] CNRS UMR144, 26 rue d'Ulm, 75248, Paris, France
[5] INSTITUT CURIE DEPARTEMENT OF SURGERY, 26 rue d'Ulm, 75248, Paris, France
fabien.reyal@curie.net

Keywords Microarrays, methylation, spatial normalisation.

DNA methyaltion is a major epigenetic modification in human cells. Illumina HumanMethylation27 BeadChip makes it possible to quantify the methylation state of 27,578 loci spanning 14,495 genes. We developed a non-parametric normalisation method to correct the spatial background noise in order to improve the signal-to-noise ratio. The prediction performance of the proposed method was assessed on 3 fully methylatted and 3 fully unmethylated samples. We demonstrate that the spatial normalisation outperforms BeadSTudio to predict the methylation state of a given locus.

Our method allows a better concentration of the beta values around their expected value allowing the possibility to better identify intermediate state such as hemi-methylation. For a l of 0,35, the global performance criterion is 80% after normalisation, 53% with BeadStudio, and 60% without normalisation (raw).

Availability and implementation: A R script and the data are available at the following address: http://bioinfo.curie.fr/projects/smethillium

Contact: smethillium@curie.fr

Improving Mosquito Genome Annotation using RNA-Seq

Gautier KOSCIELNY[1], Daniel HUGHES[1], Karyn MEGY[1], Derek WILSON[1], Daniel LAWSON[1] and Paul KERSEY[1]

[1] EMBL - EBI, Wellcome Trust Genome Campus, CB10 1SD, Hinxton, United Kingdom
{koscieln, dsth, kmegy, dwilson, lawson, pkersey}@ebi.ac.uk

Keywords transcriptomics, genome annotation, high-throughput mRNA sequencing, RNA-Seq, alternative splice isoforms, EST clustering, Anopheles gambiae, Aedes aegypti, vector human pathogens, VectorBase.

VectorBase ([1], http://www.vectorbase.org) is an NIAID-funded Bioinformatic Resource Center focused on invertebrate vectors of human pathogens. VectorBase annotates and curates vector genomes providing a web accessible integrated resource for the research community. Currently, VectorBase contains genome information for three mosquitoes: *Aedes aegypti*, *Anopheles gambiae* and *Culex quinquefasciatus*, the body louse *Pediculus humanus*, the tick *Ixodes scapularis* and the triatomine bug *Rhodnius prolixus*.

VectorBase partners and the vector community have provided a significant amount of second generation sequencing datasets produced at low cost. Incorporating this data to inform and improve gene models present a novel challenge. One example is the transcriptional profiling studies of *Aedes* and *Anopheles* species using high-throughput RNA-Seq technologies. We have developed two pipelines to analyze RNA-Seq data from 454 pyrosequencing and Illumina sequencing platforms.

For the *Anopheles gambiae* transcriptome, we have analyzed a total of several million 454 reads from different experimental conditions. In this particular case, we have adopted a *de novo* assembly strategy to construct a set of transcript fragments used to improve the annotation of the UTR regions of the protein-coding genes. To cluster the 454 reads into contigs, we have used MIRA [2], a multi-pass DNA sequence data assembler for genome and EST projects.

For *Aedes aegypti*, we have started to work on a collection of Illumina experimental datasets containing either unpaired or paired-end reads of various lengths ranging from 36-bp to more than 100-bp depending on the sequencing platform, and representing in total hundreds of millions of reads. To date, different strategies have been employed in the analysis pipeline, depending on the type of Illumina run and length of the sequences. For instance, to map short reads back to the *Aedes* genome and identify the splicing junctions, we have used BWT-based aligners like Bowtie [3] and hash-based aligners like GSNAP [4] or GMAP [5] for longer reads. Gene models were reconstructed using existing software including Cufflinks [6] and Scripture [7] (for paired-end reads only), and a customized procedure we have developed to detect known and new transcripts from short-read alignments. Transcripts detected from different methods and experiments will be merged to build a consensus set of gene models (with alternative splice isoforms) and will help to complement existing annotations of both coding and non-coding regions of the genome. All these changes in the genome annotation will be reflected in the forthcoming releases of VectorBase.

[1] D. Lawson, P. Arensburger, P. Atkinson, et al., VectorBase: a data resource for invertebrate vector genomics. *Nucleic Acids Res.*, 37:D583-7, 2009.

[2] B. Chevreux, T. Pfisterer, B. Drescher, et al., Using the miraEST Assembler for Reliable and Automated mRNA Transcript Assembly and SNP Detection in Sequenced ESTs. *Genome Res.*, 14: 1147-1159, 2004.

[3] B. Langmead, C. Trapnell, M. Pop and S. L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10:R25, 2009.

[4] T. D. Wu and S. Nacu, Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26: 873-881, 2010.

[5] T. D. Wu and C. K. Watanabe, GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21: 1859-1875, 2005.

[6] C. Trapnell, B. A. Williams, G. Pertea et al., Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28, 511–515, 2010.

[7] M. Guttman, M. Garber, J. Z. Levin, et al., Ab initio reconstruction of cell type–specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnology*, 28, 503–510, 2010.

Importance des Banques de Séquences pour la Métagénomique

Léa SIEGWALD[1], Frédéric TEXIER[1] et Christine HUBANS-PIERLOT[1]

[1] GENOSCREEN, 1, rue du Professeur Calmette, 59000, Lille, France
{lea.siegwald, frederic.texier, christine.hubans}@genoscreen.fr

Keywords Métagénomique, Séquençage haut-débit, Biodiversité, Banques de données, Méthodologies d'analyse, Taxonomie.

Les analyses métagénomiques actuelles se basent sur la comparaison entre des séquences non identifiées et une banque de données de séquences connues, afin de déterminer le contenu taxonomique d'un échantillon d'intérêt. On peut identifier une séquence inconnue si elle est retrouvée dans la banque au-delà d'une certaine similarité. L'arbre taxonomique représentatif des organismes en présence est ainsi dessiné [1].

La métagénomique ciblée se concentre sur une région génétique précise suffisamment variable pour être un marqueur d'identification. Cette région doit être encadrée de régions flanquantes universelles pour permettre la sélection d'amorces d'amplification. On ne séquence ainsi pas tout le contenu génétique d'un échantillon, mais seulement cette région préalablement amplifiée.

De ce fait, le choix de la banque de données de référence est une décision cruciale, puisqu'elle influence directement l'interprétation des résultats. Une telle banque doit représenter une taxonomie la plus exhaustive possible : un taxon non présent dans la banque sera impossible à révéler dans l'échantillon qui y est comparé. La banque ne doit pas être redondante afin de ne pas surreprésenter certains taxons, et doit contenir des séquences de qualité et longueur suffisante pour éviter de mauvaises identifications.

Il existe déjà des banques publiques de certaines régions dont nous nous servons dans le cadre de nos analyses ; par exemple la banque SILVA [1] (ADN ribosomique, couramment utilisé chez les bactéries) ou encore l'ITS2 database [2] (région entre l'ADNr 5,8S et l'ADNr 28S, marqueur très discriminant chez les champignons). Ces banques sont construites en extrayant des séquences des banques généralistes selon différents critères : SILVA se base sur des mots-clefs et des profils de séquences, tandis que l'ITS2 database utilise des similarités structurales entre séquences. Toutefois, nous devons faire face à plusieurs problèmes inhérents à la construction de ces banques et aux séquences qui y sont comparées.

En effet, nous avons constaté des erreurs d'identification, ou encore l'absence de taxons pourtant validés expérimentalement dans l'échantillon. Nous avons par conséquent fait une comparaison sur la banque nr du NCBI, afin d'avoir un maximum d'identifications. La quantité accrue de ces dernières a augmenté le taux d'erreurs, du à la mauvaise qualité ou annotation de certaines. Malgré de nombreux tests de création de banques spécifiques à partir de nr, les résultats n'étaient pas satisfaisants, puisque de nombreuses séquences ne sont pas annotées, et n'ont donc pas été sélectionnées.

En outre, nous utilisons pour certains projets d'autres marqueurs génétiques pour lesquels il n'existe pas de banque de référence à l'heure actuelle. Nous avons ainsi développé une nouvelle méthode de construction automatique de banques de séquences mieux adaptées aux analyses métagénomiques ciblées, quelle que soit la région choisie.

Références

[1] L. Siegwald, F. Texier and C. Hubans-Pierlot, Development and optimization of metagenomic analyses, *Journées Ouvertes Biologie Informatique Mathématiques (poster session)*, Montpellier, 2010.

[2] E. Pruesse, C. Quast, K. Knittel, B. Fuchs, W. Ludwig, J. Peplies, and F. O. Glöckner, SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.*, 35:7188-7196, 2007.

[3] J. Schultz, T. Müller, M. Achtziger, P.N. Seibel, T. Dandekar and M. Wolf, The internal transcribed spacer 2 database - a web server for (not only) low level phylogenetic analyses. *Nucleic Acids Res.*, 34:704-707, 2006.

RNA-seq without a Reference Genome: a Comparison of the Mapping and the Assembly Approaches

Marie-Christine CARPENTIER[1], Delphine CHARIF[2], Janice KIELBASSA[1], Vincent LACROIX[1], Marie-France SAGOT[1] and Fabrice VAVRE[1]

[1] Laboratoire de Biométrie et Biologie Evolutive, UMR5558 CNRS, 43 boulevard du 11 Novembre 1918, 69622 Villeurbanne, France
{marie-christine.carpentier, janice.kielbassa, vincent.lacroix, marie-france.sagot, fabrice.vavre}@univ-lyon1.fr

[2] Institut Jean-Pierre Bourgin, UMR1318 INRA-AgroParisTech, Centre de Versailles-Grignon Route de St-Cyr (RD10), 78026 Versailles, France
delphine.charif@versailles.inra.fr

Keywords RNA-seq. De novo assembly.

Recent advances in sequencing technologies now enable to sequence transcriptomes. In principle, the analysis of the reads obtained from an RNA-seq experiment should enable to identify and quantify all transcripts present in an RNA sample, or to assess if the transcripts are differentially expressed, when comparing two biological conditions. One of the main novelty of RNAseq is that no a priori knowledge of the transcripts is required. RNA-seq may therefore be applied both to model or non model species.

When a reference genome is available (model species), reads from an RNA-seq experiment are directly mapped to the genome and give direct access to genes and their expression levels. If no reference genome is available (non model species), de novo assembly of reads can be used to reconstruct the transcriptome for further analyses.

In order to assess how much one can trust the results obtained by de novo assembly, we studied the level of confidence one can have in the assembled transcripts. To do this, we used RNA-seq single end libraries from Drosophila melanogaster (under two differents biological conditions with two biological replicats for each). As the genome of this species is available, we were able to apply both the mapping and the assembly approach. The mapping was carried out using TopHat [3], whereas assembly was performed using Velvet [1] and Oases (unpublished).

For each approach, we identified and quantified genes expression. For the mapping approach quantification was done by htseq. For the assembly approach, per gene counts were obtained by summing the reads that composed all the predicted transcripts for a given locus.

We then assessed if the detected genes were differentially expressed across biological conditions with the DESeq R package [2]. Next, taking the mapping approach as the gold standard, we assessed the sensitivity and the specificity of the assembly approach.

References

[1] D. Zerbino and E. Birney, Velvet: Algorithms for De Novo Short Read Assembly Using De Bruijn Graphs. *Genome Research*, 18:821-829, 2008.
[2] S. Anders and W. Huber, Differential expression analysis for sequence count data. *Genome Biology*, 11:R106, 2010.
[3] C. Trapnell, L. Pachter, SL. Salzber, TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25 (9):1105-1111, 2009.

Exploration of Host Immune Responses to Pathogens through an Analysis of Gene Expression

Timothee CEZARD[1], Seanna MCTAGGART[2], Desiree ALLEN[2], Marian THOMSON[1], Urmi TRIVEDI[1], Mark BLAXTER[1] and Tom LITTLE[2]

[1] Genepool, University of Edinburgh, Ashworth Laboratories West main road, Edinburgh, EH93JT, United Kingdom
tcezard@staffmail.ed.ac.uk

[2] Little Lab, University of Edinburgh, Ashworth Laboratories West main road, Edinburgh, EH93JT, United Kingdom
smctagga@staffmail.ed.ac.uk

Keywords Daphnia, Host, Parasite, Immunity, RNA-Seq, Expression.

Coevolving host-pathogens are characterised by genetic specificity, where the outcome of the host–pathogen interaction (e.g., the probability of infection, pathology, or parasite transmission success) depends on the specific pairing of host and pathogen genotypes. In invertebrates, however, immune system pathways studied so far are not capable of such fine-tuned specificity[1,2]. To identify the genes and pathways responsible for these interactions, we exposed three *Daphnia* host genotypes to two naturally-infecting bacterial strains in a fully factorial design including unexposed host controls.

Each condition was replicated three times and the complete transcriptome of the host was sequenced using Illumina RNA-seq technology. Quality controlled reads were mapped to the genome, duplicate reads were flagged, and differential gene expression was assessed. Gene enrichment analysis was carried out on all significantly up and down regulated genes. Allelic specific expression was determined by detecting heterozygous SNPs in individual genotypes, and testing for changes in allelic imbalance between control and treated samples.

We show that host genotypes differ dramatically in the genes they express in response to the invading pathogen, both in terms of absolute gene expression, as well as allele-specific response. The genes identified in our analysis are the first candidates to be implicated in the host pathogen specific response in Daphnia.

[1] Y. Carton, F. Frey, and Nappi, Genetic determinism of the cellular immune reaction in Drosophila melanogaster. *Heredity,* 69:393-399, 1992.

[2] H. Agaisse and N. Perrimon, The roles of JAK/STAT signaling in Drosophila immune responses. *Immunol. Rev.,* 198:72-82 2004.

EMA - A R package for Easy Microarray data Analysis

Nicolas SERVANT[1,2,3,*], Eléonore GRAVIER[1,2,3,4], Pierre GESTRAUD[1,2,3], Cécile LAURENT[1,2,3,6,7,8], Caroline PACCARD[1,2,3], Anne BITON[1,2,3,5], Isabel BRITO[1,2,3], Jonas MANDEL[1,2,3], Bernard ASSELAIN[1,2,3], Emmanuel BARILLOT[1,2,3] and Philippe HUPÉ[1,2,3,5]

[1] Institut Curie, Paris F-75248, France
{nicolas.servant, pierre.gestraud, cécile.laurent, caroline.paccard, anne.biton, isabel.brito, jonas.mandel, emmanuel.barillot, philippe.hupe}@curie.fr
{eleonore.gravier, bernard.asselain}@curie.net
[2] INSERM, U900, Paris F-75248, France
[3] Ecole des Mines ParisTech, Fontainebleau, F-77300 France
[4] Institut Curie, Departement de Transfert, Paris F-75248, France
[5] CNRS, UMR144, Paris F-75248, France
[6] CNRS, UMR3347, Orsay F-91405, France
[7] INSERM, U1021, Orsay F-91405, France
[8] Université Paris-Sud 11, Orsay F-91405, France
[*] Contact author
ema-support@curie.fr

Keywords Microarray analysis.

The increasing number of methodologies and tools currently available to analyse gene expression microarray data can be confusing for non specialist users. Based on the experience of biostatisticians of Institut Curie, we propose both a clear analysis strategy and a selection of tools to investigate microarray gene expression data. The most usual and relevant existing R functions were discussed, validated and gathered in an easy-to-use R package (**EMA** [4]) devoted to gene expression microarray analysis.

Removing noise and systematic biases is performed using the most famous techniques for Affymetrix GeneChip normalisation. The data are then filtered to both reduce the noise and increase the statistical power of the subsequent analysis. Exploratory approaches based on R packages such as **FactoMineR** [3], or **mostclust** [1] and classically used to find clusters of genes (or samples) with similar profiles are also offered. Supervised approaches, as Significance Analysis of Microarrays (**siggenes** package [5]) approach or ANOVA functions, are proposed to identify differentially expressed genes (DEG) and functional enrichment of the DEG list is assessed based on **GOstat** package [2].

The package includes a vignette which describes the detailed biological/clinical analysis strategy used at Institut Curie. Most of the functions were improved for ease of use (fewer command lines, default parameters tested and chosen to be optimal). Relevant, enhanced and easy-to-interpret text and graphic outputs are offered. The package is available on The Comprehensive R Archive Network repository.

[1] A. Bertoni and G. Valentini, Model order selection for bio-molecular data clustering. *BMC Bioinformatics*, 8 Suppl 2:S7, 2007.

[2] S. Falcon and R. Gentleman, Using GOstats to test gene lists for GO term association. *Bioinformatics*, 23:257–258, 2007.

[3] S. Le, J. Josse, and F. Husson (2008). Factominer: an R package for multivariate analysis. *Journal of statistical software*, 25:1–18, 2008.

[4] N. Servant, G. Eleonore, P. Gestraud, C. Laurent, C. Paccard, A. Biton, I. Brito, J. Mandel, B. Asselain, E. Barillot, and P. Hupé, Ema - a R package for easy microarray data analysis. *BMC Research Notes*, 3:277, 2010.

[5] V. G. Tusher, R. Tibshirani, and G. Chu, Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA*, 98:5116–21, 2001.

A Fast Ab Initio Method for Predicting miRNA Precursors in Genomes

Sébastien TEMPEL[1] and Fariza TAHI[1]

Laboratoire IBISC, Université d'Evry-Val d'Essonne/Genopole, 523, Place des Terrasses, 91000 Evry, France
{sebastien.tempel, fariza.tahi}@ibisc.univ-evry.fr

Keywords microRNA, miRNA precursor, ab intio prediction, genome-wide search.

MicroRNAs (miRNAs) are non-coding RNAs with only 21-25 nt in sequence length that are present in all sequenced higher eukaryotes ([1]). miRNA genes are cleaved into a 40-940 nt long precursor of miRNA sequences (pre-miRNAs). Pre-miRNAs, structured as hairpins, are transported into the cytoplasm and are cleaved into mature miRNA ([1]). They are involved as negative regulators of gene expression by binding to specific mRNA targets ([1]). Bioinformatics methods that predict pre-miRNAs can be divided into three approaches: comparative genomics, homology-based approaches and *ab initio* approaches. Comparative genomics and homology-based approaches cannot detect miRNAs of unknown families and/or miRNAs with no close homologous in genomes. Furthermore, comparative approaches do not work on new genomes that do not have a closely related sequenced species. Ab-initio methods are needed to predict new miRNAs in genomes. In our knowledge, there are very few ab initio algorithms that search for pre-miRNA structures in whole genomes and all are specific to one or some genomes.

We present a new ab initio method, called miRNAFold, for predicting pre-miRNA structures in any genome. Our method consider a sliding window of a given size L sufficiently long to contain a pre-miRNA. In a first step, we search for long exact Watson-Crick stems which verify some criteria. In a second step, we extend the selected stem in order to get the longest symmetrical non-exact Watson-Crick stem verifying some criteria. This longest symmetrical non-exact stem can correspond to a large portion of a pre-miRNA. Possible pre-miRNA hairpins are then searched for in the subsequence associated to the selected symmetrical non-exact stem. At each step, several selection criteria are used, corresponding to several features observed on the exact stems, the symmetrical non-exact stems and the hairpins. Some of these criteria, for example ΔG; ratio A, U, C and G, are also used in ([2,4]). Because a miRNA hairpin can present some of these features but not all, an exact stem, a symmetrical non-exact stem or an hairpin is selected when a certain percentage of the criteria are verified. This percentage is a parameter which could be set by the user.

We compared our algorithm miRNAFold with RNALFold ([3]) which searches in genomic sequences for all possible non-coding RNA secondary structures including hairpins. We thus compared the hairpins predicted by RNALFold with the ones predicted by our algorithm miRNAFold. We used RNALFold software in version 1.8.4. downloaded from the Vienna RNA Package (www.tbi.univie.ac.at/RNA/) and it was run with its default parameters. We used a sliding window of 150 nt for each of thr two software. We tested miRNAFold and RNALFold on the human, mouse, zebrafish and sea squirt genomic sequences. Each sequence contains a cluster of several known miRNAs. miRNAFold was run with a threshold of 70% for the minimum percentage of verified criteria. miRNAFold has better sensitivity and selectivity results than RNALFold on the human, mouse, zebrafish and sea quirt genomic sequences. Moreover miRNAFold is the fastest algorithm. Our average time execution is 57 seconds for a sequence of 1 million of nucleotides, when RNALFold has an average time execution of 5 minutes and 46 seconds. miRNAFold is then almost 6 times faster than RNALFold.

miRNAFold is available at http://EvryRNA.ibisc.univ-evry.fr/

References

[1] D. Bartel, MicroRNAs: genomics, biogenesis, mechanism and function. *Cell*, 116:281-197, 2004.

[2] S.A. Helvik, O.J. Snove and P. Saetrom, Reliable prediction of Drosha processing sites improves microRNA gene prediction. *Bioinformatics*, 23:142-149, 2007.

[3] I.L. Hofacker, B. Priwitzer and P.F. Stadler, Prediction of Locally Stable RNA Secondary Structures for Genome-Wide Surveys. *Bioinformatics*, 20:186-190, 2004.

[4] A. Sewer, N. Paul, P. Landgraf, A. Aravin, S. Pfeffer, M.J. Brownstein, T. Tuschl, E. van Nimwegen and M. Zavolan, Identification of clustered microRNAs using an ab initio prediction method. *BMC Bioinformatics*, 6:267, 2005.

Characterizing Novel Non-coding Transcripts in Eukaryotic Genomes Using RNA-seq Data

Marc DESCRIMES[1,2], Zohra SACI[1,2] , Chun-Long CHEN[3], Maxime WÉRY[1], Maud SILVAIN[3], Antonin MORILLON[1], Claude THERMES[3] and Daniel GAUTHERET[2]

[1] Institut Curie, UMR 3244, CNRS, 26 Rue d'Ulm, 75005, Paris, France

[2] IGM, UMR8621 CNRS, Université Paris-Sud, Bât 400, 91405, Orsay Cedex, France

daniel.gautheret@u-psud.fr

[3] CGM, UPR3404, CNRS, Bat 24-26, Ave de la Terrasse, 91198 Gif sur Yvette Cedex, France

Keywords non-coding RNA, unstable RNA, human, yeast.

Since high density DNA chips provided the first global view of the human transcriptome [1], it has become clear that most of the non protein-coding regions that comprise 98% of our genome are under active transcription. This initial finding now extends to other model eukaryotes, including the simplest ones such as single-cell yeasts. Many questions remain unanswered about this extensive transcription. Can we discriminate functional from background transcripts? Can we identify precise transcript boundaries, and classify transcripts based on their expression, processing or conservation? Answering these questions will require considerable amounts of experimental evidence gathered from multiple conditions. In the meantime, our view of the non-coding transcriptome will remain largely deficient. It is revealing that most sequencing centers still distribute genome annotations limited to CDS (coding sequence) coordinates. Basic information such as transcript boundaries remains beyond reach for most genomes. In this poster, we present how our laboratories address issues of non-coding transcript annotation at several levels, using RNA-seq experiments carried out in human and yeast.

First, we questioned the effect of RNA library preparation protocols on transcript annotation. RNA-seq protocols use either enzymatic or chemical RNA fragmentation for sequencing. We show that these methods significantly affect the accuracy of transcript reconstruction. One of the methods produces more irregular transcript coverage, which favors artifacts such as artificial dissociated transcripts within the same locus.

Second, we have been screening transcriptome variations in mutants of the RNA degradation pathway in yeast. In a recent study (Fig. 1) the AM lab identified a novel class of non coding transcripts that are normally degraded by the Xrn1 nuclease. These RNAs, termed XUTs (Xrn1-sensitive Unstable non-coding Transcripts) constitute a novel class of transcripts that further expand the repertoire of cryptic transcripts [2]. Other mutant yeasts are now under scrutiny, including one producing small interfering RNAs through expression of an ectopic argonaute/dicer system.

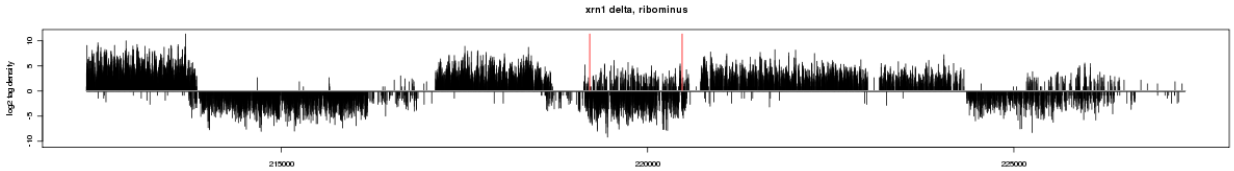


Figure 1. Strand-specific mapping of RNA-seq reads in the vicinity of the *ARG1* gene (red lines) in a *S. cerevisiae* *Δxrn1* mutant [2]. The *ARG1* gene shows unusual antisense expression (bottom).

Third, we are now performing meta-analysis of human RNA-seq data in order to refine our understanding of spliced non-coding transcripts. There are several thousand such transcripts in human, but only a few have proposed functions. We show that public RNA-seq data is now sufficient to enable a comprehensive analysis of these transcripts and their variations.

[1] E. Birney, R.G. Anindya Dutta, R.G. Thomas et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447:799-816, 2007.

[2] E. VanDijk et al. XUTs are Xrn1-sensitive non-coding regulatory transcripts in yeast. *Nature*. 2011 *in press*.

Screening Bacterial Regulatory RNAs and their Targets Using Evolutionary Profiles

Alban OTT[1], Anouar IDALI[1], Antonin MARCHAIS[2] and Daniel GAUTHERET[1]

[1]IGM, UMR8621 CNRS, Université Paris-Sud, Bât 400, 91405, Orsay Cedex France
daniel.gautheret@u-psud.fr

[2] Swiss Federal Institute of Technology, Biology Dept., Universitatstrasse 2, 8092 Zurich, Switzerland

Keywords sRNA targets, ncRNA, Bacteria.

Bacterial small RNAs (sRNAs) regulate target messenger RNAs by mediating their degradation or translation block. A bacterial genome may contain up to a few hundred sRNAs, and each sRNA can target up to dozens of mRNAs. Few actual sRNA-mRNA interactions are experimentally demonstrated and the total number of targeted mRNAs is unknown. In the absence of large scale data for these interactions, a global view of the RNA-RNA interaction network remains out of reach.

The NAPP (Nucleic Acid Phylogenetic Profile) pipeline uses all available bacterial genomes to identify ncRNAs and mRNA displaying similar patterns of presence/absence across species [1]. We showed NAPP is efficient for both sRNA prediction and the identification of functional clusters of sRNAs and mRNAs [2]. Here we will present a web interface that allows biologists to access NAPP prediction of sRNAs and other non-coding RNAs in about 1000 bacterial genomes, together with functional information derived from the analysis of NAPP clusters. The NAPP database is available at http://rna.igmors.u-psud.fr/toolbox/

In a parallel study, we analyzed the conservation profiles of mRNA sequences in order to observe potential signatures of targeting by sRNAs. A recent article [3] shows that, in sRNAs, the region targeting the mRNA is more conserved than other regions. To investigate what happens on the mRNA side, we computed conservation profiles of all *E. coli* mRNAs by comparison with 1070 bacterial species. We found evidence that known sRNA targets present a conservation signature in their 5' untranslated region that distinguishes them from other mRNAs (Fig. 1). This is consistent with the observation that most documented sRNA-mRNA pairs interact in this region, around the ribosome binding site (RBS). Using a clustering analysis, we identify groups of mRNAs with similar conservation profiles that may be enriched in novel sRNA targets. We present an analysis of these clusters in terms of mRNA functions and putative associated sRNAs.

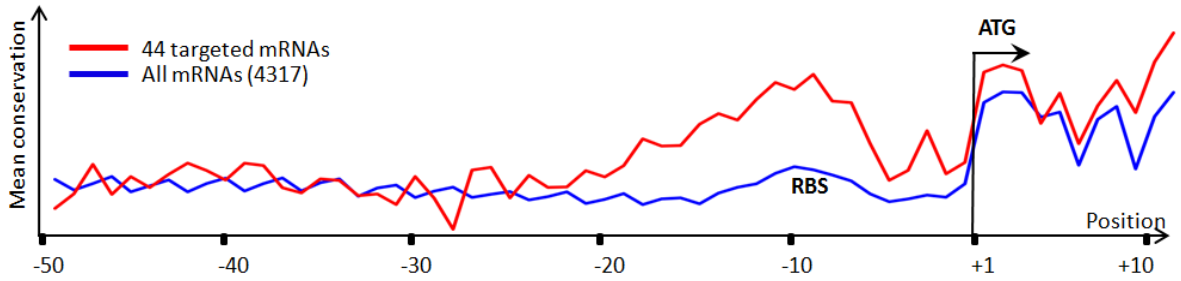


Figure 1. Average conservation profiles of *E. coli* mRNAs in a 60-nt window around the start codon (ATG). mRNAs targeted by sRNAs show manifest conservation peaks at and upstream of the ribosome binding site (RBS).

[1] A. Marchais, M. Naville, C. Bohn, P. Bouloc and D. Gautheret. Single-pass classification of all noncoding sequences in a bacterial genome using phylogenetic profiles. *Genome Res*. 19:1084-92, 2009.

[2] A. Marchais, S. Duperrier, S. Durand, D. Gautheret and P. Stragier. CsfG, a family of sporulation-specific, small non-coding RNA highly conserved in endospore formers. *RNA Biol*. in press, 2011.

[3] A. Peer, and H. Margalit. Accessibility and evolutionary conservation mark bacterial small-RNA target-binding regions. *J Bacteriol*. 193:1690-701, 2011.

Annotation of Non-coding RNA in Vibrio Using RNA-seq Data

Claire Toffano-Nioche, Claire Kuchly, Ngoc An Nguyen, Philippe Bouloc, Daniel Gautheret and Annick Jacq

IGM, UMR8621 CNRS, Université Paris-Sud, Bât 400, 91405, Orsay Cedex, France
annick.jacq@u-psud.fr

Keywords sRNA, ncRNA, Bacteria, Vibrio.

The Vibrio genus comprises a number of important human and animal pathogens. Among genes that confer virulence to these bacteria, small regulatory RNAs (sRNAs) appear to play a significant role [1]. These sRNA genes, as well as other non-coding RNAs (ncRNAs) such as cis-acting RNAs and antisense RNAs, are difficult to characterize as they are devoid of a specific sequence signature. Most known ncRNAs in Vibrio are derived from genetic studies or medium-throughput analyzes. Here we attempt to expand the catalog of Vibrio ncRNAs using high-throughput RNA-seq analysis of *Vibrio splendidus* and comparative genome analysis.

Our RNA-seq experiment involved total RNA extraction from exponentially growing *V. splendidus* and treatment by the exoribonuclease Terminator for rRNA depletion and enrichment in primary transcripts. Sequencing on the Illumina GA IIx platform (36 nt length), and subsequent mapping using the Bowtie program [2] produced 4.5 million uniquely mapping reads. We implemented computational pipelines for detecting cis-acting, trans-acting and antisense ncRNAs from these mapping data. Our pipelines rely on the S-MART program (M. Zytnicki, unpub.) that performs transcript reconstruction and facilitates the comparison of mapped reads with existing annotation based on user-defined parameters.

After parameter optimization based on a set of 30 RFAM RNAs [3], we predict 943 ncRNAs in the *V. splendidus* genome, including 584 long 5' UTRs that may contain cis-regulatory RNAs, 282 sRNAs and 77 cis-encoded antisense RNAs. Manual curation validates about 82% of these RNA candidates. We compared the RNA-seq-derived transcripts to RNAs predicted by bioinformatics alone using the SIPHT [4] and NAPP [5] systems, both based on conservation analysis. The overlap with SIPHT prediction is relatively low (~30% for sRNAs), suggesting that (i) a significant fraction of in silico-predicted RNAs are either false positives or expressed in specific conditions and (ii) a significant fraction of RNA-seq-supported transcripts are not conserved and may represent functional transcripts of recent emergence or high mutation rate. To further analyze the emergence of novel RNAs in the Vibrio genus, we present a comparison of our *V. splendidus* results with results of medium throughput screens performed on other Vibrio species [6].

We thank Matthias Zytnicki from INRA-URGI Versailles for assistance with the S-MART program and the high throughput sequencing platform of IMAGIF (www.imagif.cnrs.fr) for their facilities and expertise.

[1] B.K. Hammer and B.L. Bassler, Regulatory small RNAs circumvent the conventional quorum sensing pathway in pandemic Vibrio cholerae. *Proc Natl Acad Sci U S A*, 104(27):11145-9, 2007.

[2] B. Langmead, C. Trapnell, M. Pop and S.L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, 10(3):R25, 2009.

[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 Res.*, 37:D136-40, 2009.

[4] J. Livny, H. Teonadi, M. Livny and M.K. Waldor, High-throughput, kingdom-wide prediction and annotation of bacterial non-coding RNAs. *PLoS One*, 3:e3197, 2008.

[5] A. Marchais, M. Naville, C. Bohn, P. Bouloc and D. Gautheret, Single-pass classification of all noncoding sequences in a bacterial genome using phylogenetic profiles. *Genome Res*., 19:1084-92, 2009.

[6] J.M. Liu, J. Livny, M.S. Lawrence, M.D. Kimball, M.K. Waldor and A. Camilli, Experimental discovery of sRNAs in Vibrio cholerae by direct cloning, 5S/tRNA depletion and parallel sequencing. *Nucleic Acids Res*., 37:e46, 2009.

Bios2mds – an R Package for Metric Multidimensional Scaling Analysis of Multiple Sequence Alignments

Julien Pele[1], Jean-Michel Becu[1] ,Hervé Abdi[2] and Marie Chabbert [1]

[1] BNMI, UMR CNRS 6214 – INSERM U771, 3 rue de Haute de reculée, Faculté de médecine, 49045, Angers, France
{julien.pele,marie.chabbert}@univ-angers.fr,jean-michel.becu@etu.univ-rouen.fr
[2] SCHOOL OF BEHAVIORAL AND BRAIN SCIENCES, University of Texas, TX 75080-3021, Dallas, USA
herve@utdallas.edu

Keywords Metric Multiple Dimensional Scaling, Sequence analysis, R package, Phylogeny.

The multiple alignment of homologous sequences can provide numerous information on the evolution and the sequence-function relationships of protein families. Most methods for the analysis of multiple sequence alignments rely on protein clustering after the building of a tree inferring phylogenetic relationships. However, multivariate analysis can also provide useful information without inferring a hierarchical structure of the data. In particular, metric multidimensional scaling (MDS) is a powerful exploratory procedure designed to identify patterns in a distance matrix [1, 2]. MDS visualizes elements into a 2D or 3D space in such a way that the distances between these elements best approximate the original distances. Applied to biological sequences, this method usefully complements phylogenetic data [3]. Moreover, MDS allows projection of supplementary elements onto a space of reference, based on the matrix of distances between reference and supplementary elements [1, 4]. MDS is thus a very useful tool to compare orthologous sequence sets.

We have developed the R package *bios2mds* to provide all the tools necessary to perform an MDS analysis from a multiple sequence alignment and to analyze the data. It allows users to build matrices of distances between aligned sequences, to analyze these matrices by MDS and the resulting data by K-means clustering. Moreover, *bios2mds* allows the projection of supplementary sequences onto a reference space and the visualization of the reference and supplementary elements, with user provided color scheme (Fig.1). Finally, data can be exported in a PDB format for 3D visualization with molecular graphics programs.

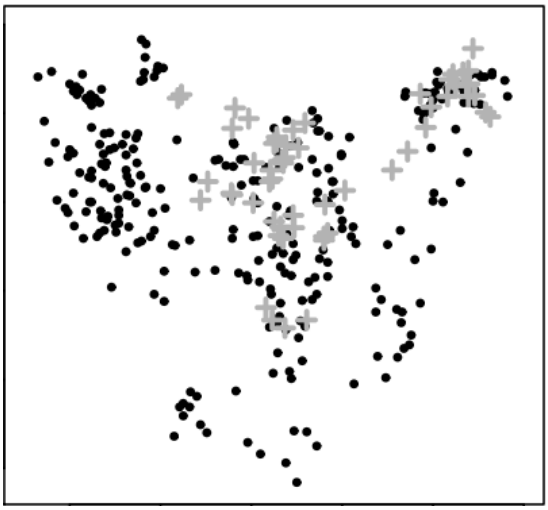


Figure 1. 2D representation of the sequence space of human GPCRs (black dots), onto which the GPCRs from *D. melanogaster* (grey crosses) are projected.

[1] H. Abdi, Metric multidimensional scaling, *Encyclopedia of Measurement and Statistics.*, 598-605, 2007.

[2] W.S. Togerson, Theory and methods of scaling. Wiley, New York, 1958.

[3] J. Pelé, H. Abdi, M. Moreau, D. Thybert and M. Chabbert, Multidimensional scaling reveals main evolutionary determinants and class A G-protein-coupled receptors, *PloS one*,april 2011.

[4] J.C. Gower, Adding a Point to Vector Diagrams in Multivaraiate Analysis *Biometrika*, 55, 582-585, 1968.

Mixing Biological Descriptors For High Throughput Metalloproteins Prediction In Bacterial Genomes

Johan ESTELLON[1], Sandrine OLLAGNIER DE CHOUDENS[2], Alain VIARI[3], Marc FONTECAVE[2] and Yves VANDENBROUCK[1]

[1] Laboratoire de Biologie à Grande Echelle, CEA, IRTSV; INSERM U1038 ; Université Joseph Fourier, 38054 Grenoble Cedex 09, France

{johan.estellon, yves.vandenbrouck}@cea.fr

[2] Laboratoire de Chimie et Biologie des Métaux, CEA, IRTSV ; CNRS, UMR5249 ; Université Joseph Fourier, 38054 Grenoble Cedex 09, France

{sandrine.ollagnier, marc.fontecave}@cea.fr

[3] Baobab team, INRIA Rhône-Alpes 38330 Montbonnot-Saint Martin, France

alain.viari@inrialpes.fr

Keywords metalloproteins identification, iron-sulfur proteins, bacterial genomes, generalized linear model, sequence analysis, bioinformatics.

Up to 40% of all proteins are known to functionally bind metals, the intrinsic metal atoms providing catalytic, regulatory and/or structural roles critical to their functions. These metalloproteins are of major importance within the three domains of life. However, current methods dedicated to identifying members of this large family within bacterial proteomes are either not suitable for large-scale screening purposes or are of relatively limited performance when no 3D structural templates are available [1]. Within this context, we developed an approach based on a generalized linear model which combines the results of different sequence analysis tools. These tools are based on the screening of protein descriptors (*e.g.* patterns, conserved domains, structural protein domains) or the building of protein profiles for remote homolog detection, each with a different scoring function. We assessed their respective predictive power towards the identification of a subset of metalloproteins, the iron-sulfur proteins (Fe-S), either separately or in combination. The linear model is trained on a dataset composed of protein sequences from the PDB70 (protein structures databank). Each protein is represented by a boolean vector that indicates the presence or absence of the 83 Fe-S descriptors we considered in this study. Five predictive models were built: four correspond to each class of descriptors and a mixed one concatenating the whole set of descriptors. Each linear model was estimated by a logistic regression procedure, allowing the selection and weighting of the most relevant Fe-S protein-predictive descriptors. We observed that descriptors based on distant homology profiles are more sensitive and less specific than those commonly used (patterns, domains), and that their inclusion in the combined model increases its global prediction quality. Then, we tested performances of each linear model on the complete genome of *Escherichia coli* K12 and noticed that the mixed model outperformed each approach considered separately.

Descriptors families	TP	TN	FP	FN	Pre.	Rec.	F_2
Patterns	53	4071	2	88	96.4%	37.6%	0.43
Conserved domains	76	4070	3	65	96.2%	53.9%	0.59
Structural domains	72	4069	4	69	94.7%	51.1%	0.56
Distant homologies	91	4012	63	48	59.1%	65.5%	0.64
Mixed model	91	4062	13	48	87.5%	65.5%	0.69

Table 1. Performances of each predictive model: four models for each class of descriptors and the mixed model which comprises the overall set of FeS descriptors. All models were trained on the PDB70 without E.coli sequences and assessed on E.coli genome. True positives (TP) and negatives (TN), false positives (FP) and negatives (FN) were determined by comparison of the prediction with Hamap annotations and literature. Recall (Rec.), is the fraction of correct predictions among all FeS proteins, while precision (Pre.) is the fraction of correct predictions among those that the algorithm believes to belong to the FeS proteins family. The F_2 measure (F_2) is the weighted harmonic mean of precision and recall; this latter metric reflects the efficiency of the model.

[1] A. Cvetkovic et al. Microbial metalloproteomes are largely uncharacterized. *Nature*, 466: 779-782, 2010.

Fine-tuning Motif Detection among ChIP on Chip DNA Fragments

Fabrice TOUZAIN[1], Théo MOZZANINO[2], Sophie SCHBATH[3] and Marie-Agnès PETIT[4]

[1] Génomique des Microorganismes, UMR7238 CNRS, Université Pierre et Marie Curie, Cordeliers, Paris, France
fabrice.touzain@courriel.upmc.fr

[2] Micalis, UMR1319 INRA, bat 222, Jouy en Josas 78352 cedex, France
theo.mozzanino@jouy.inra.fr

[3] Mathématique Informatique et Génome, INRA, Jouy en Josas 78352 cedex, France
sophie.schbath@jouy.inra.fr

[4] Micalis, UMR1319 INRA, bat 222, Jouy en Josas 78352 cedex, France
marie-agnes.petit@jouy.inra.fr

Keywords DNA position weight matrix, motif exceptionality, motif enrichment.

Besides promoters, bacterial genomes are also rich in small (10-30 bp in length), dispersed and repeated (20-80 copies) DNA motifs that serve to maintain or structure the genome [1]. These are progressively being uncovered in the last years, thanks to the Chip on chip technique, when the structuring/organizing protein is already known, or also sometimes with pure statistical means [2]. Chip on chip is a powerful technique allowing to identify collections of DNA fragments bound by a given protein at a given time in the living cell. A bioinformatics analysis is usually required to deduce, from the collection of DNA fragments given as an output of the technique (usually 1-2 kb in length), the DNA motif that is common to all fragments and may be the target of the protein under study. Motif search can be completed by enumeration of all motifs and ranking their exceptionality compared with a random Markovian model of the sequence under study, using programs such as R'mes. Another approach for motif detection is based on multiple alignment with programs such as Bioprospector or Gimsan. In this last case, the output is a set of weight position matrices describing motifs of a given length (Gimsan) or two motifs separated by a gap (Bioprospector, conceived for bacterial promoter detection). However, they usually propose more than one single candidate motif, and another criterium, such as the enrichment of the motif among the precipitated fragments, relative to the rest of the genome, would be of a great help to assist motif prediction. In most cases reported however, such an enrichment factor is not calculated, and authors remain silent about the procedure that lead them to the appropriate motif prediction.

This poster investigates methods for estimating the enrichment of a DNA motif, when described by a weight position matrix. To test these methods, a set of data was taken from the literature. The Chip on chip results obtained for four proteins structuring the bacterial chromosome, MatP, SlmA, Noc and Ram were collected. The two first proteins are encoded by *Escherichia coli*. MatP constrains the 800 kb TER macrodomain of *E. coli*, and permits its proper segregation after replication is completed. SlmA covers a large 3 Mb region around the origin of replication, and prevents the formation of a septum, so that septation occurs preferentially at mid-cell. Noc is the functional equivalent of SlmA in *Bacillus subtilis*, and Ram binds a region around the origin of replication of *B. subtilis*, to facilitate chromosome packing into the spore.

The difficulty for estimating a motif enrichment with weight position matrices is due to the fact that such matrices do not permit to count motif occurrences in a genome, they only provide a probability that a given position in the genome corresponds to the motif. We compare two methods for estimating enrichment in precipitated fragments, that both detect efficiently the motifs tested. We also intend to use this approach for de novo detection of motifs still hidden in bacterial genomes.

[1] F. Touzain, M.-A. Petit., S. Schbath, and M. El Karoui, DNA motifs that sculpt the bacterial chromosome. *Nature Reviews Microbiology* **9**: 15-26, 2011.

[2] R. Mercier, M.-A. Petit, S. Schbath, S. Robin, M. El Karoui, F. Boccard and O. Espeli, The MatP/matS site specific system organizes the Terminus region of the E. coli chromosome into a Macrodomain. *Cell* **135**: 475-485, 2008.

Interaction Profile of Small Inhibitors Complexed with Falcipain-2 and Falcipain-3 Plasmodial Cysteine Proteases

Priscila DA SILVA FIGUEIREDO CELESTINO[1], Diego ENRY BARRETO GOMES[2] and Pedro GERALDO PASCUTTI[3]

[1,3] Laboratório de Modelagem e Dinâmica Molecular (LMDM), UFRJ, Av. Carlos Chagas Filho, 373/ D30, Ilha do Fundão, Rio de Janeiro, 21941-902, Brazil.
{pfigueiredo,pascutti}@biof.ufrj.br

[2] Laboratoire Bioinformatique, Modélisation et Dynamique Moléculaire (BiMoDyM), ENS Cachan, 61, avenue du Président Wilson 94235 Cachan cedex, France
dbarreto@ens-cachan.fr

Keywords malaria, cysteine-protease, docking, molecular dynamics.

It is estimated that 500 million cases of malaria occur each year, leading to 1 million of deaths, mainly caused by the *Plasmodium falciparum* specie [1]. *P.falciparum* cysteine proteases Falcipain-2 (FP2) and Falcipain-3 (FP3) act in the hemoglobin degradation pathway, the parasite's main source of aminoacids. The use of cysteine proteases inhibitor interrupts the hemoglobin degradation and some of them lead to cure of the disease in infected mice [2]. Many FP2 or FP3 inhibitors have been described, this work focused on the complexes with some of the classes described as potent and specific inhibitors: a vinyl-sulfone peptide-based (v1b) [3]; three peptidomimetics, with a pyridone ring scaffold (v5b) [3] and benzodiazepine scaffold (et2b and et4c) [4,5]; and two non-peptide inhibitors (des4 and zhu2k) [6,7]. Our goal is to characterize the enzimes-ligand interactions to support the rational design of new compounds. To accomplish that molecular docking and molecular dynamics simulations were performed for the complexes FP2/FP3-inhibitors.

The analysis of the intermolecular contact area and the hydrogen bond showed that the introduction of non-peptidic scaffold in the backbone of the peptidomimetic inhibitors did not interfere with the stabilization of the complexes. It was found highly prevalent hydrogen bonds involving the inhibitors backbone and critical residues of the active site of the enzymes, but in higher number for the complexes with FP2, suggesting that in general the inhibitors are more specific to FP2. The exception was v1b inhibitor. The non-peptidic zhu2k showed highly prevalent bonds with both enzymes, suggesting that it can be a candidate for common inhibition. The analysis of the binding energy confirmed the better interaction of v1b with FP3. Smallest inhibitors as v5b, des4 and zhu2k also showed small energy values for this enzyme. For FP2 the peptidomimetic et2b and et4c together with the non-peptidic zhu2k showed the smallest en energy values of interaction. Based on the intermolecular contact area and the hydrogen bond network we also performed an analysis of the chemical groups components derived from the inhibitors, highlighting the individual portions that would best fit each three of the four subsite cavities of both FP2 and FP3 active sites.

[1] S. H. I. Kappe, A. M. Vaughan, J. A. Boddey and A. F. Cowman. That Was Then but This Is Now: Malaria Research in the Time of an Eradication Agenda. *Science*, 328 (5980):862-866, 2010.

[2] S. Soni, S. Dhawan, K. M. Rosen, M. Chafel, A. H. Chishti and M. Hanspal, Characterization of Events Preceding the Release of Malaria Parasite from the Host Red Blood Cell. *Blood Cells Mol. and Dis.*, 35(2):201-211, 2005.

[3] E. Verissimo, N. Berry, P. Gibbons, M. L. S. Cristiano, P. J. Rosenthal, J. Gut, S. A. Ward and P. M. O'Neill, Design and Synthesis of Novel 2-Pyridone Peptidomimetic Falcipain 2/3 Inhibitors. *Bioorg. & Med. Chem. Lett.*,18(14):4210-4214, 2008.

[4] R. Ettari, E. Nizi, M. E. Di Francesco, M.-A. Dude, G. Prade, R. Vicik, T. Schirmeister, N. Micale, S. Grasso and M. Zappalà, Development of Peptidomimetics with a Vinyl Sulfone Warhead as Irreversible Falcipain-2 Inhibitors. *J Med Chem*, 51(4):988-996, 2008.

[5] R. Ettari, N. Micale, T. Schirmeister, C. Gelhaus, M. Leippe, E. Nizi, M. E. Di Francesco, S. Grasso and M. Zappalà, Novel Peptidomimetics Containing a Vinyl Ester Moiety as Highly Potent and Selective Falcipain-2 Inhibitors. *J Med Chem*, 52(7):2157-2160, 2009.

[6] P. V. Desai, A. Patny, J. Gut, P. J. Rosenthal, B. Tekwani, A. Srivastava and M. Avery, Identification of Novel Parasitic Cysteine Protease Inhibitors by Use of Virtual Screening. 2. The Available Chemical Directory. *J Med Chem*, 49(5):1576-1584, 2006.

[7] J. Zhu, T. Chen, L. Chen, W. Lu, P. Che, J. Huang, H. Li, J. Li and H. Jiang, 2-Amido-3-(1h-Indol-3-Yl)-N-Substitued-Propanamides as a New Class of Falcipain-2 Inhibitors. 1. Design, Synthesis, Biological Evaluation and Binding Model Studies. *Molecules*, 14(1):494-508, 2009.

Structure-based Classification of the Plant Non-specific Lipid Transfer Protein Superfamily Towards its Functional Characterization

Cécile FLEURY[1], Marie-Françoise GAUTIER[1], Franck MOLINA[2], Frédéric DE LAMOTTE[1] and Manuel RUIZ[1]

[1]UMR AGAP CIRAD/INRA, avenue Agropolis, 34398, Montpellier, Cedex 5, France
{cecile.fleury, marie-françoise.gautier, frederic.de_lamotte, manuel.ruiz}@cirad.fr

[2] SysDiag, UMR3145 CNRS/Bio-Rad, 1682 rue de La Valsière, CS 61003, Montpellier, Cedex 4, France
franck.molina@sysdiag.cnrs.fr

Keywords comparative modeling, structure-function relationships, nsLTP superfamily.

The non specific Lipid Transfer Proteins (nsLTPs) show large variations in their sequences, biological roles, quaternary associations and the nature of bound hydrophobic ligands. However, they share a conserved cysteine pattern which plays an important role in the structural scaffold [1]. Besides, they are involved in a large number of biological processes relative to plant development and defense. For these reasons, the nsLTP superfamily constitutes an interesting case of study to validate a method to investigate protein structure-function relationships. Eight hundreds mature amino acid sequences belonging to more than 100 plant species have been selected and submitted to comparative phylogenic, structural, and functional analysis [2, 3, 4]. Using the evolutionary trace method [5], the observation of structurally equivalent positions allowed identifying evolutionarily important residues potentially involved either in the structural integrity or in the ligand binding diversity of the nsLTPs.

[1] F. Boutrot, N. Chantret and M.F. Gautier, Genome-wide analysis of the rice and Arabidopsis non-specific lipid transfer protein (nsLtp) gene families and identification of wheat nsLtp genes by EST data mining. *BMC Genomics*, 9:86, 2008.

[2] A. Sali and T.L. Blundell, Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol.*, 234:779-815, 1993.

[3] C. Fleury, V. Moreau, L. Felicori, S. Pérès, P.S.L. Beirão and F. Molina, *in preparation*.

[4] A.R. Ortiz, C.E. Strauss and O. Olmea, MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci.*, 11:2606-2621, 2002.

[5] O. Lichtarge, H. Bourne and F.E. Cohen, An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol.*, 257:342-358, 1996.

Modélisation d'ADN en Épingle à Cheveux avec l'Approche Mésoscopique "Biopolymer Chain Elasticity" (BCE), RMN, et Dynamique Moléculaire

Jean A. H. COGNET[1], Meriem BAOUENDI[1], Guillaume P. H. SANTINI[2], Sotiris MISSAILIDIS[3], Edith HANTZ[4] and Catherine HERVE DU PENHOAT[4]

[1] Laboratoire Acides Nucléiques et Biophotonique, ANBioPhi, Université Pierre et Marie Curie FRE 3207 CNRS, 4 place Jussieu, T32-33 4e, 75252, PARIS, Cedex 05, France

jean.cognet@upmc.fr

[2] Laboratoire d'Informatique de Paris-Nord, Université Paris 13 UMR 7030 CNRS, 99 av J-B Clément, 93430, VILLETANEUSE, France

[3] Chemistry Department, The Open University, MILTON KEYNES, MK7 6AA, UK,

[4] Laboratoire CSPBAT, Université Paris 13 UMR 7244 CNRS, 74 rue Marcel Cachin, 93017, BOBIGNY, France

Keywords Modélisation moléculaire mésoscopique, épingle à cheveux d'ADN, théorie de l'élasticité des barres minces, RMN.

Un objectif majeur de la modélisation biomoléculaire est de calculer et de prédire les conformations et les propriétés des biopolymères (ADN, ARN, protéines). La principale difficulté provient de la taille des molécules et de la complexité des conformations possibles à l'échelle atomique. Nous présentons ici les deux premières applications d'une méthode permettant de rechercher et d'obtenir les conformations qui correspondent à des minima énergétiques à différentes échelles de taille.

La résolution de la conformation de l'ADN par RMN reste toujours difficile parce qu'expérimentalement sous déterminée, même à haute sensibilité (cryoprobe 750 MHz RMN). La région du squelette sucre-phosphate la plus récalcitrante depuis les premières résolutions (~1990) dans les épingles à cheveux d'ADN est précisément la région du coude ou "sharp turn". Pour obtenir des conformations de haute qualité pour deux séquences d'ADN, représentantes de deux principales catégories de boucles, 5'-d(GCGAAAGC)-3', 5'-d(…CCTTTGG…)-3', nous avons utilisé une méthode de modélisation mésoscopique - déformation globale du squelette de l'ADN à l'aide de la théorie de l'élasticité des barres minces - appelée BCE. La déformation d'une hélice d'ADN-B canonique génère à l'échelle globale de plusieurs nucléotides la conformation en épingle à cheveux de référence la moins déformée qui satisfait les contraintes expérimentales RMN. Puis nous avons exploré les différentes combinaisons d'angles de torsion qui restaient possibles ou indéterminées par mécanique moléculaire avec AMBER pour interpréter les données RMN de ^{31}P. Grâce à cette méthode de modélisation à l'échelle globale, mésoscopique, puis atomique, nous obtenons une conformation qui satisfait non seulement toutes les contraintes RMN, mais qui permet aussi d'initier une dynamique moléculaire reproduisant toutes les observations avec une seule contrainte de plissement d'un sucre dans le premier cas, et sans aucune contrainte dans le second, chose impossible autrement.

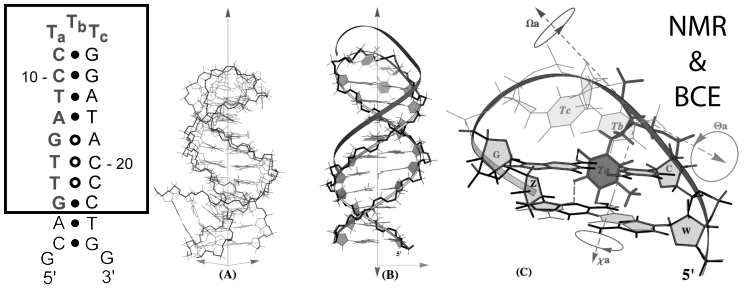


Figure 1. Application à la résolution d'un aptamère d'ADN anti MUC1 comportant la boucle TTT en épingle à cheveu.

[1] G.P.H. Santini, J.A.H. Cognet, D. Xu, K. K. Singarapu & C. Hervé du Penhoat, Nucleic acid folding determined by mesoscale modeling and NMR spectroscopy: solution structure of d(GCGAAAGC). *J. Phys. Chem. B,* 113:6881-6893, 2009.

[2] M. Baouendi, J.A.H. Cognet, C.S.M. Ferreira, S. Missailidis, J. Coutant, M. Piotto, E. Hantz & C. Hervé du Penhoat, Solution structure of a truncated anti-MUC1 DNA aptamer determined by mesoscale modeling and NMR), *en préparation.*

Protein-peptide HADDOCKing

Mikael TRELLET[1], Adrien S.J. MELQUIOND[1] and Alexandre M.J.J. BONVIN[1]

[1]Computational Structural Biology, Bijvoet Center for Biomolecular Research, Utrecht University, Padualaan 8, 3584CH Utrecht, The Netherlands

mikael.trellet@gmail.com

Keywords Protein-peptide interactions, molecular docking, HADDOCK.

Protein-protein interactions govern all cellular mechanisms that underlie life. Peptide-mediated interactions, which contribute up to 40% of the protein interactome, play a prominent role in regulatory processes such as signal transduction and protein trafficking. Given their prevalence and the broad repertoire of biological processes they mediate, peptides are associated to many human diseases and cancer, such as HIV. Short peptide stretches are also known to block existing protein-protein interactions, making them promising leads for drug design developments.

The short size of these peptides combined with their highly flexible nature results in small and very versatile binding protein-peptide interfaces that are therefore difficult to predict. Nevertheless, recent efforts in molecular docking show that this technique can be successfully applied to study protein-peptide interactions [1], even when the structure of the peptide is missing. In this work, we report on the modelling of protein-peptide complexes using our in-house flexible docking program, HADDOCK [2]. HADDOCK distinguishes itself from other docking software by its unique data-driven approach where experimental or bioinformatic information is directly used in the docking search to find near-native solutions. The performance of HADDOCK for protein-peptide modelling was benchmarked on a set of 103 protein-peptide complexes [3] covering the diversity of protein-peptide interfaces.

Our initial results show that near-native (bbRMSD < 2Å) structures can be obtained for 90% of the dataset starting from a fully extended conformation of the peptide and the bound conformation of the protein. Even when starting from the unbound form of the protein an impressive 80% success rate is observed! Analysis of the results suggests that desolvation energy is less important than for the modelling of protein-protein complexes.

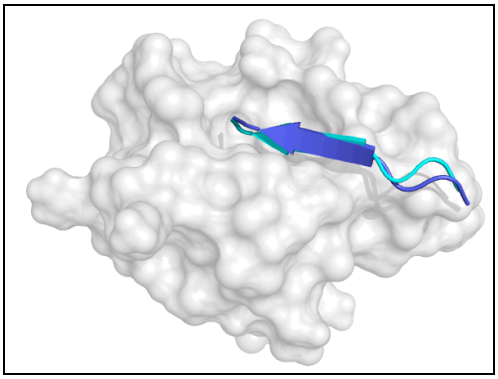


Figure 1. 1N7F (PDZ domain) - 0.59 Å. In dark blue, the bound conformation; in cyan, the best solution given by HADDOCK starting from an extended conformation of the peptide.

[1] B. Raveh, N. London and O. Schueler-Furman, Sub-angstrom modeling of complexes between flexible peptides and globular proteins. *Proteins*, 78(9):2029–2040, 2010.

[2] C. Dominguez, R. Boelens and A. Bonvin, HADDOCK: a protein-protein docking approach based on biochemical and/or biophysical information. *J. Am. Chem. Soc.*, 125:1731–1737, 2003.

[3] N. London, D. Movshovitz-Attias, O. Schueler-Furman, The structural basis of peptide-protein binding strategies. *Structure*, 18:188-199, 2010.

The Plasticity of TGF-beta Signaling

Géraldine CELLIÈRE[1], Georgios FENGOS[1] and Dagmar IBER[1]

Computational Biology Group, ETH Zurich D-BSSE, Mattenstrasse 26, Basel, 4058 Basel, Suisse
geraldine.celliere@gmail.com

Keywords TGF-beta signaling, computational modeling, parameter screening.

The family of TGF-beta ligands is large and its members are involved in many different important signaling processes. These signaling processes strongly differ in type with TGF-beta ligands eliciting sustained, transient, and possibly oscillatory responses. We are interested how this signaling network can exhibit these different behaviors. Several differential equation-based models for the TGF-beta signaling network have been developed [1-3] that focus on different aspects of the TGF-beta signaling network (receptor dynamics, the shuttling between the cytoplasm and the nucleus, and the negative feedback via the I-Smad). We chose the most simple model for further investigation.The response to TGF-beta is a transcriptional activity, here monitored as the nuclear concentration of R-Smad/Co-Smad complexes. Biologically meaningful ranges where determined for each parameter (rate constants) and 10^6 simulations of the response to a step increase in TGF-beta where generated and classified into sustained (14.8%), transient (2.2%), or oscillatory response (306 simulations).

A comparison of parameter sets leading either to transient or sustained responses indicates that a transient response requires a quick import of R-Smad complexes into the nucleus and a strong feedback via I-Smad. Similarly, a comparison between transient and oscillatory response highlights the importance of the binding rate of TGF-beta to its receptor. Only when it is low enough, oscillations can appear. Especially when TGF-beta ligands act as morphogen we expect the response to be proportional to the ligand concentration. To obtain such a behavior both the affinity of TGF-beta for it receptor and the affinity of I-Smad for the receptor (sequestration) must be rather low. The TGF-beta pathway is known to exhibit different behaviors over time and in different cell types. Figure 1 shows that there are parameter ranges for which the response can easily switch between sustained and transient responses when the initial concentrations of R-Smad, Co-Smad or receptor are changed.

We conclude that the TGF-beta networks appears to be designed for great regulatory flexibility and that cellular protein concentrations offer a powerful point of control. Protein concentrations, unlike kinetic parameters, can easily be modified by a cell.

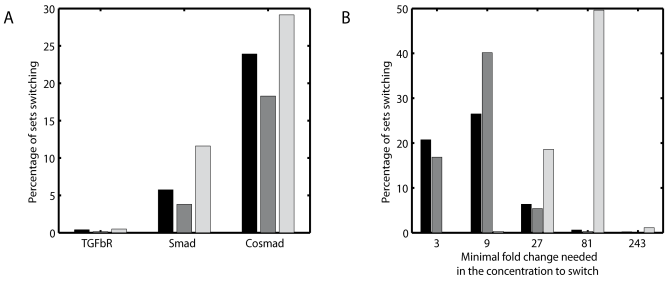


Figure 1. Cellular protein concentrations can define the TGF-beta response. (A) Percentage of parameter sets that can switch their qualitative response when TGFb-Receptor, R-Smad or Co-Smad concentrations are increased or decreased 100 fold. The parameters are chosen in the initial ranges (black), in more restricted ranges (grey) or in even more restricted ranges (light grey). **(B)** Minimal change in TGF-Receptor (black), R-Smad (grey), or Co-Smad (light grey) needed to allow the switch, when parameters are chosen in the most restricted range (corresponding to the light grey column in A).

References

[1] P. Melke, H. Jönsson, E. Pardali, P. ten Dijke, and C. Peterson, A rate equation approach to elucidate the kinetics and robustness of the TGF-beta pathway. *Biophysical journa*l, 91:4368-80, 2006.

[2] B. Schmierer, A.L. Tournier, P. a Bates, and C.S. Hill, Mathematical modeling identifies Smad nucleocytoplasmic shuttling as a dynamic signal-interpreting system. *Proceedings of the National Academy of Sciences of the United States of America*, 105:6608-13, 2008.

[3] J.M.G. Vilar, R. Jansen, and C. Sander, Signal processing in the TGF-beta superfamily ligand-receptor network. *PLoS computational biology*, 2:e3, 2006.

Meta-Prediction of Amyloidogenic Fragments Using Logistic Regression

Anthony TALVAS[1,3], Christian DELAMARCHE[1] and Mathieu EMILY[2,3]

[1] Laboratoire Interactions Cellulaires et Moléculaires, UMR6026 CNRS, Campus Beaulieu, 35042 Rennes, France
anthony.talvas@etudiant.univ-rennes1.fr, christian.delamarche@univ-rennes1.fr
[2] Université Rennes 2, Place du recteur Henri Le Moal, 35043 Rennes, France
Mathieu.Emily@univ-rennes2.fr
[3] IRMAR, UMR6625 CNRS, Campus Beaulieu, 35042 Rennes, France

Keywords Amyloid proteins, Meta-prediction, Neurodegenerative diseases.

Background. Protein aggregation in an insoluble fibrillar form is involved in amyloidosis, a heterogeneous group of diseases, such as Alzheimer's, Huntington's disease, type 2 diabetes...

Short sequences, named "hotspots", play a key role in the ability of proteins to aggregate into ordered fibrillar structures. Over the last few years, various methods have been developed in the literature to detect these "hotspots" in proteins [1,2,3,4]. Existing approaches produce different aggregation indexes and profiles by exploiting several data, including *in vitro* experiments on synthetic peptides, amino acid properties, conformation space and/or 3D-structures. A recent work [5] showed the complementarity of the published methods and highlighted that a combined predictor (or meta-predictor) might improve the predicting preformances.

In this work, we proposed a new meta-predictor of amyloidosis based on a logistic regression. We estimated the best linear combination of four published indexes (Salsa [1], Pafig [2], FoldAmyloid [3], Waltz [4]) as a predictor of the probability for a fragment to be amyloidogenic.

Results. We compared the overall accuracy and Area Under the ROC Curve (AUC) of our meta-predictor to previously proposed methods. The comparison was performed using a recently published validated dataset [4], composed by 116 hexapeptides known to induce amyloidosis and by 162 hexapeptides that do not induce amyloidosis. Table 1 summarizes the performances of the five compared methods: Salsa [1], Pafig [2], FoldAmyloid [3], Waltz [4] and our proposed meta-predictor.

	Salsa	Pafig	FoldAmyloid	Waltz	Our meta-predictor
Accuracy	0.69	0.69	0.62	0.77	**0.84**
AUC	0.79	0.82	0.70	0.85	**0.89**

Table 1. Comparison of four recently published methods with our proposed predictor in terms of accuracy and AUC.

The results show that our meta-predictor outperforms the other methods with a cross-validated accuracy of 0.84 and an AUC of 0.89.

Conclusion and perspectives. Based on complementary methods we propose a meta-predictor for amyloidogenic proteins. Our meta-predictor has been proved to be efficient and fast enough to screen all possible hexapeptides. Such exhaustive search will give insight into potential amino acid sequences associated with amyloidosis, providing new perspectives in the understanding of neurodegenerative diseases.

References

[1] S. Zibaee, O.S. Makin, M. Goedert and L.C. Serpell, A simple algorithm locates beta-strands in the amyloid fibril core of alpha-synuclein, Abeta, and tau using the amino acid sequence alone. *Protein Sci.* 16:906-918, 2007.

[2] J. Tian, N. Wu, J. Guo and Y. Fan, Prediction of amyloid fibril-forming segments based on a support vector machine. *BMC Bioinformatics* 30;10, 2009

[3] S.O. Garbuzynskiy, M.Y. Lobanov and O.V. Galzitskaya, FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence. *Bioinformatics* 26:326-332, 2010.

[4] S. Maurer-Stroh, M. Debulpaep, N. Kuemmerer, M. Lopez de la Paz, I.C. Martins, J. Reumers, K.L. Morris, A. Copland, L. Serpell, L. Serrano, J.W. Schymkowitz and F. Rousseau, Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nat Methods.* 7:237-242, 2010.

[5] A. Le Béchec, A. Talvas, E. Rio, M. Emily, C. Garnier and C. Delamarche, A large scale comparison of predicted amyloigogenic regions using several published methods. *Conference Jacques Monod*, 2010.

Modeling gamma-Cytokine Signalisation from Molecule to Cell

Pascal BOCHET[1], Blanche TAMARIT[1], Louis JONES[2] and Thierry ROSE[1]

[1] Institut Pasteur, Unité d'Immunogénétique Cellulaire, 25 rue du Dr Roux, 75724, Paris, Cedex 15, France
[2] Institut Pasteur, Pôle Informatique, Logiciels et Bases de données, 28 rue du Dr Roux, 75724, Paris, Cedex 15, France
{pascal.bochet,blanche.tamarit,louis.jones,thierry.rose}@pasteur.fr

Keywords receptor, signaling, modeling, protein complex, lipid raft, cytoskeleton.

Gamma-cytokines (interleukin- (IL-) 2, 4, 7, 9, 15 and 21) regulate the differentiation, the homeostasis and the activation of T lymphocytes. These IL bind at subnanomolar concentration to their respective receptors and signal through common main signaling pathways but play markedly divergent roles in lymphoid biology *in vivo*.

These signaling events take place within subcellular compartments where signaling complex associated to the receptor are formed. Within these domains the diffusion of signal molecules is restricted, increasing local concentrations and modulating response times. In addition active transport along components of the cytoskeleton replaces passive diffusion and accelerate the response further. Finally slow cascades of reactions in solution are replaced by an efficient network of surface-bound transport and reactions on the complexes.

We study the formation of the complexes induced by IL-7 on human CD4 T lymphocytes: their composition is analyzed by mass spectrometry and immunoprinting, their size and architecture by fluorescence correlated spectroscopy and transmission electron microscopy, their structure by scanning electron microscopy (EM) and cryoEM, their integrative function by raster-imaging correlated spectroscopy (RICS) and particle tracking from confocal and STED microscopy.

We use these data to reconstruct single particles and simulate the dynamics of their distribution and their function in an entire cell. We use these time-resolved simulations to test hypothesis of signal transduction mechanisms by comparing virtual and experimentally observed kinetics. So far we have focused our simulation on the effect on signaling rates of receptor localization in rafts and of microtubule transport of signaling intermediates.

We will show how we combine data from biochemical and biophysical measurements together with simple modelisation of essential features. In human primary T4 lymphocytes this will allow us to assess the influence of the compartimentalisation and the formation of complexes on the time course and the dynamics of the response to IL.

References

[1] T. Rose, A. Pillet, V. Lavergne, B. Tamarit, P. Lenormand, J.C. Rousselle, A. Namane and J. Theze, Interleukin-7 compartmentalizes its receptor signaling complex to initiate CD4 T lymphocyte response.*J Biol Chem*, 285:14898-14908, 2010.

Cross-Species Metabolic Pathways Comparison – Focus on Mouse, Human and Chicken Lipid Metabolism

Charles Bettembourg[1-2], Christian Diot[2] and Olivier Dameron[1]

[1] Modélisation conceptuelle des connaissances biomédicales, U936 INSERM, 2 av. L. Bernard, 35000 Rennes, France
{charles.bettembourg, olivier.dameron}@univ-rennes1.fr

[2] Génétique Animale, UMR598 INRA, 65 rue de St Brieuc, 35000 Rennes, France
christian.diot@rennes.inra.fr

Keywords pathways, semantic similarity, ontology, data-mining, lipids, cross-species.

Lipid assimilation in animals is a major agronomic challenge. Furthermore, the human associated lipid diseases (i.e. obesity) is an important public health concern in industrialized countries. Lipid metabolism relies on some genetic factors which are mostly conserved among species. Thus, it is necessary to qualify and to quantify the differences and the similarities to compare the lipid metabolism between species.

There is a need for an automatic method of metabolic pathways comparison between species. In a first step, we compare the structure of the pathways to identify common and species-specific reactions. This automatic comparison is based on data from the main pathways databases (Reactome, KEGG, BioCyc, WikiPathway...). Indeed, the low level of consistency, comprehensiveness and compatibility between these databases [1] forces us to use them all. Next, we quantify the semantic similarities and differences using the Gene Ontology annotations of the genes products present in each step. Gene Ontology annotations are compared using a semantic similarity measure according to the Gene Ontology rules [2] by extending Wang's method [3]. A thesis work has started in oct. 2010 to develop this method. Further steps will address generalization to the comparison for pathways and validation.

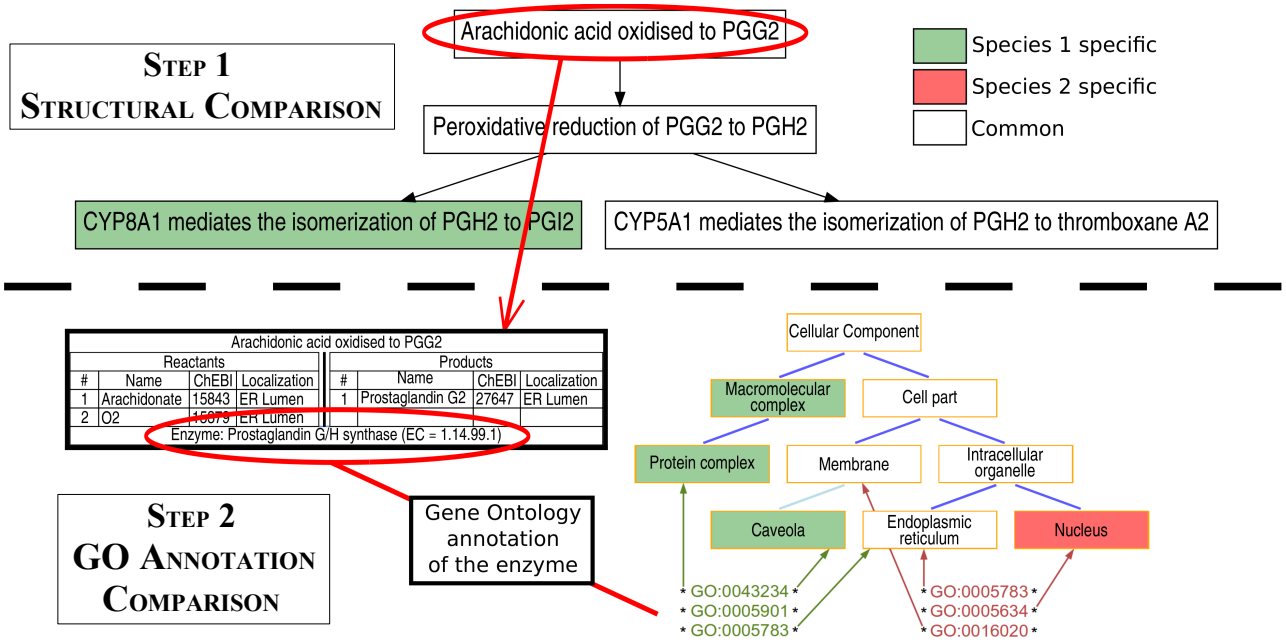


Figure 1. Step 1: Identification of common and species-specific steps with different colors. Step 2: Computation of similarities and specificities using Gene ontology annotations of the genes products.

[1] D. Soh, D. Dong, Y. Guo and L. Wong, Consistency, comprehensiveness, and compatibility of pathway databases. *BMC Bioinformatics*, 11:449, 2010.

[2] S.Y. Rhee, V. Wood, K. Dolinski and S.F. Draghici, Use and misuse of the gene ontology annotations. *Nat Rev Genet.,* 9(7):509-15, 2008.

[3] J.Z. Wang, Z. Du, R. Payattakool, P.S. Yu and C.-F. Chen, A new method to measure the semantic similarity of GO terms. *Bioinformatics*, 23:1274-1281, 2007.

FragMixer: A Modular Framework for (Phospho)Peptide Identification from Multiple MS/MS Fragmentation Modes

Véronique HOURDEL[1], Mathias VANDENBOGAERT[2], Olivia JARDIN-MATHE[2], Jean BIGEARD[3], Delphine PFLIEGER[4] and Benno SCHWIKOWSKI[2]

[1] Institut Pasteur, Plate-forme de Protéomique, 28 rue du Docteur Roux, 75015 Paris.
{veronique.hourdel, mathias.vandenbogaert, benno.schwikowski} @pasteur.fr
[2] Institut Pasteur, Laboratoire de Biologie Systémique, CNRS URA 2171, 25 rue du Docteur Roux, 75015 Paris.
[3] URGV Plant Genomics, INRA/CNRS/Université d'Evry Val d'Essonne, rue Gaston Crémieux, 91057 Evry.
[4] Laboratoire Analyse et Modélisation pour la Biologie et l'Environnement, UMR CNRS 8587
Université d'Evry Val d'Essonne 91025 Evry.
delphine.pflieger@univ-evry.fr, jean.bigeard@evry.inra.fr

Keywords Bioinformatics, LC-MS/MS, Software pipeline, Fragmentation mode, Phosphopeptides.

Post-translational modification (PTM) of proteins is a key mechanism for the regulation of cellular processes. In any cell, a fraction of the expressed proteins is likely to be subject to PTMs that modify their activity, subcellular localization, stability, etc… In this context, the most frequently studied PTMs are phosphorylations. Liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS) is a powerful approach for the analysis of complex protein mixtures, allowing identification of proteins together with their possible PTMs [1]. The sensitivity and the speed with which mass spectrometers can characterize the content of a protein sample have significantly improved over the past years, allowing more and more spectra to be recorded and identified. Recent mass spectrometers like the LTQ-Orbitrap also allow combining several fragmentation modes (e.g., MS2, MSA, ETD, HCD…) for fine-tuning the identification process [2]. Besides, the importance of robust statistical data analysis of MS/MS spectra has become evident. Searches in decoy databases and estimation of the FDR have become standard elements of the validation process. Yet, reliable identification of PTMs and PTM sites by database matching algorithms remains a difficult task. When studying complex phosphopeptide mixtures, typically enriched on IMAC or TiO_2 resins, it remains common practice to manually validate the putatively identified phosphorylated sequences. The main obstacle for computational approaches currently lies in the automatic localization of the exact phosphorylation site within the sequence. Because different fragmentation modes can provide complementary information, combining two different scans on each phosphopeptide may improve its identification. We developed FragMixer to help users validate the MS/MS data obtained on phosphopeptide samples analyzed by acquiring one or two fragmentation scans on every precursor. The pipeline starts with the output of the commonly used search engine Mascot [3]. Peptides are filtered by specifying permissible ranges on different scores (Mascot identity or homology score, a user-defined arbitrary score or a score threshold automatically determined from the specified FDR). Phosphosite localization is predicted using the Mascot Delta-Score (MD-Score) [4] which has recently been shown to be a reliable indicator of correct/uncertain phosphosite localization. Taking as input the Mascot results obtained with one MS fragmentation mode (MS2-only, MSA-only) or with the combination of two modes (MS2/MSA or MSA/ETD), the pipeline incorporates a number of decision rules to automatically classify the identified (phospho)peptides and position the phosphorylations at one precise or several putative sites. We designed the tool with the possibility for manual validation. In addition, statistical measures allow the end user to assess the overall effectiveness of different fragmentation modes in terms of the number of peptide and protein identifications.

[1] J Villén, SA Beausoleil and SP Gygi, Evaluation of the utility of neutral-loss-dependent MS3 strategies in large-scale phosphorylation analysis. *Proteomics*, Nov;8(21):4444-52, 2008.

[2] PJ Ulintz, AK Yocum, B Bodenmiller, R Aebersold, PC Andrews and AI Nesvizhskii, Comparison of MS(2)-only, MSA, and MS(2)/MS(3) methodologies for phosphopeptide identification. *J Proteome Res*, Feb;8(2):887-99, 2009.

[3] DN Perkins, Pappin DJ, Creasy DM and Cottrell JS, Mascot: Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis,* Dec;20(18):3551-67*,* 1999.

[4] MM Savitski, S Lemeer, M Boesche, M Lang, T Mathieson, M Bantscheff and B Kuster, Confident phosphorylation localization site using the Mascot Delta Score. *Mol Cell Proteomics*, Feb;10(2):3830(1-12), 2011.

Patho-Genes.org : Collecte et Analyse des Amorces de PCR utilisées pour la Détection des Micro-organismes Pathogènes

Julien GARDÈS[1], Dipankar BACHAR[1] et Richard CHRISTEN[1]

[1] Université de Nice & CNRS UMR6543. Centre de Biochimie, Parc Valrose, 06108, Nice. France
julien.gardes@unice.fr

Mots-clés Amorces PCR, PCR, détection, pathogènes, site web.

La détection d'organismes pathogènes et le diagnostic très précoce (avant l'apparition de symptômes) font partie des enjeux de la biologie actuelle. Pour cela, le recours à la biologie moléculaire, en particulier la PCR, permet d'analyser et de détecter des gènes cibles. Cependant la sensibilité et la spécificité de la réaction PCR sont tributaires de la bonne conception des amorces. On admet généralement que ces amorces sont bonnes si (1) leur température d'hybridation (Tm) est supérieure à 55°C, (2) qu'elles sont spécifiques du gène et de l'espèce donnés, et (3) qu'elles s'hybrident à tous les allèles connus du gène. Il est donc essentiel d'avoir accès à toutes les séquences du gène et toutes les amorces PCR déjà publiées, ainsi qu'au maximum d'informations relatives à ces deux types de données (annotations fonctionnelles, ontologie, Tm, bibliographie...). C'est dans cette perspective que nous avons développé une procédure semi-automatique permettant de collecter, trier et organiser ces données au sein d'un site internet : www.patho-genes.org.

Nous avons déjà appliqué cette procédure sur la totalité des gènes annotés des principales bactéries d'intérêt biodéfense. Pour commencer, chaque gène est trié par ordre décroissant du nombre de séquences par simple comptage des noms de gène présent dans les entrées EMBL. Puis, pour chaque gène, une recherche par similarité et par mots-clefs des séquences nucléiques est effectuée. Les articles pertinents sont collectés automatiquement au format pdf, pour permettre d'extraire les amorces PCR publiées puis de vérifier leur efficacité sur la totalité des allèles du gène considéré. En parallèle, la méthode récupère, pour chaque gène, toutes les informations associées comme le nom des souches bactériennes, des liens vers les bases de données d'ontologie et des annotations fonctionnelles.

Pour une meilleure lisibilité, nous avons uniformisé les annotations des gènes : chaque gène est défini par un seul nom de gène et un seul nom de protéine. Cette nomenclature est conservée entre les espèces de telle manière qu'un même gène porte le même nom quelle que soit l'espèce (par exemple : gyrA/DNA gyrase A). Les séquences nucléiques des gènes peuvent être téléchargées au format fasta, afin d'être directement utilisables dans les logiciels de conception d'amorces. Il est aussi possible de rechercher un gène à partir d'amorces grâce à un système de requête en ligne. Le logiciel utilisé pour cela a été développé au laboratoire et permet l'utilisation du code IUPAC pour les positions dégénérées (contrairement à BLAST par exemple).

Ces travaux ont permis de montrer dans le cas des gènes de pathogénicité de Vibrio cholerae que (1) seulement un tiers des amorces PCR publiées sont "bonnes" selon les critères mentionnés ci-dessus pour la détection, et (2) que la date de publication et le nombre de citations d'une amorce ne sont pas des facteurs permettant d'estimer leur qualité.

Dans le temps imparti avant la soutenance de ma thèse, nous voulons étendre les ressources de notre site internet à l'ensemble des bactéries biodéfenses de classe B et C, afin de fournir aux équipes de biologistes de la DGA un support d'information complet pour mener à bien et faciliter leurs recherches.

Remerciements

Ces travaux ont bénéficié d'une bourse de thèse financée par la Direction Générale de l'Armement (DGA). Nous tenons à remercier M. Gilles Vergnaud et M. Vincent Ramisse de la DGA, ainsi que M. Manfred Höfle du Helmholtz Centre for Infection Research en Allemagne et Mme Carla Pruzzo de l'Université de Gênes en Italie.

YeastIP: a Database for Identification and Phylogeny of Hemiascomycetous Yeasts

Stéphanie WEISS[1, 2], Franck SAMSON[3], David NAVARRO[4] and Serge CASAREGOLA[1, 2]

[1] INRA, UMR1319 Micalis, CIRM-Levures, F-78850 Thiverval-Grignon, France

[2] AgroParisTech, UMR Micalis, F-78850 Thiverval-Grignon, France

{stephanie.weiss, serge.casaregola}@grignon.inra.fr

[3] INRA, UR 1077 Mathématique Informatique et Génome (MIG), Domaine de Vilvert, F-78352 Jouy-en-Josas Cedex, France

franck.samson@jouy.inra.fr

[4] INRA, UMR-1163 Biotechnologie des Champignons Filamenteux, ESIL, 13288 Marseille, France

david.navarro@esil.univmed.fr

Keywords Taxonomy, phylogeny, hemiascomycetous yeasts, nucleotidic sequences database.

The taxonomy of hemiascomycetous yeasts has greatly evolved in the genomic era. Genomic studies yielded large amount of sequences, which were subsequently used to improve yeast phylogeny with multigenic analyses. This led and is still leading to frequent species name changes. Currently, information on yeast taxonomy associated to sequences can be found in various databases, which are not up to date or inexact. Finally, sequences of interest for taxonomy are diluted in environmental and typing sequences and are difficult to retrieve. Within the frame of the FP7 european project EMbaRC (European Consortium of Microbial Resources Centres), the construction of a relevant database with verified nucleotidic sequences and an interface allowing the users to obtain correct identification, taxonomy and phylogeny of yeasts species, was undertaken.

The YeastIP database was constructed using MySQL and HTML/PHP/JavaScript. Before their introduction in the database, sequences from generalist databases like GenBank/EMBL were screened through an expert selection: priority was given to type strains and to the most relevant markers for phylogeny. Sequence quality was also checked: length of the sequence and the presence of undefined nucleotides were verified. At the moment, YeastIP contains more than 3500 sequences, representing more than 750 species and a choice of up to 10 markers per species, when available in generalist databases.

The interface was developed to propose a maximum of relevant information and choices to search the database. First, an authentication tool using the Blast program was implemented to allow users to compare their sequences of interest to the YeastIP database. Second, a tool allows the database search for sequences *via* taxonomy or keyword. A file containing all relevant information for each sequence is available i.e. species current name and synonyms, origin of the sequence. Both tools allow the retrieval of selected sequences in Fasta file, or the display of a table showing the available markers for each selected sequences and species or group of species. This table will guide the users for the choice of markers to sequence in order to perform phylogenic analysis. In this step, strains and markers can be selected to construct a concatenation file, retrievable in Fasta format or viewable as a phylogenetic tree via the Phylogeny.fr website. Users can also add their own sequences to the file and obtain phylogeny with their sequences and strains of interest. The YeastIP database will be updated in collaboration with the CBS, the major collection of yeast strains in Europe.

YeastIP is a tool, which facilitates the retrieval of taxonomic information and guides the users to obtain a robust phylogenic analysis. This work is linked to another part of the EMbaRC project, consisting in producing new sequences, which will feed the YeastIP database. YeastIP put the emphasis on multigenic analysis to improve good practice in hemiascomycetous yeasts phylogeny, and could be extended in a future work to all fungal species.

IMGT/HighV-QUEST 2011

Eltaf ALAMYAR[1], Véronique GIUDICELLI[1], Patrice DUROUX[1] and Marie-Paule LEFRANC[1]

[1] IGH, UPR CNRS 1142, 141, rue de la Cardonille, 34396, Montpellier, Cedex 05, France
{eltaf.alamyar,giudicel,patrice.duroux,marie-paule.lefranc}@igh.cnrs.fr

Keywords Next-Generation Sequencing, NGS, immunoglobulin, antibody, T cell receptor, expressed immune repertoires, IMGT-ONTOLOGY.

The analysis of expressed repertoires of antigen receptors - immunoglobulins (IG) or antibodies and T cell receptors (TR) - represents a huge challenge for the study of the adaptive immune response in normal and disease-related situations, such as viral infections. To answer that need, IMGT®, the international ImMunoGeneTics information system® (http://www.imgt.org) has developed IMGT/HighV-QUEST [1]. IMGT/HighV-QUEST is devoted to the analysis of large repertoires of IG and TR sequences that result from Next Generation Sequencing technologies. IMGT/HighV-QUEST, a high throughput version of IMGT/V-QUEST, analyses up to 150.000 sequences per run. It identifies the IG and TR variable (V), diversity (D) and joining (J) genes and alleles by alignment with the germline IG and TR gene and allele sequences of the IMGT reference directory. It describes the V-REGION mutations and identifies the hot spot positions in the closest germline V gene. It integrates IMGT/JunctionAnalysis for a detailed analysis of the V-J and V-D-J junctions, and IMGT/Automat for a full V-J- and V-D-J-REGION annotation. The analysis is based on the IMGT-ONTOLOGY concepts of description, classification and numerotation.

IMGT/HighV-QUEST uses two different systems of HPC resources at CINES, and a local computational server, in order to perform the analysis of submitted sequences by a standalone version of IMGT/V-QUEST. Since the management of many analyses of thousands of sequences is a challenging task, IMGT/HighV-QUEST manipulates a local database for the local analysis queue, and in order to manage the jobs, the tasks are split into three independent layers. The web service layer is responsible for providing user interaction facilities and adding new analyses in the local queue. The scheduled tasks layer (also called background layer) includes all core logical functionalities of IMGT/HighV-QUEST. Background operations such as selection of a resource, dispatching of analyses, monitoring the running jobs, preparation of results of the completed analyses are performed in this layer. Finally the computational resources layer is where the real analysis of user sequences is performed. The analysis results of IMGT/HighV-QUEST comprise a set of text files which include 11 files in CSV format equivalent to the eleven sheets of the 'Excel files' of IMGT/V-QUEST and, for each analysed sequence, the 'Detailed view' that allows one to visualize the individual detailed results. These result files are archived in a single ZIP file that is downloaded by the user.

Since its availability in October 2010, more than 41 million sequences have been submitted and 118 users have registered to IMGT/HighV-QUEST (11/05/11). The jobs required 17,000 computational hours of resources and generated about one terabyte of results data. More than three quarters of the sequences were submitted by users from USA, the others being submitted by users from EU for most, but also from China, Japan, Australia, Canada, Korea and Venezuela. New functionalities have been developed that comprise the introduction of statistical analysis on the results of the batch to help the user in estimating the reliability of the results. Statistics are performed on results selected as ‘1 copy’ (redundancies are enregistered but not treated), and with quality criteria (identification of a single gene/allele, known functionality, REGION length, absence of IMGT/V-QUEST warnings regarding the CDR1-IMGT and CDR2-IMGT lengths or the percentage of identity). These statistics include the frequency of gene expression and of CDR3-IMGT length; they also report the number of identical CDR3-IMGT sequences and the number of sets of CDR3-IMGT with identical nucleotides (nt) and amino acid (AA) sequences. These functionalities, which have been set up in a first step for the human TR, are particularly useful to evaluate the significance of the results of a batch.

Acknowledgments: this work was granted access to the HPC resources of CINES under the allocation 2010-036029 made by GENCI (Grand Equipement National de Calcul Intensif).

[1] E. Alamyar, V. Giudicelli, P. Duroux and M.-P. Lefranc, IMGT/HighV-QUEST: A high-throughput system and web portal for the analysis of rearranged nucleotide sequences of antigen receptors - High-throughput version of IMGT/V-QUEST, *Proceedings of the 11th Journées ouvertes en Biologie, Informatique et Mathématiques (JOBIM)*, Montpellier, P27 pp. 156, 2010.

MODIM : Model-Driven Data Integration for Mining

Birama NDIAYE[1], Emmanuel BRESSO[1, 2], Malika SMAÏL-TABBONE[1], Michel SOUCHET[2] and Marie-Dominique DEVIGNES[1]

[1] LORIA, UMR7503 CNRS-Nancy Université et INRIA Nancy Grand-Est, Campus Scientifique, BP239, 54503, Vandoeuvre les Nancy, Cedex, France

{birama.ndiaye, bressoem, malika, michel.souchet, devignes}@loria.fr

[2] Harmonic Pharma, Espace Transfert INRIA, 615 rue du Jardin Botanique, 54600 Villers les Nancy, France

Keywords data integration, relational databases, distributed data sources, data mining.

The success of a knowledge discovery process relies on expert-guided iterative design and selection of datasets to be mined. In the post-genomic era this crucial issue is made difficult due to the multiplicity of data sources, the volume of the data they contain, their heterogeneity and frequent update. Integrated information systems exist to select datasets from most public databases (*e.g.*, SRS, BioMart, BioWarehouse). However it is frequently desirable to build or modify datasets according to a data model representing the domain knowledge and reflecting a particular point of view [1]. A dedicated integrated database constructed on such a data model can facilitate the knowledge discovery process because it becomes easier to iteratively design and obtain appropriate datasets for mining.

The "Model-driven Data Integration for Mining" (MODIM) system is a generic approach for automating data collection and integration according to a dedicated relational data model in order to conduct data mining experiments. In this approach, users' requirements and expertise play the central role while fastidious repetitive tasks are automated.

The MODIM architecture relies on three interactive modules (i) database, (ii) task configuration, and (iii) task enactment. Once a relational data model is available for a given data mining problem, the corresponding database can be created through the MODIM database module. A set of tasks is then configured with the MODIM task configuration module in order to populate the database. Each task is specific of an input query (for example a protein identifier in a database) and is composed of subtasks, one for each data source visited (most often represented by a URL query). Configuring a subtask schematically consists in describing how to recognize, in the result page, which item to collect and where to store this output data in the database. Once validated thanks to several testing functionalities, the MODIM tasks are stored and can be edited later. Finally data collection and integration can be launched through the MODIM task enactment module which can upload a file with all desired input queries. It should be noted that the concept of subtask in MODIM is not limited to the querying of distant web databases. For example, transformation scripts can also be invoked and executed if their output is needed in the database for the mining purpose.

A test example will be presented in the domain of drug discovery aimed at analyzing drug side effects. Drug properties pertaining from the DrugBank and SIDER databases, MedDRA vocabulary, and MeSH thesaurus are collected and integrated in a consistent database using the MODIM software. The conduction and interpretation of iterative data mining experiments are facilitated thanks to this dedicated database.

The MODIM approach is related to various efforts made today for offering model-driven solutions for biological data integration [2, 3]. MODIM is a light domain-independent system which should simplify analysis workflows in biology.

[1] S. Yilmaz, P. Jonveaux, C. Bicep, L. Pierron, M. Smaïl-Tabbone and M.D. Devignes, Gene-disease relationship discovery based on model-driven data integration and database view definition. *Bioinformatics*, 25:230-6, 2009.

[2] D. Smedley, M.A. Swertz, K. Wolstencroft, G. Proctor, M. Zouberakis, J. Bard, J.M. Hancock and P. Schofield, Solutions for data integration in functional genomics: a critical assessment and case study. *Brief Bioinform.* , 9:532-44, 2008.

[3] J.A. Vizcaíno, F. Reisinger, R. Côté and L. Martens, PRIDE and "Database on Demand" as valuable tools for computational proteomics. *Methods Mol Biol.*, 696:93-105, 2011.

SM2PH-kb: Data Warehouse Intelligence for the Integrated Study of Human Structural Mutation to Phenotypes Relationships

Tien-Dao LUU[1], Ngoc-Hoan NGUYEN[1], Jean MULLER[1,2], Luc MOULINIER [1] and Olivier POCH[1]

[1] Departement de Biologie et Génomique Structurales, Institut de Génétique et de Biologie Moléculaire et Cellulaire, UMR7104, 1 rue Laurent Fries, 67404 Illkirch Cedex, France

{luudao, nguyen, jmuller, moumou, olivier.poch}@igbmc.fr

[2] Laboratoire de Diagnostic Génétique, Centre Hospitalier Universitaire Strasbourg, Nouvel Hôpital Civil, 1 place de l'hôpital, 67000 Strasbourg, France

Keywords genotype-phenotype relationship, Single Nucleotide Polymorphisms, human genetic disease, mutation impact, knowledge discovery, Inductive Logic Programming.

Understanding the effects of genetic variation on the phenotype of an individual is a major goal of biomedical research, especially for the development of diagnostics and effective therapeutic solutions.

Here we present SM2PH-kb (from Structural Mutation to Pathology Phenotypes in Human), upgraded from SM2PH-db [1], a second generation of our data warehouse intelligence designed to investigate structural and functional impacts of missense mutations and their phenotypic effects in the context of human genetic diseases.

Our data warehouse mines and regroups up-to-date heterogeneous interconnected information, including data retrieved from biological databases (GenBank, RefSeq, Uniprot, PDB, Gene Ontology, NCBI Taxonomy, Interpro, ELM), 'omics' data (transcriptomics from GEO or from in house-developed expression data relational databases (GxDB), interactomics from the STRING database), variant data (dbSNP, Uniprot), disease data (GeneCard, OMIM), pathways (KEGG) and data generated from a Sequence-Structure-Evolution Inference in Systems-based approach, such as multiple alignments, 3D structural models and multidimensional (physicochemical, functional, structural and evolutionary) characterizations of mutations. At time of writing (Avril 2011), SM2PH-kb holds a total of 2,460 human proteins related to monogenic diseases with 62,454 missense mutations, among which 26,373 are considered as disease-causing and 36,081 as nonpathogenic. The data warehouse is automatically updated every 2 months thanks to the Decrypthon computation grid and is publicly accessible online at http://decrypthon.igbmc.fr/sm2ph.

We applied Inductive Logic Programming to automatically extract knowledge about deleterious/neutral mutations. This will guide human experts to improve our understanding of the relationships between physico-chemical and evolutionary features and deleterious mutations. SM2PH-kb provides a web service to predict the pathogenicity of submitted missense variations (2).

SM2PH-kb has been organized to give the user a robust infrastructure associated with interactive analysis tools supporting in-depth study and interpretation of the molecular consequences of mutations, with the more long-term goal of elucidating the chain of events leading from a molecular defect to its pathology.

[1] A. Friedrich, N. Garnier, N. Gagnière, H. Nguyen, L. P. Albou, V. Biancalana, E. Bettler, G. Deléage, O. Lecompte, J. Muller, D. Moras, J. L. Mandel, T. Toursel, L. Moulinier and O. Poch, SM2PH-db: an interactive system for the integrated analysis of phenotypic consequences of missense mutations in proteins involved in human genetic diseases, *Human Mutation*, vol. 31, pp. 127-135, 2010.

[2] D. Luu, H. Nguyen, A. Friedrich, J. Muller, L. Moulinier and O. Poch, Extracting Knowledge from a Mutation Database Related to Human Monogenic Disease Using Inductive Logic Programming. *Proceedings of the International Conference on Bioscience, Biochemistry and Bioinformatics*, Singapore, IEEE Catalog Number: CFP1134M-PRT. ISBN: 978-1-4244-9388-3, 2011.

A Three-dimensional Modeling Software for Group-wise Data Integration and Analysis of Spatial Distributions in Biological Imaging

Eric BIOT[1], Jasmine BURGUET[1] and Philippe ANDREY[1,2]

[1] Institut Jean-Pierre Bourgin, UMR1318 INRA-AgroParisTech,
INRA, Centre de Versailles-Grignon, Route de Saint-Cyr (RD 10), 78026 Versailles, France
{eric.biot, jasmine.burguet, philippe.andrey}@versailles.inra.fr
[2] UPMC, Univ Paris 06, France

Keywords 3D imaging, 3D modeling, spatial normalization, density estimation.

We describe and make a practical demonstration of Free-D, a software we are developing to integrate imaging data acquired on multiple individuals and to quantify biological organizations at the group level.

Protein intracellular localization and trafficking is essential to many biological processes in living cells. Cellular imaging techniques such as confocal microscopy are tools of choice to visualize, within the three-dimensional cellular space, the distribution of vesicles transporting labeled proteins. However, comparing spatial distributions of vesicle positions in different physiological, biological and experimental conditions requires integrating imaging data from different cells into common representations amenable to quantitative spatial analyses. This represents a challenging task because individual data are affected by biological and experimental variability in both cell size and shape.

To address this issue, we have adopted a spatial modeling strategy and have developed a collection of algorithms for reconstructing graphical 3D models from image stacks [3,5], for registering and averaging individual models [6], for spatially normalizing individual data into average models [1] and for generating 3D statistical maps of point densities [4].

To make these tools practical and available for biologists, we have developed Free-D, an integrated, multi-platform, freely distributed 3D modeling software. A first version of the software, offering tools for image segmentation, registration, and model reconstruction, has been previously presented and applied to neuroimaging data [2]. We describe here the latest release, which includes dedicated tools for processing multiple stacks, for registering and averaging 3D models, and for the non-linear spatial normalization of individual data and their integration into average, standardized 3D models. This version additionally offers tools for the quantitative analysis of 3D models and spatial statistical tools for building representative density maps of point populations. More generally, the software can be used for the group-wise analysis of biological structures at the histological, cellular or subcellular scales. We believe this software should be of interest for studies in quantitative biological imaging and, more generally, for large-scale approaches in Systems Biology.

References

[1] P. Andrey, E. Maschino, and Y. Maurin, Spatial normalisation of three-dimensional neuroanatomical models using shape registration, averaging, and warping. In *Fifth IEEE International Symposium on Biomedical Imaging (ISBI'08): From Nano to Macro*, pages 1183–1186, Paris, May 14-17 2008.

[2] P. Andrey and Y. Maurin, *Free-D*: an integrated environment for three-dimensional reconstruction from serial sections. *J. Neurosci. Methods*, 145(1–2):233–244, 2005.

[3] E. Biot, E. Crowell, H. Höfte, Y. Maurin, S. Vernhettes, and P. Andrey, A new filter for spot extraction in N-dimensional biological imaging. In *Fifth IEEE International Symposium on Biomedical Imaging (ISBI'08): From Nano to Macro*, pages 975–978, Paris, May 14-17 2008.

[4] J. Burguet, P. Andrey, O. Rampin, and Y. Maurin, Three-dimensional statistical modeling of neuronal populations: illustration with spatial localization of supernumerary neurons in the locus coeruleus of quaking mutant mice. *J. Comp. Neurol.*, 513(5):483–495, 2009.

[5] J. Burguet, P. Mailly, Y. Maurin, and P. Andrey, Reconstructing the three-dimensional surface of a branching and merging biological structure from a stack of coplanar contours. In *Eighth IEEE International Symposium on Biomedical Imaging (ISBI 2011): From Nano to Macro*, 2011. To appear.

[6] E. Maschino, Y. Maurin, and P. Andrey, Joint registration and averaging of multiple 3D anatomical surface models. *Comput. Vis. Image Underst.*, 101(1):16–30, 2006.

Eoulsan: a Cloud Computing-Based Framework facilitating High Throughput Sequencing Analyses

Laurent JOURDREN[1], Maria BERNARD[1], Marie-Agnès DILLIES[2] and Stéphane LE CROM[1]

[1] École normale supérieure, Institut de biologie de l'ENS, UMR8197 CNRS, U1024 INSERM,
46 rue d'Ulm, 75005, Paris, France
{laurent.jourdren, maria.bernard, stephane.lecrom}@ens.fr

[2] Plate-forme Transcriptome et Epigénome, Institut Pasteur
28 rue du Dr. Roux, 75724 Paris cedex 15, France
marie-agnes.dillies@pasteur.fr

Keywords High throughput sequencing, RNA-Seq, MapReduce, Cloud computing.

A new generation of high throughput sequencing [1] machines has been released in the past few months, allowing this technology to spread over the scientific community and exponentially increasing the amount of sequence data, making the analysis process to be the bottleneck of any sequencing experiment [2]. As an answer to the need for an adapted but expensive hardware, the cloud computing approach [3] is an economic and scalable solution compared to large computer infrastructure investment that will be used occasionally.

Here we present Eoulsan, an open source framework based on the Hadoop implementation of the MapReduce algorithm and dedicated to high throughput sequencing data analysis on distant computers. With Eoulsan users can easily set up a cloud computing cluster, automate the complete analysis of several samples at once and select among various analysis solutions available. We first implemented Eoulsan to work on the differential analysis of transcript expression. This workflow runs in 5 steps: quality control filtering, mapping, expression calculation, normalization and differential analysis. All information available on the experimental design is gathered in one text file inspired from the one of the limma R package for microarray analysis [4]. All the options needed to run the workflow are gathered in one XML file that allows for the usage of plugin programmed by external developers. We demonstrate the modularity and scalability of this workflow by performing mapping of different RNA-Seq experiments using several softwares: BWA, Bowtie, SOAP2. We assess the analysis duration and cost with various type and number of servers (call instances) using Amazon Web Services (AWS) cloud computing facilities. We show that once a minimal number of instances has been selected, the cost is linear with the number of instances booked. This is achieved through the full parallelization of the mapping and transcript expression estimation steps and allows for the optimization of either speed and cost of the analysis with no risk to fall in a suboptimal configuration in order to speed up the data analysis process. Finally, we show that running times performed with Eoulsan evolve linearly with the increase of the amount of data using from 188 to 752 million total reads.

To conclude, our framework provides from standalone workstation to cloud computing clusters an integrated and flexible solution for high throughput sequencing data analysis from reads alignment to the list of significant differentially expressed transcripts. With its modular structure and parallel data processing, Eoulsan is ready to fulfill the challenges coming from the massive increasing of data amount and the new applications of sequencing technologies.

[1] ER. Mardis, A decade's perspective on DNA sequencing technology. *Nature*, 10:198-203, 2011.

[2] E. Pennisi, Human genome 10th anniversary. Will computers crash genomics? *Science*, 331:666-668, 2011.

[3] EE. Schadt, MD. Linderman, J. Sorenson, L. Lee and GP. Nolan, Computational solutions to large-scale data management and analysis. Nat. Rev. Genet., 11:647-657, 2010.

[4] GK. Smyth, Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol.*, 3, 2004

RENABI GRISBI - Infrastructure Distribuée pour la Bioinformatique

Christophe BLANCHET[1], Clément GAUTHEY[1], Christophe CARON[2], Olivier COLLIN[3], Stéphane DELMOTTE[4], Tiphaine MARTIN[5], Aurélien ROULT[3], Franck SAMSON[6] and Bruno SPATARO[4]

[1] Infrastructure Distribuée pour la Biologie, IDB IBCP FR3302 CNRS, 7 passage du Vercors, 69007, Lyon, France
{christophe.blanchet, clement.gauthey}@ibcp.fr

[2] ABiMS, FR2424 CNRS-UPMC, Station Biologique, Place G. Teissier, 29680, Roscoff, France,
christophe.caron@sb-roscoff.fr

[3] GenOuest, CNRS IRISA - UMR6074, Campus de Beaulieu, 35042 RENNES Cedex - France,
{ocollin, aurelien.roult}@irisa.fr

[4] LBBE, UMR5558 CNRS, 43 bd du 11 novembre 1918, 69622 VILLEURBANNE cedex, France
{stephane.delmotte, bruno.spataro}@univ-lyon1.fr

[5] LaBRI, UMR5800 CNRS, 351 cours de la Libération, 33400, Talence, Cedex, France
tiphaine.martin@labri.fr

[6] MIGALE, Mathématique, Informatique et Génome, INRA, Jouy-en-Josas, France
franck.samson@jouy.inra.fr

Keywords Bioinformatique, Infrastructure distribuée, Calcul scientifique.

La Bioinformatique fait face à un déluge de données souvent complexes, de nature très hétérogène, et provenant de nombreuses sources. En conséquence, l'échelle des analyses s'est déplacée du gène/protéine au génome/protéome complet, d'une voie métabolique à la modélisation des réseaux. Pour analyser ces données, les logiciels existants sont nombreux avec des modes opératoires variables (E/S, mémoire). Ainsi cette masse de données génère un besoin d'une infrastructure informatique distribuée pour répondre aux besoins des biologistes/bioinformaticiens. A propos des données, les besoins concernent l'accès, depuis n'importe quel site, aux banques internationales maintenues à jour, l'accès à un espace personnel et à un espace commun (groupe, projets…). En terme de calcul, les besoins relèvent de la distribution de ces calculs (avec optimisation de l'utilisation des ressources informatiques et rapidité d'exécution), du déploiement des logiciels communément utilisés, ainsi que le développement et le déploiement de workflows et pipelines. Et finalement les interfaces doivent permettre la gestion des accès aux différents ressources, et surtout être faciles à utiliser et puissantes.

L'infrastructure GRISBI (www.grisbio.fr) est une initiative conjointe entre plusieurs centres de Bioinformatique, collaborant dans le cadre du réseau national RENABI (www.renabi.fr) : PRABI, RENABI-GO, RENABI-SO, RENABI-NE et APLIBIO. La couche logicielle de l'infrastructure GRISBI s'appuie sur le logiciel de grille européen gLite (www.glite.org), en collaboration avec l'Institut des Grilles du CNRS, et d'autres développements bioinformatiques nationaux comme BioMaj [1], le gestionnaire de mise à jour des banques de données biologiques. Les ressources disponibles représentent près de 900 processeurs et 26 To au sein d'une infrastructure où l'accès et les mécanismes sont unifiées. L'infrastructure initialement composée de 8 sites, s'est étendue récemment avec l'arrivée de la plateforme RENABI-SO GenoToul. Un des objectif est, à terme, d'accueillir les autres plateformes bioinformatiques nationales qui le souhaitent. Dans ce sens, le comité technique a défini un ensemble de recommandations et de procédures pour l'intégration d'une nouvelle plateforme. Les ressources de l'infrastructure GRISBI ont été également étendue par la participation de 3 mésocentres de calcul pluridisciplinaires (CRI Univ. Lille 1, IPHC Strasbourg et M3PEC Bordeaux), grâce aux mécanismes d'organisation virtuelle de gLite.

L'utilisation de l'infrastructure par la communauté relève de différents domaines comme l'analyse NGS, la génomique comparative, la biologie des systèmes, la prédiction de fonction de protéines ou les interactions moléculaires. Notamment, un pilote de mise en oeuvre d'expériences NGS est en cours avec une collaboration entre certains partenaires du projet GRISBI. Son objectif est d'évaluer les solutions proposées par l'infrastructure GRISBI avec un cas d'usage réaliste sur des données maitrisées: l'analyse du génome complet de la levure *Saccharomyces cerevisiae* avec les logiciels du moment (BWA, Abyss et Ray). L'accès à l'infrastructure est ouvert à l'ensemble de la communauté et passe par la collaboration avec une des plateformes RENABI, qui sert d'interlocuteur et de guide pour l'utilisation des ressources

Remerciements: Delphine Naquin, Christelle Eloto, Pierre Gay, Daniel Jacob, Didier Laborie, Nouredine Melab, Alexis Michon, Frédéric Plewniak, les GIS IBISA et FranceGrilles.

[1] O. Filangi, Y. Beausse, A. Assi, L. Legrand, J.M. Larré, V. Martin, O. Collin, C. Caron, H. Leroy and D. Allouche, BioMAJ: A flexible framework for databanks synchronization and processing. *Bioinformatics*, 24:1823-1825, 2008.

Virtualisation of Bioinformatics Applications on Cloud Infrastructure

Christophe BLANCHET[1] and Charles LOOMIS[2]
[1] Infrastructure Distribuée pour la Biologie, IBCP FR3302 CNRS, 7 passage du Vercors, 69007, Lyon, France
christophe.blanchet@ibcp.fr
[2] LAL, UMR8607 CNRS, Bât. 200 BP 34, lieu, 91898, Orsay, Cedex, France
charles.loomis@lal.fr

Keywords Bioinformatics, Virtualization, Elastic computing, Cloud computing.

Today, the bioinformatics community is facing a deluge of data. Several experimental technologies have been improved in such a way that obtaining data is easy. The challenge is to be able to analyze these data with the relevant applications. For example, sequencing a whole genome has became usual with the new technologies called Next Generation Sequencing (NGS). Many projects are working on the genome sequence of different organisms, thus continuously providing new sequences for analysis. Some bioinformatics algorithms like BLAST, FastA or ClustalW are used intensively for that analysis and usually classified as data-intensive. They are processing gigabytes of data stored in flat-file databases like UNIPROT, EMBL or PDBseq on a shared filesystem. To give an insight to this challenge, we have built two virtual bioinformatics appliances in the context of the StratusLab project (EU-FP7, www.stratuslab.org). They are "Biological databases repository" and "Bioinformatics compute node", and they provide bioinformaticians with a repository appliance maintaining up-to-date international reference databases, then made available through shared filesystem in destination to bioinformatics cloud nodes with pre-installed bioinformatics software.

Bioinformaticians need access from any compute node to international reference databases recording biological resources such as protein or gene sequences and associated data, protein structures or complete genomes. The 2011 edition of the Nucleic Acids Research "Database" issue [1] lists 1330 carefully selected molecular biology databases. We have built a virtual appliance that acts as a proxy between the internet where all the reference databases are published and the cloud instances that will compute the bioinformatics analyses. To import and maintain the required biological databases, we use the BioMaJ system. This virtual machine stores the data in files organized from a root directory `/biodb' that is being exported with NFS protocol in read-only mode to all the bioinformatics computing machines of the cloud.

Distributing the computation is also an important requirement because bioinformatics applications could require very different resources depending on the analysis to perform: multiple alignments of sequences, genome assembling or intensive protein sequence comparison. Biologists and bioinformaticians are combining regularly multiple software packages to analyze their data. They used these software for their intensive processes from Web portals, or with shell commands or scrips written in interpreted languages. Regarding the computations and the virtual machines, the main requirements are related to satisfying the software dependencies and very different behaviors of biological applications in terms of CPU and memory. Some applications only require one CPU but with a lot of memory (96 or 128MB) whereas others require lot of CPUs that are accessed through MPI mechanism. We have built a virtual machine with bioinformatics software pre-installed with the help of a script system, called `bioapps' developed in our team. This tool downloads the application package from the reference site, compiles and installs the binary on the machine. Because bioinformatics applications require access to reference data to process their analyses, this bioinformatics compute appliance is linked to the biological databases repository appliance, and mounts the exported volumes containing the biological databases.

The usage of cloud for Bioinformatics has to be connected with public bioinformatics infrastructures like the French Bioinformatics Network RENABI (www.renabi.fr) and especially its grid infrastructure GRISBI (www.grisbio.fr). The adoption of clouds for bioinformatics applications will be strongly correlated to the capability of cloud infrastructures to provide ease-of-use and access to reference biological data and applications. In that sense, StratusLab is collaborating with RENABI to help solving the requirements from the Bioinformatics community.

[1] M. Y. Galperin and G. R. Cochrane, The 2011 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. Nucl. Acids Res., 39 (suppl 1):D1-6, 2011.

Utilisation d'une Grille de Calcul (GRISBI) pour le Traitement de Données NGS

Florence Maurier[2], Alexis Groppi[2], Tiphaine Martin[1], Aurélien Barré[2], Delphine Naquin[3], Aurélien Roult[3], Clément Gauthey[4] and Christophe Blanchet[4]

[1] LaBRI, UMR5800 CNRS, 351 cours de la Libération, 33400, Talence, Cedex, France
tiphaine.martin@labri.fr

[2] Université de Bordeaux, CBiB, 146 Rue Léo Saignat, Centre de Génomique Fonctionnelle Bordeaux, 33076, Bordeaux, Cedex, France
{alexis.groppi, aurelien.barre, florence.maurier}@u-bordeaux2.fr

[3] EPI Symbiose, INRIA/ IRISA, Campus de Beaulieu, 35042, Rennes, Cedex, France
{delphine.naquin,aurelien.roult}@irisa.fr

[4] Infrastructure Distribuée pour la Biologie, IDB IBCP FR3302 CNRS, 7 passage du Vercors, 69007, Lyon, France
{christophe.blanchet, clement.gauthey}@ibcp.fr²

Keywords NGS, Assemblage *de novo*, Grille de calcul, Calcul intensif, Gestionnaire de workflow.

Les approches de séquençage de nouvelle génération ont permis et permettent désormais d'obtenir très rapidement des séquences de génome entier. Cependant, l'obtention de séquences « de haute qualité », totalement assemblées, sans manque ni ambigüité reste problématique. En effet, dans le cas de séquençage *de novo* l'étape de l'assemblage est très coûteuse en temps de calcul. La bioinformatique est confrontée à une explosion de ces nouvelles données. Le stockage et analyse des données devient difficile sur des serveurs locaux, et il est donc nécessaire de passer à d'autres infrastructures. Les grilles de calculs peuvent permettre de résoudre ce défi. Cependant, ces infrastructures de calculs ne sont pas facilement accessibles aux utilisateurs novices et en particulier aux biologistes.

Pour accompagner et répondre à de telles problématiques de biologie à grande échelle, l'infrastructure distribuée GRISBI (Grille Support pour la Bioinformatique) [1] a été initiée dans le cadre du réseau ReNaBi. L'objectif est de permettre la mutualisation des ressources de plates-formes de bioinformatiques nationales (dont le CBIB et PRABI) pour la communauté bioinformatique nationale et européenne.

Afin d'évaluer la faisabilité et l'apport d'une grille de calcul telle que RENABI GRISBI pour ce type d'analyse, nous avons défini plusieurs cas d'utilisations et testé plusieurs solutions logicielles. Nous avons choisi d'utiliser une source de donnée maitrisées : le génome complet de la levure *Saccharomyces cerevisiae* [2]. A partir de ces séquences chromosomiques nous avons généré des jeux de lectures artificiels simulant des séquençages de type Illumina et de type Roche 454 grâce au logiciel MetaSim [3]. A partir de ces jeux d'essai, nous avons réalisé des alignements sur le génome de référence avec le logiciel Burrows-Wheeler Aligner (BWA) [4] et des assemblages *de novo* avec les logiciels Abyss [5] et Ray [6].

Nos premiers résultats démontrent la faisabilité et l'apport considérable de l'utilisation de cette infrastructure de calcul pour ces traitements. En effet, d'une part les temps de calculs sont réduits par rapport à des serveurs locaux de capacité moyenne, mais il est également possible de lancer en simultané plusieurs analyses en faisant varier les paramètres en vue d'optimisation des résultats dans le cas d'un assemblage *de novo*. La suite de ce travail va consister maintenant à intégrer ces traitements au sein d'un gestionnaire de workflow comme ERGATIS [7]. L'utilisateur pourra alors lancer une chaine de traitement via l'interface web du gestionnaire de workflow qui enverra sur la grille les traitements nécessitant le plus de ressources.

[1] http://www.grisbio.fr
[2] http://db.yeastgenome.org/cgi-bin/seqTools
[3] http://ab.inf.uni-tuebingen.de/software/metasim/
[4] http://bio-bwa.sourceforge.net/
[5] http://www.bcgsc.ca/platform/bioinfo/software/abyss
[6] http://sourceforge.net/apps/mediawiki/denovoassembler/
[7] http://ergatis.sourceforge.net/

Listes et Index

Liste des conférences invitées

Liste des présentations orales

Liste des présentations industrielles

Liste des affiches

Table des matières

– Avant-Propos –

– Contributions –

–Listes et Index–

Index des contributeurs

Journées Ouvertes de Biologie, Informatique et Mathématiques

Institut Pasteur, Paris, 28 juin – 1er juillet 2011