JOBIM 2012

13E JOURNÉES OUVERTES EN BIOLOGIE, INFORMATIQUE ET MATHÉMATIQUES

3-6 JUILLET 2012

RENNES

François Coste
et Denis Tagu

Journées Ouvertes de Biologie, Informatique et Mathématiques

IV+xiv+496 pages.

Impression: 25 juin 2012

Réalisation, conception du logo: Pierre Peterlongo
Mise en page: Pierre Peterlongo, avec le support d'Alban Mancheron
Couverture: Candice Bachelet (Glad)
Éditeur: Inria Rennes - Bretagne Atlantique, Campus universitaire de Beaulieu 35042 Rennes Cedex
ISBN-13 978-2-7261-1301-1

Préface

Depuis 12 ans, JOBIM s'est imposé dans le paysage de la communauté francophone en bioinformatique, biostatistiques et biomathématiques, avec son rendez-vous annuel de plusieurs centaines de personnes. Les disciplines de JOBIM représentent un des meilleurs exemples de réussite d'interfaçage de compétences, de pluridisciplinarité, dans un souci de participer à la compréhension du fonctionnement du vivant, dans sa complexité et diversité structurale, dynamique et fonctionnelle. Ces couplages entre informatique/statistique/-mathématique et biologie se font maintenant dans un contexte florissant, voire débordant, de la biologie à haut-débit dont la révolution technologique n'est pas encore terminée : chaque année voit apparaître de nouvelles technologies de séquençage plus rapides, plus performantes et moins chères ; sans oublier les révolutions à venir dans d'autres domaines de la biologie visant à l'automatisation d'acquisition de données autres que les acides nucléiques comme les protéines, les métabolites, les images biologiques fixes et bientôt dynamiques.

Les approches bioinformatiques doivent faire face à de nouveaux défis scientifiques afin de traiter les données brutes et leur donner un sens interprétable par l'esprit humain ! De beaux challenges en perspective pour acquérir, stocker, analyser, partager et modéliser ces données... C'est ainsi qu'à JOBIM s'exposent et se discutent des résultats scientifiques en génomique, informatique, statistiques et mathématiques, concernant aussi bien la description des génomes, l'étude de leur fonctionnement, de leur dynamique et de leur évolution.

Cette année, nous accueillons sept conférenciers invités David B. Searls, Martin Vingron, Ivo Hofacker, Pierre Baldi, Hugues Roest Crollius, Bertil Schmidt et Toni Gabaldón que nous remercions vivement pour leur participation. Au programme également : 33 articles longs ont été sélectionnés (sur les 60 soumissions dans cette catégorie) et feront l'objet d'une présentation orale. Enfin, 112 contributions seront exposées sous forme de poster et discutées lors de ces journées. Nous souhaitons remercier ici chaleureusement les relecteurs pour leur travail sur les 158 soumissions reçues et les retours nombreux fournis aux auteurs. De nouveau en 2012, deux jeunes scientifiques se verront récompensés par l'attribution de prix à l'intention de doctorants ou post-doctorants. Par ces deux prix, nous souhaitons apporter la reconnaissance à des travaux prometteurs et encourager nos jeunes talents à poursuivre leurs recherches dans la bio-informatique.

Un grand merci à tous les membres des comités de programme et d'organisation sans oublier l'ensemble des bénévoles qui ont largement œuvré à la réussite de Jobim 2012, placé sous l'égide de la Société Française de Bio-Informatique (S FBI). Nous remercions également tous nos partenaires industriels et institutionnels, particulièrement Inria et l'université de Rennes 1 pour l'organisation et l'accueil offerts à Jobim.

Bienvenue à toutes et à tous !

Pour le comité de Programme :
François Coste, Inria Rennes Bretagne Atlantique
Denis Tagu, Inra Rennes

Pour le comité d'organisation :
Claire Lemaitre, Inria Rennes Bretagne Atlantique
Pierre Peterlongo, Inria Rennes Bretagne Atlantique

Comité d'organisation

Claire Lemaitre et Pierre Peterlongo

Elisabeth Lebret
Marie-Noelle Georgeault
Olivier Collin
Raluca Uricaru

Et l'aide indispensable des nombreux bénévoles de l'équipe
Symbiose (Dyliss+GenOuest+GenScale)

Comité de programme

François Coste et Denis Tagu

Benjamin Audit
Grégory Batt
Anne Bergeron
Philippe Besse
Christophe Blanchet
Jérémie Bourdon
Laurent Bréhélin
Anne-Claude Camproux
Hubert Charles
Hélène Chiapello
Eric Coissac
Erwan Corre
François Coste
Florence D'Alché-Buc
Etienne Danchin
Hidde de Jong
Gilbert Deléage
Sébastien Duplessis
Pascal Ferraro
Christine Froidevaux
Olivier Gascuel
Christine Gaspin
Daniel Gautheret
Mathieu Giraud
Christophe Hitte
Vincent Lacroix
Dominique Lavenier
Claire Lemaitre
Frédérique Lisacek
Claudine Médigue
Karyn Megy
Jacques Nicolas
Cédric Notredame
Grégory Nuel
Aida Ouangraoua
Etienne Paux
Guy Perrière
Pierre Peterlongo
Charles Pineau
Yann Ponty
Anne Poupon
Eric Rivals
Manuel Ruiz
Sophie Schbath
Hervé Seitz
David James Sherman
Anne Siegel
Thomas Simonson
Denis Tagu
Claude Thermes
Denis Thieffry
Julie Thompson
Hélène Touzet
Pierre Tuffery
Yves Vandenbrouck
Jean-Philippe Vert
Stéphane Vialette
Alain Viari

Relecteurs additionnels

Florian Alberto
Cristian Chaparro
Rayan Chikhi

Thérèse Commes
Damien Éveillard
Delphine Flatters

Morgan Magnin
Michel Petitjean
Mikaël Salson

Partenaires

Organisation et soutien

Partenaires académiques

Partenaires industriels

Types de contributions

Hormis les résumés proposés par les 7 conférenciers invités et nos 5 partenaires industriels, les contributions présentées dans ce recueil sont de trois types :

- **Articles originaux** : (8 à 10 pages) : réservés aux résultats originaux non publiés par ailleurs. Ce type de contributions donne lieu à une présentation orale lors de la conférence.
- **Résumés étendus** (2 à 8 pages) : présentation de résultats récents, éventuellement déjà soumis ou acceptés ailleurs. Ce type de contributions donne lieu à une présentation orale lors de la conférence.
- **Résumés courts** (1 à 2 pages) : concernant des travaux récents ou en cours, publiés ou non. Ce type de contributions donne lieu à une présentation sous forme de poster lors de la conférence.

Sommaire

— Avant-Propos —

— Contributions —

—Listes et Index—

Session 1: Molecules and Sequences

Conférence invitée

David B Searls

Independent Consultant, Philadelphia, United States

Molecules, Languages, and Automata

Molecular biology is replete with linguistic metaphors, from the language of DNA to the genome as "book of life". Certainly the organization of genes and other functional modules along the DNA sequence invites a syntactic view, which can be adopted for purposes of pattern-matching search via parsing. This in turn has led to the development of novel grammar formalisms specially adapted to the biological domain. It has also been shown that folding of RNA structures is neatly expressed by grammars that require expressive power beyond context-free on the Chomsky hierarchy, an approach that has been conceptually extended with other grammar formalisms to the much more complex structures of proteins. Grammars and their cognate automata have even been adopted to describe evolutionary processes and algorithms for their reconstruction via sequence alignment, and indeed the analogy between the evolution of species and of languages (first noted by Darwin) has been exploited by applying bioinformatics tools to human languages as well. Processive enzymes and other "molecular machines" can also be cast in terms of automata, and thus of grammars, opening up new possibilities for the formal specification, modeling, and simulation of biological processes, and perhaps tools useful in the fields of DNA computing and nanotechnology. This talk will review linguistic approaches to molecular biology, and perspectives on potential future applications of grammars and automata in this field.

Expressive Pattern Matching with Logol. Application to the Modelling of -1 Ribosomal Frameshift events

Catherine BELLEANNÉE[1], Olivier SALLOU[1] and Jacques NICOLAS[1]

Irisa/Inria/Université de Rennes1, campus de Beaulieu, 35042 Rennes Cedex France
{Catherine.Belleannee,Jacques.Nicolas,Olivier.Sallou}@irisa.fr

Abstract *The current practice of pattern matching tools and the gap that may be observed with the actual modelling needs of people analysing genome structures clearly demonstrates the need for higher level languages to describe and search for these structures in genomic sequences. It appears necessary to offer new tools allowing to build more expressive models of families of biological sequences, on the basis of their content and structure. This article presents Logol, a new application designed to achieve pattern matching in possibly large sequences with realistic biological motifs. Logol consists in both a language for describing patterns, and the associated parser for effectively scanning sequences (RNA, DNA or protein) with such motifs. The language, based on an high level gramatical formalism, allows to express flexible patterns (with misparings and indels) composed of both sequential and structural elements (such as repeats or pseudoknots). A web page on the GenOuest BioInformatics Platform* http://www.genouest.org/ *gives access to the Logol application. It includes an interface for graphically drawing the motif model and an interface to display the resulting matches within the targetted pattern.*
Logol is presented through an illustrative application using a quite intricate motif model, which is the detection of -1 ribosomal frameshifting events in messenger RNA sequences.

Keywords Pattern Matching, Sequence Modelling, String Variable Grammar, Frameshift event, Expressive Motif, Pseudo-Knot, Complex Pattern

Recherche d'instances de motifs expressifs avec Logol. Application à la modélisation d'événements de *frameshift -1*

Résumé *L'état de la pratique des outils de reconnaissance de motifs et l'écart qui peut être observé avec les besoins réels de modélisation des personnes en charge de l'analyse des structures génomiques montrent clairement le besoin de langages de plus haut niveau pour décrire et rechercher ces structures dans les séquences génomiques. Il apparaît ainsi nécessaire de proposer de nouveaux outils permettant de définir des modèles expressifs de familles de séquences biologiques, modèles basés à la fois sur le contenu et la structure des séquences. Cet article présente Logol, une application de reconnaissance de motifs conçue pour analyser des séquences potentiellement grandes avec des motifs biologiques réalistes. Logol est constitué d'un langage de description de motifs et de la suite logicielle associée, permettant de réaliser effectivement l'analyse de séquences (d'ADN, ARN ou protéines) avec ces motifs. Le langage, basé sur un formalisme grammatical de haut niveau, permet d'exprimer des motifs flexibles (autorisant substitutions et indels) composés à la fois d'éléments de séquences et de structures (tels que des répétitions ou des pseudo-noeuds). La suite logicielle est accessible sur le web, sur la plateforme bioinformatique GenOuest* http://www.genouest.org/*. Elle contient notamment deux interfaces, l'une pour dessiner graphiquement le modèle de motif et la seconde pour afficher les résultats comme des instances de ce modèle.*
Logol est présenté au travers d'une application illustrant les concepts utiles via l'utilisation d'un modèle de motif assez riche. Il s'agit de la détection d'événements de décalage de phase en -1 dans les ARN messagers.

Mots-clés Reconnaissance de motifs, Modélisation de séquences, Grammaire à variables de chaînes, événements de décalage de cadre en -1, motifs expressifs, pseudo-noeuds, pattern complexes

1 Introduction

La reconnaissance de motifs consiste, étant donné un motif et une séquence d'entrée, à trouver les occurrences du motif dans la séquence. En biologie, il existe un besoin croissant de recherche de motifs variés afin d'aider les biologistes à analyser leurs séquences. Il peut s'agir de rechercher toutes les instances exactes d'une chaîne dans une banque de protéines, de rechercher les instances approchées d'un transposon dans un génome entier, de localiser des pseudo-noeuds dans des séquences d'ARN etc. Ainsi, ces dernières années, de nombreux outils ont été conçus pour répondre à un certain nombre de ces objectifs, et beaucoup d'entre eux sont utiles et utilisés. Il apparaît cependant qu'un besoin subsiste d'un outil de modélisation de séquences générique, ouvert (i.e. évolutif et non spécialisé sur la reconnaissance d'un type de motif, ni sur l'analyse d'un type de séquence) et réellement opérationnel. C'est dans cette perspective que nous proposons Logol, un nouvel outil général de reconnaissance de motifs dont les objectifs sont à la fois d'être le plus expressif possible (dans le sens où il vise à traduire des modèles complexes, déjà imaginés ou émergeants) et raisonnablement efficace (dans le sens où il doit être utilisable en pratique sur de grandes séquences, de par ses performances et son ergonomie).
Après avoir présenté un panorama de travaux existants en reconnaissance de motifs, nous présentons les bases de Logol et ses principaux constituants, pour ce qui concerne à la fois le langage de motif et l'outil d'analyse de séquences. Une exploration plus détaillée de certains éléments de Logol est ensuite effectuée au travers l'observation d'un exemple d'application de Logol : l'aide à la détection d'événements de "frameshift-1" dans les ARN messagers. Le modèle concerné ici est particulièrement riche puisqu'il contient des informations de séquence (telles que la détection de codons start et stop, ou de "fenêtres glissantes") et de structure (recherche de pseudo-nœuds, calage de phases de lecture), et amène à investiguer la composition de segments (pourcentage de GC) ou le type d'appariement utilisé dans les tiges-boucles (Watson-Crick avec ou sans Wobble).

2 Reconnaissance de motifs

De nombreux outils existent pour effectuer de la recherche de motifs, répondant à un large spectre d'objectifs. Nous dressons ici un panorama -non exhaustif- de ces outils et de leurs objectifs pour situer Logol dans cette perspective.

2.1 Outils généraux

Certains des outils de reconnaissance de motifs sont conçus pour analyser les différents types de séquences (ADN, ARN ou protéines) et ont une expressivité "ouverte", dans le sens où ils ne sont pas dédiés à la recherche d'un type de motif particulier. Ce sont donc les outils les plus *généraux*. Parmi ceux-là, certains ont pour objectif principal d'être le plus *efficace* possible. C'est particulièrement le cas pour Vmatch[15] qui offre une suite logicielle pour résoudre efficacement une grande étendue de tâches de reconnaissance. L'efficacité est obtenue grace à un précalcul sur les séquences d'entrée qui génère une structure d'index. L'index, un "tableau de suffixes amélioré", référence toutes les sous-chaînes des séquences analysées. L'indexation peut réduire de façon considérable le temps nécessaire à la localisation des motifs, surtout si le motif contient des sous-chaînes discriminantes. De nombreux outils plus spécifiques basent leurs recherches sur la suite Vmatch (e.g. recherche de "tandem-repeats" ou de retrotransposons-LTR [15]). Un autre outil général, Biogrep[13], a été conçu avec l'objectif de *reconnaître rapidement de nombreux motifs simples* (plus de 100) contre les banques de séquences biologiques. Biogrep utilise POSIX, format standard d'expressions régulières étendues, et peut répartir la tâche de reconnaissance de motifs sur un certain nombre de processeurs.

Parmi les outils généraux, certains autres tendent à être le plus *expressif* possible. Une contribution essentielle à cet objectif est due à D. Searls, qui a posé les fondations d'une recherche dans le domaine. Il a été le premier à superviser des développements permettant aux utilisateurs de concevoir des grammaires puis analyser des séquences génomiques avec celles-ci [23,7,22]. Il a introduit un nouveau type d'objet dans les grammaires algébriques, la *variable de chaîne*, qui permet d'exprimer élégamment la notion de copie (directe ou inverse). Il a mis en œuvre le formalisme résultant, appelé SVG pour "String Variable Grammars", dans l'outil Genlang[7]. La copie directe (*e.g.* `X...X`) permet de rechercher deux copies d'une même chaîne inconnue, avec éventuellement une indication sur la taille de la chaîne. La copie inverse quant-à elle (*e.g.*

`X...` ∼`X`) permet de rechercher une chaîne et son complément inverse biologique, ce qui permet d'exprimer des palindromes biologiques tels que les tige-boucles (`Stem, Loop,` ∼`Stem`) ou les pseudo-nœuds (`Stem1, Loop1,Stem2, Loop2,` ∼`Stem1, Loop3,`∼`Stem2`). Genlang, Stan[18] (développé dans notre équipe), Patscan[8] et Patsearch[19] sont des outils de cette famille. Grâce aux variables de chaînes et à des composants additionnels, ces langages permettent de mêler facilement dans les modèles des informations de séquence et de structure.

2.2 Outils dédiés

Il est impossible ici de fournir une vue complète du foisonnement d'outils spécifiques qui ont été mis à disposition des bioanalystes. Certains sont spécifiques à une famille de séquences ou à un type de motif. L'outil le plus connu qui soit *dédié aux proteines* est ScanProsite[6], où les motifs sont basés sur les expressions régulières, recherchées soit dans une base de données précalculée soit par l'algorithme PS-SCAN[11].

Un grand nombre d'outils est *dédié aux séquences d'ARN*, du fait de la nécessité d'explorer les très complexes structures d'ARN, particulièrement chez les ARN non codants. Par exemple, RNAmotif[16], RNAbob[9], Hypasearch[12,24] et Palingol[5] permettent de décrire les motifs comme une succession de tiges et de boucles, avec généralement la possibilité de choisir le type d'appariement entre l'appariement standard Watson-Crick (`A-U`, `G-C`) ou l'appariement avec Wobble (`A-U`, `G-C`, `G-U`). Les motifs peuvent aussi contenir des informations de séquence qui doivent être présentes dans certaines parties des tiges ou des boucles. RNAmotif[16] est probablement le plus populaire de ces outils. Un nouvel outil de cette famille, Structator[17], améliore significativement le temps d'analyse grâce à une structure d'index adaptée à l'analyse de palindromes, les "tableaux affixes".

L'outil Locomotif[20] a presque la même expressivité (un peu moins) que les précédents mais il est conçu pour répondre à des objectifs supplémentaires intéressants. Un premier est de proposer une *conception graphique du motif*. Le motif est décrit graphiquement et l'analyseur correspondant est dérivé du graphique de façon complètement automatique. L'éditeur permet de dessiner des structures secondaires, par composition de tiges et de boucles, annotées avec des informations de séquence et de taille. De plus, un second objectif est de ne renvoyer qu'un seul résultat *le meilleur résultat d'après un modèle thermodynamique*.

3 Langage Logol : Quelle expressivité ?

3.1 A la base : Modèle grammatical

• **Grammaires à Variables de Chaînes** Ayant vocation à être un langage général et expressif, et permettant d'exprimer naturellement la notion de structure, le langage Logol a pris pour base les grammaires à variables de chaînes (grammaires SVG) définies par D.Searls [23,7,22] et introduites en section 2.1. Comme on l'a déjà évoqué, si l'expression de motifs ou de gaps est accessible dès le niveau des langages réguliers (*e.g.* PROSITE [6]), les grammaires SVG permettent de plus l'expression des palindromes (nécessaires pour traduire tiges-boucles et pseudo-nœuds) accessible à partir des langages algébriques, ainsi que la notion de répétition (duplication d'une sous-chaîne) qui nécessite une expressivité encore supérieure, et qui est captée par les variables de chaînes. Ainsi, les grammaires SVG et celles de Logol se situent au-delà des grammaires algébriques, dans la classe que A. Joshi appelle "faiblement contextuelle" (*midly context sensitive languages*) [14]. Si Logol a construit ses bases autour des grammaires SVG, le langage a été ensuite largement étendu avec l'objectif de le rendre le plus adapté possible à l'expression de motifs biologiques réalistes. Nous en décrivons ici les principaux constituants.

• **Premières grammaires Logol** La grammaire Logol suivante permet de rechercher toutes les instances du motif "aaa" dans une séquence :

```
mod1()==>"aaa"
mod1()==*> SEQ1
```

La dernière ligne est le point d'entrée de la grammaire, appelé "rule" dans le modèle gaphique. Elle indique quel est le modèle, ici `mod1()`, à rechercher dans la séquence. Les autre lignes donne la définition du modèle (i.e. du motif). Le motif recherché par la deuxième grammaire est constitué de deux copies d'une même chaîne,

de taille comprise entre 5 et 8, les copies étant distantes l'une de l'autre de 1 à 10 caractères :
`mod2() ==> X1:#[5,8], .*:{#[1,10]}, X1.`
`mod2()==*> SEQ1`
Le modèle `mod2()` se lit de la manière suivante : X1 désigne une variable ; toute chaîne constituée de 5 à 8 caractères peut être instance de X1. '`.*`' désigne l'espaceur ('gap'). Il a pour contrainte une taille comprise entre 1 et 10 caractères. Après le gap, on attend une seconde occurrence de X1. Ainsi, **ACUGGC**CCGACUGGC**ACUGG**
est une instance de ce motif sur la séquence d'entrée UUCAG*ACUGGCCCGACUGGCACUGGC*CAC.
Voici une autre façon d'exprimer ce motif :
`mod2() ==> X1:{#[5,8],_IX1}, .*:{#[1,10]}, ?IX1`. Ici, l'instance de X1 est sauvegardée (par _) dans une variable (notée ici IX1). Après un gap de 1 à 10, on souhaite retrouver la même chaîne IX1 (que l'on rappelle par `?`). La deuxième version de `mod2()` est ici inutilement lourde. Elle permet cependant d'introduire une notion qui sera utilisée par la suite, la notion de mémorisation d'instance. En effet, les instances d'une variable ne sont pas forcément des copies exactes (cf la section suivante), et ce procédé de nommage explicite (ici `_IX1`) permet de distinguer les instances entre elles. Plus généralement, ce mécanisme permet de sauvegarder, pour s'y référer plus loin, les instances concrètes de n'importe quelle partie de modèle.

3.2 Principaux composants du langage

• **Copies non exactes** Les séquences génomiques évoluent à travers un processus de duplication entrainant des mutations ou des erreurs. Les variations élémentaires (d'un caractère) entre un modèle et son instance sont prises en compte par deux compteurs de coût : le compteur de *substitutions* et le compteur *indel* d'insertion/délétions. En pratique, les substitutions sont définies par une contrainte dite de contenu : `$[m,n]` avec `m` et `n` des entiers. Cette contrainte autorise de `m` à `n` substitutions. Une contrainte de substitution peut également s'exprimer sous la forme d'un pourcentage de substitutions autorisées : `p$[m,n]` dans ce cas, `n` représente le pourcentage maximum de substitutions autorisé (et `m` le minimum). En reprenant l'exemple précédent, on peut donc autoriser une substitution dans la seconde occurrence de X1 en le précisant dans le modèle de la manière suivante : `X1:{#[5,8],_IX1}, .*:{#[1,10]} , ?IX1:$[0,1]` Les indels sont définis de façon similaire, par `$$[m,n]` et `p$$[m,n]`. Ainsi, `"aaaa" ::$$[0,1]` accepte les chaînes `aaaa` (pas d'indel), `aaa` (une délétion) ou `aaaaa` (une insertion).

Voici un exemple pour compléter le propos sur le nommage des instances des variables de chaîne :
`X1:{#[5,8],_IX1},.*:{#[1,7]}, ?IX1: {_IX2}:{$[1,1]}, .*:{#[1,7]}, ?IX2:{$[1,1]}`
Ce modèle permet de chercher trois instances d'une même chaîne qui dériveraient successivement l'une de l'autre (e.g `IX1 = aaaaa`, `IX2 = aaaca` et `IX3 = agaca`). L'individualisation des instances peut ainsi contribuer à traduire la notion d'évolution dans les séquences.

• **IUPAC** L'alphabet *IUPAC* définit des caractères ambigus (*e.g.* R désigne un A ou un G dans l'ADN). L'utilisation de cet alphabet est autorisée dans les modèles Logol, et contribue à la prise en compte de la variabilité élémentaire dans les séquences.

• **Morphismes** Un morphisme est une fonction applicable à une chaîne. Certains morphismes sont prédéfinis. Ainsi `"wc"` transforme une chaîne d'ARN en son complémentaire, en appliquant la complémentarité "Watson Crick" qui transforme A en U, G en C et inversement. Ainsi le motif `("wc" "ACUGGC")` représente la chaîne `"UGACCG"`. Un autre morphisme prédéfini est le morphisme inverse, noté −, qui renverse une chaîne. Ainsi `(-"UGACCG")` représente la chaîne `"GCCAGU"`. Ces deux morphismes combinés permettent de représenter le *complémentaire inverse* d'une chaîne, et donc de modéliser les palindromes biologiques que sont les tiges-boucles. Le motif suivant décrit par exemple une tige-boucle, dont la longueur de la tige varie entre 5 et 11, de la boucle entre 1 et 10, et dont l'appariement "Watson Crick" de la tige n'est pas forcément parfait, pouvant contenir jusqu'à deux substitutions et un indel (*i.e.* une insertion ou une délétion)
`STEM5:{#[5,11],_IS5}, .*:{#[1,9]}, -"wc" ?IS5 :{$[0,2],$$[0,1]}`
Ainsi, le contenu de "STEM5" (la tige aller, *i.e.* située du coté 5' de la séquence) est sauvegardé lors de l'analyse Logol. La partie "STEM3" (la tige retour) a pour contenu le complément inverse de STEM5 précédemment sauvegardé (donc `-"wc" ?IS5`), à deux substitutions et un indel près. L'utilisateur peut par ailleurs définir ses propres morphismes.

• **Compositions des segments** Pour exprimer la composition des séquences, telle que l'hydrophobicité d'une zone de protéines, ou la richesse en GC d'un fragment d'ARN, Logol propose l'expression de "contraintes d'alphabet" qui vérifient le taux de présence de certaines lettres dans la séquence. Ainsi, `X1:{#[2,43]}:{%gc :` désigne un segment de 2 à 43 caractères ayant au moins cinquante pour cent de GC.

• **Contraintes de contenu négatives** Les contraintes de contenu négatives permettent d'exclure certaines valeurs dans un motif. Elles s'expriment avec le symbole `!`. Ainsi, `("aaa"| "ttt"), !"ga":{#[2,2]}` désigne une chaîne constituée de cinq caractères, les trois premiers étant trois `a` ou trois `t`, et les deux suivants étant tout sauf le mot `ga`.

• **Répétitions** Les répétitions en tandem, qui sont les copies successives d'une même entité, constituent des structures fréquentes dans les séquences génomiques. Pour modéliser des structures de ce type, Logol propose un constructeur de répétitions, `repeat`, qui gère un compteur d'occurrences. Son format standard est
`repeat(<entité>,<distance>)+<nombre d'occurrences>`.
Ainsi par exemple, `repeat("acgt",[0,3])+[7,38]` indique la répétition de l'entité `"acgt"` de 7 à 38 fois, avec un espacement autorisé d'au plus trois caractères entre deux répétitions.
Un autre format possible, `repeat(<entité>;<distance>)+<nombre d'occurrences>` (*i.e.* avec `;` à la place de `,`), indique que les occurrences successives peuvent être chevauchantes.

• **Analyses multiples** Parmi les caractéristiques importantes des séquences biologiques se situe la cohabitation de structures alternatives dans une même séquence. Les chevauchements de gènes, par exemple, sont des configurations fréquentes. Logol permet de modéliser de telles situations, en indiquant les modèles alternatifs au niveau de la règle principale de la grammaire (`==*> SEQ1`). Pour être acceptée par la grammaire, la séquence doit contenir une instance de chaque modèle alternatif.

```
mod1() ==> "YVCPFDGCNK"
mod2() ==> "NKLKSHIL"
mod1().mod2()==*> SEQ1
```

Ainsi, la grammaire ci-dessus accepte des séquences qui contiennent les deux chaînes `"YVCPFDGCNK"` et `"NKLKSHIL"`, indépendemment de leurs positions respectives, chevauchantes ou non. On peut passer des paramètres entre 2 motifs alternatifs, par exemple pour caler les positions respectives de certains éléments.

• **Structuration des modèles** Les modèles, dès lors qu'ils sont un peu complexes, sont amenés à être structurés. Ceci se fait naturellement dans la mesure où le formalisme grammatical est particulièrement adapté à l'écriture de modèles hiérarchiques.

```
mod2()==> repeat(("K"|"R"|"L"),[0,0])+[8,12]
mod1()==>"CVC", .*:{#[3,8]}, mod2()
```

• **Vues et portée des contraintes** Les contraintes (de contenu, de taille...) peuvent être posées sur différentes parties du modèle. On peut les appliquer à des entités élémentaires telles qu'une chaîne ou une variable, comme vu sur les exemples précédents, ou bien à un ensemble d'entités possèdant elles-mêmes leurs contraintes individuelles.
Si l'ensemble des éléments est contigu dans la séquence, on parle de *vue*. Syntaxiquement, la vue est définie par des parenthèses. Sur l'exemple suivant, les instances des trois variables X1,X2,X3 peuvent compter chacune jusqu'à dix caractères, mais la contrainte supplémentaire sur la vue (X1,X2,X3) impose que la totalité du segment soit inférieure à vingt caractères

```
(X1:{#[1,10]}, X2:{#[1,10]}, X3:{#[1,10]}) : {#[8,20]}
```

Il est également possible de poser des contraintes sur une collection d'éléments non contigus du modèle (par exemple sur les deux segments constituant la tige d'une tige-boucle). Ces contraintes sont alors placées dans un module global spécifique (le "panneau de contrôle").

4 Analyseur Logol : Quelles fonctionnalités ?

• **Construction du modèle** Le modèle Logol peut se construire de deux façons, grammaticalement via un éditeur de texte[2] ou graphiquement via une interface web[3]. L'interface graphique permet une abstraction de la syntaxe et une visualisation plus intuitive de ce que représente le modèle (les boucles par exemple) tout

en produisant un modèle grammatical. Des exemples des deux types de modèles sont présents dans cet article. L'outil accepte en entrée chacun de ces modèles indifféremment mais l'outil d'analyse offre des fonctionnalités supplémentaires avec le modèle graphique, comme la visualisation d'un match sur le modèle graphique originel.

• **Fonctionnement de l'analyse syntaxique** Le logiciel effectue son analyse en quatre étapes. La première étape convertit le modèle en un script Prolog. en affineant au mieux les recherches en fonction du modèle d'entrée (utilisation des propriétés de taille d'une variable connue *a posteriori* par exemple). La seconde étape découpe les séquences d'entrée, si possible, afin de paralléliser les traitements localement ou sur un cluster. La recherche peut également être exécutée en parallèle sur plusieurs CPU localement. La troisème est l'étape principale, celle d'analyse. Le programme va exécuter le programme Prolog généré pour rechercher le motif dans la séquence donnée en entrée. Cette recherche s'effectue en lecture gauche-droite . Basiquement, deux types de recherche vont être effectués, locale ou distante, pour trouver un élément du motif. La recherche distante est utilisée pour chercher un élément connu (avec ou sans erreurs) dont la position n'est pas connue. Pour cela plusieurs possibilitées sont offertes par le logiciel mais la solution la plus performante est l'utilisation de VMatch[15]. La recherche locale recherche un élement du motif quelconque (connu ou non) à la position courante, de manière récursive. Le programme va sauvegarder dans le contexte d'exécution les éléments du motif au fur et à mesure de ses correspondances avec la séquence jusqu'à ce qu'une équivalence complète soit trouvée. La correspondance est alors enregistrée comme un match. En cas d'échec, un retour arrière est effectué pour tester la condition suivante. La sauvegarde de chacun des éléments permet à la fois de tester des conditions sur des ensembles (condition de taille sur un ensemble d'éléments) et de fournir dans les résultats le détail de la correspondance entre l'élément trouvé et le motif. Une transformation est éventuellement appliquée à l'élément en cours avant de tester sa correspondance (complément inverse, matrice de transformation, ...). Enfin, la dernière étape regroupe les résultats, les reformate éventuellement et compresse les fichiers dans une archive zip.

• **Lancement de l'analyse** Le logiciel peut être exécuté en ligne de commande ou via une interface web. L'interface web permet d'exécuter le programme avec un nombre d'options réduit et propose une liste de banques de données disponibles (*i.e.* les banques accessibles sur la plateforme GenOuest). L'utilisateur peut ainsi lancer la reconnaissance de motif sur des séquences publiques, ou bien fournir sa propre liste de séquences. En ligne de commande, un fichier d'options permet de définir les préférences de l'utilisateur et d'entrer des valeurs par défaut pour ses recherches. De nombreuses options sont disponibles pour affiner la recherche ou limiter le nombre de résultats.

• **Format des résultats** Les résultats fournis par le programme sont par défaut sous la forme de fichiers XML archivés dans un fichier zip. Le programme permet également d'exporter les résultats au format FASTA et GFF. Il y a un fichier XML par séquence d'entrée. Chaque fichier contient la référence de la séquence d'entrée, le modèle utilisé (graphique ou non) et l'ensemble des résultats. Chaque résultat contient le détail de la correspondance avec le motif de départ, c.a.d conserve la même structure que le modèle (boucles, sous-modèles) avec les positions de la correspondance, les erreurs trouvées (nombre de substitutions et d'indels), le sous-motif d'origine... Ces détails permettent à l'utilisateur de réaffiner le modèle au besoin. Une interface web d'analyse permet de lister l'ensemble des résultats pour un fichier, d'afficher sous forme d'arbre un résultat, d'afficher sous forme statistique le nombre de résultats par branche du motif et d'afficher un résultat dans l'éditeur de modèle en colorant le chemin effectué.

5 Modélisation des décalages de phase en -1, ou 'frameshift -1'

Le recodage est un processus biologique qui peut intervenir lors de la traduction d'ARN messagers. Il génère une modification de la lecture traditionnelle du message génétique par le ribosome et permet ainsi la production de deux protéines distinctes à partir d'une même séquence d'ARN initiale. Parmi les événements de recodage se situe le "décalage de phase en -1" ("frameshift -1"). Le décalage de phase se fait par glissement du ribosome d'un nucléotide en amont au niveau de la "fenêtre glissante" `X XXY YYZ`. Le premier `X` est alors lu deux fois. Ainsi en phase 0, les codons `ABX XXY YYZ CDE ...` sont décodés par le ribosome alors qu'en phase -1 se sont les codons `ABX` **XXX YYY ZCD ...** qui sont décodés (le premier X est ainsi traduit 2 fois). La structure typique favorisant un événement de frameshift-1, qu'on appellera par la suite "motif F-1", est

représentée en Fig. 1. Elle est constituée successivement : d'un codon start, d'un certain nombre de codons, d'une fenêtre glissante (un heptamère) `XXXYYYZ` situé en phase -1, d'un "espaceur" de quelques nuléotides, et d'une structure secondaire. La structure secondaire est l'obstacle sur lequel vient buter le ribosome pendant qu'il traduit l'heptamère, et qui peut alors le faire reculer d'un nucléotide, provoquant le décalage de phase. La stucture secondaire typique est un "pseudo-nœud de type H", constitué de deux tiges-boucles imbriquées.

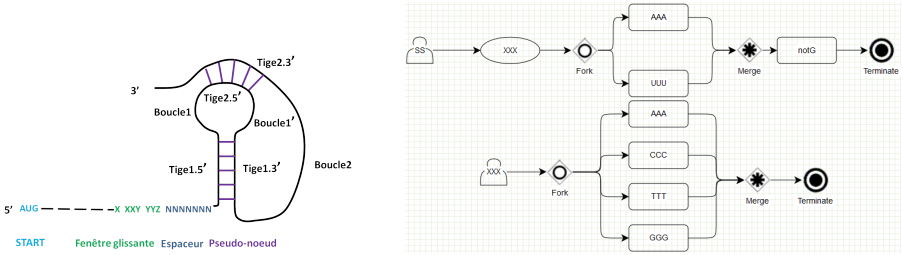


Figure 1. Fig1.A : Struture typique, appelée ici "motif F-1", favorisant un décalage de phase en -1. Fig1.B : modèle graphique Logol pour la fenêtre glissante du motif F-1

De nombreux outils existent pour tenter de détecter de potentiels sites de frameshift-1 [10], mais la détection reste un sujet de recherche actif, car le motif F-1 n'est pas universel (les caractéristiques de l'heptamère, du spacer, de la structure secondaire ne sont pas identiques d'un organisme à l'autre) et la détection de pseudo-nœuds est un problème difficile. Beaucoup de ces méthodes procèdent par filtres successifs, comme le fait par exemple KnotInFrame[25], un des outils le plus avancé sur le sujet. KnotInFrame détecte dans un premier temps tous les heptamères `XXXYYYZ`. Ensuite, il recherche de potentiels pseudo-nœuds en aval de ces motifs, au moyen d'une procédure de repliement d'ARN élaborée à cet usage, "pknotsRG-fs".

5.1 Modèle Logol

La complexité du motif F-1 en fait un bon candidat pour explorer l'expressivité de Logol. Sa modélisation nécessite d'utiliser un certain nombre de principes du langage tels que la multi-analyse, les contraintes de contenu négatif, les répétitions de motifs ou la recherche de palindromes biologiques. Nous en détaillons ici les éléments importants.

5.1.1 Calage de phase et multi-analyse de deux ORF chevauchants Parmi les caractéristiques structurelles indispensables à la survenue d'un événement de frameshift-1, certaines concernent le calage de phase. La traduction standard, "en phase 0", doit s'effectuer sur une séquence possédant une "cadre ouvert de lecture" : un codon start (`AUG`), suivi d'un certain nombre de codons (groupe de trois nucléotides qui va être traduit en un acide aminé), puis d'un codon stop (`UGA`,`UAG` ou `UAA`) qui arrête la traduction. L'ensemble des codons à traduire ne doit pas contenir le codon stop. La traduction avec frameshift-1 quant-à elle démarre sur le même start, puis à partir de la fenêtre glissante recule d'un nucléotide, ce qui conduit à continuer à avancer trois par trois mais "en phase-1", avant de s'arrêter sur un codon stop, situé de ce fait en phase -1. Pour qu'une séquence d'ARN ait le potentiel pour engendrer un événement de frameshift-1, il doit donc s'y trouver d'une part un cadre ouvert de lecture (un start suivi d'un nombre suffisant de codons non-stop puis un stop) et d'autre part un recalage de phase en -1 (le même start suivi à une distance suffisante d'un stop en phase -1).
Cette vérification correspond à faire en Logol une *analyse multiple* pour reconnaître les deux motifs alternatifs, "ORF" et "ORFminus", avec un *passage de paramètre* entre les deux modèles pour les caler sur le même start. Le modèle ORF contient ainsi un *repeat* qui accepte jusqu'à 300 codons non stop (les valeurs numériques sont inspirées de la littérature), soit : `repeat(notstop(),[0,0])+[0,300]` où le modèle `notstop` est un modèle, construit à partir d'une *vue*, qui accepte une chaîne de taille 3 qui ne soit pas un `stop`, pour cela une *contrainte de contenu négative* est posée sur la vue. Certains éléments du modèle graphique sont présentés en Fig. 2.

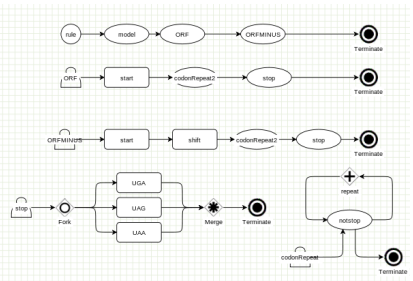


Figure 2. Modèle Logol graphique de calage de trames : 2 motifs alternatifs ORF et ORFminus ; existence de stops en phase 0 et en phase -1

5.1.2 Fenêtre glissante et espaceur Le motif F-1 contient trois compartiments principaux : la fenêtre glissante, l'espaceur et la structure secondaire. Le modèle Logol proposé pour la *fenêtre glissante* respecte le consensus établi : c'est un motif héptamérique de la forme XXXYYYZ, qui doit être positionné en phase-1, où X est un nucléotide quelconque répété 3 fois, Y est la base A ou U répétée trois fois, et Z est différent de G (cf Fig. 1.B). L'*espaceur* est réalisé par un simple gap (de taille ici inférieure à dix).

5.1.3 Pseudo-nœuds Comme évoqué précédemment la structure secondaire la plus efficace pour les frameshift-1 est le pseudo-nœud de type H (deux tiges-boucles imbriquées, cf Fig. 1.A) , même si elle n'est pas la seule possible (elle consiste parfois en une simple tige). C'est donc celle que nous avons modélisée ici.
Une vision graphique des différentes parties d'un pseudo-nœud est donnée en figure Fig. 3. `STEM15` désigne la tige 1 aller (côté 5'), `STEM13` désigne la tige 1 retour (côté 3'). `STEM25` et `STEM23` désignent les deux éléments de la deuxième tige. `L1A` ,`L1B` et `L2` sont les différents éléments des boucles.
Une première modélisation a été réalisée suivant cette structure de base, avec les valeurs numériques de la grammaire Logol suivante

```
STEM15:{#[4,16],_IS15},.*:{#[1,5]}, STEM25:{#[3,8],_IS25},.*:{#[0,4]},
     -"wc" ?IS15 :{$[0,4]},.*:{#[4,40]},-"wc" ?IS25 :{$[0,2]}
```

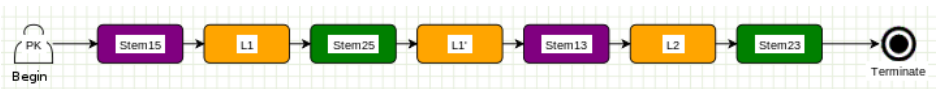


Figure 3. Premier modèle Logol pour représenter un pseudo-nœud

Notre démarche de validation de ce modèle sur des données réelles (cf section 5.2), nous a amené à le raffiner considérablement (cf Fig. 4). Le modèle final utilise une grande variété d'éléments de langage de Logol. Nous avons ainsi introduit la prise en compte du pourcentage de GC dans les tiges, la dissociation des deux nucléotides aux extrémités des tiges -pour interdire des mismatchs à cet endroit, ou encore l'acceptation de l'appariement non canonique "wobble" (G-U) dans certaines parties de tiges [21] (dans notre modèle, l'appariement "wcw" est utilisé aux extrémités de tige et "wc" en partie centrale).

5.2 Une première validation du modèle

Pour pouvoir tester la pertinence de notre modèle de frameshift-1, et l'affiner, nous avons élaboré un jeu de séquences de test[21]. Ce jeu est constitué en premier lieu de séquences connues pour produire des événements de frameshift-1. Nous avons choisi pour cela les trente séquences avérées ("validated -1 frameshift") de la base de référence Recode2 *http ://recode.genetics.utah.edu*. Le jeu de données est complété par des séquences aléatoires obtenues ainsi : chacune des trente séquences du jeu de données de référence est aléatoirement

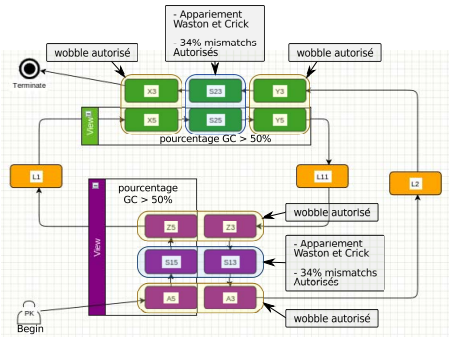


Figure 4. Modèle Logol final pour représenter le pseudo-nœud du Motif F-1

mélangée cent fois par le logiciel Shuffleseq *http ://emboss.bioinformatics.nl/cgi-bin/emboss/shuffleseq*, ce qui a pour intérêt de conserver les longueurs de séquence et un pourcentage nucléotidique constant pour chaque set constitué. On dispose ainsi de trente séquences "positives" (dans lesquelles on tente de retrouver, au bon emplacement, un motif F-1) et trois cents séquences "négatives" (dans lesquelles on souhaite retrouver le moins possible d'instances du motif F-1). Ces travaux de validation [21] nous ont amené à effectuer des comparaisons entre les prédictions de pseudo-nœuds faites par Logol et celles réalisées par "DotKnot", un logiciel de repliement de structures nouées *http ://dotknot.csse.uwa.edu.au/*, et ils ont permis de faire évoluer de façon importante notre modélisation des pseudo-nœuds, en ajustant les paramètres : appariement wobble, pourcentage de gc et taux de mésappariement acceptés dans les tiges (cf section 5.1.3). Au final, sur les séquences de référence de Recode2, le modèle Logol trouve une centaine de matchs par séquence, parmi lesquels se situe généralement le match Recode2. En appliquant a posteriori un calcul de score (sur la qualité des tiges) pour trier ces matchs, on localise la zone de frameshift officiel dans 20 cas sur 30.

• **Temps de calcul** Pour donner un ordre d'idée des temps de calcul, l'analyse de la plus grosse séquence de référence (30K) avec le modèle Logol final dure 1mn30s (sur Intel X5550, 144Go RAM) - versus réponse immédiate sur le site KnotinFrame. L'analyse du génome complet de Bacillus subtilis (`str 168 NC_000964` 4215606 bp) produit 7000 matchs en 2h. KnotinFrame ne peut pas analyser cette séquence.

6 Conclusion

Logol a été conçu pour initier la construction d'un outil de reconnaissance de motifs générique qui permette d'exprimer et de rechercher dans les séquences moléculaires les structures existantes typiques du mécanisme du vivant. C'est ainsi qu'après avoir développé un pemier outil, Stan[18], et d'en avoir perçu les limites, nous avons posé les bases d'un nouveau langage expressif et évolutif autour du principe des grammaires à variables de chaînes. Logol est actuellement opérationnel (il est par exemple utilisé pour rechercher les tiges-boucles des CRISPRS pour alimenter la base CRISPI[26]), et accessible sur la plateforme Genouest[1]. Le fait que le langage s'avère assez bien adapté pour traduire le modèle complexe des structures de frameshift-1, alors qu'il n'a pas été conçu à cette intention, nous semble un signe encourageant sur la généricité des éléments de langage de Logol. Ceci n'exclut pas que Logol reste un langage ouvert, amené à s'adapter aux nouveaux besoins d'expressivité, car sa finalité est de permettre d'exprimer des motifs complexes et réalistes, tels qu'ils sont découverts par les biologistes. Parmi les ajouts majeurs à envisager, afin de pouvoir prendre en compte l'extrême variété des interactions moléculaires possibles au sein de la cellule, se situe la prise en compte dans un même modèle de différents types de données (génomes, ARN et protéines) d'un même organisme.

Logol est accessible sur le site internet de la plateforme bioinformatique Genouest [1].

• **Remerciement** à la fondation Rennes1 qui a financé un "semestre pour l'innovation" à C.B. en 2011.

Références

[1] URL - Genouest bioinformatics platform : http ://www.genouest.org/.

[2] URL - Logol Designer and language tutorial : http ://training.genouest.org/claroline/claroline/learnpath/navigation/vie

[3] URL - Logol Designer graphical interface : http ://webapps.genouest.org/logoldesigner/.

[4] C. BELLEANNÉE et J. NICOLAS : Logol : Modelling evolving sequence families through a dedicated constrained string language. Research Report 6350, INRIA, 11 2007.

[5] B. BILLOUD, M. KONTIC et A. VIARI : Palingol : a declarative programming language to describe nucleic acids' secondary structures and to scan sequence database. *Nucleic Acids Res*, 24(8), avr. 15 1996.

[6] E. de CASTRO, C. J. A. SIGRIST, A. GATTIKER, V. BULLIARD, P. S. LANGENDIJK-GENEVAUX, E. GASTEIGER, A. BAIROCH et N. HULO : Scanprosite : detection of prosite signature matches and prorule-associated functional and structural residues in proteins. *Nucleic Acids Research*, 34(suppl 2):W362–W365, july 2006.

[7] S. DONG et D. B. SEARLS : Gene structure prediction by linguistic methods. *Genomics*, 23(3):540–551, 1994.

[8] M. DSOUZA, N. LARSEN et R. OVERBEEK : Searching for patterns in genomic data. *Trends in Genetics*, 13(12): 497–498, dec 1997.

[9] S. EDDY : Rnabob : a program to search for rna secondary structure motifs in sequence databases. 1996.

[10] A. E. FIRTH, M. BEKAERT et P. V. BARANOV : Computational resources for studying recoding. *In* J. F. ATKINS et R. F. GESTELAND, éds : *Recoding : Expansion of Decoding Rules Enriches Gene Expression*, vol. 24 de *Nucleic Acids and Molecular Biology*, p. 435–461. Springer New York, 2010.

[11] A. GATTIKER, E. GASTEIGER et A. BAIROCH : Scanprosite : a reference implementation of a prosite scanning tool. *Applied Bioinformatics*, 1(2):107–108, 2002.

[12] S. GRAF, D. STROTHMANN, S. KURTZ et G. STEGER : HyPaLib : a Database of RNAs and RNA Structural Elements defined by Hybrid Patterns. *Nucleic Acids Res.*, 29(1):196–198, 2001.

[13] K. JENSEN, G. STEPHANOPOULOS et I. RIGOUTSOS : Biogrep : A multi-threaded pattern matcher for large pattern sets. 2002.

[14] A. K. JOSHI, K. VIJAY-SHANKER et D. WEIR : The convergence of midly context-sensitive grammars. *In* S. M. SHIEBER et T. WASOW, éds : *The Processing of Natural Language Structure*, p. 31–81. MIT Press, Bosto, MA, 1991.

[15] S. KURTZ : The vmatch large scale sequence analysis software.

[16] T. J. MACKE, D. J. ECKER, R. R. GUTELL, D. GAUTHERET, D. A. CASE et R. SAMPATH : Rnamotif, an rna secondary structure definition and search algorithm. *Nucleic acids research*, 29(22):4724–4735, nov 2001.

[17] F. MEYER, S. KURTZ, R. BACKOFEN, S. WILL et M. BECKSTETTE : Structator : fast index-based search for rna sequence-structure patterns. *BMC Bioinformatics*, 12(1):214, 2011.

[18] J. NICOLAS, P. DURAND, G. RANCHY, S. TEMPEL et A.-S. VALIN : Suffix-tree analyser (stan) : looking for nucleotidic and peptidic patterns in chromosomes. *Bioinformatics*, 21(24):4408–4410, 2005.

[19] G. PESOLE, S. LIUNI et M. DSOUZA : Patsearch : a pattern matcher software that finds functional elements in nucleotide and protein sequences and assesses their statistical significance. *Bioinformatics*, 16(5):439–450, 2000.

[20] J. REEDER, J. REEDER et R. GIEGERICH : Locomotif : from graphical motif description to rna motif search. *Bioinformatics*, 23(13):i392–i400, 2007.

[21] A. ROCHETEAU et C. BELLEANNÉE : Recherche d'éléments structurés dans les génomes par modèles logiques. Rapport de recherche PI-1994, Dyliss - Inria - Irisa, avr. 2012.

[22] D. B. SEARLS : String variable grammar : A logic grammar formalism for the biological language of DNA. *Journal of Logic Programming*, 24(1 & 2):73–102, 1995.

[23] D. B. SEARLS et S. DONG : A syntactic pattern recognition system for DNA sequences. *In* C. R. CANTOR, H. A. LIM, J. FICKETT et R. J. ROBBINS, éds : *Proceedings 2nd International Conference on Bioinformatics, Supercomputing, and Complex Genome Analysis*, p. 89–101. World Scientific, 1993.

[24] D. STROTHMANN, S. A. GRÄF, S. KURTZ et G. STEGER : The syntax and semantics of a language for describing complex patterns in biological sequences. Rap. tech., Universität Bielefeld, Technische Fakultät, Arbeitsgruppe Praktische Informatik, août 2000.

[25] C. THEIS, J. REEDER et R. GIEGERICH : Knotinframe : prediction of -1 ribosomal frameshift events. *Nucleic Acids Research*, 36(18):6013–6020, 2008.

[26] C. VROLAND, C. BELLEANNÉE et J. NICOLAS : Recherche et annotation des structures de CRISPR dans l'ensemble des génomes procaryotes. Rapport de recherche PI-1986, Symbiose - Inria - Irisa, nov. 2011.

SA-Mot: a new web server for the extraction of motifs of interest from protein loops

SA-Mot web server

Leslie REGAD[1,2], Juliette MARTIN[3], Colette GENEIX[1,2]and Anne-Claude CAMPROUX[1,2]

[1] INSERM, U973, F-75205 Paris, France
[2] Univ Paris Diderot, Sorbonne Paris Cité, UMRS 973, MTi, F-75205 Paris, France
{leslie.regad, colette.geneix,anne-claude.camproux}@univ-paris-diderot.fr
[3] University Lyon 1, Univ Lyon, France; CNRS, UMR 5086; Bases Moléculaires et Structurales des Systèmes Infectieux, IBCP 7 passage du vercors, F-69367, France
{juliette.martin}@ibcp.fr

Abstract *The detection of functional motifs is an important step for the determination of protein function. We present here a new web server SA-Mot (Structural Alphabet - Motif) for the extraction and location of structural motifs of interest from protein loops. SA-Mot is based on the "structural word" notion to extract all structural motifs from loop structures. Then, SA-Mot provides a description of these structural motifs using statistics, sequence and structural parameters. This description allows the user to easily locate some loop regions that are important for the protein folding and function. The SA-Mot web server is available at* http://sa-mot.mti.univ-paris-diderot.fr.

Keywords Loop functional motifs, loop structural motifs, structural words, structural alphabet, webserver

SA-Mot: un server web pour l'extraction de motifs d'intérêt dans les boucles protéiques

SA-Mot serveur web

Résumé *La recherche de motifs fonctionnels est une étape importante pour la détermination de la fonction des protéines. Nous présentons ici un nouveau serveur web nommé SA-Mot (Alphabet Structural - Motif) pour l'extraction et la localisation des motifs structuraux d'intérêt extraits des boucles protéiques. SA-Mot utilise la notion de mots structuraux pour extraire tous les motifs structuraux des boucles protéiques. Ensuite, SA-Mot donne une description de ces motifs en termes de paramètres statistiques, de composition en séquences et de structure. Cette description permet à l'utilisateur de localiser facilement des régions d'intérêt pour le repliement et la fonction des protéines. Le serveur web SA-Mot est disponible à l'adresse suivante :* http://sa-mot.mti.univ-paris-diderot.fr.

Mots-clés Motifs fonctionnels, Motifs structuraux extraits des boucles, Mots structuraux, Alphabet structural, serveur web

1 Introduction

Protein structures can usually be broken down into their component secondary structures: α-helices, β-strands and loops. Unlike α-helices and β-strands, protein loops were initially seen as random coils, because their sequences and structures are highly variable. But the ever-increasing availability of protein structures in the Protein Data Bank (PDB) allowed extensive analyzes of protein loops which suggested a more complex view. For example, Panchenko *et al.* [17] concluded that even longer loop regions cannot be defined as irregular conformations or random coils. Several classifications of short and medium loops have been developed [6,16,4,28,8,13], according to the type and structure of flanking secondary structures, and the length and

geometry of loops. These classifications have revealed the existence of recurrent, amino-acid dependent loop conformations.

Loop regions play a role in protein function[9]. They may be involved in the active sites of enzymes [12] or in binding sites [24,26,25,10]. The classification of protein loops has then been used to investigate the link between protein loops and function [7]. This study showed that loops contain structural motifs involved in the functional sites of proteins.

Previously, we have developped a method to analyze protein loop structures by extracting structural motifs of 7 residues from protein loops [22] without using structural alignment of proteins. This method is based on the structural alphabet HMM-SA, that is a collection of 27 structural prototypes of four residues called structural letters (SLs), permitting the simplification of all 3D protein structures into one-dimensional (1D) sequences of SLs [5]. We have shown that structural words, four successive SLs, extracted from SL-sequences of loop structures correspond to clusters of seven-residue fragments with similar structures and amino-acid preferences [22]. Thus, the extraction of recurrent structural words from loop SL-sequences allows a rapid extraction of structural motifs from loop structures.

Recently, using the structural word notion, we have developed a method to analyze the link between the structural words extracted from protein loops [22] and protein functions [21]. Then, we computed the over-representation of each structural word in SCOP superfamilies [14], using the sofware SPatt [15]. This allows us to distinguish two types of over-represented structural words: (i) motifs shared by several functional families and (ii) motifs specific to one or few functional families. These motifs actually correspond to functional sites such as the binding sites of small ligands.

Coupling our 2 previous works [22,21], we have recently developed a web server, named SA-Mot (Structural Alphabet - Motif, `http://sa-mot.mti.univ-paris-diderot.fr`) [23]. This webserver allows the users to analyze protein loop structures by extracting structural motifs important both for structure and function of protein. The extraction of structural motifs of interest is based on two steps: (i) the extraction of all structural words from loops without pairwise comparisons of fragments [22] and (ii) the description of these structural words [22,21] that allows the identification of structural motifs of interest: recurrent and non random structural motifs, structural motifs with strong structural and amino-acid sequence conservation and structural motifs likely to be involved in protein folding or function. Using these information extracted for a protein target, the user can easily locate loop regions that are important for the protein folding and function.

In the following, we present rapidly the concepts, input and output of SA-Mot. Then, we illustrate two concrete search cases in which SA-Mot could be of interest: (i) to analyze the deformation involved by the binding of a ligand in a protein target and (ii) to help the determination of the function of uncharacterized proteins, i.e. protein, for which the structure is known but not the function.

2 Concept of SA-Mot server

SA-Mot [23] is a new web server that enables the user to examine loops of a single protein structure in order to identify and locate structural motifs of interest that have been documented in other structures. Thus, the first step was the creation of a database that contains structural motifs of interest extracted from loop structures.

2.1 Construction of a database of structural motifs of interest extracted from protein loops

Structural motifs of interest were extracted using a two step protocol: (i) extraction of all structural motifs from protein loops based on the structural alphabet HMM-SA [5] (for more details, see [22]), and (ii) description of structural motifs [22,21].

Protein dataset We used a set of 4 911 non redundant (less than 50% sequence identity) protein structures. These proteins have a resolution better than 2.5 Å, longer than 30 residues and are classified into SCOP classification [14].

Structural alphabet HMM-SA and structural motif extraction HMM-SA is a library of 27 SLs, describing four successive residues, established after a geometric classification of protein fragments. Using HMM-SA, a 3D protein structure of n residues is simplified into a 1D sequence of $(n-3)$ SLs, in which each SL describes a four-residue geometry [5]. Each 3D structure of our dataset was encoded into SL sequences, and only SL sequence of loops were retained [20]. Each simplified loop was split into overlapping words (four consecutive SLs, named structural words), which correspond to structural motifs of seven residues [22]. From the 90,811 loops, named loop data set, 25,304 different structural words were extracted, describing the conformation of 238,158 seven-residue fragments.

Parameters for the description of structural words All structural words extracted from loops are described using different parameters in order to select structural words of interest. Thus, for each word, we determine :

- its occurrence: the number of times it is seen in the loop data set,
- its structural conservation by computing the α-carbon RMSD between seven-residue fragments encoded by the same structural words [21],
- sequence conservation [21],
- statistical over-representation in the loop data set [22],
- statistical over-representation in protein families [21] (represented by SCOP superfamilies [14]). This score allows the extraction of structural motifs specific to a protein family that are probably linked to structural or functional implication.

All the structural words and their values for each parameters were stored in a database.

Different types of structural words of interest These parameters allows the selection of different types of structural words of interest :

- **Rare words**: correspond to words with a low number of occurrences (< 5). It has been shown that rare structural words are linked to structural flexibility and regions with uncertain coordinates [22].
- **Recurrent words**: correspond to words with an high number of occurrences (≥ 30). These words are found in a lot of proteins, which suggests that can be structural motifs with a key role in proteins [22].
- **Structural words presenting strong structural and/or sequence conservation**: The sequence and structural conservation of these structural words could result from an evolutionary pressure since they are located in important protein sites [22].
- **Over-represented words in loop dataset**: These structural words seem correspond to non random structural motifs that could be crucial for proteins [22] with weak structural variability and strong amino-acid specificities.
- **Ubiquitous words**: correspond to structural words over-represented in several superfamilies with different folds and functions. These words are important for protein structures.
- **Functional candidate words**: correspond to structural words highly over-represented in one or few superfamilies with similar functions. This words could be likely to be involved in functional sites. It has been shown that some functional candidate words correspond to structural motifs involved in binding or active sites [21].

2.2 Extraction of motifs of interest from loops of a new protein target

To locate motifs of interest from a new protein structure, SA-Mot translates the target protein structure into the HMM-SA space, that results in a SL-sequence. This SL-sequence is splitted into 4-SL words, that correspond to 7-residue fragments. Then, for each SL-word a request to the structural motifs of interest database is performed in order to know if the word is a motif of interest.

Input data is a protein 3D structure : either the PDB formatted coordinate file [2] or the PDB code of the protein target. Input PDB files can contain several chains.

As output, SA-Mot provides two tables: the first contains the counts of extracted structural words of interest. The second is an interactive table that allows the identification of structural words of interest and their characteristics: the positions and amino-acid sequence and values for all computed parameters (see Section 2).

3 Using SA-Mot

Search of a binding site and analysis of the deformation of this binding site

We have shown that some structural words of interest correspond to binding site [21]. For example, the structural words UODO, DODQ and RUDO correspond to nucleotide, calcium and SAM/SAH-binding sites, respectively. Thus, the identification of the structural word UODO in a SL-sequence corresponding to a protein structure allows the prediction of a nucleotide-binding site. In the same way, if a SL-sequence contains the structural word DODQ, one can predict a calcium-binding site in the protein structure. Thus, the SA-Mot webserver can be used to identify and locate binding sites.

To illustrate this use, we use a homoserine kinase bound to AMP (chain C, pdb code: 1h72). The loop-residues involved in the interaction with the AMP molecule are residues 62, 63, 87, 92, 94, 96, 98, 101 according to the Ligplot software [27]. SA-Mot webserver locates in this binding site the following structural words PIKK, IKKU, KKUO, KUOD, UODO, ODOD, DODG at positions 86-98 and PBZQ at positions 57-63. We observe that this binding site contains the word UODO, which has been characterized as functional candidate by SA-Mot because it is over-represented in the superfamily 'P-loop containing nucleoside triphosphate hydrolases' (SCOPid = 52540). This word has been previously identified as corresponding to a nucleotide-binding site [21].

The protein homoserine kinase is also crystallized on apo-form (pdb code : 1fwk). Figure 1 presents the structures of the 2 proteins, where the different words are colored. In the apo structure, SA-Mot identifies the following structural words PIKK, IKKU, KKUO, KUOD, UODO, ODOD, DODG at positions 86-98 and PWZQ word at positions 57-63. Thanks to the presence of UODO word, we can predict that the protein 1fwk contains a nucleotide-binding site. The comparison between the structural words containing in the apo and holo forms of the binding site allows us to analyze the deformation implied the binding of AMP molecule. This structural deformation is translated by the change of word PWZQ to PBZQ. These two structural words are structurally close with a $C_{alpha} - RMSD = 0.14$ Å. Thus, we can conclude that the binding of the AMP molecule on the homoserine kinase induce a very weak deformation, which was rapidly identified by SA-Mot web server.

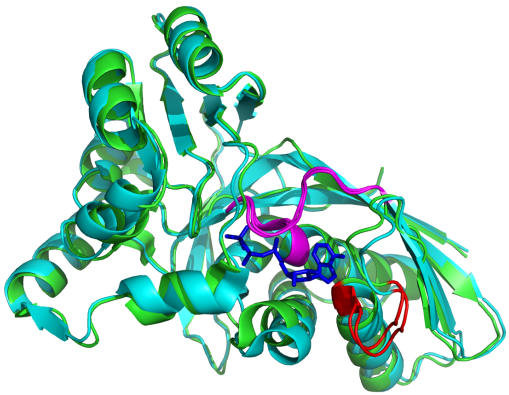


Figure 1.
Superimposition of two homoserine kinases. The structure colored in green correspond to a bounded form (pdb code: 1h72), one colored in cyan is an apo form (pdb code: 1fwk_C). In blue is represented AMP molecule, in red the word PZCD and in magenta the words PIKKUODOD.

Using SA-Mot to analysis an uncharacterized protein

Our target is an uncharacterized protein, for which a structure has been solved but without clearly established funtion: structure 2rhm, a putative putative kinase from *Chloroflexus aurantiacus*. SA-Mot is used to locate motifs of interest in protein loops of the protein target. A part of the results are presented in Figure 2. First, chain B contains rare structural words (PBEP, BEPR, EPRH) at positions 130-141 suggesting that it might be a flexible region. Then, we located 19 structural words that are recurrent and over-represented in the loop data set and present a weak structural variability and some amino-acid specificities. This suggests that they

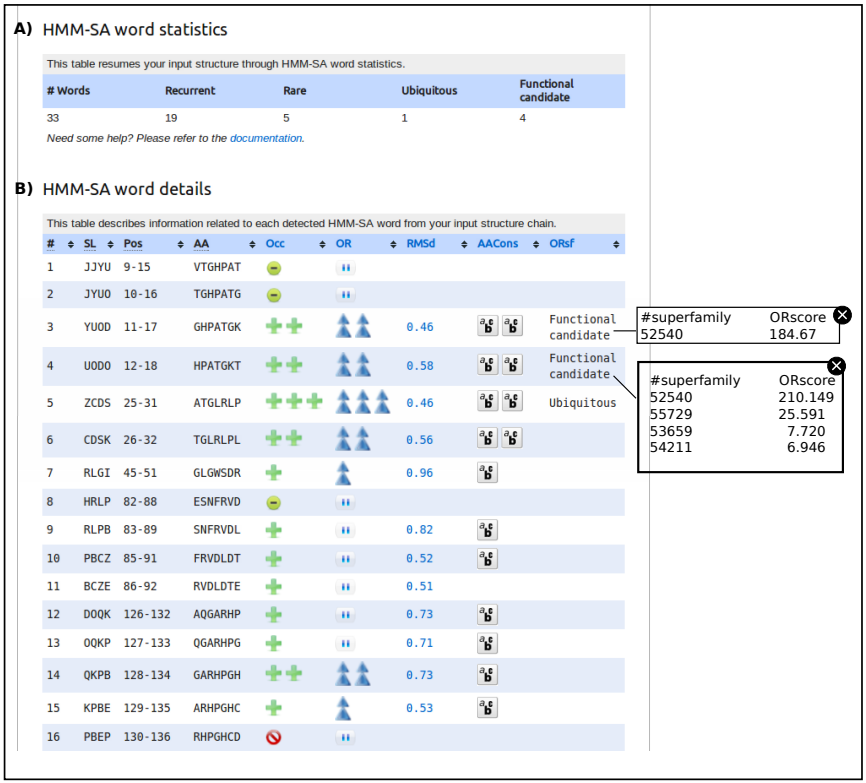


Figure 2. **(A)** Counts of structural words of interest extracted from the studied chain. **(B)** structural word description. For each structural word (column *SW*), the table provides its position and amino-acid sequence in the studied chain (columns *Pos* and *AA*), its occurrence and over-representation in the loop data set (columns *Occ* and *OR*), its structural and sequence conservation (columns *RMSd* and *AACons*) and its over-representation in SCOP superfamilies (column *ORsf*). By clicking on icons, users can access to a contextual window displaying parameter informations. For example , the list of SCOP superfamilies where a structural word is over-represented.

correspond to structural motifs involved in crucial regions in the protein target. One of these motifs of interest, ZCDS, located at positions 25-31, corresponds to an ubiquitous word clearly associated with beta-turns [21]. Chain B of the protein target contains also four functional candidate words: YUOD, UODO, CDSK, OZGB. For two structural words, CDSK and OZGB, no available annotation was found to confirm their functionality. OZGB is located at positions 165-171 and is over-represented in the "Trypsin-like serine proteases" superfamily (SCOP id=50494), suggesting that this structural motif is important for these proteases. It is surprising that a putative kinase shares a structural motif specific to trypsin-like serine proteases. Structural word CDSK is located at

positions 26-32 and is over-represented in the superfamily "YVTN repeat-like/Quinoprotein amine dehydrogenase" (SCOP id=50969). This suggests that this structural word is important for these dehydrogenase proteins. Overlapping structural words YUOD and UODO, located at positions 11-18, are strongly over-represented in the superfamily "P-loop containing nucleotide triphosphate hydrolases" (SCOPid=52540). These structural words have been identified as structural motifs with residues involved in nucleotide-binding sites [21].

SA-Mot results allow the location of ATP-binding site, that is the functional site needed for this activity. SA-Mot also locates two other structural motifs of interest, including one that seems to be important for the C-terminal region and seems be involved in another function than kinase activity.

4 Comparison between SA-Mot and other web servers

There are some web servers dedicated to the extraction of structural motifs from protein structures. Some of them focus on the extraction of structural motifs conserved accross protein cores to identify structural motifs important for the protein structures, such as MegaMotifBase [19]. Other web servers allow the extraction of structural motifs involved in protein function such as Webfeature [11], SitePredict [3] and FunClust [1]. SA-Mot is the only web server that provides informations about both structural motifs involved in the structure and function of the protein. Moreover, contrary to these servers SA-Mot extracts structural motifs without pairwise alignment of the structures.

There are two types of methods allowing the extraction of functional motifs: methods that focus on the extraction of structural motifs specific to a functional site and consist in learning the SMs of known functional sites: Webfeature [11] and SitePredict [3]. The difference between these website and SA-Mot is that SA-Mot can be used without knowledge about the location of binding site. Other methods such as FunClust [1] look for conserved structural motifs in proteins with the same function, that is the same principle than SA-Mot. However, FunClust webserver can be used only to extract functional motifs from a set of non homologous proteins with the same function but can not be used to predict a previously identified motif in a new protein.

5 CONCLUSION

SA-Mot allows the users to easily identify structural motifs of interest such as those involved in the structural and sequence redundancy and those with a putative role in protein structure or function. Thus, in contrast to classical methods, SA-Mot does not focus only on the detection of a binding site associated with a ligand, but explores all crucial structural motifs for proteins such as motifs of interest for protein folding or structural motifs involved in active site or REPEAT regions [21]. The second advantage of SA-Mot compared to other local methods is that the learning of motifs of interest is not based on the knowledge of functional sites. Thus, SA-Mot is able to propose novel structural motifs putatively important for the protein function.

During the analysis of the uncharacterized protein 1h72 using SA-Mot, we observed that some structural words of interest are overlapping. This means that the strucutral conservation can extend to more than 7 residues. This suggests that one extension of this work could be the extraction of structural motifs of interest of different lengths

Moreover, SA-Mot only runs on proteins for which the 3D structure has been resolved but it could be extended to circumstances where only the sequence is known. For this last case, we are currently developing a method to predict the structural words of interest directly from amino acid sequences. Then it will be possible to integrate this new method into the SA-Mot server to detect structural motifs of interest from protein loops using either the protein structure or the protein sequence.

References

[1] G. Ausiello, P. Gherardini, P. Marcatili, A. Tramontano, A. Via, and M. Helmer-Citterich. Funclust: a web server for the identification of structural motifs in a set of non-homologous protein structures. *BMC Bioinformatics*, 9:S2, 2008. [PubMed:18387204] [PubMed Central:PMC2323665] [doi:10.1186/1471-2105-9-S2-S2].

[2] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucl Acids Res*, 28:235–242, 2000. [PubMed:10592235] [PubMed Central:PMC102472] [doi:10.1093/nar/28.1.235].

[3] A. Bordner. Predicting small ligand binding sites in proteins using backbone structure. *Bioinformatics*, 24(24):2865–2871, 2008. [PubMed:18940825] [PubMed Central:PMC2639300] [doi:10.1093/bioinformatics/btn543].

[4] D. F. Burke, C. M. Deane, and T. L. Blundell. Browsing the sloop database of structurally classified loops connecting elements of protein secondary structure. *Bioinformatics*, 16:513–19, 2000.

[5] A. C. Camproux, R. Gautier, and T. Tufféry. A hidden Markov model derivated structural alphabet for proteins. *J Mol Biol*, 339:561–605, 2004. [PubMed:15147844] [doi:10.1016/j.jmb.2004.04.005].

[6] L. E. Donate, S. D. Rufino, L. H. Canard, and T. L. Blundell. Conformational analysis and clustering of short and medium size loops connecting regular secondary structures: a database for modeling and prediction. *Protein Sci.*, 5(12):2600–2616, 1996.

[7] J. Espadaler, E. Querol, F. X. Aviles, and B. Oliva. Identification of function-associated loop motifs and application to protein function prediction. *Bioinformatics*, 22:2237–2243, 2006.

[8] N. Fernandez-Fuentes, A. Hermoso, J. Espadaler, E. Querol, F. X. Aviles, and B. Oliva. Classification of common functional loops of kinase super-families. *Proteins*, 56(3):539–555, 2004.

[9] J. S. Fetrow. Omega loops: nonregular secondary structures significant in protein function and stability. *FASEB J*, 9:708–717, 1995.

[10] A. Golovin and K. Henrick. Msdmotif: exploring protein sites and motifs. *BMC Bioinformatics*, 9:312–312, 2008.

[11] I. Halperin, D. Glazer, S. Wu, and R. Altman. The feature framework for protein function annotation: modeling new functions, improving performance, and extending to novel applications. *BMC Genomics*, 9:S2, 2008. [PubMed:18831785] [PubMed Central:PMC2559884] [doi:10.1186/1471-2164-9-S2-S2].

[12] L. N. Johnson, E. D. Lowe, M. E. Noble, and D. J. Owen. The eleventh datta lecture. the structural basis for substrate recognition and control by protein kinases. *FEBS Lett*, 430:1–11, 1998.

[13] W. Li, Z. Liu, and L. Lai. Protein loops on structurally similar scaffolds: database and conformational analysis. *Biopolymers*, 49:481, 1999.

[14] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247:536–540, 1995. [PubMed:7723011].

[15] G. Nuel. S-spatt: simple statistics for patterns on Markov chains. *Bioinformatics*, 21:3051–3052, 2005.

[16] B. Oliva, P. A. Bates, E. Querol, F. X. Aviles, and M. J. Sternberg. An automated classification of the structure of protein loops. *J Mol Biol*, 266:814–830, 1997.

[17] A. R. Panchenko and T. Madej. Structural similarity of loops in protein families: toward the understanding of protein evolution. *BMC Evol Biol*, 5:10, 2005.

[18] G. Pugalenthi, P. N. Suganthan, R. Sowdhamini, and S. Chakrabarti. MegaMotifBase: a database of structural motifs in protein families and superfamilies. *Nucl Acids Res.*, 36:D218–221, 2008. [PubMed:17933773] [PubMed Central:PMC2238926] [doi:10.1093/nar/gkm794].

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

[20] L. Regad, J. Martin, and A. C. Camproux. Dissecting protein loops with a statistical scalpel: functional implication of structural motifs. *BMC Bioinformatics*, 12:247, 2011.

[21] L. Regad, J. Martin, G. Nuel, and A. C. Camproux. Mining protein loops using a structural alphabet and statistical exceptionality. *BMC Bioinformatics*, 11:75, 2010. [PubMed:20132552] [PubMed Central:PMC2833150] [doi:10.1186/1471-2105-11-75].

[22] L. Regad, A. Saladin, J. Maupetit, C. Geneix, and A. Camproux. Sa-mot: a web server for the identification of motifs of interest extracted from protein loops. *NAR*, 39:W203–W209, 2011.

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

[24] D. Stuart, K. Acharya, N. Walker, S. Smith, M. Lewis, and D. Phillips. Lactalbumin possesses a novel calcium binding loop. *Nature*, 324:84–87, 1986.

[25] A. Via, F. Ferre, B. Brannetti, A. Valencia, and M. Helmer-Citterich. Three-dimensional view of the surface motif associated with the p-loop structure: cis and trans cases of convergent evolution. *J. Mol. Biol.*, 303(4):455–465, 2000. [PubMed:11054283] [doi:10.1006/jmbi.2000.4151].

[26] A. Wallace, R. Laskowski, and J. Thornton. Ligplot: a program to generate schematic diagrams of protein-ligand interactions. *Protein Eng.*, 8(2):127–134, 1995.

[27] 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. *J Mol Biol*, 289:1469–1490, 1999.

The sequence space of G-protein-coupled receptors:
Implications for molecular modeling

Jean-Michel BECU[1], Julien PELE[1], Hervé ABDI[2] and Marie CHABBERT[1]

[1] Laboratoire BNMI, UMR CNRS 6214 – INSERM U1083, Faculté de médecine, 49045 ANGERS, France
marie.chabbert@univ-angers.fr

[2] School of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson, TX 75080-3021, USA
herve@ut.dallas.edu

Abstract *The recent resolution of several structures of G-protein-coupled receptors (GPCRs) has led to a breakthrough in the understanding of the structure-function of these receptors that represent major pharmaceutical targets. However, the structure of most GPCRs is not resolved yet and drug design still relies heavily on homology modeling. The quality of a model strongly depends on the capability to determine the best template(s). Here, to gain information on the relative similarity between human GPCRs, we analyze their sequence space by multidimensional scaling (MDS), using the R package bios2mds that we have developed. In the MDS determined sequence space, the receptors cluster into four groups that are characterized by specific proline patterns. The position of the structurally resolved receptors in the GPCR sequence space provides hints for the choice of the best template(s) for molecular modeling. This is exemplified by the molecular modeling of the μ opioid receptor with templates from each group and comparison of the resulting models with the crystal structure of this receptor.*

Keywords GPCR, sequence space, multidimensional scaling, molecular modeling.

1 Introduction

Class A G-protein-coupled receptors (GPCRs) form the largest family of transmembrane receptors in the human genome. They are involved in the cellular response to external stimuli such as hormones and neurotransmitters and in vision, smell and taste. They participate in numerous physiological functions and diseases. They are thus very important pharmaceutical targets, representing up to 25% of available drugs [1]. Recently, the structures of several GPCRs have been resolved, both in the active and inactive states (reviewed in [2,3]). These structures represent a major advance in the understanding of the structure and mechanism of action of GPCRs. However, the structure of most GPCRs has not been resolved yet and drug design still relies heavily on homology modeling. The quality of a molecular model by homology strongly depends on the structural similarity with the template(s). The determination of the best template(s) depends on our understanding of the relationships between the receptors. To gain insight into theses relationships, we analyzed class A GPCRs by multidimensional scaling (MDS) [4]. MDS is a multivariate technique that transforms a distance matrix into points in a low dimensional scale, such as the relative distances between these points best approximate the distance matrix [5]. We show that the sequence space resulting from the MDS analysis provides insights into the main mechanisms that drove the evolution of GPCRs and helps choose template(s) for reliable molecular modeling.

2 Methods

We developed *bios2mds*, an R package that performs multidimensional scaling (MDS) analysis and includes several visualization tools of the resulting sequence space [6,7]. *Bios2mds* builds distance matrices based on either sequence identity or a broad range of amino acid substitution matrices and analyzes them by MDS. Specific functions allow the coloring of the resulting sequence space with user-provided classifications. The *bios2mds* package is available at the CRAN (Comprehensive R Archive Network) under the GNU General Public License. The GPCR sequence space was built from the MDS analysis of a set of 283 aligned sequences of non-olfactory class A GPCRs from *H. sapiens* [4]. Receptors were clustered by both K-means and visual inspection.

Molecular modeling was carried out with the 9v8 version of the MODELLER program [8]. The human μ opioid receptor was modeled by homology with (1) the chemokine receptor CXCR4 (PDB:3ODU), (2) rhodopsin (PDB:1U19), (3) the adrenoceptor β2AR (PDB:2RH1) and the sphingosine 1-phosphate receptor 1, S1PR1 (PDB:3V2Y). For each template, 50 models were generated and refined by simulated annealing, then the model with the lowest Objective Function [8] was selected. The Cα atoms of the transmembrane region of these models were superimposed with the equivalent positions in the crystal structure of the μ opioid receptor (PDB:4DKL) and the root mean square deviations (rmsd) of these atoms computed using the SUPERPOSE function of MODELLER. PYMOL (http://www.pymol.org) was used for molecular graphics analysis and figure preparation.

3 Results and discussion

3.1 The GPCR sequence space

The MDS analysis of a multiple sequence alignement (MSA) allows the visualization of the sequences in a low dimensional space, the so-called sequence space, which corresponds to the first three principal components of the cross-correlated distance matrix derived from the MSA [5]. In the sequence space, the sequences are represented by points whose relative distance best approximates the initial distances in the alignment.

The three-dimensional mapping of human GPCRs indicates that the receptors have a radial organization and cluster along a few privileged directions (Fig. 1). Their distribution leads to a straightforward classification into four groups that are intermediate between the class and the sub-family levels [4]. The central group (G0) is composed of the peptide receptors (PEP), the opsins (OPN) and the melatonin (MTN) receptors. Group G1 includes the somatostatin/opioid (SO), chemotactic (CHEM) and purinergic (PUR) receptors. Group G2 includes the amine (AMIN) and adenosine (AD) receptors. Group G3 includes the glycoprotein hormone (LGR), prostaglandin (PTG) and Mas-related receptor (MRG) sub-families along with the melanocortin, sphingosine-1 phosphate and cannabinoid receptors (MEC).

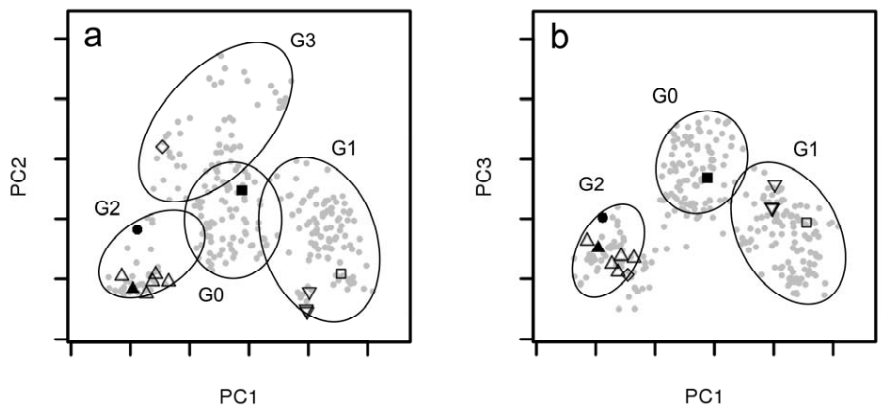


Figure 1. Sequence space of human GPCRs. The receptors are projected onto the planes determined by the first and second components (a) and the first and third component (b) of the sequence space resulting from the MDS analysis. The open symbols indicate receptors with known structures for the inactive state (up triangles: amine receptors, down triangles: opioid receptors, square: CXCR4, diamond: sphingosine-1 phosphate receptor 1). The closed symbols indicate receptors with known structures for both the active and inactive states (up triangle: β2 adrenergic receptor, circle: adenosine 2A receptor, square: rhodopsin). The spanning ellipses indicate the receptor groups obtained by MDS.

A previous phylogenetic study based on maximum parsimony [9] led Fredriksson and coworkers to propose a classification of class A GPCRs into four groups, labeled from α to δ. Comparison of these groups with those obtained by MDS reveals significant differences. The α-group includes the AMIN, AD, MEC, PTG, OPN and MTN receptors, the β-group includes only PEP receptors, the γ-group includes the SO and CHEM receptors whereas the δ-group includes the MRG, LGR and PUR receptors.

3.2 Sequence analysis

Sequence analysis of the GPCR clusters indicates that (1) a proline residue at position 2.58 in transmembrane helix 2 (TM2) is the hallmark of G1 receptors and (2) the absence of proline in both TM2 and TM5 is characteristic of G3 receptors [4]. *Bios2mds* allows the visualization of these patterns in the GPCR sequence space (Fig. 2). The sequence space can thus be interpreted in terms of evolutionary pathways. The first pathway (group G1) corresponds to the differentiation of P2.58 receptors that was initiated by a deletion in TM2 in an ancestor of PEP receptors [10]. This deletion has been experimentally evidenced by the crystal structure of CXCR4 [11] and modifies the bulged structure of TM2 into a kinked structure. Following this deletion, the P2.58 receptors evolved and led to the SO, CHEM and PUR sub-families by divergent evolution [10]. The second pathway (group G2) is related to the differentiation of amine and adenosine receptors. The third pathway (group G3) corresponds to correlated mutations (p-value < 10^{-10}) of proline residues in TM2 and TM5, in independent GPCR sub-families [4]. The correlated mutations of these two residues that are 25 Å apart in the receptor structures indicate long range interactions and suggest that these receptors might have evolved a specific mechanism related to the initial mutation of either the TM2 or the TM5 proline.

This study evidences the major role of proline residues in the evolution of GPCRs and points towards structural and functional constraints, specific of each MDS groups. In addition, this study provides hints to the differences in receptor clustering obtained by the MDS and the maximum parsimony approaches. In particular, some PUR receptors have lost the proline residue at position 5.50 in TM5. The clustering of the PUR receptors with the MRG and LGR receptors observed by Fredriksson [9] might be related to an over importance of this position as an informative site in the maximum parsimony approach. On the other hand, our clustering of the PUR receptors with the SO and CHEM receptors is in agreement with their common P2.58 pattern and with our previous phylogenetic studies of P2.58 receptors [10].

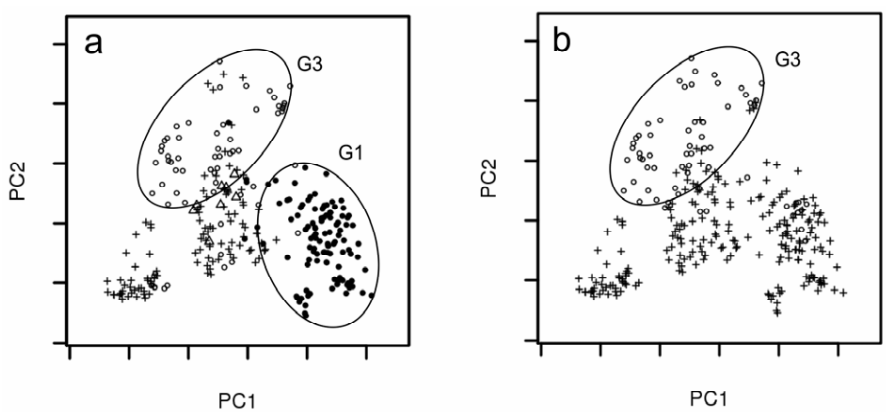


Figure 2. Proline patterns in TM2 (a) and TM5 (b) of human GPCRs. In (a), receptors with a proline residue at position 2.58, 2.59 and 2.60 are indicated by closed circles, crosses and triangles, respectively. Receptors with no proline in TM2 are indicated by open circles. In (b), receptors with and without a proline residue at position 5.50 are indicated by crosses and open circles, respectively.

It must be emphasized that the MDS analysis is based on average sequence identities and not on phylogenetic models based on the evolution of specific sites. Nevertheless, the clustering obtained by MDS, which may result from divergent or convergent evolution, enlightens the importance of specific sequence patterns in GPCR evolution. The group G1 provides an example of divergent evolution, whereas detailed analysis of the group G3 points towards convergent or parallel evolution [4].

3.3 Implications for molecular modeling

The position of the thirteen receptors whose structure is presently resolved [2,3,12] can be visualized in the GPCR sequence space (Fig. 1). Most of these receptors belong to group G2 that includes the amine receptors and the phylogenetically related adenosine receptors. Group G1 has four representatives, three highly related opioid receptors (SO sub-family) and the chemokine receptor CXCR4, whereas groups G0 and G3 have a single representative, the rhodopsin and S1PR1, respectively.

In order to find the templates best adapted to model a receptor with unknown structure, we modeled the recently resolved μ opioid receptor [13] with different templates from each MDS group: rhodopsin, CXCR4, β2AR and S1PR1. Then, we superposed the transmembrane domain of these models with the X-ray structure of the μ opioid receptor and compared the resulting rmsd of the Cα atoms (Table 1).

MDS group	Template	Rmsd (Å)
G0	rhodopsin	2.3
G1	CXCR4	1.8
G2	β2AR	2.8
G3	S1PR1	3.4

Table 1. Rmsd between the crystal structure of the μ opioid receptor and molecular models of the same receptor obtained by homology with different templates. The rmsd was calculated from the positions of the Cα atoms of the transmembrane region.

Clearly, the best model of the μ opioid receptor which belongs to group G1 is obtained by homology modeling with the chemokine receptor CXCR4, which belongs to the same group, as a template. The second best model is obtained with rhodopsin, which is located in the central group G0, as a template. The quality of the models with templates in groups G2 and G3 is significantly lower. In addition, it is worth to note that the rmsd computations are based on the Cα atoms only and do not take into account the deletion of one residue in TM2 observed in receptors belonging to group G1 [10].

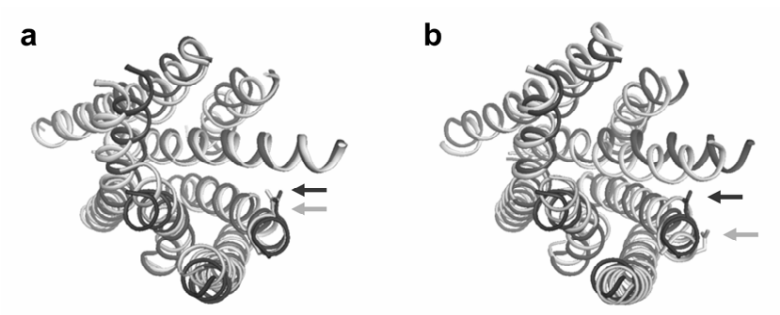


Figure 3. Crystallographic structure of the μ opioid receptor (PDB entry: 4DKL) superimposed on molecular models of the same receptor (dark and light grey, respectively). Models were computed by homology with (a) the chemokine receptor CXCR4 (PDB entry: 3ODU) and (b) rhodopsin (1U19) with MODELLER. The arrows indicate the side chain of Asn-127. Ribbon view of the seven TMH bundle from the extracellular side.

The comparison of the X-ray structure of the μ opioid receptor with the best two models (Fig. 3) shows the superiority of the model obtained with the CXCR4 template as compared to the rhodopsin template, in particular for modeling the minor binding pocket, composed of helices 1-3 and 7 [14]. The C-terminal part of TM2 is correctly modeled with the CXCR4 template, as indicated by the positioning of Asn-127 towards the ligand binding pocket, whereas it is located on the receptor surface with the rhodopsin template. This residue is important for ligand binding [15]. The quality of the model obtained with the rhodopsin template, and other templates with a bulged TM2, might however be improved by modeling a deletion in the TM2 bulge [10,16], in order to avoid the frame shift observed in Fig. 3b.

4 Conclusions

Bios2mds provides the tools necessary to perform the MDS analysis of multiple sequence alignments and to visualize the resulting sequence space. This representation provides useful insights into the evolutionary history of a protein family and its structure-function relationships. Applied to GPCRs, the analysis of the sequence space reveals the major role of the TM2 and TM5 proline residues in the evolution of these receptors with structural and functional implications.

The clustering of GPCRs into four groups helps select templates for molecular modeling. As evidenced by modeling the μ opioid receptor by homology with different templates and comparing with the crystal structure, the choice of a template from the same group clearly improves the quality of the model. Rhodopsin provides the second best choice. This might be related to its central location in the sequence space. Caution is required for modeling G3 receptors. In the crystal structure of S1PR1 [17], TM2 and TM5 are straight and do not show any bulged structure. Whether such a structure is also present in the other receptors from group G3 remains to be determined.

Acknowledgements

This work was supported by institutional grants from CNRS, INSERM and University of Angers and by a grant from the Agence Nationale de la Recherche (ANR-11-BSV2-026). We thank NEC Computers SA (Angers) for the kind availability of a multi-processor server. JP was supported by a fellowship from Conseil Général de Maine-et-Loire. JMB was supported by studentships from CHU of Angers and CNRS.

References

[1] J.P. Overington, B. Al-Lazikani and A.L. Hopkins, How many drug targets are there? *Nat Rev Drug Discov, 5*:993-996, 2006.

[2] V. Katritch, V. Cherezov and R.C. Stevens, Diversity and modularity of G protein-coupled receptor structures. *Trends Pharmacol Sci, 33*:17-27, 2012.

[3] B.K. Shoichet and B.K. Kobilka, Structure-based drug screening for G-protein-coupled receptors. *Trends Pharmacol Sci, 35*:268-272, 2012.

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

[5] H. Abdi, Metric multidimensional scaling, in Salkind, N.J. (ed.), *Encyclopedia of Measurement and Statistics*, Sage, Thousand Oaks (CA), pp 598-605, 2007.

[6] J. Pelé, J.M. Bécu, H. Abdi, and M. Chabbert, *Bios2mds*: an R package for comparing orthologous protein families by metric multidimensional scaling analysis. *BMC Bioinformatics*, *in press*.

[7] J.M. Bécu, J. Pelé, P. Rodien, H. Abdi and M. Chabbert, Structural evolution of G-protein-coupled receptors: A sequence space approach. *Methods Enzymol, in press*.

[8] A. Sali and T.L. Blundell, Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol, 234*:779-815, 1993.

[9] R. Fredriksson, M.C. Lagerstrom, L.G. Lundin and H.B. Schioth, The G-Protein-Coupled Receptors in the Human Genome Form Five Main Families. Phylogenetic Analysis, Paralogon Groups, and Fingerprints. *Mol Pharmacol,*

63:1256-72, 2003.

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

[11] B. Wu, E.Y. Chien, C.D. Mol, G. Fenalti, W. Liu, V. Katritch, R. Abagyan, A. Brooun, P. Wells, F.C. Bi, D.J. Hamel, P. Kuhn, T.M. Handel, V. Cherezov and R.C. Stevens, Structures of the CXCR4 chemokine GPCR with small-molecule and cyclic peptide antagonists. *Science, 330*:1066-1071, 2010.

[12] A.A. Thompson, W. Liu, E. Chun, V. Katrich, H. Wu, E. Vardy, X.P. Huang, C. Trapella, R. Guerrini, G. Calo, B.L. Roth, V. Cherezov and R.C. Stevens, Structure of the Nociceptin/Orphanin FQ receptor in complex with a peptide mimetic. *Nature*, *in press*.

[13] A. Manglik, A.C. Kruse, T.S. Kobilka, F.S. Thian, J.M. Mathiesen, R.K. Sunahara, L. Pardo, W.I. Weis, B.K. Kobilka, S. Granier, Crystal structure of the μ-opioid receptor bound to a morphinan antagonist. Nature, *in press*.

[14] M.M. Rosenkilde, T. Benned-Jensen, T.M. Frimurer and T.W. Schwartz, The minor binding pocket: a major player in 7TM receptor activation. *Trends Pharmacol Sci*, *31*:567-74, 2010.

[15] M. Minami, T. Nakagawa, T. Seki, T. Onogi, Y. Aoki, Y. Katao, S. Katsumata and M. Satoh, A single residue, Lys108, of the delta-opioid receptor prevents the mu-opioid-selective ligand [D-Ala2,N-MePhe4,Gly-ol5]enkephalin from binding to the delta-opioid receptor. *Mol Pharmacol*, *50*:1413-22. 1996.

[16] M. Chabbert, H. Castel, J. Pele, J. Deville, R. Legendre and P. Rodien, Evolution of class A G-protein-coupled receptors: Implications for molecular modeling. *Curr Med Chem*, *19*:1110-8. 2012.

[17] M.A. Hanson, C. B. Roth, E. Jo, M.T. Griffith, F.L. Scott, G. Reinhart, H. Desale, B. Clemons, S.M. Cahalan, S.C. Schuerer, M.G. Sanna, G.W. Han, P. Kuhn, H. Rosen, R.C. Stevens. Crystal structure of a lipid G protein-coupled receptor. *Science*, *17*:851-5. 2012.

Session 2: Regulation

Conférence invitée

Martin VINGRON

Max-Planck-Institut für molekulare Genetik, Germany

Computational Regulatory Genomics

Genome sequence encodes not only genes but also the regulatory relationships among genes. Thus, the time and spatial patterns of gene expression are also encrypted in the DNA sequence. In order to unravel this other genetic code, regulatory genomics attempts to integrate functional genomics data with sequence data. This talk will summarize several approaches developed in our group, starting with a biophysically motivated method for prediction of transcription factor binding sites. Main applications are the identification of tissue specific transcription factors and the prediction of regulatory changes due to SNPs. Further, the talk will describe some indications that the division of promoters into two classes with high and low CpG contents, respectively, is of functional importance and helps in understanding mammalian promoters. In fact, the two classes of promoters display different features when it comes to binding site usage and tissue specific regulation. The dichotomy is further supported by an analysis of histone modifications in the promoters. Taken together, we interpret this as indication that different regulatory mechanisms govern transcription in these two classes of promoters.

Session 3_A : RNA

Development of a pipeline for *Scyliorhinus canicula* miRNA identification from NGS data

Wilfrid CARRÉ[1], Pierre PERICARD[1] , Erwan CORRE[1] , Christophe CARON[1] and Sylvie MAZAN[2]

[1] ABiMS, FR2424 CNRS-UPMC, Station Biologique, Place Georges Teissier, 29680 Roscoff, France
{wilfrid.carre, pierre.pericard, erwan.corre, christophe.caron} @sb-roscoff.fr

[2] Développement et Evolution des Vertébrés, UMR 7150, CNRS-UPMC, Station Biologique, Place Georges Teissier, 29680 Roscoff, France
sylvie.mazan@sb-roscoff.fr

Abstract *Within the context of Dogfish (Scyliorhinus canicula) genome sequencing and the de novo transcriptome analysis, small RNA libraries on different stages and tissues have been made and sequenced. We present here the pipeline developed for the analysis and the identification of the Dogfish microRNAs (miRNAs) from these sequencing data. This study is the first complete miRNA analysis conducted in chondrichthyans, with both NGS and genomic data. The pipeline developed used different miRNA identification strategies in order to evaluate robustness of the prediction.*

Keywords *miRNA, Dogfish, Scyliorhinus canicula, chondrichthyan, NGS*

Le ou les auteur(s) ne souhaite(nt) pas que ce document soit diffusé en ligne

Résumé *Dans le cadre du séquençage du genome de la Roussette (Scyliorhinus canicula) et de l'analyse transcriptomique de novo, des banques de petits ARN à différents stades et tissus ont été réalisées puis séquencées. Nous présentons ici le pipeline développé pour l'analyse et l'identification des microARN (miARN) de la Roussette à partir de ces données de séquençage. Cette étude constitue la première analyse complète de miroARN réalisée chez les chondrichtyens, avec à la fois des données NGS et des données génomiques. Le pipeline mis en place utilise différentes stratégies d'identification de miARN afin d'évaluer la robustesse des prédictions.*

Mots-clés *microARN, Roussette, Scyliorhinus canicula, chondrichtyen, NGS*

1 Introduction

The increasing availability of metazoan genomes gives us a better understanding of evolutive tendencies in the different phylogenetic branches. Within this context, the genome sequencing of the dogfish has been initiated. In parallel, a de novo transcriptomic approach has been carried. The availability of a high coverage and complete chondrichthyan genome will constitute an important step for the reconstruction of the gnathostome ancestral genome.

To complete the transcriptome analysis, small RNA libraries on different stages and tissues have been constructed and sequenced in order to identify dogfish microRNAs (miRNAs). miRNAs constitute a large class of non-coding genes that plays an important role in post-transcriptional regulation of gene expression. The pipeline we have developed here uses different existing approaches to identify miRNAs in order to get the most reliable predictions.

BoostSVM: A miRNA classifier with high accuracy using boosting SVM

Van Du TRAN, Benjamin ZERATH, Sébastien TEMPEL, Farida ZEHRAOUI and Fariza TAHI

IBISC - IBGBI, 23, Bd de France, 91034 EVRY, France
{tvdtran,bzerath,stempel,zehraoui,tahi}@ibisc.univ-evry.fr

miRNAFold: A fast ab-initio method for searching for miRNA precursors in whole genomes

Sébastien TEMPEL[1] and Fariza TAHI[1]

IBISC - IBGBI, 3ème étage, 23, Bd de France 91034 EVRY, France
{sebastien.tempel, fariza.tahi}@ibisc.univ-evry.fr

Abstract *miRNAs are small non coding RNA structures which play important roles in biological processes. Finding miRNA precursors in genomes is therefore an important task, where computational methods are required. With the new generation of sequencing techniques, it is important to have fast algorithms that are able to treat whole genomes in acceptable times. We developed an algorithm, called miRNAFold, based on an original method where an approximation of miRNA hairpins are first searched, before reconstituting the pre-miRNA structure. The approximation step allows a substantial decrease in the number of possibilities and thus the time required for searching. Our method was compared with CID-miRNA, miRPara and VMir. It gives in almost all cases better sensitivity and selectivity. miRNAFold is 60 to 6600 times faster than other methods. miRNAFold is available at http://EvryRNA.ibisc.univ-evry.fr/.*

Keywords miRNA, miRNA precursors, *ab-initio* prediction, miRNA search, hairpins

1 Introduction

MicroRNAs (miRNAs) are non-coding RNAs which are only 21-25 nt in sequence length and are present in all sequenced higher eukaryotes [1]. 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 downregulated [2]. The precursors of miRNA sequences (pre-miRNAs) are structured as a hairpin.

Since the detection of pre-miRNAs by experimental techniques is difficult, expensive and requires a large amount of time, computational methods represent the first step in pre-miRNA identification. These methods can be divided into three approaches: comparative genomics, homology-based approaches and *ab-initio* approaches. Comparative genomics and homology-based approaches cannot detect pre-miRNAs of unknown families and/or pre-miRNAs with no close homologues in genomes. Furthermore, comparative approaches do not work on new genomes that do not have a closely related species sequenced. Ab-initio methods are needed to predict new pre-miRNAs in genomes. Almost all existing *ab-initio* algorithms use an early step secondary structure predictor like RNAFold [4] or UNAFold [5]. Different methods and filters are then applied for predicting pre-miRNAs. We can classify the *ab-initio* methods into three categories:

- Methods that take as input a pre-miRNA candidate sequence and classify it as true or false pre-miRNA.
- Methods that take as input a genomic sequence and some other informations in order to predict (several) pre-miRNAs in the given sequence.
- Methods that are completely *ab-initio*, since they take as input a genomic sequence only (without any other information) and then search for all possible pre-miRNAs occurring in the sequence.

In the first category, we have Triplet-SVM [6], mir-KDE [7], miPred [8] and microPred [9]. Triplet-SVM and miPred are algorithms that classify real and pseudo pre-miRNAs using, respectively, a support vector machine (SVM) and a random forest prediction model. mir-KDE transforms the secondary structure into a vector of 33 features that are then estimated with a relaxed variable kernel density estimator (RVKDE) [10]. microPred classifies human pre-miRNAs using only a SVM with 48 features [9]. In the second category, we have miR-abela [11], and MIReNA [12]. miR-abela predicts new pre-miRNAs that are close to a given known pre-miRNA. In case of MIReNA, it is necessary to enter approximative positions of pre-miRNAs. To our knowledge, there are very few *ab-initio* algorithms of the third category that search for pre-miRNAs without

any given additional information. There are CID-miRNA [13], miRPara [14], miRPred [15], miRANK [16], Virgo [17] and VMir [18]. However, only CID-miRNA, miRPara and VMir are available.

With the new generation of genome sequencing technologies, it is nowadays important to have *ab-initio* automatic methods for quickly analyzing the newly sequenced genomes, and an important aspect of this analysis is the prediction of pre-miRNAs. In this article, we present a new *ab-initio* method (belonging to the third category), called miRNAFold, for predicting pre-miRNAs in any genome. We developed an algorithm that, given a genomic sequence, searches directly for pre-miRNA hairpins occurring in that sequence. It targets more precisely pre-miRNA structures by taking into account their characteristics, in order to (i) better select the true pre-miRNAs and (ii) reduce the search time. The main idea is to first search for a long hairpin stem, which is considered as an anchor allowing to predict the hairpin structure. miRNAFold was tested on an artificial sequence and on several real genomic sequences. It was compared with CID-miRNA [13], miRPara [14] and VMir [18]. We show in this article that our algorithm predicts successfully almost all known pre-miRNAs in genomic sequences of different species. It gives better or at least similar sensitivity and selectivity than CID-miRNA, miRPara and VMir. We also show that our algorithm is very fast; it takes less than 30 s to process a 1 Mb sequence, while VMir takes 30 min, miRPara about 20 h, and CID-miRNA about 55 h.

2 Materials and Methods

2.1 Pre-miRNA features

Our first objective was to find features of pre-miRNAs. For this purpose, we downloaded the last version of miRBase database (Release 17, April 2011) that contains 16 772 pre-miRNAs (10) and we studied the pre-miRNAs contained in this database. We then observed several characteristics:

2.1.1 Pre-miRNA hairpins contain long stems We observed that pre-miRNAs are almost always composed of at least one long exact stem. An exact stem is a couple of subsequences (p, p') such that:

$$\begin{array}{ll} (i) & |p| = |p'| = m \\ (ii) & p[k]\ R_c\ p'[m-k+1], \qquad \forall k, 1 \leq k \leq m \end{array}$$

where R_c is the relation of complementarity between nucleotides: AR_cU, GR_cC and GR_cU. In other terms, an exact stem is a succession of base pairings A-U, C-G and G-U.

We observed that all pre-miRNA hairpins of miRBase [3] have at least one exact stem of length greater or equal to 5. And as we can see in Fig. 1A, the longest exact stem in pre-miRNA hairpins of miRBase is often between 5 and 10 nt.

2.1.2 Pre-miRNA hairpins are symmetric We also observed that most pre-miRNAs either have very few bulges or bulges of one side almost compensate with bulges of the other side (i.e. there is a similar number of nucleotides on both sides of the hairpin from the terminal loop to the extremities). Fig. 1B shows the number of hairpins decreases when the gap increases. In all, 90% of pre-miRNAs have less than 3 nt in excess on one side. In other words, pre-miRNAs do not form a 'curved' hairpin but form an 'almost straight' hairpin.

2.1.3 Pre-miRNA hairpins can be approximated by a non-exact stem In almost all pre-miRNAs, there is a non-exact stem. A non-exact stem is composed of a succession of exact stems separated by symmetrical loops such that the size of each symmetrical loop is less than the length of the exact stems surrounding it. We define a symmetrical loop as follows:

$$\begin{array}{ll} (i) & |p| = |p'| = m \\ (ii) & p[k]\ \overline{R_{nc}}\ p'[m-k+1], \qquad \forall k, 1 \leq k \leq m \end{array}$$

where $\overline{R_{nc}}$ is the relation of non complementarity between nucleotides. The size of a symmetrical loop is the number of unpaired nucleotides on one side of the loop.

A non-exact stem forms an important part of the structure (Fig. 1C). This percentage corresponds to the ratio of the size of the non-exact stem size and the size of the hairpin. More than 75% of pre-miRNAs in miRBase have a non-exact stem that represents at least 40% of their length (Fig. 1C). 91.5% of pre-miRNAs have symmetrical loops whose length ranges from 1 to 3 nt (Fig. 1D).

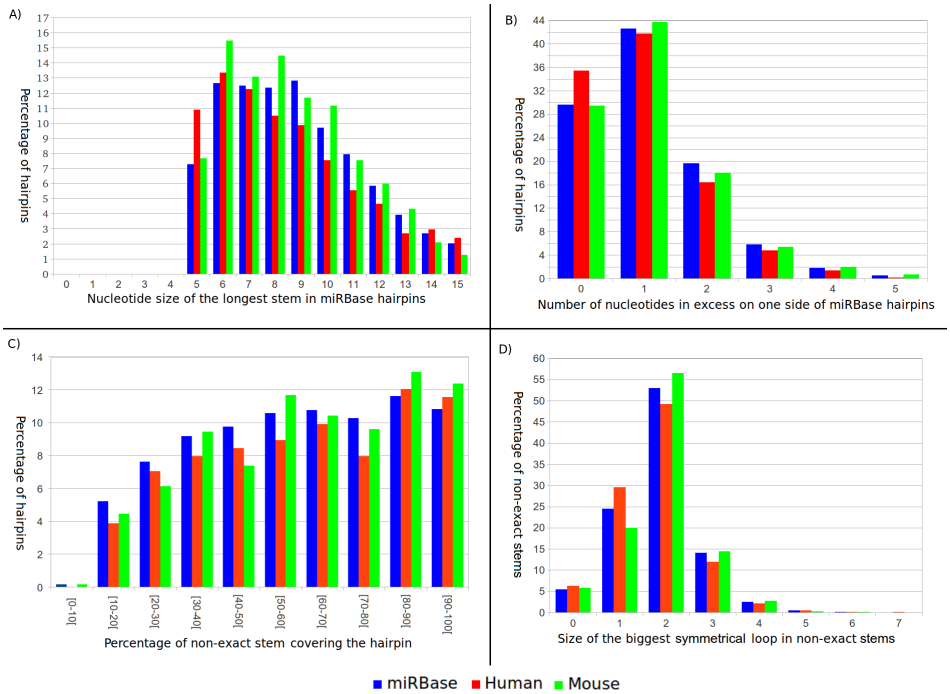


Figure 1. (A) Percentage of pre-miRNA hairpins in human genome, mouse genome and in all miRBase, in function of the length of their longest stem. (B) Percentage of pre-miRNAs having an excess of nucleotides on one side of the hairpin. A gap of zero corresponds to a same number of nucleotides on both sides. (C) Percentage of pre-miRNAs in miRBase in function of the percentage of nucleotides covered by a non-exact stem. (D) Percentage of the biggest symmetrical loop in a non-exact stem from miRBase.

2.1.4 Other pre-miRNA features By studying pre-miRNAs of miRBase, we observed several other characteristics. We split the features into two categories: global characteristics that are present in all species in miRBase and local characteristics that are species dependent. For example, we observed in all species that the longest stems are composed of at least three base pairings and have a ratio of GU base pair always lower than 33.33%. For the species-dependent features, we used some usual characteristics like the hairpin size, the minimum free energy (MFE) and the ratio of A, C, G and U nucleotides. We also calculated some characteristics from Helvik et al. [20] and van der Burgt et al. [21] like the ratio of G-U and G-C base pairings and the ratio of G over C.

2.2 Our approach

Our goal was to develop an algorithm which is able to find efficiently pre-miRNAs in whole genomes in an acceptable time. For this purpose, we adopted the following approach, which was motivated by the different observations we made on miRBase pre-miRNAs.

We consider a sliding window of a given size L sufficiently long to contain a pre-miRNA, in which we search for pre-miRNA hairpins. In a first step, we search for long exact stems that verify some criteria, so they are considered as anchors of possible hairpins. In a second step, we extend the selected stems in order to get the longest non-exact stems verifying some criteria. Each selected non-exact stem can be considered as a good approximation of a pre-miRNA hairpin, and gives the hairpin position. Possible pre-miRNA hairpins are then searched considering the middle position of the non-exact stem as the middle position of the hairpin. Hairpins verifying some criteria are then selected. Thus, our approach consists of three main steps applied on each window subsequence:

1. search for longest exact stems;
2. extend the selected stems and select the longest non-exact stems;
3. predict the secondary structure of the hairpins corresponding to the selected non-exact stems.

At each step, several selection criteria are used, corresponding to several features observed on the pre-miRNA hairpins of miRBase and on their exact stems and non-exact stems. There are 12 criteria for the longest exact stem, 17 criteria for the longest non-exact stem and 26 criteria for the hairpin. Because a pre-miRNA can present some of these features but not all, an exact stem, a non-exact stem or a hairpin is selected when a certain percentage of the criteria are verified. This percentage is a parameter which could be set by the user (set by default to 70%). After the three steps, the sliding window is shifted by 10 nt. The overlapping sequence between two sliding windows allows the algorithm to find the complete secondary structure of the hairpin.

2.3 The algorithm

Given a genomic sequence of any size, for each subsequence delimited by the sliding window, a triangular base pairing matrix M is built such as:

$$M(i,j) = \begin{cases} M(i-1,j-1)+1 & \text{if M(i) and M(j) form a basepair} \\ 0 & \text{otherwise} \end{cases}$$

The algorithm performs then the three following main steps.

2.3.1 Longest exact stem searching Longest stems verifying a certain percentage of criteria are searched for in the matrix. The 10 longest stems verifying a certain percentage of criteria [set by default to 70% (see 'Results' section or given by the user)] are then selected. For example in Fig. 2A, three stems (surrounded by blue) are selected. The 12 exact stem criteria are the size of the exact-stem, the MFE, the size of the terminal loop, the percentage of A, C, G and U, the number of consecutive A, C, G and U and the ratio of GU pairings in the stem.

2.3.2 Longest non-exact stem searching When an exact stem is selected, it is used as an 'anchor' for finding a non-exact stem. The extension of the exact stem is done by considering only the diagonal containing the exact stem. For example, in Fig. 2A, a non-exact stem is indicated by red and corresponds to two exact stems. Once a non-exact stem is extended, the 17 non-exact parameters are calculated: the total length, the number of exact stems composing it, the MFE, the length of the terminal loop, etc. The stem is selected if it meets a certain percentage of these parameters. Each selected non-exact stem is considered as a good approximation of a pre-miRNA hairpin, and gives the hairpin position. Possible pre-miRNA hairpin is then searched for in the subsequence associated to the selected non-exact stem.

2.3.3 Hairpin formation The hairpins are predicted from selected non-exact stems. The anchor of the considered non-exact stem is positioned in the matrix, and then is extended inwards left and right (Fig. 2B, light green areas) on different diagonals, in order to allow here bulges and non-symmetrical internal loops. The hairpins determined from the matrix must verify a set of criteria in order to be selected (26 criteria): length, MFE, size of the terminal loop, ratio of base pair GU and GC, etc. Only hairpins where the percentage of verified criteria is higher than a certain percentage are selected.

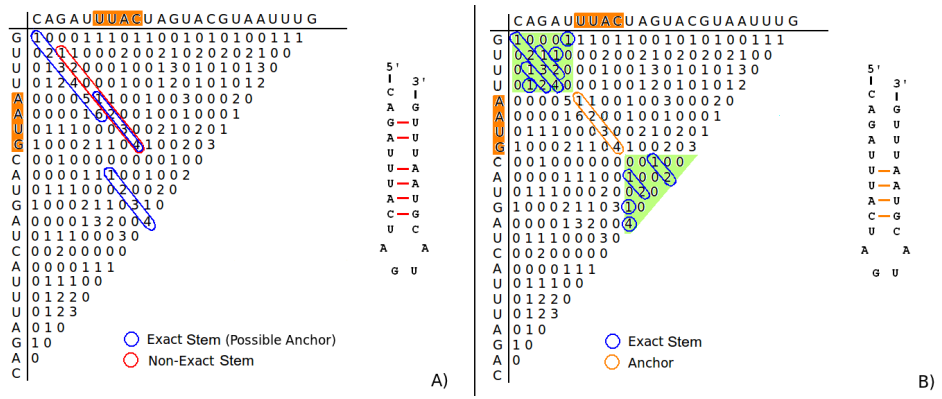


Figure 2. (A) Example of a matrix for searching for exact and non-exact stems in a given genomic subsequence. Three longest stems are selected (surrounded by a blue circle). One of the three stems has been extended to a non-exact stem (surrounded by a red circle). (B) Search for hairpins. The anchor (surrounded by an orange circle) of the non-exact stem shown in (A) is positioned in the matrix, and then is extended in left and in right (green areas) on different diagonals, in order to allow bulges and internal loops.

3 Results and Discussion

3.1 Tested algorithms

We compared our algorithm to existing algorithms of the same *ab-initio* category (the third category as described in Introduction). We considered CID-miRNA, miRPara and VMir for our tests. All have their binaries available and take as input the genome sequence in a FASTA format.

CID-miRNA uses a sliding window for parsing the genomic sequence. On each subsequence delimited by the window, it uses a Cocke-Younger-Kasami (CYK) parser to build the most likely secondary structures and uses a classification tree to determine if the secondary structure is a pre-miRNA. miRPara also uses a sliding window and on each subsequence uses first UNAFold [5] for predicting the secondary structure of the pre-miRNA candidate. miRPara calculates 77 parameters and uses a SVM classifier to select or not the candidate. VMir also uses a sliding window and on each subsequence delimited by the window, performs RNAFold [4] to predict the secondary structures and calculates a score for RNAFold hairpins using several parameters like the size, the number of copies and the number of sliding windows where a same hairpin is detected.

For the three *ab-initio* software CID-miRNA, VMir and miRNAFold, we used a sliding window of 150 nt. We also considered their default parameters. In order to evaluate and compare the tested programs, we used the standard measures of sensitivity and selectivity (specificity). miRNAFold was run with a default threshold of 70% for its parameter of minimum percentage of verified criteria.

In order to evaluate and compare the tested programs, we used the measures of sensitivity and selectivity (specificity). The sensitivity measures the capability of the software to find known pre-miRNAs. The selectivity represents the probability that a predicted hairpin corresponds to a pre-miRNA. The sensitivity and the selectivity are given by the following equations:

$$Sensitivity = 100 \cdot \frac{TP}{TP+FN} \qquad Selectivity = 100 \cdot \frac{TP}{TP+FP}$$

where TP (True Positives) is the number of known pre-miRNAs predicted, FN (False Negatives) is the number of known pre-miRNAs not predicted, and FP (False Positives) is the number of wrong pre-miRNAs predicted.

3.2 Results on an artificial sequence

An artificial sequence was created by the concatenation of human mRNAs and the insertion of 100 human pre-miRNAs. The mRNA sequences came from the Human genome (build 37.2) of the NCBI Website (www.ncbi.nlm.nih.gov) and the pre-miRNAs came from miRBase database (release 17) (www.mirbase.org).

An important parameter of miRNAFold is the minimal percentage of criteria that must be verified at each step of the algorithm. To select this important parameter, we ran miRNAFold on the artificial sequence. The results of sensitivity and selectivity obtained by miRNAFold, CID-miRNA, miRPara and Vmir are given in Table 1. Because the artificial sequence contains 100 pre-miRNAs, the sensitivity shown in Table 1 corresponds to the number of pre-miRNAs correctly predicted (true positives).

	Sensitivity	Selectivity	Time
CID-miRNA	97	11.72	90m49s
miRPara	97	9.7	5m24s
VMir	28	1.32	2m32s
miRNAFold_{60}	**98**	18.96	0.88s
miRNAFold_{70}	97	19.17	0.84s
miRNAFold_{80}	96	**22.91**	**0.76s**

Table 1. Prediction results obtained on an artificial sequence by miRNAFold, CID-miRNA, miRPara and Vmir. miRNAFold was run with different values for the parameter of minimal percentage of verified criteria: 60%, 70% and 80%.

CID-miRNA, miRPara, VMir, miRNAFold_{60}, miRNAFold_{70} and miRNAFold_{80} have predicted respectively 97, 97, 28, 98, 97 and 96% of the real pre-miRNAs. miRNAFold was always more selective than the other methods whatever the considered criteria percentage. For example, miRNAFold_{70} found at least 12 times less false pre-miRNAs than VMir and five times less than miRPara. As expected, the lower the percentage parameter value considered in miRNAFold, the higher the sensitivity and the higher the percentage, the higher the selectivity.

In the following, we set to 70% the default value of this parameter as it represents the percentage giving a good compromise between the sensitivity and the selectivity.

3.3 Results on real data

To test our method, we considered four real genomic sequences in human, mouse, zebrafish and sea squirt genomes. We chose these genomes as each of them present a large cluster of known miRNAs:

- The human chromosome 19 has a cluster of 50 pre-miRNAs (from position 54,169,933 to 54,485,651).
- The mouse chromosome 2 has a cluster of 71 pre-miRNAs (from position 10,388,290 to 10,439,906).
- The zebrafish chromosome 4 has a cluster of 50 pre-miRNAs (from position 34,353,975 to 34,481,435).
- The sea squirt chromosome 7q has a cluster of 46 pre-miRNAs (from position 5,400,066 to 6,168,570).

For each of these genomes, we extracted from NCBI Website the sub-sequence that includes the considered pre-miRNAs cluster.

3.3.1 Sensitivity and selectivity results Table 2 shows the sensitivity and selectivity results obtained with CID-miRNA, miRNAFold, miRPara and VMir in each of the considered sequences.

	Sensitivity				Selectivity			
	Human	Mouse	Zebrafish	Sea squirt	Human	Mouse	Zebrafish	Sea squirt
CID-miRNA	38	29. 58	19.30	28.26	0.69	0.82	0.75	**10.88**
miRPara	98	**98.59**	47.37	58.7	**0,93**	5.34	1.4	5.86
VMir	**100**	88.73	84.21	**100**	0.56	2.93	1.35	5.29
miRNAFold_{70}	**100**	**98.59**	**94.74**	91.30	0.89	**7.71**	**2.60**	7.98

Table 2. Sensitivity and selectivity of CID-miRNA, miRNAFold, miRPara and VMir

miRNAFold has a better or similar sensitivity than other methods in three of the four genomic sequences. CID-miRNA is the least sensitive of the four programs. For the four genomic sequences, the sensitivity of CID-miRNA is lower than 40%. Vmir has the higest sensitivity of the compared methods. Only Vmir has a higher sensitivity than miRNAFold in sea squirt sequence : VMir found all known pre-miRNAs while our algorithm missed four pre-miRNAs (91.30%). miRPara has a good sensitivity in human and mouse sequences but the sensitivity dropped under 50% for the other sequences. Finally, miRNAFold is the only algorithm giving a sensitivity always greater than 90% for all tested sequences. Unlike CID-miRNA and miRPara, it gives homogeneous and stable sensitivity results whatever the genomic sequence.

miRNAFold gives the best results for the mouse and the zebrafish genomes and the second best results for the human and the sea squirt genomes.

To summarize, miRNAFold has better sensitivity and selectivity results than CID-miRNA, miRPara and Vmir on the mouse and zebrafish sequences. In human genomic sequence, miRPara has a slightly better selectivity than miRNAFold, but it has a lower sensitivity. In sea quirt genomic sequence, VMir predicts only four supplementary known pre-miRNAs compared with miRNAFold but VMir also predicted 344 false supplementary hairpins. CID-miRNA has better selectivity in this genome sequence but missed 33 pre-miRNAs when miRNAFold missed only four pre-miRNAs.

3.3.2 Running time With the increase of sequencing of large genomes, the running time is an important evaluation parameter of pre-miRNA searching algorithms. To compare the run time of CID-miRNA, miRNAFold, miRPara, and VMir, we considered subsequences of 1 million of nucleotides from the Human, Mouse, Zebrafish and Sea squirt genomes, respectively, containing the clusters considered above.

Experiments were performed on a Linux machine equipped with an Intel Core Duo 2 T6600 of 2.2 GHz and 4 GB of RAM. The execution time of the three programs on the four sequences is given in Table 3.

	Human	Mouse	Zebrafish	Sea squirt	Average
CID-miRNA	54h58m	54h48m	54h40m	55h29m	55h08m
miRPara	20h12m	19h47m	19h40m	19h25m	19h46m
VMir	30m	30m	30m	30m	30m
miRNAFold_{70}	**0m25s**	**0m22s**	**0m29s**	**0m24s**	**0m25s**

Table 3. Execution time of the algorithms CID-miRNA, miRNAFold, miRPara and VMir for predicting pre-miRNAs in genomic sequences of 1 million of nucleotides each. The values of miRNAFold was rounded to the second. The values of CID-miRNA, miRPara, and Vmir were rounded to the minutes.

miRNAFold is the fastest algorithm. Our average execution time is 25 s for a sequence of 1 million of nucleotides when VMir, the second fastest algorithm, has an average time of 30 min. miRNAFold is almost 60 times faster than VMir, about 2400 times faster than miRPara and 6600 times faster than CID-miRNA.

4 Availability and Implementation

miRNAFold takes in input a genomic sequence in a Fasta format. The size of the sliding window is a parameter which could be set by the user (by default equal to 150 nt). Another important parameter is the percentage of criteria that must be verified at each step of the algorithm. The value of this parameter is set by default to 70% and the user can vary it between 0% and 100%. miRNAFold was implemented using the C++ language. The software can be used through the web server: http://EvryRNA.ibisc.univ-evry.fr.

5 Conclusion

We presented here an original *ab-initio* method called miRNAFold, which allows a fast search for miRNA precursors in genomes. This method first searches for the position of pre-miRNAs by approximating their structure before deducing the final structure. The interest of this first step is to reduce the run time. We obtain better (or similar) sensitivity and selectivity results than other existing methods but with an average running

time at least 60 times faster than the fastest tested algorithm, i.e. Vmir. On the tested sequences, miRNAFold takes less than 30 s for a sequence of 1 million length, when VMir takes 30 min, miRPara takes about 20 h and CID-miRNA more than 55 h. Our method is the only one that permits whole genome analysis.

One of our further work is to optimize and adapt our code for using it on HPC solutions, and more precisely on GPU solutions, in order to make it much faster for whole genomes.

Acknowledgements

This work was supported by the Council of Essonne Region (Pôle System@tic, OpenGPU project).

References

[1] D. Bartel, MicroRNAs: genomics, biogenesis, mechanism and function. *Cell*, 116:281-197, 2004.

[2] Y. Lee, M. Kim, J. Han, K. Yeom, S. Lee, S. Baek and V. Kim, microRNA genes are transcribed by RNA polymerase II. *EMBO J.*, 23:4051-4060, 2004.

[3] S. Griffiths-Jones, H.K. Saini, S. van Dongen and A.J. Enright, miRBase: tools for microRNA genomics. *Nucleic Acids Res.*, 36:D154-D158, 2008.

[4] I.L. Hofacker, W. Fontana, P.F. Stadler, S. Bonhoeffer, M. Tacker, P. Schuster, Fast Folding and Comparison of RNA Secondary Structures. *Monatshefte f. Chemie*, 125:167-188, 1994.

[5] N.R. Markham and M. Zuker, M, DINAMelt web server for nucleic acid melting prediction. *Nucleic Acids Res.*, 33:W577-W581, 2005.

[6] C. Xue, F. Li, T. He, G.P. Liu, Y. Li, X. Zhang, Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics*, 6:310, 2005.

[7] D.T. Chang, C.C. Wang and J.W. Chen, Using a kernel density estimation based classifier to predict species-specific microRNA precursors. *BMC Bioinformatics*, 9:12, 2008.

[8] P. Jiang, H. Wu, W. Wang, W. Ma, X. Sun and Z. Lu, MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Res.*, 35:W339-344, 2007.

[9] R. Batuwita and V. Palade, microPred: effective classification of pre-miRNAs for human miRNA gene prediction. *Bioinformatics*, 25:989-995, 2009.

[10] S.N.G. Kwang Loong and S.K. Mishra, Unique folding of precursor microRNAs: Quantitative evidence and implications for de novo identification. *RNA*, 13:170-187, 2007.

[11] 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.

[12] A. Mathelier and A. Carbone, MIReNA: finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data. *Bioinformatics*, 26:2226-34, 2010.

[13] S. Tyagi, C. Vaz, V. Gupta, R. Bhatia, S. Maheshwari, A. Srinivasan and A. Bhattacharya, CID-miRNA: A web server for prediction of novel miRNA precursors in human genome. *Biochemical and Biophysical Res. Comm.*, 372:831-834, 2008.

[14] Y. Wu, B. Wei, H. Liu, T. Li and S. Rayner, MiRPara: a SVM-based software tool for prediction of most probable microRNA coding regions in genome scale sequences. *BMC Bioinformatics*, 12:107, 2011.

[15] M. Brameier and C. Wiuf, Ab initio identification of human microRNAs based on structure motifs. *BMC Bioinformatics*, 8:478, 2007.

[16] Y. Xu, X. Zhou and W. Zhang, MicroRNA prediction with a novel ranking algorithm based on random walks. *Bioinformatics*, 24:50-58, 2008.

[17] S. Kumar, F.A. Ansari and V. Scaria, Prediction of viral microRNA precursors based on human microRNA precursor sequence and structural features. *Virol J*, 6:129, 2009.

[18] A. Grundhoff, C.S. Sullivan and D. Ganem, A combined computational and microarray-based approach identifies novel microRNAs encoded by human gamma-herpesviruses. *RNA*, 12:733-750, 2006.

[19] T. Joachims, Making large-scale support vector machine learning practical. *MIT Press*, 11:169-184, 1999.

[20] S.A. Helvik, O.J. Snove and P. Saetrom, Reliable prediction of Drosha processing sites improves microRNA gene prediction. *Bioinformatics*, 23:142-149, 2007.

[21] A. van der Burgt, M.W.J.E. Fiers, J.P. Nap and R.C.H.J. van Ham, In silico miRNA prediction in metazoan genomes: balancing between sensitivity and specificity. *BMC Genomics*, 10:204, 2009.

Session 3_B : Protein Structure

Exploring and Predicting Allosteric Communication

Elodie LAINE, Christian AUCLAIR and Luba TCHERTANOV

LBPA, UMR 8113 CNRS - ENS de Cachan, LabEx LERMIT, 61, avenue du Président Wilson, 94235 Cachan, France
{elodie.laine, christian.auclair, luba.tchertanov}@lbpa.ens-cachan.fr

Impact of the D802V Mutation on the Structure and Dynamics of the CSF-1R Tyrosine Kinase Receptor

Priscila DA SILVA FIGUEIREDO CELESTINO[1,2], Elodie LAINE[1] , Pedro Geraldo PASCUTTI[2] and Luba TCHERTANOV[1]

[1] ByMoDyM, LBPA, CNRS, École Normale Superiéure de Cachan, LabEx LERMIT, 61 Av. du Président Wilson, 94235, Cachan cedex, France.

{pdasilva, elodie.laine, luba.tchertanov}@lbpa.ens-cachan.fr

[2] Laboratório de Modelagem e Dinâmica Molecular, Instituto de Biofísica Carlos Chagas Filho – Universidade Federal do Rio de Janeiro, 373 Av. Carlos Chagas Filho, 21941-902, Rio de Janeiro, Brazil.

{pascutti@biof.ufrj.br}

Abstract *Protein kinases are important mediators of signal transduction. The colony stimulating factor-1 receptor (CSF-1R) belongs to the type III receptor tyrosine kinases (RTKs) family, along with PDGF-α and β, KIT and FLT3. Type III RTKs overexpression, mainly caused by point mutations, is associated with several types of inflammatory diseases and human cancers. The activating mutation D816V/H of KIT was shown to alter the conformational equilibrium of the kinase toward the active form and to compromise the efficacy of imatinib, first line treatment against gastrointestinal tumors. The equivalent mutation D802V in CSF-1R also confers imatinib resistance. In this study, we characterized the impact of D802V mutation on the structure and dynamics of CSF-1R by molecular dynamics simulations. The comparative analysis of our data with those obtained for KIT D816V/H mutants allowed us to contribute to establish a sequence-structure-dynamics-function relationship for type III RTKs.*

Keywords molecular dynamics, tyrosine kinase, oncogenic mutation, resistance

Impact de la Mutation D802V sur la Structure e la Dynamique du Récepteur Tyrosine Kinase CSF-1R

Mots-clés dynamique moléculaire, tyrosine kinase, oncogénique mutation, résistance

1 Introduction

Receptor tyrosine kinases (RTKs) are involved in many signaling processes essential for organism development and adult homeostasis. CSF-1R is the cell surface receptor for the macrophage colony-stimulating factor-1 (CSF-1) [1] and it is part of RTK type III family, that also includes PDGFR-α and β, the stem cell factor receptor (KIT) and the FMS-like tyrosine kinase 3 (FLT3) [2]. RTKs III share a common fold that includes an extracellular portion to which polypeptide ligands bind, a single-pass transmembrane helix and an intracellular juxtamembrane region (JMR), followed by a tyrosine kinase domain (TKD).

The TKD domain is composed of a two-lobe structure: N-lobe, consisted of a twisted five-stranded anti-parallel β sheet adjacent of an α helix (αC) and the C-lobe, predominantly α-helical. Located on the C-lobe, the activation loop (A-loop) is a long flexible polypeptide that contains two β strands when in its inactive state. The A-loop N-terminal end contains three conserved residues (Asp-Phe-Gly) that make alternative interactions within the TKD depending on its "on" (active) or "off" (inactive) switch states, which is achieved by rotation of the main chain φ torsion angle of the conserved Asp. In the inactive form, the DFG motif is blocking the ATP binding site and the A-loop is folded back over the C-lobe. Also, the side chain of the A-loop tyrosine is enclosed into the kinase active site and binds as a pseudo substrate, making hydrogen bonds with highly conserved residues, maintaining the inactive conformation [3,4,5].

The mechanism of kinase autoinhibition by the JMR domain is indeed particular for the type III RTKs, distinct from other kinase structures [6]. The autoinhibited crystallographic structures of KIT, CSF-1R and

FLT3 have revealed that the JMR makes extensive interactions with the TKD. Its N-terminal end occupies the area of the extended A-loop in the active conformation, hence blocking the movement of the DFG motif and the A-loop to its active form position. Comparative analysis of the available structures of KIT in its inactive and active states [4] evidenced that kinase activation involves extensive rearrangement of the A-loop, JMR and the αC helix. Upon phosphorylation of tyrosine residues on the JMR and the A-loop, the restraints would be released and JMR departure would allow the DFG and the A-loop to reach its active open conformation, accessible for ATP and protein substrates binding [6].

Point mutations identified in the A-loop, lead to constitutive activation of the receptors, altering the conformational equilibrium of the kinase toward the active form, which in turn, compromises the efficacy of the inhibitors that target the inactivate state [7]. For instance, mutations of D835 in FLT3 have been reported in 7% of acute myelogenous leukemia cases [8] and KIT D816V/H mutations were identified clinically in patients resistant to imatinib and sunitinib treatment, first and second-line treatment drugs for advanced gastrointestinal tumors [9]. The transforming mutation D802V in CSF-1R also confers resistance to imatinib [10]. Although CSF-1R is not commonly mutated in cancer [11], CSF-1R overexpression has been associated in a number of chronic inflammatory diseases and several types of human cancers [12-17]. The deleterious effect of the Asp substitutions may derive from the ability of this negatively charged residue to stabilize a small positively charged α-helical dipole. Consequently, nonpolar substitutions would destabilize this helical structure and could promote the transition of the A-loop into its active structure [18].

Recently, Laine et al. [19] and Chauvot de Beauchene et al. [20] have characterized the long range effect of the D816V/H mutations on KIT receptor, manifested in the form of a structural reorganization of the JMR that facilitates its departure from the TKD. The present study aimed at characterizing the role of the equivalent mutation D802V on the structure and dynamics of the CSF-1R. These comparative analyses contribute to the understanding of (i) the auto-activation process of type III RTK and (ii) the resistant mutation structural and functional role.

2 Methods

Wild type (WT) and D802V mutant models were built using Modeller [21] based on the crystallographic structure of the autoinhibited CSF-1R (PDB ID: 2OGV) in order to reconstruct some missing parts of the crystal and also obtain an accurate side chain orientation of the Valine substitution, in the case of the mutant. After, we performed molecular dynamics simulations (MD) of the WT and the D802V mutant forms of CSF-1R. Two independent runs were carried out for 50 ns for each form of the receptor using the parameter set 99SB from the AMBER force field [22] inside the GROMACS [23] package, version 4.5. Principal Component Analysis (PCA) and hydrogen bonds prevalence calculations were performed with the software tools included in GROMACS. Single point free energy calculations were performed on the equilibrated conformations of the two simulated protein forms according to the MM-GBSA method implemented in AMBER 9 [24].

3 Results and discussion

Our simulations first showed a local perturbation pattern on the A-loop of the D802V mutant (Fig. 1A), manifested in the form of a structural destabilization of the small positively charged helical dipole (residues 803-806) that contribute to the stabilization of the A-loop inactive conformation. Such perturbation was also put in evidence in KIT D816V/H mutants [19-20]. We also observed a two times reduction on the hydrogen bond prevalence between Y809 (A-loop) and D778 (catalytic loop), a key interaction that was also weakened in KIT D816V mutant [19]. In addition, PCA analysis of the trajectories indentified a high amplitude mode related to the unfolding of the A-loop (Fig. 1B). These observations indicate that the local impact of the mutation D802V on CSF-1R structure and dynamics is similar to that of the mutations D806V/H on KIT.

Although CSF-1R and KIT TKD show great structural similarity [5], the JMR sequence diverges between the two proteins (Fig. 1C). This difference is reflected in the JMR secondary structure of the wild type receptors: CSF-1R has larger beta-sheets (2 extra residues on the first beta-sheet), compared to KIT. Data obtained for KIT D816V/H mutants [19-20] showed that JMR becomes well-structured and it departs from the C-lobe of the protein towards to an axial position, whereas the structure of the JMR remains rather floppy in WT. By contrast in our simulations, the JMR well-ordered secondary structure remains highly stable for

both the WT and D802V mutant proteins. Also, the position of the JMR is similar in the two forms. Comparing KIT and CSF-1R JMR electrostatic potential profiles, we observed that the JMR residues are more complementary to the surface of the TKD in CSF-1R than in KIT, since CSF-1R has a more negative electrostatic surface and contains more positive charged residues on the JMR (Fig. 1C, 1D).

CSF-1R single point free energy calculations of the binding of the JMR to the TKD showed that the difference between D802V and WT is +33.44 kcal/mol. This result clearly indicates that the affinity of the JMR is unfavorable in the mutant, compared to the wild type. A comparative detailed analysis of the non-covalent contacts between the TKD and two JMR elements, namely the buried region (JM-B) and the switch motif (JM-S), showed that several hydrogen bonds established with the N-lobe are weakened upon mutation, while the interactions with the C-lobe were maintained. Consequently, although we do not observe major structural rearrangement nor repositioning of the CSF-1R JMR upon D802V mutation, the attachment of the JMR to TKD is weakened in the CSF-1R mutant compared to the wild type, as was observed for KIT. The different nature of the JMR in the two receptors could explain that the effects of the D802V mutation on CSF-1R structure and dynamics are more subtle that those of KIT D816V/H mutations.

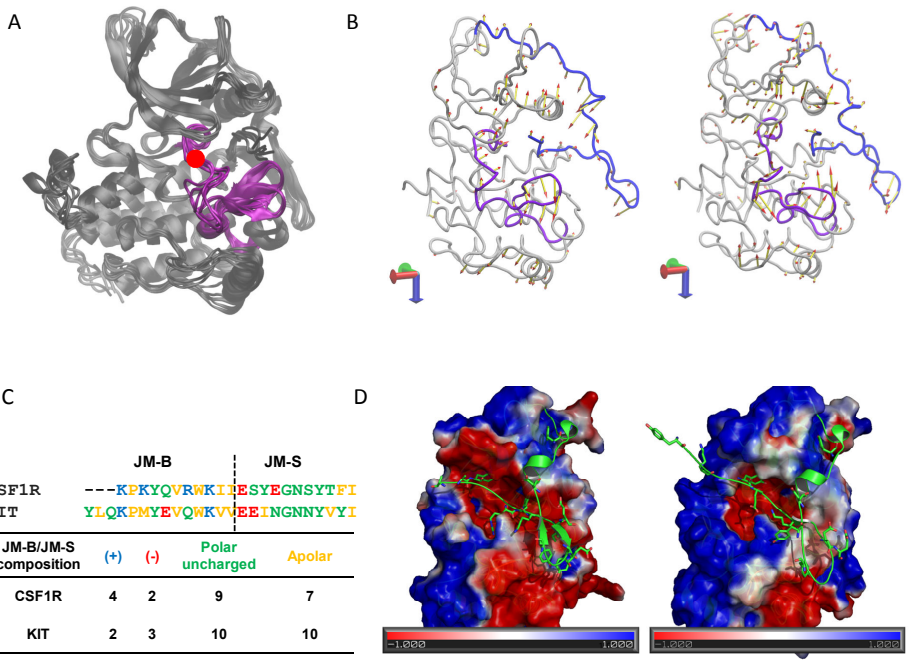


Figure 1. Molecular data of CSF-1R and KIT. (A) Superposition of consecutive trajectory frames from the CSF1F D802V simulations. The A-loop is highlighted in pink with the point mutation in red. (B) PCA analysis derived from the simulations trajectories for the WT (left) and D802V (right) forms of receptor. The highest amplitude mode is represented in both cases. (C) Sequence alignment of the JM-B and JM-S portions of CSF-1 and KIT receptors, based on the crystal structures, colored by the aminoacids properties, with a description of the JM-B and JM-S residue composition on the table below. (D) Electrostatic potential surface of CSF1R (PDB ID: 2OGV) and KIT (PDB ID: 1T45) TKD are shown on the left and on the right, respectively. JMR polar residues are represented as green sticks in both receptors.

4 Conclusions and perspectives

Our results indicate that the sequence variability of the JMR in type III RTKs is associated with differential impacts of the equivalent resistant mutations on the structure and dynamics of the receptors. The more subtle effects of D802V mutation on the JMR of CSF-1R compared to D816V/H of KIT could be the sign of a less efficient activating role of CSF-1R D802V compared to KIT D816V/H. No data are available for *in vitro* WT and D802V CSF-1R activation rate. Further analysis should be performed to investigate the role of the mutation in the resistance to imatinib.

MD simulations on the WT and D835V FLT3 structures are currently in progress. Our future goal is to gather the results for these three members of RTK III in order to relate the differences in sequence and structure with the dynamics and function of this receptor family. Also, this work will provide insights into the development of novel rational drug-design strategy for cancer treatment.

References

[1] C. J. Sherr, The Fms Oncogene. *Biochimica Et Biophysica Acta*, 948(2): 225-243, 1988.

[2] D. R. Robinson, Y. M. Wu and S.F. Lin, The protein tyrosine kinase family of the human genome. *Oncogene*, 19(49): p. 5548-5557, 2000.

[3] J. Griffith, J. Black, C. Faerman, L. Swenson, M. Wynn, F. Lu, J. Lippke, and K. Saxena, The structural basis for autoinhibition of FLT3 by the juxtamembrane domain. *Molecular Cell*, 13(2):169-178, 2004.

[4] C. D. Mol, D. R. Dougan, T. R. Schneider, R. J. Skene, M. L. Kraus, D. N. Scheibe, G. P. Snell, H. Zou, B. C. Sang and K. P. Wilson, Structural basis for the autoinhibition and STI-571 inhibition of c-Kit tyrosine kinase. *Journal of Biological Chemistry*, 279(30): 31655-31663, 2004.

[5] M. Walter, I. S. Lucet, O. Patel, S. E. Broughton, R. Bamert, N. K. Williams, E. Fantino, A. F. Wilks and J. Rossjohn, The 2.7 angstrom crystal structure of the autoinhibited human c-Fms kinase domain. *Journal of Molecular Biology,* 367(3): 839-847, 2007.

[6] M. Huse and J. Kuriyan, The Conformational Plasticity of Protein Kinases. *Cell*, 2:275-282, 2002.

[7] K. S. Gajiwala, J. C. Wu, J. Christensen, G. D. Deshmukh, W. Diehl, J. P. DiNitto, J. M. English, M. J. Greig, Y. A. He, S. L. Jacques, E. A. Lunney, M. McTigue, D. Molina, T. Quenzer, P. A. Wells, X. Yu, Y. Zhang, A. H. Zou, M. R. Emmett, A. G. Marshall, H. M. Zhang and G. D. Demetri, KIT kinase mutants show unique mechanisms of drug resistance to imatinib and sunitinib in gastrointestinal stromal tumor patients. *Proceedings of the National Academy of Sciences of the United States of America*, 106(5):1542-1547, 2009.

[8] Y. Yamamoto, H. Kiyoi, Y. Nakano, R. Suzuki, Y. Kodera, S. Miyawaki, N. Asou, K. Kuriyama, F. Yagasaki, C. Shimazaki, H. Akiyama, K. Saito, M. Nishimura, T. Motoji, K. Shinagawa, A. Takeshita, H. Saito, R. Ueda, R. Ohno and T. Naoe, Activating mutation of D835 within the activation loop of FLT3 in human hematologic malignancies. *Blood*, 97(8): 2434-2439, 2001.

[9] M. J. Frost, P. T. Ferrao, T. P. Hughes and L. K. Ashman, Juxtamembrane mutant V560GKit is more sensitive to Imatinib (STI571) compared with wild-type c-kit whereas the kinase domain mutant D816VKit is resistant. *Molecular Cancer Therapeutics*, 1(12): 1115-1124, 2002.

[10] J. R. Taylor, N. Brownlow, J. Domin and N. J. Dibb, FMS receptor for M-CSF (CSF-1) is sensitive to the kinase inhibitor imatinib and mutation of Asp-802 to Val confers resistance. *Oncogene*, 25: 147-151, 2006.

[11] J. T. Reilly, Class III receptor tyrosine kinases: Role in leukaemogenesis. *British Journal of Haematology*, 116(4): 744-757, 2002.

[12] I. K. Campbell, M. J. Rich, R. J. Bischof and J. A. Hamilton, The colony-stimulating factors and collagen-induced arthritis: exacerbation of disease by M-CSF and G-CSF and requirement for endogenous M-CSF. *Journal of Leukocyte Biology*, 68(1): 144-150, 2000.

[13] T. Rajavashisth, J. H. Qiao, S. Tripathi, J. Tripathi, N. Mishra, M. Hua, X. P. Wang, A. Loussararian, S. Clinton, P. Libby and A. Lusis, Heterozygous osteopetrotic (op) mutation reduces atherosclerosis in LDL receptor-deficient mice. *Journal of Clinical Investigation*, 101(12): 2702-2710, 1998.

[14] N. Kirma, R. Luthra, J. Jones, Y. G. Liu, H. B. Nair, U. Mandava and R. R. Tekmal, Overexpression of the colony-stimulating factor (CSF-1) and/or its receptor c-fms in mammary glands of transgenic mice results in hyperplasia and tumor formation. *Cancer Research*, 64(12): 4162-4170, 2004.

[15] E. Y. Lin, V. Gouon-Evans, A.V. Nguyen and J. W. Pollard, The macrophage growth factor CSF-1 in mammary gland development and tumor progression. *Journal of Mammary Gland Biology and Neoplasia*, 7(2):147-162, 2002.

[16] H. O. Smith, P. S. Anderson, D. Y. S. Kuo, G. L. Goldberg, C. L. Devictoria, C. A. Boocock, J. G. Jones, C. D. Runowicz, E. R. Stanley and J. W. Pollard, The Role of Colony-Stimulating Factor-1 and Its Receptor in the Etiopathogenesis of Endometrial Adenocarcinoma. *Clinical Cancer Research*, 1(3): 313-325, 1995.

[17] B. M. Kacinski, CSF-1 and its receptor in breast carcinomas and neoplasms of the female reproductive tract. *Molecular Reproduction and Development*, 46(1): 71-74, 1997.

[18] N. J. Dibb, S. M. Dilworth and C. D. Mol, Switching on kinases: oncogenic activation of BRAF and the PDGFR family. *Nature Reviews Cancer*, 4:718-727, 2004.

[19] E. Laine, , I. C. de Beauchene, D. Perahia, C. Auclair and L. Tchertanov, Mutation D816V Alters the Internal Structure and Dynamics of c-KIT Receptor Cytoplasmic Region: Implications for Dimerization and Activation Mechanisms. *Plos Computational Biology*, 7(6), 2011

[20] I. C. de Beauchene, E. Laine, C. Auclair, L. Tchertanov, Structural, dynamic and thermodynamic effects of KIT mutations: a computational multi-approach study. *8th EBSA European Biophysics Congress*, Budapest, pp. 103-103, 2011.

[21] M. A. Marti-Renom, A. C. Stuart, A. Fiser, R. Sanchez, F. Melo and A. Sali, Comparative protein structure modeling of genes and genomes. *Annual Review of Biophysics and Biomolecular Structure*, 29: 291-325, 2000.

[22] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg and C. Simmerling, Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins-Structure Function and Bioinformatics*, 65(3): 712-725, 2006.

[23] D. Van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, GROMACS: Fast, flexible, and free. *Journal of Computational Chemistry*, 26(16): 1701-1718, 2005.

[24] D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang and R. J. Woods, The Amber biomolecular simulation programs. *Journal of Computational Chemistry*, 26(16): 1668-1688, 2005.

Structural determinants of Raltegravir specific recognition by the HIV-1 Integrase

Rohit Arora and Luba Tchertanov

BiMoDyM, LBPA, CNRS -ENS de Cachan, LabEx LERMIT, 61 Avenue du Président Wilson 94235 Cachan, France

{rohit.arora,luba.tchertanov}@lbpa.ens-cachan.fr

Abstract *The Raltegravir (RAL) induced mutations in the HIV-1 Integrase can be attributed to its structure and conformational flexibility. To study this phenomenon, we 1) characterized the structural and conformational properties of RAL; 2) tried to understand the main factors contributing to RAL recognition by the targets - IN, IN/vDNA complex, vDNA, and their impact to the inhibition mechanism. We aim to establish the factors that contribute to recognition processes: flexibility and structural features (H-bond donors and acceptors, chelating center and complementarities with the targets) of RAL and flexibility of target(s).*

Keywords molecular modeling, dynamic simulations, docking, HIV-1 IN, vDNA, Raltegravir

1 Introduction

Integrase (IN) is a key enzyme in the HIV replication process. It catalyses the covalent insertion of the viral DNA (vDNA) into the host genome. IN represents an attractive target for antiretroviral drugs and has been the object of intensive pharmacological research. The IN inhibitors were developed to block either 3'-Processing (3'-P) or Strand-Transfer (ST) reactions – the two steps of integration process. Raltegravir (RAL), the first IN inhibitor approved for AIDS treatment, specifically inhibits the ST activity. Like other antiretroviral inhibitors, RAL develops a resistance effect which was associated with amino acid substitutions following three distinct pathways that involve either the N155H, either Q148R/H/K or Y143H/R/C primary mutation [1, 2]. We have previously proposed a possible mechanism of RAL inhibition and the basic principles of RAL-induced resistance [3].To contribute further in the understanding of inhibition mechanisms and mechanisms of resistance to antiretroviral inhibitors, we studied (i) conformational features of RAL that shows a high configuration ability and conformational flexibility and (ii) RAL recognition by the viral targets – IN and vDNA.

2 Methods

The intrinsic conformational features of RAL have been investigated in the gas phase, in the water solution and the solid state. For the gas phase, the *ab initio* calculations were performed at the Hartree Fock (HF) level of theory using *Jaguar* application of Maestro package [4]. Geometry optimization and energy calculations were carried out for all RAL configurations (Figure, Top). Following the geometry optimization, the RAL conformations were generated by relaxed scan of all four RAL configurations. Fragment-based analysis using the crystallographic data from the Cambridge Structural Database (CSD) was performed to characterize RAL structure in the solid state. To study RAL in water solution, explicit Molecular Dynamics (MD) simulations were performed using GROMACS 4.5 package [5] over 3 trajectories encompassing 10 ns each.

Further, RAL in the different conformational and configurational states was docked onto the 3D models representing IN and vDNA targets before and after 3'-P reaction, which were generated as we reported previously [6].

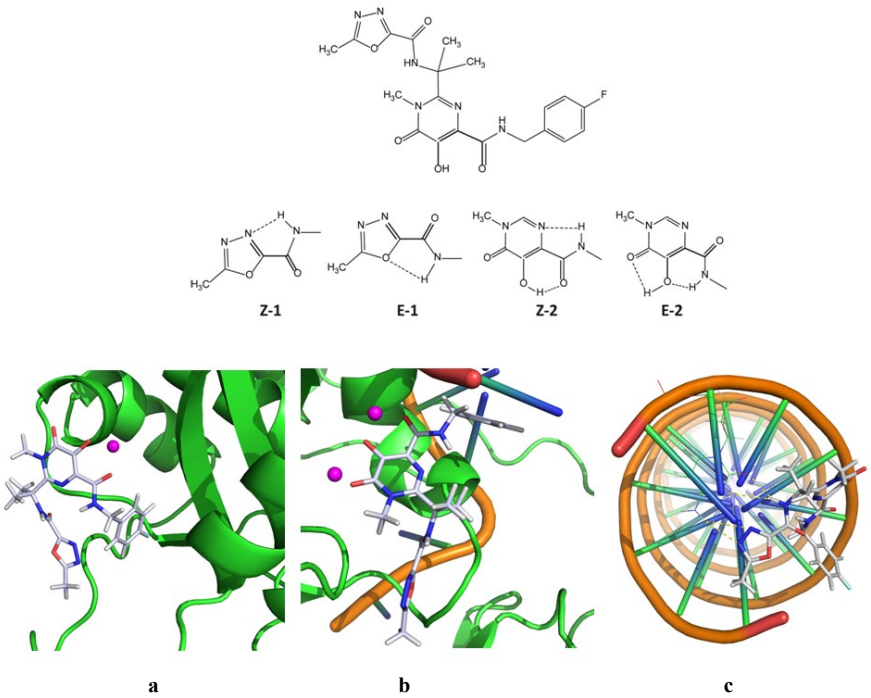


Figure. The structure and binding of RAL. Top: E- and Z-isomerism of two RAL pharmacophores stabilized by intramolecular H-bonds. Bottom: RAL binding obtained by docking of inhibitor into **(a)** the HIV-1 IN, **(b)** IN•vDNA complex and **(c)** the cleaved vDNA.

3 Results

RAL represents a highly flexible polydentade ligand based on two pharmacophores, capable to hit more than one target in HIV-1. Like the other IN inhibitors of the ST reaction, RAL was designed to bind IN by chelating the metal cation or cations present in the active site. Indeed, Mg^{2+} plays an important role in the catalysis mediated by integrase, and this has led to the incorporation of Mg-chelating functions into the rational design of IN inhibitors. These inhibitors, mainly based on a β-ketoenol fragment, bind to the IN by chelating the metal cation or cations present in the active site [7]. Many compounds of this family effectively inhibit the retroviral replication blocking specifically the ST reaction, and consequently are the INSTIs [8]. RAL, the unique integrase specific clinically used drug, is a member of INSTI family [9].

The structural fragment-based analysis indicated a higher probability of RAL, in solid state, to be in the E, Z-configuration that is stabilized by strong intramolecular H-bonding. Our study of various RAL isomers by the energy optimization and relaxed scan evidenced that the relative stability of the four RAL isomers may be represented as follows: Z-1/Z-2 ≥ E-1/Z-2 >> E-1/E-2 >> Z-1/E-2. These results indicate that the Z/Z configuration of RAL is the more energetically favorable in the gas phase with only small difference between the Z/Z and E/Z isomers. Probing the RAL chelating functions taking into consideration the isomerism effect we evidenced that carbonylamino hydroxypyrimidinone in the Z configuration is the best chelating agent respective to metal cations. This finding agrees well with the X-ray data, representing RAL complexed with the Foamy virus intasome [10]. Explicit MD simulations of Z/Z-isomer of RAL revealed a high conformationa

Molecular docking of RAL in the distinct conformations onto different viral targets (the unbounded IN he vDNA and the IN•vDNA complex) evidenced that the RAL binding with the IN•vDNA complex is energetically more favorable and stable as compared to RAL interaction with the unbound IN. Consequently he IN•vDNA is the most appropriate RAL target (Figure 1 b). The docking of RAL onto the uncleaved and the cleaved vDNA (the terminal GT nucleotides at 3'-end were removed) indicates to a specific RAL recognition by the cleaved vDNA. We evidenced that RAL docked onto uncleaved vDNA has no contacts with vDNA and t is positioned in the minor groove. In contrast, RAL docked onto the cleaved vDNA is located in the region o he removed dinucleotides (Figure 1 c). The best score corresponds to the Z/Z isomer of RAL with all chelating centres oriented outside of the vDNA helix. Such RAL orientation is stabilized by the two parallel strong H bonds with the unpaired cytosine (C), characterising high and specific affinity of the RAL – C recognition. This ype of interactions is identical to stabilizing interactions in the bases pair G-C. Consequently, RAL play the role of removed guanidine.

Our *in silico* study describes the structure of the unbound RAL in the different states – in the gas phase, in the solution and in the solid state, – and represents the first characterization of the RAL configurational and conformational features. Description of RAL binding with the putative viral targets – the unbounded IN, the viral DNA and the IN•vDNA complex – contributes in further understanding of the RAL inhibition mechanism and allows to distinguish between the different steps in specific recognition of RAL by the viral targets.

We speculate that in the initial step of inhibition, RAL is recognized specifically by the cleaved vDNA Further, the tightly bonded RAL•vDNA complex associates with IN to form an inhibited RAL•vDNA•IN complex which was characterized for the prototype foamy virus (PFV) IN complex.

Acknowledgements

This work was supported by the Ecole Normale Supériere (ENS) de Cachan, the Centre National de la Recherche Scientifique (CNRS) and the funding by SIDACTION

References

[1] Delelis, O., Malet, I., Na, L., Tchertanov, L., Calvez, V., Marcelin, A. G., Subra, F., Deprez, E., & Mouscadet, J. F. (2009). The G140S mutation in HIV integrases from raltegravir-resistant patients rescues catalytic defect due to the resistance Q148H mutation. *Nucleic Acids Research 37*, 1193-1201, doi:DOI 10.1093/nar/gkn1050.

[2] Delelis, O., Thierry, S., Subra, F., Simon, F., Malet, I., Alloui, C., Sayon, S., Calvez, V., Deprez, E., Marcelin, A. G.,Tchertanov, L., & Mouscadet, J. F. (2010a). Impact of Y143 HIV-1 Integrase Mutations on Resistance to Raltegravir In Vitro and In Vivo. *Antimicrobial Agents and Chemotherapy 54*, 491-501, doi:DOI 10.1128/AAC.01075-09.

[3] Mouscadet, J. F., Arora, R., Andre, J., Lambry, J. C., Delelis, O., Malet, I., Marcelin, A. G., Calvez, V., & Tchertanov, L. (2009a). HIV-1 IN alternative molecular recognition of DNA induced by raltegravir resistance mutations. *Journal of Molecular Recognition 22*, 480-494, doi:DOI 10.1002/jmr.970.

[4] Maestro, version 9.1, Schrödinger, LLC, New York, NY, 2010.

[5] Hess, B., Kutzner, C., van der Spoel, D. & Lindahl, E. (2008). GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *Journal of Chemical Theory and Computation 4* (3), 435–447, doi:DOI 10.1021/ct700301q

[6] Ni, X., Abdel-Azeim, S., Laine, E., Arora, R., Osemwota, O., Marcelin, A.-G., Calvez, V., Mouscadet, J.-F. & Tchertanov, L. In silico and in vitro Comparison of HIV-1 Subtypes B and CRF02_AG Integrases Susceptibility to Integrase Strand Transfer Inhibitors. (Accepted in *Advances in Virology*)

[7] Kawasuji, T., Fuji, M., Yoshinaga, T., Sato, A., Fujiwara, T., and Kiyama, R. (2006) A platform for designing HIV integrase inhibitors. Part 2: a two-metal binding model as a potential mechanism of HIV integrase inhibitors. *Bioorg.Med.Chem*. 14(24), 8420-8429.

[8] Espeseth AS, Felock P, Wolfe A, Witmer M, Grobler J, Anthony N, Egbertson M, Melamed JY, Young S, Hamill T, Cole JL, Hazuda DJ.(2000) HIV-1 integrase inhibitors that compete with the target DNA substrate define a unique

[9] Grinsztejn, B., Nguyen, B. Y., Katlama, C., Gatell, J. M., Lazzarin, A., Vittecoq, D., Gonzalez, C. J., Chen, J., Harvey, C. M., and Isaacs, R. D.(2007) Safety and efficacy of the HIV-1 integrase inhibitor raltegravir (MK-0518) in treatment-experienced patients with multidrug-resistant virus: a phase II randomised controlled trial. *Lancet* 369(9569), 1261-1269.

[10] Hare, S.,Vos, AM.,Clayton, RF.,Thuring, JW.,Cummings, MD. & Cherepanov, P (2010). Molecular mechanisms of retroviral integrase inhibition and the evolution of viral resistance. *Proc Natl Acad Sci USA* 107(46), 20057-20062, doi:DOI 10.1073/pnas.1010246107

Session 3_C : Genes

Importance of family size and function in the fate of duplicated genes in the protein-protein interactome of *Arabidopsis thaliana*

Justin WHALLEY[1], Etienne BIRMELÉ[1] and Carène RIZZON[1]

[1] LABORATOIRE Statistique and genome, UMR8071 CNRS, 23 bvd de France, 91037, France
Justin.whalley@genopole.cnrs.fr

Abstract. Gene duplication is readily accepted as a primary mechanism for generating organismal 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 because their interacting context is taken into account. Such network data, especially protein-protein interactions (PPIs) can now be generated in a high throughput manner in different organisms, notably for *Arabidopsis thaliana* which is an excellent model organism for studying the evolution of duplicate genes.

Results from yeast protein interaction networks suggest a rapid evolution of the interactions with the probability of an interaction to get lost getting higher with time since the duplication event [2]. Wagner's model predicts also that duplicates loss interactions with sequence divergence. Some results are not in agreement with this model: duplicability, that is the proportion of the number of unique types of gene constituted of duplicated genes in the total number of unique types of genes, and its link with connectivity can be different according to the organism. For example, human hubs -that are highly connected genes- and mouse essential proteins that are involved in development are preferentially encoded by duplicated genes, while other categories of essential mouse genes can be both singletons and duplicates [3 for review]. Moreover, in *E. coli*, yeast, fly and human it has been shown that the origin and conservation of a gene significantly correlates with properties of the encoded protein in the PPI network [3]. It is well known that duplicate genes are biased in function compared to single genes [1 for review]. In some vertebrate and invertebrate genomes, it has been shown that family size is linked with gene function [4]. In yeast protein connectivity and duplicability vary with gene function [5].

Our purpose is to tackle the following question: are the protein-protein interactions driving the evolution of duplicate genes in a time-dependent manner in *A. thaliana*, as in yeast [2]? Can the maintenance of a pattern in a genome after duplication be explained in *A. thaliana* by the function of the involved genes or the family size they belong to? Can we characterize the fate of duplicate genes in the PPI network of *A.*

Keywords: duplicated genes, *Arabidopsis thaliana*, protein-protein interactions, functionalization

1 Introduction

Gene duplication is readily accepted as a primary mechanism for generating organismal complexity. Fate of duplicate genes is by far mainly loss. When both copies are maintained in the genome it is supposed to be mainly either by subfunctionalization, neofunctionalization or gene dosage. In subfunctionalization, each of the two copies of the gene performs only a subset of the functions of the ancestral single copy gene. Neofunctionalization implies that one of the two genes possesses a new, selectively beneficial function that was absent in the population before duplication. In gene dosage, both copies are maintained because it is selectively advantageous [1 for review]. 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 because their interacting context is taken into account. Such network data, especially protein-protein interactions (PPIs) can now be generated in a high throughput manner in different organisms, notably for *Arabidopsis thaliana* which is an excellent model organism for studying the evolution of duplicate genes.

Results from yeast protein interaction networks suggest a rapid evolution of the interactions with the probability of an interaction to get lost getting higher with time since the duplication event [2]. Wagner's model predicts also that duplicates loss interactions with sequence divergence. Some results are not in agreement with this model: duplicability, that is the proportion of the number of unique types of gene constituted of duplicated genes in the total number of unique types of genes, and its link with connectivity can be different according to the organism. For example, human hubs -that are highly connected genes- and mouse essential proteins that are involved in development are preferentially encoded by duplicated genes, while other categories of essential mouse genes can be both singletons and duplicates [3 for review]. Moreover, in *E. coli*, yeast, fly and human it has been shown that the origin and conservation of a gene significantly correlates with properties of the encoded protein in the PPI network [3]. It is well known that duplicate genes are biased in function compared to single genes [1 for review]. In some vertebrate and invertebrate genomes, it has been shown that family size is linked with gene function [4]. In yeast protein connectivity and duplicability vary with gene function [5].

Our purpose is to tackle the following question: are the protein-protein interactions driving the evolution of duplicate genes in a time-dependent manner in *A. thaliana*, as in yeast [2]? Can the maintenance of a pattern in a genome after duplication be explained in *A. thaliana* by the function of the involved genes or the family size they belong to? Can we characterize the fate of duplicate genes in the PPI network of *A.*

GO2PUB PubMed Query Tool Based on Semantic Expansion of Gene Ontology Terms, a Lipid Metabolism Case Study

Charles BETTEMBOURG[1,2], Christian DIOT[2], Anita BURGUN[1] and Olivier DAMERON[1]

[1] UMR936 INSERM, Université de Rennes 1, 2 av. Léon Bernard, F-35043 Rennes, France
{charles.bettembourg, anita.burgun, olivier.dameron}@univ-rennes1.fr
[2] UMR1348 INRA, Agrocampus Ouest, 65 Route de Saint-Brieuc, F-35042 Rennes, France
{charles.bettembourg, christian.diot}@rennes.inra.fr

Abstract *As PubMed grows, literature searches become more complex and time-consuming. Automated search tools with good precision and recall are thus necessary. We developed GO2PUB to address this demand. Our goal was to use the knowledge from Gene Ontology (GO), its annotations and to follow the true path rule for performing semantic expansion. GO2PUB enriches PubMed queries based on selected GO terms and keywords. It processes the results and displays the PMID, title, authors, abstract and bibliographic references of the articles. Gene names, symbols and synonyms that have been generated as extra keywords from the GO terms are also highlighted. Two experts manually assessed the relevance of GO2PUB, GoPubMed and PubMed on three queries about lipid metabolism. Experts agreement was high (kappa=0.88). GO2PUB returned 69 % of the relevant articles, GoPubMed: 40 % and PubMed: 29 %. 36 % of the relevant articles were returned only by GO2PUB, 17 % only by GoPubMed and 14 % only by PubMed. We also generated 20 queries based on random GO terms with a granularity similar to those of the first three queries and compared the proportions of GO2PUB and GoPubMed results. These were respectively of 70 % and 38 %. The comparison of GO2PUB, based on semantic expansion, with GoPubMed, based on text-mining techniques, showed that both tools are complementary. GO2PUB is available at http://go2pub.genouest.org*

Keywords Gene ontology, Semantic expansion, Query enrichment, PubMed

1 Background

Due to the permanent growth of data, information retrieval becomes an increasingly difficult task. PubMed is the most comprehensive public database of biomedical literature. It comprises more than 21 million entries for biomedical literature [1]. There are now 4 million more articles than 5 years ago [2]. A well defined query is important for retrieving as many relevant articles as possible with as few irrelevant ones as possible when exploring or when keeping up with a domain. To reach this goal the development of automatic tools helping the users build such complex queries is needed.

PubMed supports MeSH-based query expansion [1]. MeSH (medical Subject Headings) terms are classical PubMed keywords used for the indexation of the articles of Medline. Some additional literature search tools have been developed [2] and evaluated [3]. These correspond to three major approaches. The first approach, exemplified by tools like SLIM [4], is based on an intuitive interface to set some filters on PubMed queries for obtaining a better precision than with the basic PubMed querying system. A good proficiency with PubMed advanced search brings similar results. The second approach developed in SEGOPubMed uses a Latent Semantic Analysis (LSA) framework. It is based on a semantic similarity measure between the user query and PubMed abstracts [5]. The authors of SEGOPubMed state that the LSA approach outperforms the other approaches when using well-referenced keywords. Unfortunately, no implementation of SEGOPubMed is currently available. Moreover, this method requires that a corpus of well-referenced keywords be constituted and maintained before the search. Such a corpus is not available either. The third approach is based on query enrichment using

1. www.ncbi.nlm.nih.gov/pubmed
2. www.nlm.nih.gov/bsd/licensee/baselinestats.html

controlled vocabularies and ontologies. These knowledge representations in which concepts are described both by their meaning and their relations to each other are useful for information retrieval [6]. QuExT performs a concept-oriented query expansion to retrieve articles associated with a given list of genes from PubMed and to prioritize them [7]. GoPubMed [8] uses a text extraction algorithm to mine PubMed abstracts with GO terms [9]. However, GO terms are linked by relationships according to strict rules conveying the semantics. Among them, the true path rule requires that every child term inherits the meaning of all of its ancestor terms. GoPubMed does not follow this rule, resulting in a loss of flexibility and power [10].

We hypothesized that genes annotated by GO terms of interest in Gene Ontology Annotation database [11] can be used as additional keywords in gene-oriented PubMed queries. In GO2PUB, we propose a new approach that considers not only the genes annotated with a GO term of interest, but also those annotated by a descendant of this GO term, complying with the true path rule. GO2PUB's user inputs a list of GO terms, one or more species, and a list of keywords. GO2PUB retrieves all the genes annotated by at least one of the GO terms. It generates a PubMed query with the names, symbols and synonyms or aliases of these genes, the species and the keywords and processes PubMed result.

We performed a qualitative relevance study on our domain of expertise using three queries related to lipid metabolism. For each query, we compared PubMed, original GoPubMed and after having manually-generated the semantic expansion of the GO term and GO2PUB results. Two experts manually determined the relevance of all the articles. We computed the precision, relative recall and F-score of GO2PUB and GoPubMed. In order to determine whether the results of the qualitative study could be generalized, we performed a study on twenty randomly-generated queries.

2 Resources and Methods

2.1 Resources

In the manuscript, we used the March 2011 version of GO for semantic expansion and of Gene Ontology Annotations database for retrieving the genes annotated by the provided GO terms. These tables allowed to build queries about 7 different species. Since June 2011, GO2PUB uses the Uniprot-GOA table instead of the species-specific tables, allowing to mine literature about over 2000 different species.

We represented the overlap of the results using Venn diagrams generated by BioVenn [14].

2.2 Methods

2.2.1 GO2PUB query building. GO2PUB creates an expanded PubMed query. Fig. 1 presents the process.

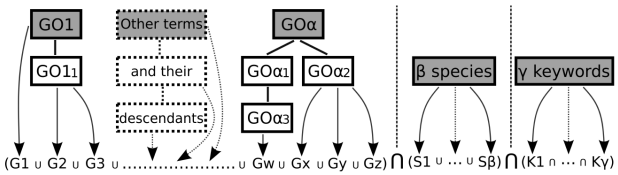


Figure 1. GO2PUB Query composition process using parameters provided by the user. Step 1: initial α GO terms (grey boxes) are enriched by their descendants. Step 2: genes (G1 to Gz) annotated by the GO terms are retrieved. Step 3: query is composed using names, symbols and synonyms of the genes, β species (S) and γ MeSH or free keywords (K).

The first part of each query involves one or more GO terms. The user can enter either the name, a synonym or the identifier of a GO term. These terms are suggested when the user starts to fill the field. For example, GO2PUB will search for "lipid biosynthetic process" if the user provides "lipogenesis". When two or more GO terms are entered, GO2PUB makes the union of them ("OR" connector). Then, the user selects one or several species using a name (common or scientific names and their synonyms are allowed) or a NCBI taxon code. In this case, the user can choose to join them (using "OR") or intersect them (using "AND"). Logical connectors

"AND" and "OR" are set by default to make the union of species and intersection of keywords, but this can be modified. Next, the user can enter additional MeSH terms to decrease the number of false positives. MeSH terms associated to the articles by PubMed are not all of same importance, some of them are "Major topic" (MAJR). We can qualify each keyword as a simple MeSH term or a Major topic. Again, the user can specify the connector between keywords. At this point, the user has built a simple GO2PUB query. We called this query [BASICq].

The system supports 3 modifications for [BASICq] for studying if minor changes bring additional relevant results. The first modification lets the opportunity to ignore MAJR qualifiers and search all keywords in PubMed [MeSH] tag. As MAJR terms are also MeSH terms, articles associated to them will still be found. We called this query [MeSHq]. The second modification replaces "AND" connectors between keywords by "OR" connectors. However, as it can return many more results with a lot of noise, all keywords in this additional query are tagged with MAJR. Species that are normally searched in MeSH are also tagged with MAJR. We called this query [ORq]. The third modification ignores MeSH and MAJR keywords, and tags species with MAJR. This option must be used carefully because it can yield several hundreds of results. It is of interest only for very narrow topics if the user does not obtain enough results with the other queries. We called this query [NOKq]. Last, GO2PUB proposes 3 additional options. The first option allows to set a lower limit on the publication year. The second option proposes an exhaustive search of the official synonyms of gene names using Entrez gene [3]. As the authors sometimes use in their articles synonyms that are absent in the GOA database, this option allows to build more complete PubMed queries in order to obtain more relevant results. The third option toggles the display of the MeSH table associated with each article.

2.2.2 Query rewriting using semantic expansion. After semantic expansion genes annotated with GO terms and those annotated by descendants are selected. Fig. 2 shows that the expansion yields five genes associated with the regulation of fatty acid metabolic process, instead of two if the true path rule is ignored.

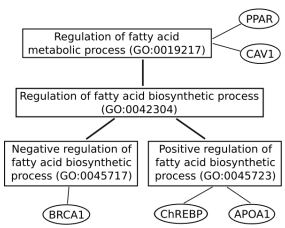


Figure 2. Keyword semantic enrichment. The two genes PPAR and CAV1 are directly annotated by the GO term "Regulation of fatty acid metabolic process" (`GO:0019217`). However, the true path rule implies that genes annotated by at least one descendant of the original term (BRCA1, ChREBP and APOA1) should also be considered.

GO2PUB retrieves all the genes annotated by each GO term, directly or indirectly through the true path rule. GO2PUB builds a query on the model "(n gene names, symbols or synonyms separated by OR) AND (m species) AND (p MeSH terms)". The name, symbol and synonyms of each gene compose the first part of the query. They will be searched in title and abstract. Species and keywords make up the second part of the query. Finally, GO2PUB submits to PubMed a query composed of genes annotated directly or indirectly (name OR symbol OR Synonym), at least one species and some MeSH terms and free keywords. GO2PUB compiles the results and displays all citations numbered and sorted by date.

3 Results

3.1 Qualitative study

In order to evaluate GO2PUB relevance and to compare it with GoPubMed, we assessed three queries (Q1, Q2 and Q3) about biological processes related to lipid metabolism and including different GO terms, species

3. http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene

and MeSH terms. As GoPubMed only considers the GO term(s) provided by the user and ignores the inheritance rules of GO, we expanded queries manually then submitted them to GoPubMed. Our GoPubMed queries were composed by the GO term(s) and all its descendants separated by "OR", plus MeSH keywords. This ensured the closest comparison possible. We have also written PubMed queries as close to ours as possible.

3.1.1 Relevance criteria We analyzed the results of GO2PUB, GoPubMed and PubMed queries according to the following criteria. We considered that a relevant article has to describe at least one gene product occurring in the chosen metabolism for the selected species, in particular its role, its interactions, and how and when it is activated. For queries Q1, Q2 and Q3, the results obtained by the different tools were mixed for a blind selection by two reviewers. This ensured that they did not know the tool(s) that retrieved the articles. The final list of relevant articles is the union of the two reviewers' lists.

3.1.2 Relevance measurement For each query and tool, we computed the precision (fraction of retrieved articles that are relevant), recall (fraction of relevant articles that are retrieved) and F-score (harmonic mean of precision and recall). Computing the real recall for each query is impossible because it would require to know all the really relevant articles available in Medline. As it is possible that some of these articles were missed by all of the three tools, recall was defined as relative to all relevant articles obtained by at least one of the tools. Most of the relevant articles were found in the intersection of the two reviewers' selections. Indeed, reviewers agreed on 35 relevant and 113 irrelevant articles while selecting separately 3 and 4 articles as relevant. We used Cohen's kappa coefficient as a statistical measure of inter-rater agreement [12]. The value obtained was 0.88, which corresponds to an almost perfect agreement [13].

3.1.3 Query Q1: Lipogenesis in chicken liver In our first reviewed query in GO2PUB, we used "Lipid biosynthetic process" (`GO:0008610`) as GO term, "Gallus gallus" as species, "Liver" as Major Topic and "Lipid Metabolism" as MeSH keyword, and we considered the articles published in the last five years. We ran query Q1 on GO2PUB using the [BASICq], [MeSHq] and [ORq]. "Lipid biosynthetic process" has 243 descendants in GO. The mean number of edges to reach the root of the ontology from this term was 3.5. GoPubMed equivalent query was *"lipid biosynthetic process"[go] AND Chickens[mesh] AND Liver[majr] AND "Lipid Metabolism"[mesh] AND last5years[time]*. This is the "standard" query for GoPubMed. We have also considered the manually-expanded version of this query by adding the descendants of "lipid biosynthetic process" separated by "OR". It should be noted that 47 of the 243 terms generated by the semantic expansion of "Lipid biosynthetic process" generated a GoPubMed "missing term error" and had to be ignored. PubMed equivalent query was "Chickens liver lipogenesis", which PubMed interpreted using MeSH terms combinations. Fig. 3A presents Venn diagrams of Q1 results obtained with PubMed, GoPubMed (after manual expansion) and GO2PUB. Although queries as similar as possible were issued to the three tools, the resulting sets of articles had little overlap (a). Most of relevant articles (b) were yielded by GO2PUB. We notice that most of the articles retrieved by at least two tools (overlaps in a) were relevant (overlaps in b).

Table 1 presents the precision, relative recall and F-score for each tool. GO2PUB had better precision and relative recall than GoPubMed and Pubmed. There was no difference between "standard" and "expanded" GoPubMed results.

	PubMed	GoPubMed (std)	GoPubMed (exp)	GO2PUB
(a) Number of results	19	16	16	24
(b) Relevant among (a)	8	5	5	13
Precision	0.421	0.313	0.313	0.542
Relative Recall	0.5	0.313	0.313	0.813
F-score	0.457	0.313	0.313	0.650

Table 1. Precision, relative recall and F-score values using PubMed, GoPubMed without (std) or with (exp) manual expansion and GO2PUB search tools for query Q1 about lipogenesis in chicken liver. Values are calculated from (a) and (b) lines using a total number of relevant results of 16.

3.1.4 Query Q2: Lipid transport in human blood In Q2, we used "Lipid transport" (GO:0006869) as GO term, "Homo sapiens" as species, "Blood" as Major Topic and "Lipid Metabolism" as MeSH keyword, and we considered the articles published in the last five years. "Lipid transport" has 109 descendants in the GO graph. The mean number of edges to reach the root of the ontology from this term was 4.3. We ran equivalent queries on GoPubMed ("standard" and "expanded" versions) and PubMed. 46 of the 109 terms generated by the semantic expansion of "Lipid Transport" generated a GoPubMed "missing term error" and had to be ignored. Fig. 3B presents the results obtained by PubMed, GoPubMed (after manual expansion) and GO2PUB for Q2. As observed for query Q1, the majority of the results were tool-specific (a). PubMed yielded 45 articles, none of which were retrieved by GO2PUB nor GoPubMed while there was an overlap between GO2PUB and GoPubMed results. Three of the four common articles between GoPubMed and GO2PUB were relevant (b). GO2PUB yielded half of GoPubMed relevant results while having an important specific relevant results set.

Table 2 gives Q2 precision, relative recall and F-score of PubMed, GoPubMed and GO2PUB. GO2PUB has a slightly lower precision than GoPubMed (standard and after manual expansion) and better relative recall and F-score. For GoPubMed, there was no difference between "standard" and "expanded" results.

	PubMed	GoPubMed (std)	GoPubMed (exp)	GO2PUB
(a) Number of results	45	9	9	16
(b) Relevant among (a)	2	6	6	10
Precision	0.044	0.667	0.667	0.625
Relative Recall	0.133	0.4	0.4	0.667
F-score	0.067	0.5	0.5	0.645

Table 2. Values of precision, relative recall and F-score using PubMed, GoPubMed without (std) or with (exp) manual expansion and GO2PUB search tools for query Q2 about lipid transport in human blood. Values are calculated from (a) and (b) lines using a total number of relevant results of 15.

3.1.5 Query Q3: Regulation of lipase activity in human cell membrane In a third query, we used "Regulation of lipase activity" (GO:0060191) as GO term, "Homo sapiens" as species and "Cell Membrane" and "Lipid Metabolism" as MeSH keywords, and we considered the articles published in the last ten years. "Regulation of lipase activity" has 35 descendants in the GO graph. The mean number of edges to reach the root of the ontology from this term was 5.25. We ran equivalent queries on GoPubMed ("standard" and "expanded" versions) and PubMed. 16 of the 35 terms generated by the semantic expansion of "Regulation of lipase activity" generated a GoPubMed "missing term error" and had to be ignored. PubMed yielded 23 articles for a query using "regulation", ("lipase" AND "activity"), "human" and ("cell" AND "membrane") as keywords. As shown in Fig. 3C (a), a larger set of results was returned by GO2PUB compared to GoPubMed (24 and 8, respectively), but most of these results are irrelevant (Fig. 3C (b)). As observed for query Q2, none of the PubMed results were retrieved by GO2PUB nor GoPubMed. Only 2 article on 23 yielded by PubMed were relevant.

Table 3 gives Q3 precision, relative recall and F-score of GoPubMed and GO2PUB. GO2PUB has a relative recall equivalent to GoPubMed's and lower precision and F-score. For GoPubMed, the "manually-expanded" results have a higher relative recall and F-score and a lower precision than the "standard" ones.

	PubMed	GoPubMed (std)	GoPubMed (exp)	GO2PUB
(a) Number of results	23	6	8	24
(b) Relevant among (a)	2	5	6	6
Precision	0.087	0.833	0.75	0.25
Relative Recall	0.182	0.455	0.545	0.545
F-score	0.118	0.588	0.632	0.343

Table 3. Values of precision, relative recall and F-score using PubMed, GoPubMed without (GoPM std) or with (GoPM exp) manual expansion and GO2PUB search tools for query Q3 about regulation of lipase activity in human cell membrane. Values are calculated from (a) and (b) lines using a total number of relevant results of 11.

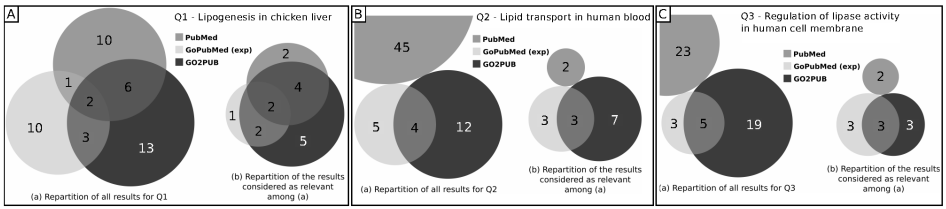


Figure 3. Comparison of the PubMed, GoPubMed and GO2PUB results for Q1 (A), Q2 (B) and Q3 (C). For each subfigure, (a) presents the repartition and intersections of results and (b) the repartition and intersections of relevant results. GoPubMed data result from manual semantic expansion.

3.2 Generalization study

We performed a generalization study on 20 randomly-generated queries based on Biological Process terms having a granularity similar to those of the three GO terms used in the qualitative study. We built queries following the pattern: "a random GO term + a species (mouse) + a publication date limit (2011) + a keyword (the GO term name)". We assumed that the granularity of a term depends on the mean length of its path to the root, and its number of descendants. Each GO term of the generalization study had a mean path length to the root between 3.5 and 5.25 edges and had between 35 and 244 descendants. As we could not add a MeSH keyword in relation with the random GO term of each query, we simply added the name of this GO term. This keyword was added in the free field for GO2PUB and without [go] tag for GoPubMed. We submitted these queries to GO2PUB and to GoPubMed. GO2PUB yielded 70 % of all results and GoPubMed 38 %. The repartition profile of the results is close to the qualitative study one.

(A) Lipids	GO:0044242		GO:0008299		GO:0008654	
Tool	GPM	G2P	GPM	G2P	GPM	G2P
(a) Number of results	4	26	9	16	25	27
(b) Relevant among (a)	3	20	1	2	5	11
(c) Total relevant	22		3		12	
(d) Common results	1		2		5	
(e) Relevant among (d)	1		0		4	
Precision	0.750	0.769	0.111	0.125	0.200	0.407
Relative Recall	0.136	0.864	0.333	0.667	0.417	0.917
F-score	0.231	0.814	0.167	0.211	0.270	0.564

(B) Other	GO:0050658		GO:0033013		GO:0006805		GO:0048284	
Tool	GPM	G2P	GPM	G2P	GPM	G2P	GPM	G2P
(a) Number of results	25	10	15	19	30	23	24	17
(b) Relevant among (a)	7	2	3	3	17	14	10	9
(c) Total relevant	9		6		26		16	
(d) Common results	3		6		7		8	
(e) Relevant among (d)	1		2		4		4	
Precision	0.280	0.200	0.200	0.158	0.567	0.609	0.417	0.529
Relative Recall	0.875	0.250	0.600	0.600	0.680	0.560	0.625	0.563
F-score	0.424	0.222	0.300	0.250	0.618	0.583	0.500	0.545

Table 4. Results' sets sizes and values of precision, relative recall and F-score using GoPubMed and GO2PUB search tools for seven random queries: three lipid-related queries (part A) and four queries about other topics (part B). Values of precision, relative recall and F-score are calculated from (a), (b) and (c) lines.

We studied the relevance of the results from seven queries picked among the twenty queries of the generalization study. As it is the domain of expertise of our reviewers, we selected the three queries concerning lipid metabolism ("cellular lipid catabolic process" [GO:0044242], "isoprenoid biosynthetic process" [GO:0008299] and "phospholipid biosynthetic process"[GO:0008654]). The Cohen's kappa coefficient was

0.9345. We picked randomly four additional queries about "RNA transport" [GO:0050658], "tetrapyrrole metabolic process" [GO:0033013], "xenobiotic metabolic process" [GO:0006805] and "organelle fusion" [GO:0048284]. The Cohen's kappa coefficient for these four queries was 0.797. Table 4 presents the number of results, precision, relative recall and F-score respectively for the three lipid-related queries and the other four queries of the generalization study.

Results are similar to those observed in the qualitative study. The resulting sets of articles had little overlap. Moreover, each tool yielded relevant results ignored by the other, with important variation of performances among queries.

4 Discussion

Our goal was to develop a tool that uses the knowledge from the Gene Ontology (GO) and its annotations for generating semantically-expanded gene-related PubMed queries. Indeed, there is no [GO] tag for a search in PubMed. We compared the results of four different queries sent to PubMed, to our tool GO2PUB and to GoPubMed, another tool that allows to query PubMed with GO terms.

The qualitative study showed that both GO2PUB and GoPubMed retrieved relevant articles ignored by PubMed. For Q1, 26 of the 35 articles (8 of 14 relevant) returned by either GO2PUB or GoPubMed were ignored by PubMed. Conversely, 9 of the 19 articles (6 out of 8 relevant) returned by PubMed were also returned by either GO2PUB or GoPubMed. For Q2 and Q3, the set of articles returned by PubMed was disjoint from both GO2PUB and GoPubMed results. The discrepancy observed between PubMed and the other tools is probably due to the absence of [GO] search field tag in PubMed.

Overall, GO2PUB had better performances than GoPubMed and PubMed, but both GO2PUB and GO-2PUB yielded relevant articles ignored by the other. The differences varied among the queries. For Q1 and Q2, GO2PUB yielded most of the relevant articles and had therefore the highest relative recall value while its precision was slightly lower than that of GoPubMed. Consequently, GO2PUB had the best F-score. For Q3, GO2PUB yielded as many relevant articles as GoPubMed but had a higher proportion of noise. GO2PUB had a slightly better relative recall than GoPubMed, but its precision was much lower. Consequently, GoPubMed had the best F-score. We can also notice that for Q3, the query expansion on GoPubMed improved its performances with a better relative recall and F-score at the cost of a small loss of precision.

GO2PUB performs a semantic expansion of the GO terms of interest complying with the GO true path rule before retrieving the corresponding genes for enriching the query. During the development of GO2PUB, we have also run queries without this expansion. We obtained empty or very small sets of results.

Using the true path rule is useful. The more descendants a GO term has, the more relevant results GO2PUB yields. GO2PUB performances decreased from Q1 to Q3. For Q1, "lipid biosynthetic process" has 243 descendants and annotates 646 genes for human and 145 genes for chicken. For Q2, "lipid transport" has 109 descendants and annotates 253 genes for human and 63 genes for chicken. For Q3, "regulation of lipase activity" has 35 descendants and annotates 168 genes for human and 18 genes for chicken. The more descendants a GO term has, the more genes it is likely to annotate, which increases the gene-related enrichment importance. Moreover, Q1 concerned chicken, which is less annotated than human. On less annotated species, the annotations focus on the major genes. This explains why GO2PUB yields a high proportion of relevant articles.

GoPubMed does not follow the true path rule. We manually expanded GoPubMed queries and compared it to GO2PUB. The added value of semantic expansion was null for Q1 and Q2, and important for Q3 (+33%). So query expansion is a built-in fonctionality in GO2PUB, and would be a valuable extension for GoPubMed. In GoPubMed results, a "missing term" error occured for 19% of the expanded set of GO terms for Q1, 42% for Q2 and 44% for Q3. We assume that the benefits of query expansion on GoPubMed might be higher by considering the articles related to these currently omitted GO terms.

Most of the results obtained by both tools are relevant (4 out of 4 for Q1, 3 out of 4 for Q2 and 3 out of 5 for Q3). So, the intersection of GoPubMed and GO2PUB results decreases noise. As each tool yields relevant articles that are ignored by the other, the union of their results also decreases silence.

In order to consolidate the qualitative study, we submitted 20 randomly-generated queries to GO2PUB and GoPubMed. Each query contained a random GO term of a granularity similar to that of the terms used in the qualitative study. The proportion of articles returned by GO2PUB was 70 %, the one of GoPubMed was 38 %. We assume that the difference of proportions between the qualitative and the generalization studies can be attributed to Q1, Q2 and Q3 being more specific because of the use of MeSH keywords. The seven analyzed queries of the generalization study presented relevances similar to those observed in the qualitative study.

5 Conclusion

GO2PUB brings relevant results ignored by GoPubMed (9 GO2PUB' specific results for Q1, 7 for Q2 and 3 for Q3) even doing a manual query expansion for GoPubMed. This confirms our hypothesis that genes annotated by GO terms of interest are useful keywords in gene-oriented PubMed queries. Conversely GoPubMed text-mining approach finds relevant articles ignored by GO2PUB (1 GoPubMed' specific result for Q1, 3 for Q2 and 3 for Q3). This demonstrates GO2PUB relevance and its complementarity with GoPubMed for our domain of interest. The generalization analysis shows that a similar profile of results is obtained using random queries, especially when using keywords for narrowing the queries. Relevances observed in the generalization study are similar to those of the qualitative study. This suggests that the results of the qualitative study can be generalized.

Acknowledgements

Biogenouest for computer support and hosting. CB was supported by a fellowship from the French ministry of research.

References

[1] Z. Lu, W. Kim, W.J. Wilbur, Evaluation of Query Expansion Using MeSH in PubMed. *Inf Retr Boston*, 12(1):69-80, 2009

[2] Z. Lu, PubMed and beyond: a survey of web tools for searching biomedical literature. *Database (Oxford)*, :baq036, 2011

[3] A. Bajpai, S. Davuluri, H. Haridas, G. Kasliwal, H. Deepti, K.S. Sreelakshmi, D.S. Chandrashekar, P. Bora, M. Farouk, N. Chitturi, V. Samudyata, K.P. ArunNehru, K. Acharya, In search of the right literature search engine(s). *Nature Precedings*, doi:10.1038/npre.2011.2101.3, 2011

[4] M. Muin, P. Fontelo, F. Liu, M. Ackerman, SLIM: an alternative Web interface for MEDLINE/PubMed searches - a preliminary study. *BMC Med Inform Decis Mak.*, 5:37, 2005

[5] B.C. Vanteru, J.S. Shaik, M. Yeasin, Semantically linking and browsing PubMed abstracts with gene ontology. *BMC Genomics*, 9(Suppl 1):S10, 2008

[6] J.B. Bard, S.Y. Rhee, Ontologies in biology: design, applications and future challenges. *Nat. Rev. Genet.*, 5:213-222, 2004

[7] S. Matos, J.P. Arrais, J. Maia-Rodrigues, J.L. Oliveira, Concept-based query expansion for retrieving gene related publications from MEDLINE. *BMC Bioinformatics*, 11:212,2010

[8] A. Doms, M. Schroeder, GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Res.*, 33:W783–W786, 2005

[9] M.A. Harris, Consortium GO, The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, 1:D258–D261, 2004

[10] S.Y. Rhee, V. Wood, K. Dolinski, S. Draghici, Use and misuse of the gene ontology annotations. *Nat. Rev. Genet.*, 9:509–515, 2008

[11] E. Camon, M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N. Harte, R. Lopez, The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res.*, 32:D262–D266, 2004

[12] J. Cohen, A coefficient of agreement for nominal scales. *Educ Psychol Meas.*, 20(1):37-46, 1960

[13] J.R. Landis, G.G. Koch, The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159-74, 1977

[14] T. Hulsen, J. de Vlieg, W. Alkema, BioVenn - a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams. *BMC Genomics*, 9(1):488, 2008

Retrospective Analysis of a Gene by Mining Texts:

The Hepcidin gene use-case

Fouzia MOUSSOUNI[1], Bertrand AMELINE-DE-CADEVILLE[1], Ulf LESER[2] and Olivier LOREAL[1]

[1] LABORATOIRE, UMR991 INSERM, 2 rue Henri Le Guilloux, CHU Pontchaillou, 35033, Rennes, Cedex 09, France
fouzia.moussouni@univ-rennes1.fr

[2] HUMBOLDT UNIVERSITÄT ZU BERLIN, Rudower Chaussee 25, 12489 Berlin, Germany.
leser@informatik.hu-berlin.de

Session 3_D : High Throughput Sequencing

Présentation Industrielle

Elodie Dubus, D Richards, R Flannery, A Kramer, J Lerman et A Kutchma

Ingenuity Systems

Rapid Prioritization and Annotation of Variants from Human Re-sequencing Studies

Biological interpretation of thousands of potentially deleterious variants is a bottleneck in extracting valuable causal insights from DNA re-sequencing studies, often requiring months of effort after completion of the reference genome alignment and variant calling steps. We have developed a fast, easy-to-use application, Ingenuity Variant Analysis (www.ingenuity.com), that leverages an extensive knowledge base of millions of expert-curated mutation and biomedical findings from the literature to empower real-time interactive filtering and rapid prioritization of variants. It enables clinical researchers to quickly zero in on the few variants that are most compelling for follow-up. Using a combination of causal analytics, genetic analysis at the variant, gene, and pathway levels, and the ability to visualize how variants impact disease progression, we will demonstrate the application of a context-rich knowledge base to discover cancer driver variants and novel causal variants for human genetic disease.

Ingenuity Systems, Redwood City, USA

Présentation Industrielle

Patricia Otten-Hernandez, L Baerlocher , J Prados , N González , W Baroni , M Osteras et L Farinelli

Fasteris

Extracting relevant information from UHTS data

Illumina ultra high-throughput sequencing (UHTS) uses massively parallel sequencing approach to generate millions of short sequences (<150 nucleotides) from biological samples. Today, on the HiSeq 2000, more than 200 mios reads per lane can be obtained. Two human genomes can be resequenced at 50x in a 2x100 run on a single 8 lanes flow cell. Applications of UHTS include whole genome de novo assembly, genome re-sequencing and detection of variants, RNA or smallRNA expression quantification, analysis of DNA-associated proteins (ChIP-SEQ), and many others.

End 2006, Fasteris was the first service provider in the world to acquire an Illumina Genome Analyzer system. Today, our company is equipped with 2 HiSeq 2000 and 1 MiSeq. Our clients are private and academic labs around the world. Nevertheless, all these data are useless without extensive bioinformatics analyses. At Fasteris, we propose standard analyses, including de novo assembly, mapping, comparative expression profiling, peak detection and variant calling. According to our client needs, we also propose custom bioinformatics processes. One of our main challenge is to be able to deliver high quality analyses at affordable cost. In our presentation, using as context some of the analyses that were performed at Fasteris, we will describe in details the pipelines developed for this purpose.

E-mail: patricia.otten@fasteris.com, ht_seq@fasteris.com

Fasteris SA, Ch. du Pont-du-Centenaire 109, CH-1228 Plan-les-Ouates, Switzerland

Présentation Industrielle

Erwan Drezen, Alain Meil, Dominique Lavenier et Patrick Durand

Korilog & Inria/IRISA GenScale team

KLAST: high-performance sequence similarity search tool

KLAST is a fast, accurate and NGS scalable sequence similarity search tool providing significant accelerations over BLAST suite algorithms. Relying on unique software architecture, KLAST takes full advantage of recent multi-core personal computers without requiring any additional hardware devices.

KLAST is a new optimized implementation of the Inria's PLAST algorithm (1), to which several improvements have been made. KLAST is fully designed to compare large query sets with large subject sets of DNA, RNA or protein sequences. As BLAST, KLAST is declined into 5 different programs: KLASTn, KLASTp, KLASTx, tKLASTx and tKLASTn. It is significantly faster than the original PLAST software, while providing comparable sensitivity to BLAST and SSearch algorithms. KLAST contains a fully integrated data-filtering engine capable of selecting relevant hits with user-defined criteria (E-Value, identity, coverage, alignment length, etc.).

KLASTp benchmarks on the black cottonwood Populus trichocarpa species have demonstrated a 24x increase in speed compared to BLAST, both tools ran on 8 cores of an Apple MacPro computer. This benchmark processes a set of 2327 proteins against a subset of the NCBI RefSeq databank (2.9 millions of proteins). Compared to BLAST and SSearch, KLAST is as sensitive and selective as these reference tools. Equivalent acceleration and quality are achieved with the other KLAST programs, i.e. KLASTx, tKLASTn, tKLASTx and KLASTn.

To provide users with an advanced sequence similarity search platform, the KLAST engine has been integrated as a plugin inside the KNIME data analysis desktop workbench. As a result, users benefit from Korilog plugin's features and from a wide range of data analysis methods coming with the KNIME platform. KLAST is available for the various releases of KNIME, from Desktop up to Cluster Execution. Command-line execution is also available.

(1) V.H. Nguyen & D. Lavenier. BMC Bio informatics 2009, 10:329

KLAST software developments are supported by Région Bretagne and Critt Santé Bretagne.

Web site: www.klast-search.com
Korilog SARL, 4 rue Gustave Eiffel, 56230 Questembert, France

Session 4 : Non-Protein World

Conférence invitée

Ivo Hofacker

Institute for Theoretical Chemistry, Austria

RNA Structures, Interactions, and Folding Kinetics

The talk will give an overview of recent advances in the computational analysis of RNA secondary structures, focusing in particular on three specific areas: (i) Classical RNA structure prediction considers only Watson-Crick type base pairings. Tertiary structures, however, contain a multitude of non-canonical pairings that determine 3D fold and tertiary interactions. I will present a recent method that include these non-canonical pairs in the prediction without sacrificing the efficiency of classical algorithms. (ii) Most RNAs function by interacting with other RNAs. The prediction of interaction targets is therefore a promising step towards understanding the function of novel non coding RNAs. Several recent methods address this problem while striking a balance between speed and accuracy. (iii) Even relatively short RNAs may exhibit high energy barriers in their folding landscape that make folding to the minimum free energy structure prohibitively slow. We will therefore present methods to predict the folding kinetics, focusing in particular at the problem co-transcriptional folding.

Computational detection and expression profiling of conserved long non-coding RNAs in the domestic dog

Thomas DERRIEN[1,2], Amaury VAYSSE[1] , Benoit HENNUY[3], Wouter COPPIETERS[3], Benoit HÉDAN[1], Catherine ANDRÉ[1], and Christophe HITTE[1]

[1] UMR6290 CNRS, Institut de Génétique et Développement de Rennes, Université Rennes1, 35000 Rennes, France,

{toma.derrien, amaury.vaysse}@gmail.com,{benoit.hedan, catherine.andre, christophe.hitte}@univ-rennes1

[2] IRISA/INRA Symbiose/GenScale team, 35000 Rennes, France, France

tderrien@irisa.fr

[3] Unit of Animal Genomics, GIGA-R & Faculty of Veterinary Medicine, University of Liège, 4000-Liège, Belgium,

{benoit.hennuy,wouter.coppieters}@ulg.ac.be

Abstract

Accumulating evidence indicates that mammalian genomes contain many thousands of non-coding RNAs (ncRNAs). Long non-coding RNAs (lncRNAs) form a class of functional non-coding RNAs that are characterized as mRNA-like ncRNA because they undergo splicing, possess poly(A) tail/signal, are greater than 200 bp and largely multi-exonic but are not translated into proteins. While thousands of lncRNAs have been annotated in human, all mammalian species do not benefit yet from an exhaustive annotation of its lncRNA repertoire. Moreover, lncRNA conservation between phylogenetic related species argue in favor of their functionality and in the maintenance of a conserved function. The domestic dog species has [illegible] [illegible] genotype/phenotype relationships.

Le ou les auteur(s) ne souhaite(nt) pas que ce document soit diffusé en ligne

We therefore developed a computational detection of conserved lncRNAs between human and dog that we further analyzed by expression profiling derived from high throughput whole transcriptomes sequencing (RNA-seq) experiments. Genome-wide computational analysis was performed with the evaluation of two factors: (i) human-dog sequence conservation and physical clustering and (ii) recurrent adjacency with flanking protein-coding genes in related mammalian genomes. Using the recent GENCODE annotation of the human genome, our analysis identified more than 5,100 conserved lncRNAs between dog and human. Then, filtering lncRNAs based on the conservation of gene-order with flanking protein-coding genes, we defined a set of 3,540 synteny-based conserved lncRNAs. In addition, expression profiling analysis using RNA-seq data sets from four tissues in two breeds confirmed that 65% of these lncRNAs are at least expressed in one tissue. Sequence conservation and RNA-seq data provide strong evidence to support the identification of canine functional long non-coding RNAs. These results will help to assess how conserved lncRNA are distributed across genomes and their evolutionary properties between human and dog. Conserved syntenic relations between coding and lncRNAs together with correlation of expression profiles will further help to predict lncRNAs and analyze their potential role of *cis*-regulator of neighboring protein-coding genes.

Keywords Long non-coding RNAs, LncRNAs, Comparative genomics, High Throughput Sequencing, RNA-seq, Dog.

1 Introduction

Non-coding RNAs in humans have been intensively investigated for annotating their gene structure, evolution and expression and for assessing their functional role [1, 2]. It is widely

NRPS toolbox for the discovery of new nonribosomal peptides and synthetases

Maude Pupin[1], Malika Smaïl-Tabbone[3], Philippe Jacques[2], Marie-Dominique Devignes[3] and Valérie Leclère[2]

[1] LIFL, UMR8020 CNRS, INRIA, Bat M3, Univ Lille Nord de France, Sciences et Technologies, 59655 Villeneuve d'Ascq cedex, France
maude.pupin@lifl.fr

[2] ProBioGEM, UPRES EA 1026, Polytech'Lille/IUT A, Av P Langevin, Univ Lille Nord de France, Sciences et Technologies, 59655 Villeneuve d'Ascq cedex,
valerie.leclere@univ-lille1.fr, philippe.jacques@polytech-lille.fr

3 LORIA (CNRS UMR7503, INRIA Nancy Grand-Est, Nancy Université), Campus scientifique, 54506 Vandoeuvre-lès-Nancy, France
{Marie-Dominique.Devignes, malika.smail}@loria.fr

Abstract *Nonribosomal peptide synthetases are huge multi-enzymatic complexes synthesizing peptides, but not through the classical process of transcription and then translation. The synthetases are organised in modules, each one integrating an amino acid in the final peptide. The modules are divided in domains providing specialized activities. So, those enzymes are as diverse as their products. We present our toolbox designed to annotate them accurately and promising results obtained on some* Burkholderia, Bacillus *and* Pseudomonas *genomes.*

Keywords database, protein annotation, nonribosomal peptides, nonribosomal synthetase.

1 Introduction

Micro-organisms are able to synthesize peptides by a pathway alternative to the central dogma, the nonribosomal peptide synthesis. Multifunctional enzymes, called NonRibosomal Peptide Synthetases (NRPSs), assemble directly monomers to produce atypical peptides harboring original physico-chemical properties that give them various properties such as antibiotic, anti-tumor, immunosuppressive or surfactant (surface-active substances such as detergents). Several nonribosomal peptides are already exploited in pharmacology or other biotechnological area, but their great potential of new drugs or bio-active compounds is underexploited.

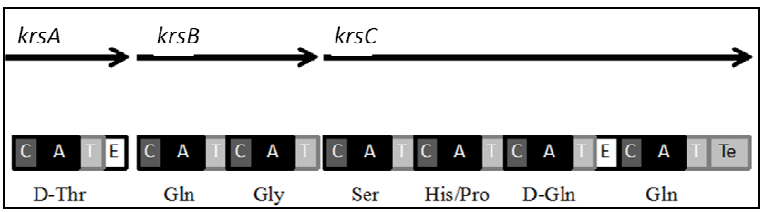


Figure 1. Scheme of a synthetase composed of 3 proteins (KrsA, KrsB and KrsC) and 7 modules. Each colored box represents a domain (C for condensation domain, A for adenylation domain, T for thiolation, E for epimerization and Te for thioesterase). The amino acid incorporated by each module is mentioned under it.

NonRibosomal Peptide Synthetases are organized in sets of catalytic domains which constitute modules containing the information needed to complete an elongation step in an original peptide biosynthesis (see Figure 1). The main catalytic functions are responsible for the activation of an amino acid residue (adenylation -A- domain), the transfer of the corresponding adenylate to the enzyme-bound 4'-phosphopantetheinyl cofactor (thiolation -T- domain) and the peptide bond formation (condensation -C- domain). The active site of the adenylation domain is specific of the incorporated amino acid. As non proteogenic amino acids or other compounds can be incorporated, we also use the term monomer. Additional

domains can lead to modification of substrates if required for peptide final structure. For example, epimerization -E- domains, tranform L-amino acids in D-amino acids. 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 T domain and, in several cases, to cyclize the peptide. To summarize, a given synthetase produces a specific peptide, with as many modules as amino acids incorporated in the final peptide. The synthetase illustrated in Figure 1 is composed of 7 modules, each incorporating the mentioned amino acid. The modules with epimerization domains transform the L-amino acid in D-amino acid.

Several bioinformatics tools [1, 2, 3, 4] were developed during the past decade to predict, from the protein sequence, the modular organization of the NRPS and the potential monomer composition of the synthesized peptidic product. The two first tools predict the modular organization of the synthetases and all the four predict the amino acids incorporated by the A-domains. As bioinformatics tools today available help predicting the genes, the produced proteins and their functions from genomes data, we can now expect to be able to predict the produced peptides and their potential activity from the NRPS protein sequence. However, one step remains difficult, which consists in the detection of the putative synthetases among the proteome of a given micro-organism. The difficulty comes from the fact that each synthetase is specific of one peptide so we cannot use a classical BLAST search to find all of them.

Our aim is to provide bio-informatics tools that help discovering new nonribosomal peptides by predicting and analyzing their synthetases from data obtained by genome sequencing or metagenomics.

2 Methods : our NRPS toolbox

Our work began with the creation of Norine [5] (http://bioinfo.lifl.fr/norine/), the unique resource dedicated to nonribosomal peptides. It provides a database with detailed annotations, including but not limited to biological activity, producing organism and the monomer structure of the peptides that deals with their non-linear 2D-structure. It provides also bio-informatics environment to analyze NRPs such as visualization or edition tools for monomer structures, statistics representations of the results, peptide search by monomer composition, structural pattern (with a list of or undetermined monomers at several positions) [6] or structure comparison.

We are now developing a complete toolbox dedicated to NRPSs and their products by integrating existing and ongoing tools. Doris, Database of nOnRIbosomal Synthetases (not yet public) contains not only synthetases automatically extracted from generalist protein databases such as UniProt, but also manually curated ones, annotated with the tools dedicated to NRPSs. A synthetase is connected to its product in the Norine database when the structure of the peptide is experimentally verified. To do so, we search the monomer composition or structural pattern predicted from the synthetase in Norine database. The results can be a perfect match with a given NRP, suggesting that we have found the synthetase of this peptide, a nearly perfect match suggesting that the produced peptide is a variant of the one stored in Norine, or a partial match suggesting either that the synthetase we have found is incomplete or that we have found a new peptide.

To complete the DORIS database, candidate NRPS are also extracted from newly sequenced genomes on the basis of sequence similarity with already described NRPS combined with manual analysis of coding sequence description as provided by automatic genome annotation. For example, we search expressions such as “NRPS”, “adenylation”, “nonribosomal” or “siderophore” among the descriptions of the studied proteome. To this aim, the MODIM (MOdel Driven Data Integration for Mining) methodology is used as it facilitates data collection and integration [7, 8]. It requires a relational database model and allows the specification of workflows for collecting data from various resources. The collected data subsequently populate the target database. Specific views can then be defined on the database for extracting datasets to be mined.

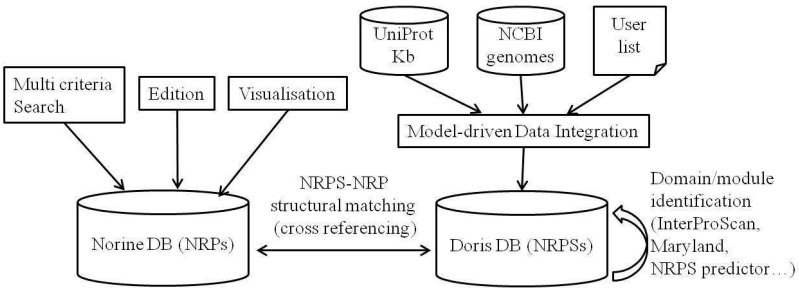


Figure 2. Scheme of the NRPS toolbox, summarizing the interaction between the tools and databases.

The association of all these tools will constitute a unique complete toolbox dedicated to NRPSs (Figure 2) and their products and will be very useful for discovering new natural antimicrobial or anti-tumoral peptides. The tools are validated with novel bacilli genomes which will be used for illustration purposes.

3 Results

The NRPS toolbox was used to implement a strategy for discovering NRPS encoded in newly sequenced genomes. We performed systematic BLAST analyses of all protein coding sequences (CDS) of a genome of interest against the database constituted by reference NRPSs stored in Doris. The best significant hits were selected. Filtering conditions were tuned manually thanks to the Korilog's KoriBlast software facility. Best hits were found with a length greater than 500 amino acids, displaying more than 5 HSPs, with the best HSP at a percentage of identity and similarity greater than 28% and 45% respectively, an E-value close to zero and a gap percentage below 10 %. The MODIM system then retrieved and integrated the annotations and positions of these CDS on the genome as well as their domain composition.

3.1 Discovery of new nonribosomal peptide synthetases

To validate our strategy, a first experiment was performed on four bacterial genomes (three *Burkholderia* chromosomes and one *Bacillus cereus* chromosome) and produced 86 BLAST hits. Domain analysis by the specialized NRPS tools provided us with NRPS domains. At this stage of the work we therefore decided to match the primary NRPS A, T, C, and Te domains with InterPro domains (see Table 1). The InterProScan localization tool was then used to retrieve start and end positions of each domain on the protein sequence. Occurrences of A, T and C domains in the same or in contiguous CDS were found for 9 protein hits delineating 15 complete NRPS elongation modules and one termination module. An example of data view obtained for *Burkholderia ambifaria* chromosome 1 is presented in Table 2.

NRPS primary domain	InterPro id	source database id
adenylation domain	IPR01007	TIGR01733
thiolation domain	IPR006163	PF00550
condensation domain	IPR001242	PF00668
thioesterase domain	IPR001031	PF00975

Table 1. Relationship between A, T, C and Te domains and InterPro domains.

Domain	Module nb	InterPro id	Hit : UniProt_ID	Domain start (aa)	Domain end (aa)	Genome Id	Locus Id
A	1	IPR010071	B1YQA7	39	449	NC_010551	BamMC406_1558
T	1	IPR006163	B1YQA7	536	600	NC_010551	BamMC406_1558
C	2	IPR001242	B1YQA7	632	929	NC_010551	BamMC406_1558
A	2	IPR010071	B1YQA7	1111	1513	NC_010551	BamMC406_1558
T	2	IPR006163	B1YQA7	1603	1663	NC_010551	BamMC406_1558

C	3	IPR001242	B1YQA7	1687	1983	NC_010551	BamMC406_1558
iprE	3	IPR010060	B1YQA7	1989	2138	NC_010551	BamMC406_1558
C	3	IPR001242	B1YQA7	2158	2451	NC_010551	BamMC406_1558
A	3	IPR010071	B1YQA7	2643	3058	NC_010551	BamMC406_1558
T	3	IPR006163	B1YQA7	3139	3201	NC_010551	BamMC406_1558
C	4	IPR001242	B1YQA8	50	350	NC_010551	BamMC406_1559
A	4	IPR010071	B1YQA8	540	946	NC_010551	BamMC406_1559
T	4	IPR006163	B1YQA8	1036	1097	NC_010551	BamMC406_1559
C	5	IPR001242	B1YQA8	1124	1425	NC_010551	BamMC406_1559
T	5	IPR006163	B1YQA8	1593	1655	NC_010551	BamMC406_1559

Table 2. View on Doris data summarizing the module signatures obtained for *Burkholderia ambifaria* chromosome 1. iprE : epimerization domain (see below)

We introduce here the concept of "module signature" which is a set of ordered protein domains always encountered in modules sharing similar function. For example the NRPS elongation module signature is composed of the three C, A, T domains (< C, A, T > signature), whereas the NRPS termination module signature is composed of C, A, T and Te domains (< C, A, T, Te > signature). When secondary domains are detected in modules associated to some specific function, enriched module signatures can be proposed (see below for modules containing an epimerisation domain).

The prediction of monomers was carried out with specialized tools [2, 3, 4] for each A domain of our 16 NRPS modules. The prediction remains impossible for 4 A domains pointing out the limits of the available tools.

3.2 Characterization of new signatures for optional domains

A unique advantage of the NRPS toolbox is the possibility, when this information is available, of matching the structure of a NRPS with the structure of the peptide it produces. This led us to investigate the domain structure of some NRPS responsible for the synthesis of peptides containing D-monomers. We thus identified two groups of such NRPS. In the first group, containing for example bacitracine and gramicidine synthetases from firmicutes, classical NRPS tools are able to detect a so-called E (epimerization) domain in modules responsible for the condensation of D-monomers. In fact, E domains are always followed by C domains in these modules. Moreover, InterProScan analysis of E domains reveals that these epimerisation domains are composed of an IPR001242 (C) domain followed by an extension of about 130 amino acids (iprE for InterPro Epimerisation domain) recognized as the IPR010060 Interpro domain. The enriched signature < C, iprE, C, A, T > can thus be defined for such modules.

The second group of NRPS (for example massetolide and arthrofactin synthetases from *Pseudomonas*) includes NRPS modules corresponding to the condensation of D-monomers but lacking iprE domains. In such modules only regular A, T and C domains are observed. No other InterPro domain is detected by InterProScan. We therefore carefully analyzed inter-domain regions searching for a yet undescribed epimerisation domain. We observed that a region of constant length of 186 amino acids is always present downstream the C domain in all modules of these NRPS. Multiple sequence alignment of 137 instances of this region (downC-186) lead to distinguishing two clusters of highly conserved sequences (downC-186, and downC-186-E). Interestingly sequences of downC-186-E cluster are always found in modules responsible for the condensation of D-monomers but not in any other module of this group of NRPS. We thus propose < C, downC-186-E, A, T > as an enriched module signature for a new type of NRPS modules responsible for the condensation of D-monomers.

4 Conclusion and perspectives

In conclusion we have shown here that our NPRS toolbox is a unique and useful resource for characterizing NRPS and further exploring the relationships between structure of NRPS modules and the type of incorporated monomers. In future work we will apply machine learning methods to refine signature description for modules associated with a given monomer. This will lead to improve peptide prediction and to better understand the function of NRPS for which no peptide is yet described. Ultimately, the NRPS

toolbox will become a precious resource for designing recombinant NRPS to produce synthetic active compounds such as novel antibiotics.

Acknowledgements

This work was supported by PPF Bioinformatique of Lille 1 University and FEDER (INTERREG IV PHYTOBIO project). We wish to thank INRIA and the CPER-Region Lorraine for their financial support , Birama Ndiaye for his help with the MODIM system. We acknowledge the contribution of Constant Denis, Nicola Gref, Jean-Philippe Monnerville and Hélène Polvèche during their student internships.

References

[1] 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)

[2] 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)

[3] 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.

[4] M. Röttig, MH. Medema, K. Blin, T. Weber, C. Rausch, and O. Kohlbacher. NRPSpredictor2 - a web server for predicting NRPS adenylation domain specificity. Nucl. Acids Res., 39:W362–W367, 2011. (http://nrps.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.

[7] B. Ndiaye, E. Bresso, M. Smail-Tabbone, M. Souchet and MD. Devignes, MODIM : Model Driven Data Integration for Mining. JOBIM, 2011 (poster).

[8] S. Yilmaz, P. Jonveaux, C. Bicep, L. Pierron, M. Smail-Tabbone and MD. Devignes Gene-disease relationship discovery based on model-driven data integration and database view definition, *Bioinformatics*, 25:230-236, 2002.

Session 5 : Proteins and Interactions

Conférence invitée

Pierre BALDI

University of California in Irvine, United states

Machine Learning Approaches in Proteomics

Over the past three decades machine learning approaches have had a profound influence on many fields, including bioinformatics. Here we will provide a brief historical perspective on machine learning and its applications to proteomics, particularly structural proteomics, and discuss why structural proteomics is important for machine learning. We will then present state-of-the art machine learning methods for predicting protein structures and structural features, from secondary structure to contact maps. We will stress and demonstrate the importance of combining supervised and unsupervised learning, and using deep and modular architectures capable of integrating information over space and "time" at multiple scales. Finally, we will describe two proteomic applications that have benefited from statistical machine learning methods: (1) the discovery of new drug leads for neglected diseases;and (2) the development of high-throughput platforms to study the immune response with applications to antigen discovery and vaccine development.

Protein-protein interaction network inference with semi-supervised Output Kernel Regression

Céline BROUARD[1], Marie SZAFRANSKI[2,1] and Florence D'ALCHÉ-BUC[1,3]

[1] IBISC, EA 4526, 23 bd de France, Université d'Évry Val d'Essonne, 91037 Évry cedex, France
{celine.brouard, florence.dalche,marie.szafranski}@ibisc.fr
[2] ÉNSIIE, 1 square de la résistance, 91025 Évry cedex, France
[3] LRI, UMR CNRS 8623, bât 650, Université Paris-Sud 11, 91405 Orsay Cedex, France

Abstract *In this work, we address the problem of protein-protein interaction network inference as a semi-supervised output kernel learning problem. Using the kernel trick in the output space allows one to reduce the problem of learning from pairs to learning a single variable function with values in a Hilbert space. We turn to the Reproducing Kernel Hilbert Space theory devoted to vector- valued functions, which provides us with a general framework for output kernel regression. In this framework, we propose a novel method which allows to extend Output Kernel Regression to semi-supervised learning. We study the relevance of this approach on transductive link prediction using artificial data and a protein-protein interaction network of S. Cerevisiae using a very low percentage of labeled data.*

Keywords Protein-protein interactions, Link prediction, Kernel methods, Operator-valued kernel, RKHS, Semi-supervised learning, Transductive learning.

1 Background

Recent years have witnessed a surge of interest for network inference in biological networks. *In silico* inference of protein-protein interaction (PPI) networks is motivated by the cost and the difficulty to experimentally detect physical interactions between proteins. It mainly relies on the assumption that some input features relative to the proteins, such as amino acids sequences, gene expressions or localizations, could provide valuable information about the presence or the absence of a physical interaction. Two main approaches are devoted to PPI network inference: supervised approaches, which aim at building a pairwise classifier able to predict if two proteins interact from labeled pairs of proteins [1,2,3,4,5], and matrix completion approaches [6,7].

Let us define $\mathcal{O}$ the set of descriptions of the proteins we are interested in. Let $f : \mathcal{O} \times \mathcal{O} \to \{0, 1\}$ be a classifier that, given the predictions of two proteins, predicts if these proteins interact or not. In this work, we have chosen to convert the binary pairwise classification task into an output kernel learning task, referred as Output Kernel Regression (OKR). As in [3,4], we assume the existence of an output kernel $\kappa_y : \mathcal{O}\times\mathcal{O} \to \mathbb{R}$ that encodes the proximities of proteins in terms of nodes in the interaction network. Then, given an approximation $\widehat{\kappa_y}$ of this output kernel, we can define a classifier f_θ by thresholding its output values:

$$f_\theta(o, o') = \mathrm{sgn}(\widehat{\kappa_y}(o, o') - \theta)\,.$$

κ_y being a positive semi-definite kernel, there exists an Hilbert space $\mathcal{F}_y$, called the feature space, and a feature map $y : \mathcal{O} \to \mathcal{F}_y$ such that $\forall (o, o') \in \mathcal{O} \times \mathcal{O}, \kappa_y(o, o') = \langle y(o), y(o')\rangle_{\mathcal{F}_y}$. In output kernel regression, we build an approximation of κ_y from the inner product between the outputs of a single input function $h : \mathcal{O} \to \mathcal{F}_y : \widehat{\kappa_y}(o, o') = \langle h(o), h(o')\rangle_{\mathcal{F}_y}$. Using the kernel trick in the output space allows one to reduce the problem of learning from pairs to learning a single variable function h with values in a Hilbert space (the output feature space $\mathcal{F}_y$).

Previous works developed tree-based Output Kernel Regression models by extending multiple output regression trees (OK3) and ensemble methods to output feature space endowed with a kernel [3,4]. In this work, we propose a novel method, which allows to extend Output Kernel Regression to the semi-supervised framework.

2 Semi-supervised Output Kernel Regression

In biology, it is often the case that labeled pairs of proteins are difficult to obtain, due to the cost and the time needed for experimental methods. However, additional description of protein properties are often available. This motivates for dealing with the problem using semi-supervised approaches. Graph-based regularization is a powerful approach to semi-supervised regression that enforces the smoothness of the function, permitting to propagate output labels over close inputs [11,12]. Belkin et al. [12] have proposed to explicitly embed such ideas into the framework of regularization within Reproducing Kernel Hilbert Space (RKHS) for real-valued functions.

In the context of OKR, the function to be learnt is not real-valued but vector-valued in the output Hilbert space. If we want to take benefit from the theoretical framework of Reproducing Hilbert Space theory (RKHS), well appropriate for regularization, we need to turn to the proper RKHS theory, devoted to vector-valued functions [9,10]. This theory requires to define an operator-valued kernel instead of a scalar input kernel. As in RKHS theory with scalar valued functions, representer theorems for different loss functions can be proven.

Let $\{(o_i, \mathbf{y}_i)\}_{i=1}^{\ell}$ be a set of labeled examples and $\{o_i\}_{i=\ell+1}^{\ell+u}$ a set of unlabeled examples. Let $\mathcal{H}$ be a RKHS with reproducing kernel $\mathcal{K}_x$, and a symmetric matrix W with positive values measuring the similarity of proteins in the input space. In this work, we propose to learn the vector-valued function h by minimizing a penalized least square cost function with a smoothness constraint :

$$\underset{h\in\mathcal{H}}{\operatorname{argmin}}\, J(h) = \sum_{i=1}^{\ell} \|h(o_i) - \mathbf{y}_i\|^2_{\mathcal{F}_y} + \lambda_1 \|h\|^2_{\mathcal{H}} + \lambda_2 \sum_{i,j=1}^{\ell+u} W_{ij} \|h(o_i) - h(o_j)\|^2_{\mathcal{F}_y} \,, \tag{1}$$

with λ_1 and $\lambda_2 > 0$.

We stated and proved a new representer theorem devoted to semi-supervised learning with a penalized least-square cost. Then, given a simple definition of the operator-valued kernel based on some input scalar kernel, we derived a close-formed solution that extends the reformulated KDE proposed by Cortes et al. [14] to the semi-supervised case. [1]

3 Experiments

3.1 Experimental protocol

The approach was evaluated in the transductive setting. For different percentage values of labeled proteins, we randomly picked a subsample of proteins as labeled examples. Labeled interactions correspond to interactions between two labeled proteins, and we searched to predict the remaining interactions. A 10% selection of labeled proteins therefore corresponds to only 1% labeled interactions. We evaluated the performance by averaging the AUC-ROC over 10 random choices of the training set.

3.2 Results

We extensively studied the behaviour of the provided model using an artificial dataset produced by sampling random graphs form a Erdős-Renyi law with different probabilities of presence of edges. The input features were obtained by applying Kernel Principal Component Analysis on the diffusion kernel associated with the graph, and using the components capturing 95% of the variance. We observe from the results obtained that the semi-supervised approach improves upon the supervised one on AUC-ROC, especially for a small percentage of labeled data (up to 10%). Based on these results one can formulate the hypothesis that supervised link prediction is harder in the case of more dense networks and that the contribution of unlabeled data seems more helpful in this case. One can also assume that using unlabeled data increases the AUCs for low percentage of labeled data. But when enough information can be found in the labeled data, semi-supervised learning does not improve the performance.

1. The details of this method are given in the publication [13].

We also illustrated our method on a protein-protein interaction network of the yeast *S. Cerevisiae* composed of 984 proteins linked by 2438 interactions. To reconstruct the PPI network, we used gene expression data, phylogenetic profiles, protein localization and protein interaction data derived from yeast two-hybrid experiments as input features [2,3,4,5,6]. For each of these features, our method compares favorably with the existing methods in the supervised setting [13]. In a second step, we experimented the method in the transductive setting using gene expression as input features. Fig. 1 reports the averaged AUC-ROC and the standard deviations for different values of λ_2 and different percentages of labeled proteins. One can see that the semi-supervised approach improves the AUC-ROC upon the supervised one (corresponding to $\lambda_2 = 0$) for all percentage values. This improvement is especially significant when the percentage of labeled proteins is low, which is usually the case in PPI network inference problems.

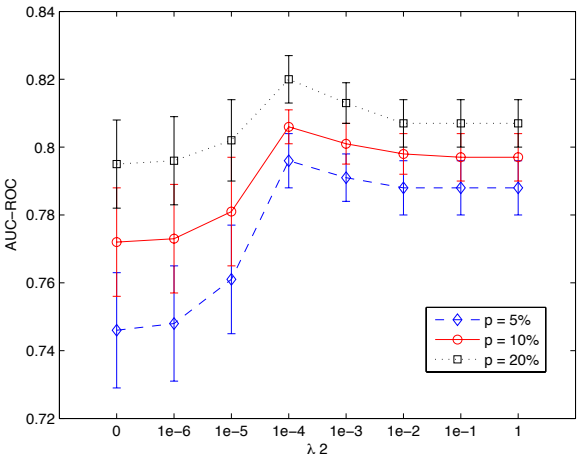


Figure 1. Averaged AUC-ROC results for the reconstruction of the Yeast PPI network from gene expression data in the supervised and semi-supervised settings. The percentage values correspond to the proportions of labeled proteins.

References

[1] A. Ben-Hur and W.S. Noble, Kernel methods for predicting protein-protein interactions, *Bioinformatics*, vol. 21, pp. 38-46, 2005.

[2] Y. Yamanishi, J.-P. Vert and M. Kanehisa, Protein network inference from multiple genomic data: a supervised approach, *Bioinformatics*, vol. 20, pp. 363-370, 2004.

[3] P. Geurts, L. Wehenkel and F. d'Alché-Buc, Kernelizing the output of tree-based methods, in *Proceedings of the 23th International Conference on Machine learning*, 2006.

[4] P. Geurts, N. Touleimat, M. Dutreix and F. d'Alché-Buc, Inferring biological networks with output kernel trees, *BMC Bioinformatics*, 8, 2007.

[5] K. Bleakley, G. Biau and J.-P. Vert, Supervised reconstruction of biological networks with local models, *Bioinformatics*, vol. 23, pp. i57-i65, 2007.

[6] T. Kato, K. Tsuda and K. Asai, Selective integration of multiple biological data for supervised network inference, *Bioinformatics*, vol. 21, pp. 2488-2495, 2005.

[7] K. Tsuda and W.S. Noble, Learning kernels from biological networks by maximizing entropy, *Bioinformatics*, vol. 20, pp. 326-333, 2004.

[8] R.I. Kondor and J.D. Lafferty, Diffusion Kernels on Graphs and Other Discrete Input Spaces, In *Proceedings of the 19th International Conference on Machine Learning*, 2002.

[9] E. Senkene and A. Tempel'man, Hilbert Spaces of operator-valued functions, *Lithuanian Mathematical Journal*, 13, pp. 665-670, 1973.

[10] C. A. Micchelli and M. A. Pontil, On Learning Vector-Valued Functions, *Neural Computation*, 17, pp. 177-204, 2005.

[11] D. Zhou, O. Bousquet, T.N. Lal, J. Weston and B. Scholkopf, Learning with Local and Global Consistency, in *Advances in Neural Information Processing Systems 16*, 2004.

[12] M. Belkin, P. Niyogi and V. Sindhwani, Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples, *Journal of Machine Learning Research*, 7, pp. 2399-2434, 2006.

[13] C. Brouard, F. d'Alché-Buc and M. Szafranski, Semi-supervised Penalized Output Kernel Regression for link prediction, *Proceedings of the 28th International Conference on Machine Learning*, 2011.

[14] C. Cortes, M. Mohri and J. Weston, A general regression technique for learning transductions, in *Proceedings of the 22nd International Conference on Machine Learning*, pp. 153-160, 2005.

Improving gene signatures by the identification of differentially expressed modules in molecular networks : a local-score approach.

Marine JEANMOUGIN[1,2], Christophe AMBROISE[1], Matthieu BOUAZIZ[1,2] and Mickaël GUEDJ[2]

[1] STATISTICS AND GENOME, UMR8071 CNRS, 23 Boulevard de France, 91037 Évry, France
{marine.jeanmougin, christophe.ambroise}@genopole.cnrs.fr
[2] PHARNEXT, 11 rue des peupliers, 92130 Issy-les-moulineaux, France
{mguedj,mbouaziz}@pharnext.com

Abstract *Discovery of gene signatures for disease prognostic and diagnostic, has become a topic of much interest during the last decade. The identification of relevant and robust signatures is seen as a major step towards a better personalized medicine. Various methods have been proposed for that purpose. However, the detection of a set of genes from microarray data is a difficult statistical problem and signatures obtained from standard tools suffer from lack of reproducibility across studies. Therefore, it is difficult to achieve a clear biological interpretation from such signatures.*
We designed an approach for the selection of functional modules (i.e. subnetworks) of genes through the integration of interactome and transcriptome data. Using a strategy based upon the local-score, a statistic that found many applications in biological sequence analysis, we aim to identify subnetworks locally enriched in genes associated with phenotypes of interest.
We proved through simulations that the resulting modules are highly reproducible. In addition the method appears to be more powerful than classical strategies of gene selection. The potential of our method to highlight relevant biological phenomena is illustrated on breast cancer data to study the Estrogen Receptor (ER) status of tumors.

Keywords gene signatures; functional modules; protein-protein interactions network; differential analysis; breast cancer.

1 Introduction

The development of high throughput genomic and genetic technologies has provided insights into the biological mechanisms underlying diseases onsets and evolutions. Through the identification of biomarker genes (or so called signatures), the aim is to improve diagnosis, prognosis and clinical decisions about treatments of a given disease. One of the most widely used approach for the identification of such signatures consists in detecting differentially expressed genes whose expression levels change between two or more experimental conditions. Such strategies imply to compute an individual score for each gene under study. This score denotes the ability of a gene to discriminate between conditions of interest (patients *versus* controls for instance).
In practice, signatures selected in comparable studies share only few genes in common. Moreover it is generally difficult to achieve a clear biological interpretation of lists of differentially expressed genes as they focus on the level of genes instead of molecular functions or biological processes. Motivated by the observation that genes causing the same phenotype are likely to interact together, we explore an approach for identifying modules, *i.e.* genes that are functionally related, rather than individual genes. The idea is to combine topological features, extracted from Protein-Protein Interaction (PPI) networks for instance, and experimental data to detect modules of connected genes associated with disease or phenotypes of interest.
In recent years, there have been several attempts to integrate knowledge on PPIs, regulatory networks or canonical pathways into gene selection strategies. One of the first approaches was described in [1] and involves a sliding window model to identify differentially expressed subnetworks. It uses the mutual information statistic to measure the association between a subnetwork expression profile and a given phenotype and select significantly differentially expressed subnetworks by comparing their discriminative potentials to those of random networks. In [2], the authors introduced an approach to extract disease-specific gene networks from both DNA microarray measurements and an initial network constructed from protein-protein and genetic interactions as

well as genes regulation information. Their strategy is to identify significant changes of correlations between two genes under various conditions. Another method developed by Haury *et al.* in [3] and based upon an extension of the Lasso regression is the graph Lasso. Instead of using a classical ℓ_1-norm penalty, they proposed a new penalty to incorporate the gene network information, leading to the selection of genes that are often connected to each other in PPI networks or pathways.

In this work, we designed a new method named DiAMS (Disease Associated Modules Selection) implemented in R, which involves a local-score strategy. By combining functional information between genes or gene products through a PPI network and measures of differential expression, the aim is to identify modules locally enriched in disease associated genes. That way, DiAMS enables the selection of candidate modules with a measure of statistical significance. We evaluated the performance of the proposed approach through simulations and we found conclusive results on power gains and stability improvements in comparison to classical approaches. The final endgame of our method is not only to make gene signatures more stable, but also more interpretable in a biological sense. To illustrate that point, we applied our approach to study the Estrogen Receptor (ER) status in breast cancer. We highlighted the relevance of the resulting signature on real data.

2 Methods

The methods section is outlined as follows: first, we introduce the DiAMS strategy dedicated to the selection of modules significantly enriched in disease associated genes. Then we detail the local-score approach for module scoring. Finally, in the last section, we describe the simulation process for evaluating the performance of DiAMS in terms of power, false-positive rate and reproducibility.

2.1 Global Approach

In the present subsection we detail the global strategy to search for functional modules presenting unexpected accumulations of genes associated to a phenotype of interest in a gene network . One of the strongest manifestation of functional relations between genes is protein-protein interactions. That is why we use PPI networks in the following to illustrate our method but other sources of information can be used, such as pathways or regulation networks. The main issue when working with biological networks lies in the impossibility of exploring the huge space of possible gene subnetworks. In [1], Chuang *et al.* strategy was to define an initial "seed", *i.e.* starting points for candidate subnetworks and to look at the effect of the addition of a gene in a module within a specified distance $d = 2$ from the seed. In practice, it leads to the identification of modules made up of only few genes. Moreover it requires to define a limited set of starting points for the algorithm. In this paper, we propose an alternative strategy which allows to screen the entire network without constraints on module sizes by converting the network into a tree structure using a clustering algorithm. This is motivated by the observation that biological graphs are globally sparse but locally dense, *i.e.* there exist groups of vertices, called communities, highly connected between them but with few links to other vertices. So, by applying a strategy of clustering which enables to obtain a hierarchical community structure we are able to capture much information about the network topology. The main advantage is that it is relatively simple to go through the tree due to the hierarchical structure instead of exploring all possible subnetworks. Several algorithms have been developed for inferring hierarchical structure from network data (see for instance Newman *et al.* [10] and Pons *et al.* [11]) and any of them can be used in the process. In the following application on breast cancer, we applied the strategy of [11] based upon random walks.
Thus the preliminary step of our approach is to convert the network structure into a relevant tree structure which constitutes the first input parameter of the method. A module is no longer defined as a subnetwork but as a subtree of the hierarchical structure (see Figure Fig. 1).

The second parameter that has to be passed to the method is a vector of scores that quantify for each gene its association to the disease. In this work the scoring function is related to the differential expression of the gene such as significant genes, *i.e.* those that are significantly differentially expressed, have a higher score than non-significant genes.

Once both the tree structure and the score vector have been defined, we search for accumulation of high-scoring genes in the tree. The strategy for the selection of significant modules can be described in the following three-step algorithm:

1. The first step consists in enumerating all the modules described by the tree structure. Thus we obtain a list of modules that will be updated during the following process.
2. The second step involves a local-score approach that we briefly describe here. The local-score statistic dedicated to module scoring is then deeply detailed in the next subsection.
   This step is an iterative algorithm: (i) each module of the list is scored by summing the scores, denoted σ_g, of all the genes that constitute it (see Fig. 1), (ii) then the highest-scoring module is identified and removed from the list of modules, (iii) finally we repeatedly applied (i) and (ii) steps until all disjoint modules have been enumerated. Thus, we obtain a list of m modules and their respective local-scores $\mathrm{L}_1, ..., \mathrm{L}_m$ such as $\mathrm{L}_1 > ... > \mathrm{L}_m$ with the i-th best module being disjoint from the preceding ith $-$ 1 best modules.

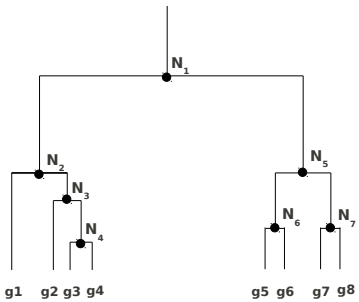


Figure 1. Module description - A module is defined as a subtree of the hierarchical structure. For instance, the module rooted at the node N_3 is composed of three leaves (*i.e.* genes). Its score is the sum of each individual gene score, σ_{g2}, σ_{g3} and σ_{g4}.

3. Given $\mathrm{L}_1, ..., \mathrm{L}_m$, the last step proposes a way to select a set of modules significantly enriched in disease associated genes. We assess the global significance of each module via Monte-Carlo simulations by permuting sample labels for each gene and computing the distribution of module scores under the null hypothesis that there is no differential expression across the sample groups. Through this permutation procedure we obtain a p-value for each module and are able to conclude about its significance of a module at a given level.

2.2 Module Scoring Through the Local-Score Statistic

In this subsection we detail the module scoring strategy based on a local-score approach. The local-score statistic is a matter of interest in biological sequence analysis and is defined as the value of a sequence with the maximal sum of scores. It found many applications in pattern identification to locate transmembrane or hydrophobic segments, DNA-binding domains as well as regions of concentrated charges (see [4,5,6]). For instance, in the context of detection of hydrophobic regions in a peptide sequence, the scoring function will assign positive or negative scores to amino acids according to the polarity of their side-chains.

We propose to extend the local-score to the detection of module of genes in a hierarchical community structure denoted $\mathcal{H}$ in the sequel. In this context we call L, the local-score statistic defined as the value of the highest-scoring module:

$$L = \max_{H \subseteq \mathcal{H}} \left(\sum_{g \in H} \sigma_g \right),$$

such as H is included in $\mathcal{H}$ if H is a subtree of $\mathcal{H}$, *i.e.* H can be obtained from $\mathcal{H}$ by deleting nodes in $\mathcal{H}$. σ_g is a scoring function of the gene g, detailed in the next paragraph. Note that a module is maximal in the sense that it can not be extended or shortened without reducing the local-score statistic.

We define the scoring function as follows: $\sigma_g = Z_g - \delta$, such as Z_g is the individual score of each gene g and δ a constant specified in the following. The individual score Z_g is based on the p-value of the `limma` t-statistic [7], which appears to be the most efficient test in the literature. Indeed it has been shown that it provides substantial power improvements compared to classical approaches such as the t-test or the Anova [8,9].
Let us represent the expression levels of the p genes by a Gaussian random vector $X = (X_1, ..., X_p) \in \mathbb{R}_p$ and let $X_{ig}^{(c)}$ be the expression level of the ith sample for gene g under condition c. The statistic used in `limma` is very similar to an ordinary t-statistic except that the estimations of gene expression variances S_g^{limma} are moderated across genes to provide robustness for the estimation of specific gene variances, shrinking possibly extreme empirical variances that one could expect from small numbers of samples:

$$t_g^{\text{limma}} = \frac{\bar{x}_{\cdot g}^{(1)} - \bar{x}_{\cdot g}^{(2)}}{S_g^{\text{limma}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

where n_1 and n_2 are the numbers of samples relative to the two conditions of interest and $\bar{x}_{\cdot g}^{(i)}$ is the average expression level for gene g under condition i. The `limma` t-statistics lead to p-values for each gene, called p_g , in the same way that ordinary t-statistics do. Given that a high score Z_g denotes a high chance of association with phenotypes of interest, the p-values need a transformation such as $Z_g = -log_{10}(p_g)$ to be used as an individual score for each gene.
A constraint of the strategy is to have negative expected individual scores, *i.e.* $\mathbb{E}(\sigma_g) \leq 0$, otherwise the module with the highest score would easily span the entire tree. In consequence, a constant δ must be subtracted to get corrected scores. Genes with a score higher than δ will improve the cumulative score of a given module whereas genes with a score below the threshold will penalize it.

2.3 Evaluation Strategy

In this part we detail the strategy adopted to evaluate DiAMS. Each evaluation criterion, namely the power, the type-I error rate and the reproducibility, are compared between our strategy and the individual scoring counterpart, `limma`.

2.3.1 Power Study Recent results suggest that genes leading to similar disease phenotypes have similar functional annotations. In other words, genes involved in the molecular mechanisms of genetic diseases interact together in functional modules. To reproduce this modular activity we simulate modules of genes under H_1 by randomly sampling nodes in the tree structure. The expression levels of the genes in modules under H_1 are assumed to change between the phenotypes of interest. Thus we simulate an expression matrix for two conditions according to the gene status: differentially expressed or not. Gene expression level $X_{ig}^{(c)}$ is drawn from a Gaussian distribution of variance $\sigma^2 = 1$, such as:

$$\begin{cases} X_{ig}^{(1)} & \sim & \mathcal{N}(\mu_g^{(1)}, \sigma^2) \\ X_{ig}^{(2)} & \sim & \mathcal{N}(\mu_g^{(2)}, \sigma^2), \end{cases}$$

where $\mu_g^{(c)}$ is the mean expression level of the gene g under condition c.
The null hypothesis, H_0, assumes that: $\{\mu_g^{(1)} = \mu_g^{(2)}\}$ while the various alternative hypotheses, H_1, assume: $\{\mu_g^{(2)} = \mu_g^{(1)} + \Delta\}$, with Δ in $\{0.5, 0.75, 1, 1.25, 1.5, 2, 3\}$. For each value of Δ, we perform $1,000$ simulations and compute the power, *i.e.* the ability to reject H_0 when it is actually false:

$$\text{Power} \quad = \quad \mathbb{P}_{H_1}(\text{rejected } H_0).$$

The p-values obtained from Monte-Carlo permutations are adjusted using the Benjamini-Hochberg procedure to control the FDR criterion at a 5% level.

2.3.2 False-Positive Rate. Using the same simulation strategy as described in the subsection 2.3.1, we assess the false-positive rate or type-I error rate. Given a statistical test, the false-positive rate, denoted α, commonly refers to the probability of rejecting reject H_0 , H_0 being true. A statistical test conducted at a significance level of 0.05 should control the false-positive rate at 5%. Thus, by simulating an entire dataset under H_0, *i.e.* $\forall g : \Delta = 0$, we evaluate the proportion of genes spuriously selected as significant. Both the limma and DiAMS false-positive rates are estimated for various sample sizes ranging from 5 to 50 samples per condition.

2.3.3 Reproducibility Study Next, we examine the agreement between signatures using a sub-sampling procedure. As described in the power study, we simulate nodes under H_1 as well as the corresponding expression matrix and compute a signature of reference. Then, we randomly sub-sample the replicates of the initial matrix with replacement and estimate the signature again. The reproducibility is calculated as the overlap between the reference signature and the signature of sub-sampled matrices. This procedure is performed for various sub-sample sizes.

3 Results

3.1 Power and False-Positive Rate Studies

Fig. 2 illustrates the results of the power analysis for both the DiAMS approach and the `limma` statistic. We plotted the mean of the power over the 1,000 simulations and its 95% confidence interval for various values of Δ, that is the difference between the mean expression levels simulated under H_0 and H_1. As expected, the curve describing the statistical power converges to 1 with increasing value of Δ. For $\Delta = 0.5$, it appears that the power is very similar for both approaches , although DiAMS is slightly more powerful. For all values of Δ in $\{0.75, 1, 1.25, 1.5, 2\}$, we observe large differences in power between the two approaches with DiAMS outperforming `limma`.

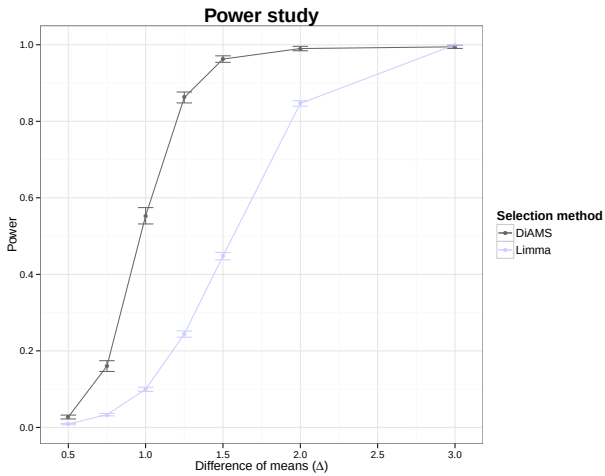


Figure 2. Power study - The mean of power values over the 1,000 simulations are calculated at a 0.05 FDR level for the DiAMS method (in dark gray) and the `limma` statistic (in light gray) and displayed according to Δ, the difference of mean expression levels under H_0 and H_1.

Fig. 3 shows the estimated false-positive rates for both selection methods and various sample sizes. It appears that the rates are similar for both approaches and they lie within the 95% confidence interval. For each sample size, `limma` and DiAMS meet the theoretical false-positive rate.

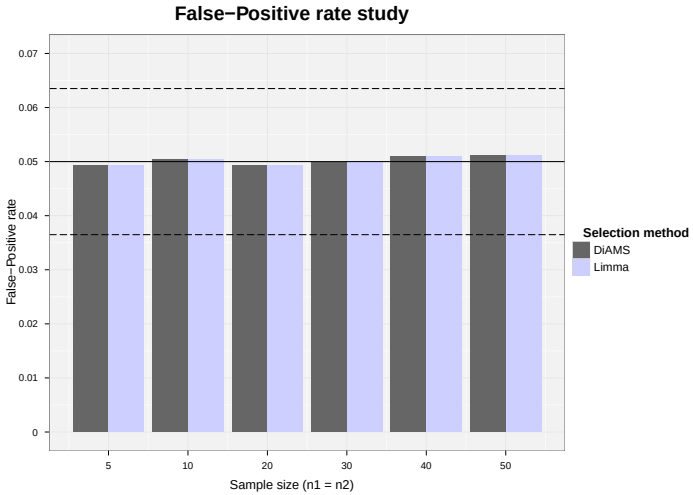


Figure 3. False-positive rate study - The estimated false-positive rate over the 1,000 simulations are displayed for the DiAMS method (in dark gray) and the `limma` statistic (in light gray) for various sample sizes such as $n_1 = n_2$. The plain black line represents the 5% level at which the tests were conducted. The dashed black lines are the 95% confidence intervals for this level.

3.2 Reproducibility Study

We also compared the stability of both gene selection methods by evaluating the reproducibility of the signatures. Fig. 4 shows the percentage of overlap between a signature of reference estimated from an initial dataset containing 50 samples for both conditions, and the signatures obtained on sub-sampled datasets. The reproducibility is performed for various sub-sample sizes and averaged over the 10,000 simulations.
For the larger sample size ($n_1 = n_2 = 50$), the initial matrix has been re-sampled with replacement. Even if the sample size is the same, meaning that the noise added to the initial dataset is relatively low, the percentage of reproducibility for `limma` is only 90% while DiAMS almost reaches 100%. All the results displayed in Fig. 4 show that `limma` is very sensitive to the noise in data while DiAMS' results appear to be more consistent. It is particularly true for small sample sizes. Indeed, for $n_1 = n_2 = 5$ the reproducibility of the signature is about 95% with the DiAMS approach while the percentage is almost null (0.3%) with the `limma` selection method. The gap remains very large for the other sample sizes and DiAMS clearly provides significantly better results than `limma` in terms of reproducibility.

3.3 Application

We applied our method to breast cancer data in order to study the Estrogen Receptor status of tumors.
Breast cancer is a heterogeneous disease comprising various subtypes defined by their amplification status of the epidermal growth factor receptor-2 gene (ERBB2) and the presence of hormone receptors. In this work we are particularly interested in the Estrogen Receptor (ER) status which is an essential parameter of the pathological analysis of breast cancer. Estrogen is a female sex hormone that may stimulate the growth of cancer by triggering particular proteins (receptors) in the damaged cells. If breast cancer cells have Estrogen Receptors,

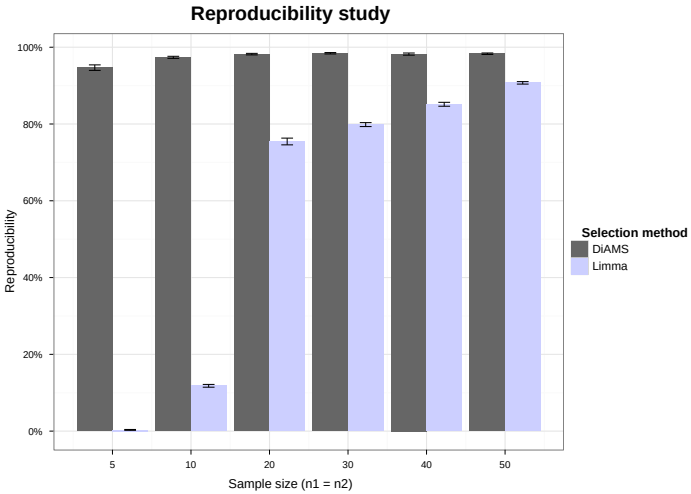


Figure 4. Reproducibility study - This bar-plot displays the results of the reproducibility analysis for which we compute the mean of the overlap between a signature of reference and signatures of sub-sampled expression matrices over 10,000 simulations. We represent the 95% confidence interval for each sample size.

the cancer is said to be ER positive (ER+). The corresponding ER+ tumors are routinely treated using anti-hormonal therapies, such as tamoxifen, an antagonist of ER, that can stop estrogen from stimulating the cells to divide and grow. ER+ tumors generally have a more favorable prognosis than tumors with estrogen negative status (ER-) that are usually more clinically aggressive. Identifying robust signatures may help to better understand the molecular mechanisms underlying the ER status. That is why we use DiAMS, to select modules of genes associated with the status of Estrogen Receptor.

The analysis used an expression dataset from Guedj *et al.* [12], which consists of the expression measures of about 54,000 probes for 537 tumors (446 ER+ vs. 91 ER-). We borrowe from the HPRD and string databases a human protein-protein interaction network. After removing duplicated PPIs and PPIs that contain proteins that are not mapped into a gene symbol, we restricted our analysis to the genes present in both the expression and PPI datasets. We finally obtained a network comprising about 600,000 interactions and 13,611 proteins.

The association of a gene to the ER status is measured using the `limma` statistic. Then, through the DiAMS approach, 27,221 initial modules (13,611 modules of size 1, *i.e.* individual genes, and 13610 modules described by the tree structure) are tested for enrichment in ER status associated genes. Finally we select 14 significant modules at a 1% FDR level, *i.e.* 159 genes. Among these modules, three of them contain more than 30 genes, while the others are made of 1 to 10 genes. We performed an over-representation analysis (see [13]) for each module to search for pathways that are more represented than expected by chance. We present the results for the five most significant modules in Table 1.

We compared our results to what is known in the literature. Both GATA3 and AGR3 are known to be associated with the ER status. Indeed a study of Voduc *et al.* [14] showed that GATA3 expression is tightly related to the ER expression and Fletcher *et al.* [15] demonstrated that the expression of AGR3 is elevated in ER-responsive breast tumors. The third modules targets the pathway of breast cancer regulation by Stathmin1, an oncoprotein which takes part in the preventive progression of ER+ tumors. Finally the fifth modules is particularly interesting as its genes are involved in PI3K/AKT signaling and Aryl Hydrocarbon Receptor (AHR)

Module	Size	Pathway
1	38	Amino-acid metabolism
2	1	GATA3 - Strong association with ER status
3	35	Breast cancer regulation by Stathmin1
4	1	AGR3 - Involved in ER-responsive breast tumors
5	7	PI3K/AKT signaling & Aryl Hydrocarbon Receptor signaling

Table 1. Over-representation analysis - In this table we show the results of the over-representation test for the five most significant modules. For each module we display its size and the pathway(s) that are significantly over-represented in its set of genes. For modules of size one we did not perform the analysis but specified which is the gene contained in the module.

signaling mechanisms, two well known pathways in breast cancer. In particular, it has been demonstrated that AHR represses the expression of the ER.

This application to the ER status of breast tumors highlights key mechanisms in cancer progression (cellular growth and proliferation through the PI3K/AKT signaling pathway) and ER-specific mechanisms (preventive progression of ER+ tumors through Stathmin1 or repression of the Estrogen Receptor by the Aryl Hydrocarbon Receptor). The DiAMS approach not only enables the selection of a gene signature but also provides insights of the molecular mechanisms underlying the ER status by identifying subnetworks rather than isolated genes.

4 Discussion

We developed a network-based approach named DiAMS for the selection of gene signatures within gene expression profiles. We demonstrated through simulations that both approaches meet the expected false-positive rate but DiAMS exhibits more powerful results than the moderated t-statistic strategy used in `limma` under the assumption of a modular activity of genes. Moreover, the signatures identified through the DiAMS approach are significantly more reproducible. Finally, the application on breast cancer data is a good illustration of the potential of our method to highlight relevant biological phenomena and shows promising results. In particular, such an approach allows to ease the interpretation of the resulting signature by providing information on molecular mechanisms through the extraction of PPI subnetworks.

However the quality and the coverage of the PPI data is one of the main limitation of this approach. In 2008, estimates of the proportion of known protein-protein interactions suggest that in human, about only 10% of interactions have been identified. Moreover, the PPI data are biased towards some particular biological interest. Indeed, some proteins are studied more extensively than others, resulting in a bias into our approach by selecting the most documented proteins. Then, gene selection approaches based on PPI networks are highly dependent on the quality and the amount of available data.

DiAMS has the advantage of being easily adjustable to various types of data. Indeed, the vector of scores passed to the method could be extracted from genetic association test for instance. Moreover the input network could be reconstructed from heterogeneous data: gene regulation information as well as genetic or protein-protein interactions.

Future works should investigate the predictive performance of our method by evaluating its ability to provide accurate classification and prediction. Moreover, we wish to test the reproducibility of signatures obtained from real datasets.

Acknowledgements

We thank Fabrice Glibert, Maurice Baudry and Ilya Chumakov for their support.

References

[1] H.Y. Chuang, E. Lee, Y.T. Liu, D. Lee, T. Ideker, Network-based classification of breast cancer metastasis. *Molecular Systems Biology*, 3:140, 2007.

[2] J. Ahn, Y. Yoon, C. Park, E. Shin, S. Park, Integrative gene network construction for predicting a set of complementary prostate cancer genes. *Bioinformatics*, 27:1846–1853, 2011.

[3] A-C. Haury, L. Jacob, J-P. Vert, Improving stability and interpretability of gene expression signatures. *Technical Report*, 2010.

[4] S. Karlin, P. Bucher, V. Brendel, S. Altschul, Statistical-methods and insights for protein and dna-sequences. *Annual Review of Biophysics and Biophysical Chemistry*, 20:175-203, 1991.

[5] V. Brendel, P. Bucher, I. Nourbakhsh, B. Blaisdell, S. Karlin, Methods and algorithms for statistical analysis of protein sequences. *Proceedings of the National Academy of Science USA*, 89:2002-2006, 1992.

[6] S. Karlin, V. Brendel, Chance and significance in protein and dna sequence analysis. *Science*, 257:39-49, 1992.

[7] G.K. Smyth, Linear models and empirical Bayes methods for assessing differential expression in microarray experiments.*Statistical Applications in Genetics and Molecular Biology*, 3, 2004.

[8] M. Jeanmougin, A. de Reynies, L. Marisa, C. Paccard, G. Nuel, M. Guedj, Should We Abandon the t-Test in the Analysis of Gene Expression Microarray Data: A Comparison of Variance Modeling Strategies. *PLoS ONE*, 5(9), 2010.

[9] C. Murie, O. Woody, A. Lee, R. Nadon, Comparison of small n statistical tests of differential expression applied to microarrays. *BMC Bioinformatics*, 10:45, 2009.

[10] A. Clauset, C. Moore, M.E.J. Newman, Hierarchical structure and the prediction of missing links in networks. *Nature*, 453:98-101, 2008.

[11] P. Pons, M. Latapy, Computing communities in large networks using random walks. *Journal of Graph Algorithms and Applications*, 10(2):191-218, 2006.

[12] M. Guedj, L. Marisa, A. de Reynies, B. Orsetti, R. Schiappa et al., A refined molecular taxonomy of breast cancer. *Oncogene*, 1-11, 2011.

[13] T. Manoli, N. Gretz, H-J. Grone, M. Kenzelmann, R. Eils, B. Brors, Group testing for pathway analysis improves comparability of different microarray datasets. *Bioinformatics*, 22:2500-2506, 2006.

[14] D. Voduc, M. Cheang, T. Nielsen, GATA-3 expression in breast cancer has a strong association with estrogen receptor but lacks independent prognostic value. *Cancer Epidemiol Biomarkers*, 17(2):365-373, 2008.

[15] G.C. Fletcher, S. Patel, K. Tyson, P.J. Adam, M. Schenker, J.A. Loader, L. Daviet, P. Legrain, R. Parekh, A.L. Harris, J.A. Terrett, hAG-2 and hAG-3, human homologues of genes involved in differentiation, are associated with oestrogen receptor-positive breast tumours and interact with metastasis gene C4.4a and dystroglycan. *British Journal of Cancer*, 88(4):579-85, 2003.

Session 6 : Evolution

Conférence invitée

Hugues Roest Crollius

Ecole Normale Supérieure, Institut de Biologie de l'ENS, IBENS, CNRS UMR8197, INSERM U1024, Paris, France

The 4^{th} dimension of Biology.
A historical perspective of biological processes

Biology is an experimental science: hypotheses aiming at a better understanding of molecular, cellular or physiological functions are tested through experiments in living models. But living models (animals, plants, microorganisms) are contemporary, and therefore only present-day biological processes may be studied directly. Yet biological processes have been shaped during hundreds of millions of years of evolution under the influence of multiple forces, including mutational events, population genetics, adaptation, genetic drift, etc. Clearly, taking into account this historical perspective to understand biology would add tremendous power to our ability to interpret present-day experimental results. Most importantly, documenting the precise historical succession of events leading to molecular interactions, cellular development and physiological equilibria observed in contemporary experiments would help us understand "why" biological processes are organised the way we see them. Unfortunately, the historical records required to achieve this have been erased, and no biological samples old and abundant enough to cover the evolution of even one biological function remain today. The only possible approach is to infer ancestral biological states from contemporary data. Genomic data is particularly suited to this undertaking because it is generally highly precise, highly accurate, highly abundant and centralised in publicly available databases. The genome also provides fundamental insights into functional properties of an organism, which in turn inform us on the likelihood that specific metabolic or developmental pathways may have existed, and on the importance of specific functions for the species. Genomes thus represent the foundation upon which many insights may be gained on ancestral biological processes, partly replacing the missing historical records mentioned above. I will present an algorithmic framework developed to reconstruct the organisation of ancestral genomes from their modern descendants. The method relies first on phylogenetic trees built from more than 50 vertebrate proteomes to infer the presence of genes in ancestors, and second on the systematic identification of conserved adjacent gene sets between modern genes to reconstruct the orders of genes in ancestral chromosomes. The method has been extensively validated using realistic simulations of genome rearrangements in vertebrate genomes. I will next describe how this information can be used to examine specific biological questions of interest. First, successive ancestral chromosomes provide a direct way of identifying branch-specific genomic rearrangements, including duplications and inversions. Ancestral gene duplications conserved in modern genomes represent a proxy for the selection of advantageous functions, such as those involved in the immune response, in sexual reproduction, or in specific metabolic pathways. Their precise mapping in genomes and during the course of evolution represent a

global resource to document the history of these adaptive events. Second, the reconstruction of ancestral coding sequences represents a new resource to observe the distribution of GC content along ancestral chromosomes, and should help answer important questions on the evolution of the compositional landscape of chromosomes. Third, ancestral vertebrate genomes reveal ways of studying ancient events such as the two whole genome duplications at the origin of this group, thus providing an unprecedented resource to examine the molecular signatures left by these founding evolutionary events.

Non-adaptive expansion of gene families

Param Priya SINGH[1], Séverine AFFELDT[1] and Hervé ISAMBERT[1]

UMR168 CNRS-UPMC, Institut Curie, 26, rue d'Ulm, 75248 Paris, France
{param-priya.singh, severine.affeldt, herve.isambert}@curie.fr

Session 7 : Systems Biology

Logical modelling of cellular decision processes with GINsim

Claudine CHAOUIYA[1,2], Aurélien NALDI[3], Lionel SPINELLI[1,4], Pedro T. MONTEIRO[2], Duncan BERENGUIER[1,4,5], Luca GRIECO[1,5,6], Abibatou MBODJ[1,5], Samuel COLLOMBET[6], Anna NIARAKIS[6], Laurent TICHIT[4,5], Elisabeth REMY[4] and Denis THIEFFRY [1,6,7]

[1]TAGC (INSERM U1090), Marseille, France

[2]Instituto Gulbenkian de Ciência, Oeiras, Portugal

chaouiya@igc.gulbenkian.pt

[3]University of Lausanne, Switzerland

[4]Institut de Mathématiques de Luminy, Marseille, France

[5]Université de la Méditerranée, Marseille, France

[6]Institut de Biologie de l'Ecole Normale Supérieure, Paris, France

thieffry@ens.fr

[7]CONTRAINTES, INRIA Paris-Rocquencourt, Le Chesney, France

Keywords: biological network, logical modelling, GINsim, regulatory graph, state transition graph.

Systems biologists are facing the difficult challenge of modelling and analysing regulatory networks encompassing numerous and diverse components and interactions. Furthermore, available data sets are often qualitative, which complicates the definition of quantitative computational models.

Logical modelling constitutes a flexible framework to build qualitative predictive models, which can be readily analysed or simulated as such, and potentially used as scaffolds to build more quantitative (continuous or stochastic) models. The dynamics of qualitative models can be represented as State Transition Graphs (STG). It is then particularly relevant to identify subgraphs corresponding to dynamical attractors (i.e. terminal strongly connected components in STGs), as well as their reachability properties. For complex networks, however, the explicit construction of STGs can be cumbersome or even intractable. This led us to develop several complementary computational strategies, which are implemented in our logical modelling software suite, GINsim [2, 8].

A first strategy consists in deducing properties directly from the model, defined as a regulatory graph (i.e. a set of regulatory components or nodes, connected by signed arcs representing regulatory interactions). Using Multi-valued Decision Diagrams to represent (multi-level) logical updating rules enabled the development of efficient algorithms for the identification of all the stable states of a model, or yet to relate specific (positive or negative) regulatory circuits with specific dynamical properties (e.g., multiple attractors or sustained oscillations) [7, 11]. The rationale consists in deducing the structure of the STG directly from the regulation graph. We are currently working on the identification of complex attractors, using recent mathematical results connecting the presence of positive or negative regulatory circuits in the regulatory graph with the occurrence of multiple attractors or dynamical cycles in the STG.

A second strategy leads to the reduction of state spaces, either by reducing the model (the regulatory graph) or directly by working on the STG. More specifically, GINsim encompasses a reduction method that essentially preserves the dynamical properties of the model, although some reachability properties could be lost [10].

More recently, we have developed an algorithm to compact STGs on the fly. The result of a simulation is compressed into a hierarchical graph, where the nodes represent connected sets of states or components, each symbolically represented by a decision diagram. Each component is labelled as either stable state (terminal nodes in the STGs), cyclic attractor (terminal strongly connected components), transient cycle(s) (non-terminal strongly connected components) or subsets of transient states, depending on the topology of the underlying transition sub-graph.

Finally, we are currently considering the application of formal methods, in particular model-checking techniques. To that end, GINsim has recently been equipped with an export facility, which enables the use of

NuSMV to query logical models [6], under different updating policies or considering fixed and varying input components.

Beside these recent developments, we are developing means for incremental, compositional verification, to analyse large logical models defined as compositions of simpler regulatory modules.

All methodological developments are motivated by and used to tackle the analysis of regulatory networks involved in the control of cell fate, including of cell proliferation, differentiation and programmed death [1, 3, 4, 5, 8, 9, 12, 13]. In this presentation, we will refer to ongoing applications to the modelling of MAPK signalling pathways or of the network controlling hematopoietic cell specification in mammals, or yet of the network controlling the specification of mesoderm in drosophila embryo. These applications are described in more details in the contributions of Grieco et al., Collombet et al., Niarakis et al., and Mbodj et al. in these proceedings.

Acknowledgements

This work has been supported grants from ANR (grant ANR-08-SYSC-003), FCT (grant PTDC/EIA-CCO/099229/2008), BELSPO (IAP BioMaGnet), and EU FP7 (Project APO-SYS).

References

[1] L. Calzone, L. Tournier, S. Fourquet, D. Thieffry, B. Zhivotovsky, E. Barillot, A. Zinovyev (2010). Mathematical Modelling of Cell-Fate Decision in Response to Death Receptor Engagement. *PLoS Computational Biology* 6: e1000702.

[2] C. Chaouiya, A. Naldi, D. Thieffry (2012). Logical modelling of gene regulatory networks with GINsim. *Methods in Molecular Biology* 804: 463-79.

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

[4] A. Fauré, A. Naldi, F. Lopez, C. Chaouiya, A. Ciliberto, D. Thieffry (2009). Modular logical modelling of the budding Yeast cell cycle. *Molecular Biosystems* 5: 1787–96.

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

[6] P.T. Monteiro, C. Chaouiya (2012). Efficient verification for logical models of regulatory networks. *Advances in Intelligent and Soft Computing* 154: 259-267.

[7] A. Naldi, Thieffry D, Chaouiya C (2007). Decision diagrams for the representation and analysis of logical models of genetic networks. *Lecture Notes in Bioinformatics* 4695: 233-47.

[8] A. Naldi, D. Bérenguier, A. Fauré, F. Lopez, D. Thieffry, C. Chaouiya (2009). Logical modelling of regulatory networks with GINsim 2.3. *BioSystems* 97: 134-9.

[9] A. Naldi, J. Carneiro, C. Chaouiya, D. Thieffry (2010). Diversity and plasticity of Th cell types predicted from regulatory network modelling. *PLoS Computational Biology* 6: e1000912.

[10] A. Naldi, E. Remy, D. Thieffry, C. Chaouiya (2011). Dynamically consistent reduction of logical regulatory graphs. *Theoretical Computer Science* 412: 2207-18.

[11] E. Remy, P. Ruet, D. Thieffry (2006). Positive or negative regulatory circuit inference from multilevel dynamics. *Lecture Notes in Control and Information Sciences* 341: 263-70.

[12] O. Sahin, H. Fröhlich, C. Löbke, U. Korf, S. Burmester, M. Majety, J. Mattern, I. Schupp, C. Chaouiya, D. Thieffry, A. Poustka, S. Wiemann, T. Beissbarth, D. Arlt (2009). Modeling ERBB receptor-regulated G1/S transition to find targets for de novo trastuzumab resistance. *BMC Systems Biology* 3: 1.

[13] L. Sánchez, C. Chaouiya, D. Thieffry (2008). Segmenting the fly embryo: logical analysis of the role of the Segment Polarity cross-regulatory module. *International Journal of Developmental Biology* 52: 1059-75.

Using large scale knowledge database about reactions and regulations to find key regulators of sets of genes

Pierre BLAVY[1,2,3], Florence GONDRET[1,2], Sandrine LAGARRIGUE[1,2],
Jaap VAN MILGEN[1,2]and Anne SIEGEL[3]

[1] INRA, UMR 1348 PEGASE, Domaine de la Prise, 35590 Saint-Gilles
[2] AgroCampus-Ouest, UMR 1348 PEGASE, 74 rue de Saint Brieuc, 35000 Rennes
[3] CNRS-Université de Rennes 1-INRIA, UMR 6074 IRISA, Campus de Beaulieu, 35042 RENNES Cedex

Abstract *Many experimental observations and known cellular mechanisms are available but large-scale analyses of such information remain difficult due to the heterogeneity of the biological mechanisms. In this context, we introduce a new automatic method to integrate this information in order to find potential key regulators of a set of molecules.*
First, we introduce the concept of "regulated reaction" to gather in a unified formalism information about reactions (consumption and prediction of molecules) and causalities (effect of the variation of a molecule on the variation of another molecule). This formalism is then modeled as a causality graph describing in a predictive way consequences of the variations of fluxes and molecules quantities. Finally, we use graph-based algorithms and biologically relevant scores to extract key regulators of a set of molecules. We validate this method by recovering among the causal graph derived from the Transpath database the main regulators of 190 groups of genes which are known to share a transcription factor according to the TRED database.

Keywords Large scale, supervised analysis, integration, key regulator

Exploitation à large échelle de bases de données de connaissances sur les réactions et les régulations afin de trouver des régulateurs clefs de groupes de gènes

Résumé *Bien que de nombreuses données expérimentales et connaissances relatives aux mécanismes cellulaires soient disponibles, leur analyse à large échelle reste difficile, notamment en raison de la diversité des mécanismes biologiques. Dans ce contexte, nous proposons une nouvelle méthode pour intégrer ces mécanismes afin d'identifier des régulateurs clefs potentiels d'ensembles de molécules.*
Dans un premier temps, nous proposons un formalisme unifié représentant des "réactions régulées" afin de réunir les informations relatives aux réactions (consommant et produisant des molécules) et aux causalités (conséquences de la variation d'une quantité de molécule sur une autre molécule). Dans un second temps, nous interprétons les régulations régulées sous la forme d'un graphe de causalités qui représente, de manière prédictive, l'effet des variations de quantités ou de flux sur les différents acteurs métaboliques et génétiques du système. Au final, nous développons une approche combinant des parcours de graphe et des scores biologiquement adéquats pour identifier les régulateurs clefs d'un ensemble de sommets connus pour être co-régulés. Nous validons notre approche en construisant un graphe de causalité à partir de Transpath, puis en recherchant les régulateurs clefs de 190 ensembles de molécules extraits de la base de données TRED dont le principal facteur de transcription est connu.

Mots-clés Large échelle, analyse supervisée, intégration, régulateur clef

1 Introduction

À grande échelle, les systèmes biologiques sont notamment décrits par deux types d'informations : d'une part, des données expérimentales sur l'expression des gènes (puce ADN, RNA-seq) [7] qui sont souvent qualitatives (ou quantitatives analysées qualitativement) et obtenues en comparant un faible nombre de conditions. Quelques études cinétiques ou concernant de nombreuses conditions expérimentales existent mais elles sont relativement rares. D'autre part, de nombreuses connaissances relatives aux mécanismes cellulaires sont présentes dans des bases de données de connaissances [17], dans la littérature ou peuvent aussi être obtenues sous forme synthétique, sur des problèmes particuliers, en questionnant les experts.

Pourtant, il est actuellement difficile d'interpréter à grande échelle les données d'expression dans le cadre des connaissances sur les mécanismes afin, par exemple, d'identifier les régulateurs d'un ensemble de molécules ou de gènes. Plusieurs raisons sont généralement invoquées pour expliquer ces difficultés : les systèmes observés par les données d'expression comportent plusieurs milliers de molécules, ce qui nécessite des approches qualitatives et à grande échelle sur un formalisme unifié des mécanismes moléculaires. Or, l'information contenue dans les bases de données n'est pas disponible de manière uniforme : les bases combinent souvent des interactions agissant à des échelles très différentes (de la transcription au métabolisme), et contiennent à la fois des mécanismes de type réactionnels (transformation d'un produit en un autre) et des mécanismes de type causaux (effet d'un produit sur un autre, sans consommation du produit initial).

Classiquement, l'analyse haut débit des données d'expression comparant quelques conditions [14,4] commence par l'identification des ensembles de gènes différentiellement exprimés entre conditions. Les groupes de gènes corrélés sont construits puis analysés à l'aide d'annotations issues d'ontologies. On obtient au final des groupes de gènes caractérisés par un ensemble d'attributs souvent relatifs à leur fonction. Ces étapes sont généralement réalisées automatiquement, y compris à l'échelle d'une puce pangénomique. Dans un second temps, une analyse plus fine est réalisée par le biologiste, soit en sélectionnant un groupe de gènes puis en faisant la bibliographie de ses éléments, soit en se concentrant sur un ensemble de gènes relatifs à son champ d'expertise. Cette partie permet de prendre en compte les connaissances précises relatives à la régulation ou à l'effet des éléments sélectionnés. Cependant, cette dernière étape est réalisée manuellement, et ne peut donc pas être appliquée à l'ensemble des éléments d'intérêts. Le travail présenté ici vise à automatiser en partie cette dernière étape.

Il existe plusieurs approches permettant d'intégrer à large échelle les connaissances relatives aux mécanismes moléculaires afin d'analyser des données qualitatives. Un premier groupe de méthodes, composé des réseaux booléens [15], de leur généralisation asynchrone [5] ou stochastique [21], des réseaux logiques généralisés [3] et systèmes d'équations différentielles linéaires par morceaux [8,20] permet d'étudier des trajectoires (*i.e.,* une succession d'états au cours du temps). Cet ensemble de méthodes est basé sur la modélisation qualitative de la dynamique des systèmes. Il s'appuie donc sur des données cinétiques portant sur les variations de quantités de produits par rapport à un ou des seuils, données qui sont encore difficiles à produire. De plus, les espaces des trajectoires possibles croissent de manière exponentielle en fonction de la taille du système, rendant impossible des études automatiques à grande échelle.

Une alternative à ces méthodes de modélisation des propriétés temporelles des systèmes biologiques consiste à se concentrer sur les causalités, c'est-à-dire les effets des variations de quantité des molécules sur le système en ne modélisant pas les phénomènes de consommation. Dans ce cas, les études se concentrent sur la propagation des effets des régulations le long d'un réseau et sur leur compatibilité avec des observations différentielles entre deux états d'un système (par exemple l'effet d'un stress environnemental ou d'une perturbation génétique). Ces approches causales sont principalement appliquées à la reconstruction de réseaux biologiques avec des méthodes statistiques, par exemple bayésiennes

[11], ou à la prédiction du comportement d'un réseau par des approches formelles [10,9]. Par construction, ces approches modélisent les causalités (*i.e.,* les conséquences de la variation d'une quantité de molécules sur la quantité d'autres molécules) ce qui les rend parfaitement adaptées à la modélisation des régulations génétiques, souvent exprimées en terme d'activation ou d'inhibition. Cependant, elles sont inadaptées pour modéliser directement les effets des réactions (*i.e.,* les consommations et productions de molécules). Pour dépasser cette limite, nous avons développé une interprétation des réactions dans le cadre des causalités.

Brièvement, nous introduisons d'abord le concept de "réaction régulée", pour représenter de manière unifiée les connaissances relatives aux réactions et aux causalités. Ce formalisme permet d'intégrer les connaissances classiquement disponibles dans les différentes bases de connaissances telles que Transpath [2], KEGG [18] ou Pathway Commons [1] [1] (c.f. partie 2.1). Dans un second temps, nous associons à toute réaction régulée un graphe de causalité. Ceci est rendu possible, sous une hypothèse de quasi-stationnarité, en introduisant différents types de sommets représentant les flux ou les quantités de molécules (c.f. partie 2.2). Cette démarche est résumée Fig. 1. Enfin, nous développons une approche s'appuyant sur des parcours du graphe de causalités pour identifier les régulateurs clefs d'un ensemble de gènes coexprimés. Cette démarche est validée en construisant l'ensemble des "réactions régulées" et le graphe de causalité associé à partir de Transpath, puis en utilisant ce dernier graphe pour rechercher les régulateurs clefs de 190 groupes de gènes connus pour être régulés par un facteur de transcription. Ces groupes sont extraits de TRED [2] [13] (c.f. partie 2.3).

2 Matériel et méthodes

2.1 Formalisme unifié des réactions et des régulations

Les formalismes existants pour représenter les mécanismes de régulation sont de deux types : d'une part, les réactions permettent de décrire les consommations et productions de molécules ; d'autre part, les causalités permettent de décrire l'effet d'une modification de la quantité de la source d'une interaction sur ses cibles sans consommer la source de l'interaction. Ces deux points de vue (réaction *Vs.* causalité) coexistent dans les banques de données portant sur les interactions biologiques.

Pour unifier ces deux points de vue, nous introduisons le concept de "réactions régulées". Ces objets sont constitués de substrats, de produits, d'activateurs, d'inhibiteurs, de modulateurs (*i.e.,* une molécule étant soit activateur, soit inhibiteur) et d'un booléen indiquant si la réaction est réversible ou irréversible. Ce formalisme très simple permet de représenter l'information utile à notre contexte à partir de la plupart des bases de connaissances existantes, et fait la distinction entre les flux de matière (substrats et produits) et les régulateurs (activateurs, inhibiteurs, modulateurs) qui influencent les vitesses des réactions sans être consommés. Sur la Fig. 1, nous décrivons les règles de transformations d'interactions causales en réactions régulées.

En pratique, l'ensemble des réactions décrites dans Transpath [16] a été extrait et codé dans ce formalisme. Les molécules ATP, ADP, *protein remanants*, NDP, NTP, sp1, phosphate, Coenzyme-A, eau et H+ ont été considérées comme jamais limitantes *a priori* et ont été retirées. De plus, afin de prendre en compte l'aspect multi-espèces des connaissances de Transpath, les réactions régulées ayant les mêmes ensembles de substrats et produits ont été fusionnées en faisant l'union de leurs régulateurs. Si au moins une réaction régulée est réversible, la fusion est considérée comme réversible. Cette règle permet de décrire l'ensemble des réactions présentes quelque soit l'espèce dans laquelle elles ont été observées. Nous avons obtenu un graphe de grande taille (c.f. Table 1) fortement connexe. Ce graphe a une structure *scale-free* [12] : un petit nombre de molécules très connectées (*i.e.,* les *hubs*) est relié à un grand nombre

[1] Pour une revue des principales bases, voir [19] et http://pathguide.org
[2] Tred est disponible sur http://rulai.cshl.edu/TRED/

de molécules très peu connectées. Privé des 1000 molécules intervenant dans le plus de réactions (*i.e.,* les principaux *hubs*), le graphe devient très faiblement connecté.

2.2 Interprétation des réactions régulées en graphe de causalité

L'ensemble des réactions régulées permet de représenter la connaissance mais il n'exprime pas directement les effets des variations de quantité de molécules. Dans la Table 2, nous introduisons des règles d'interprétation de réactions régulées en graphes de causalité à l'aide de noeuds distincts pour les flux (noeuds : "*disponibilité en substrat*" et "*vitesse de réaction*") et les quantités de molécules (noeuds de "*quantité*"). Ces règles sont une interprétation qualitative des coefficients d'élasticité [6] sous l'hypothèse quasi-stationnaire. Par la suite elles sont utilisées pour propager les effets des régulations sans prendre en compte les effets globaux de la dynamique du système. On obtient ainsi une surapproximation des comportements possibles du système.

Lors de la construction du graphe de causalités associé à Transpath, nous avons fait l'hypothèse de considérer toutes les réactions comme irréversibles. En effet, la majeure partie des réactions réversibles de Transpath correspond à des inhibitions par formation de complexes pour lesquelles le sens complexe→substrats peut être négligé, car aucune autre réaction ne produit le complexe.

Afin de pouvoir faire facilement le lien entre les observations sur les molécules et les noeuds du graphe de causalité, on associe à chaque noeud de type "*quantité*" et à chaque noeud de type "*disponibilité en substrat*", la molécule à laquelle ils se rapportent.

Cette étape induit une augmentation dans la taille du graphe (nombre de noeuds $\times$ 1,4, nombre d'arêtes $\times$ 4,9) en raison de l'introduction des différents types de noeuds et des nombreux liens lorsque plusieurs réactions concernent la même molécule (c.f. Table 1).

2.3 Recherche de régulateurs clefs

Dans le cadre de cette étude, nous avons développé une méthode pour proposer un ensemble de molécules qui sont ordonnées par leur capacité à réguler (directement ou indirectement) un autre ensemble de molécules cibles. Ces cibles sont définies *a priori* en prenant par exemple un *cluster* de gènes coexprimés issu d'une analyse à haut débit par puces à ADN, ou un groupe de métabolites d'intérêt issu d'une synthèse bibliographique.

La méthode prend comme entrée un ensemble de réactions régulées $\mathcal{R}$ et un ensemble de *molécules* cibles $\mathcal{S}$. Elle se base sur le calcul sous contraintes de la fermeture transitive du graphe de causalité, afin d'évaluer pour chaque noeud, les noeuds dont il peut expliquer la variation. Elle est décrite par les étapes ci-dessous, dans lesquelles le terme *molécule* désigne toujours une molécule dans un ensemble de réactions régulées et le terme *noeud* un sommet d'un graphe de causalité.
(**1**) Le but de cette étape est de sélectionner dans l'ensemble des réactions régulées $\mathcal{R}$ un sous-ensemble en lien avec les cibles. Pour cela, nous sélectionnons les réactions régulées de $\mathcal{R}$ situées à moins de 3 réactions de l'ensemble des *molécules* cibles $\mathcal{S}$ sans passer par les *hub* de $\mathcal{R}$. Le graphe de réactions régulées obtenu est noté $\mathcal{R}(\mathcal{S})$. Comme indiqué partie 2.1, une *molécule* est considérée comme un *hub* si elle fait partie des 1000 premières *molécules* impliquées dans le plus de réactions de $\mathcal{R}$.
(**2**) L'ensemble des réactions régulées sélectionnées $\mathcal{R}(\mathcal{S})$ est converti en graphe de causalité noté $\mathcal{C}(\mathcal{S})$ à l'aide des règles décrites Fig. 2. Au cours de cette conversion, une relation d'*association* est définie entre chaque *molécule* M_1 de $\mathcal{R}(\mathcal{S})$ et les *noeuds* : *disponibilité en* M_1 et *quantité de* M_1 dans $\mathcal{C}(\mathcal{S})$.
(**3**) A chaque *noeud* N_1 du graphe $\mathcal{C}(\mathcal{S})$, nous associons deux ensembles de *molécules* $\mathcal{U}^+_{N_1}$ et $\mathcal{U}^-_{N_1}$. L'ensemble $\mathcal{U}^+_{N_1}$ contient la molécule M_2 si la propagation de l'effet d'une augmentation de N_1 au travers d'un chemin entre N_1 et un noeud N_2 *associé* à M_2 est cohérente avec les variations observées le long

de ce chemin. Ces ensembles sont calculés par la procédure[3] ci-dessous illustrée Fig. 2.

- **Initialisation** Pour toute *molécule* M_1 et tout *noeud* N_1 *associé* à M_1, si M_1 augmente ou n'est pas observée alors $\mathcal{U}^+_{N_1}$=$\{M_1\}$, sinon $\mathcal{U}^+_{N_1}$=$\emptyset$.
- **Chemins cohérents** Pour tout *noeud* N_1 et toute molécule *molécule* M_3, M_3 est ajoutée à $\mathcal{U}^+_{N_1}$ si a) il y a cohérence avec les observations relatives à N_1 et b) il y a cohérence avec au moins un *noeud* N_2 sucesseur de N_1.
  - a) est vraie si : la molécule M_1 *associée* à N_1 augmente ou si M_1 n'est pas observée.
  - b) est vraie pour N_2 un sucesseur de N_1 :
    - si : $N_1 \overset{+}{\longmapsto} N_2$ et $\mathcal{U}^+_{N_2}$ contient M_3
    - ou si : $N_1 \overset{-}{\longmapsto} N_2$ et $\mathcal{U}^-_{N_2}$ contient M_3
    - ou si : $N_1 \overset{?}{\longmapsto} N_2$ et ($\mathcal{U}^+_{N_2}$ contient M_3 ou $\mathcal{U}^-_{N_2}$ contient M_3)
- **Propagation** Tant qu'au moins un ensemble $\mathcal{U}^+_{N_1}$ ou $\mathcal{U}^-_{N_1}$ a été mis à jour pour un sommet N_1, le calcul des chemins cohérents est répété. Ceci permet de prendre en compte les régulations indirectes.

(**4**) Afin d'étendre le concept précédent aux molécules du graphe de réactions, nous calculons pour chaque *molécule* M_1 de $\mathcal{R}(\mathcal{S})$ *associée* à l'ensemble de noeuds N deux ensembles : $\mathcal{V}^+_{M_1} = \bigcup_{\forall a \in N} U^+_a$ et $\mathcal{V}^-_{M_1} = \bigcup_{\forall a \in N} U^-_a$. Les ensembles $\mathcal{V}^+_{M_1}$ et $\mathcal{V}^-_{M_1}$ contiennent les molécules cibles (directes et indirectes) de M_1.
(**5**) Pour chaque molécule Ma de $\mathcal{R}(\mathcal{S})$, et chaque cas c (c=+ si Ma augmente, - sinon) nous calculons trois scores :

- le score de **couverture** égal au nombre de molécules appartenant à la fois à $\mathcal{V}^c_{Ma}$ et aux cibles $\mathcal{S}$.
- le score de **spécificité** égal à un moins la probabilité d'obtenir un nombe de molécules cibles régulées $\geq$ à ce qui est attendu au hasard. Cette probalité est estimée par un test hypergéométrique.
- le score de **couverture spécifique** égal au produit des deux scores précédents.

En ordonnant les différentes molécules en fonction de ces scores, nous sommes capables de fournir au biologiste une liste ordonnée de candidats ; les meilleurs étant ceux de score maximal.

2.4 Validation de la méthode

TRED est une base de connaissances décrivant un ensemble de facteurs de transcription associés à leur cibles, obtenue chez l'homme, la souris et le rat. Un script a été développé pour extraire à partir de l'interface web des couples "facteurs de transcription – cibles de ce facteur". Seuls les couples avec un nombre de cibles supérieur à 5 ont été retenus. Moins de 5 cibles est considéré comme une information bien trop faible pour justifier une étude à haut débit. 190 couples ont été extraits.

Ces couples sont ensuite utilisés pour valider notre méthode. Nous utilisons la méthode de recherche de régulateurs clefs décrite précédemment afin de proposer un ensemble de candidats. Nous considérons ensuite un test comme un succès si le facteur de transcription connu se retrouve parmi les 50 premiers gènes candidats. Le nombre de 50 a été choisi car il correspond à ce qu'un biologiste peut facilement analyser manuellement dans un temps raisonnable.

Dans un second temps, nous avons analysé manuellement les candidats proposés par notre méthode pour 19 cibles de PPARα choisies par expertise manuelle (ACSL1, ACSL3, ACSL4, ACSL5, ACSL6, CD36, ACADVL, ACADL, ACAA2, CPT1A, CPT1B, 3HCDH, 3KACT, HADHA, ETFDH, HMGCS2, FADS1, FADS2, SCD).

[3] La procédure est décrite pour $\mathcal{U}^+_{N_1}$. La procédure pour $\mathcal{U}^-_{N_1}$ est similaire, il suffit d'échanger dans la description $\mathcal{U}^+_{N_1}$ avec $\mathcal{U}^-_{N_1}$ et de remplacer augmentation par diminution.

3 Résultats et discussion

Lorsque notre méthode est appliquée entièrement : le score de couverture spécifique a un taux de succès de 58%. Ce résultat démontre l'efficacité de notre méthode pour proposer automatiquement au biologiste des régulateurs clefs potentiels à partir d'un ensemble de cibles. Lorsque le résultat des tests est calculé aléatoirement, mais que les sous-ensembles du graphe de réaction sont calculés à partir des bonnes cibles, 9% des tests sont des succès, ce taux est dû à la relative pertinence de notre méthode de sélection de sous-graphes. Lorsque les noms des cibles décrites dans TRED sont échangés au hasard, ce score tombe à 2%, ce qui nous assure que notre méthode produit bien des candidats spécifiques des cibles, et non une liste de régulateurs ubiquitaires.

Pris séparément, les scores de couverture et de spécificité donnent des résultats moins bons que leur produit. Ceci signifie qu'un bon candidat doit à la fois bien couvrir les molécules cibles, sans en réguler de nombreuses autres à coté. Dans l'avenir, on peut imaginer optimiser la méthode en testant différentes pondérations des scores de spécificité et de couverture.

Parmi les 80 échecs, 7 s'expliquent car le facteur de transcription connu n'est pas présent dans le graphe total des réactions régulées, un problème dû soit à un mauvais lien entre les identifiants de Transpath et ceux de TRED, soit à un manque de connaissance dans Transpath.

Une analyse manuelle de quelques échecs nous a permis de mettre en évidence des ensembles de cibles très vastes (plus de 100 cibles) contenant un grand nombre de cibles peu spécifiques du facteur de transcription auxquelles elles sont associées dans TRED. Il n'est donc pas du tout surprenant de retrouver parmi les régulateurs candidats de ces cibles de nombreuses molécules ayant une pertinence biologique mais différentes de la solution du test. De plus, de nombreux facteurs de transcription ont un très vaste nombre de cibles (dont la plupart sont inconnues ou absentes des bases), ce qui rend la réalisation de tests difficile et explique probablement certains échecs. Au final, la méthode développée reste pertinente pour rechercher des régulateurs clefs potentiels.

Par exemple, en utilisant le score de couverture, l'analyse des gènes candidats pour expliquer les cibles de PPARα (c.f. partie 2.4) fait ressortir un ensemble de 28 gènes qui expliquent 9 ou 10 cibles. Parmi eux, on retrouve comme attendu des gènes liés à la voie de signalisation des PPAR (PPARα,PPARγ, PPARδ, RXRα, NCOA1, PPARGC1A, FADS1 et ALOX15). La présence de RXRα et de ALOX15 s'explique par l'implication des acides gras dans la voie de signalisation de PPARα (divers acides gras ont un score de couverture de 9), ce qui illustre l'intérêt d'inclure dans le système les métabolites. De plus, on retrouve de nombreux régulateurs ubiquitaires du métabolisme de l'énergie (insuline et NRIP1) de la réponse hormonale (CREB1, VDR, NR2F1, NRIP1) et du cycle cellulaire (SP1,SP3,SRF). Parmi les gènes restants, 3 sont impliqués dans le métabolisme des lipides mais le lien avec les cibles n'est pas trivial et 6 ont un lien qui reste à élucider.

Conclusion et perspectives

Nous avons développé un formalisme permettant de représenter de manière simple et unifiée les informations courantes décrivant les mécanismes cellulaires impliqués dans les voies de régulation et dans le métabolisme. Nous avons ensuite proposé une méthode pour interpréter ce formalisme sous la forme d'un graphe de causalité en distinguant, par des noeuds différents, les concepts de flux, de vitesse de réaction et de quantités de molécules. Au final, nous avons démontré sur un jeu de test que ce formalisme est pertinent pour proposer automatiquement au biologiste un ensemble de régulateurs clefs d'un groupe de molécules d'intérêt.

Dans l'avenir, nous souhaitons tout d'abord améliorer la méthode de recherche de régulateurs candidats en développant de nouveaux scores, ou en prenant en compte de manière fine l'effet combiné de plusieurs molécules, puis utiliser les scores développés afin d'analyser des jeux de données réels tels que les clusters de gènes identifiés dans les études transcriptionnelles haut débit.

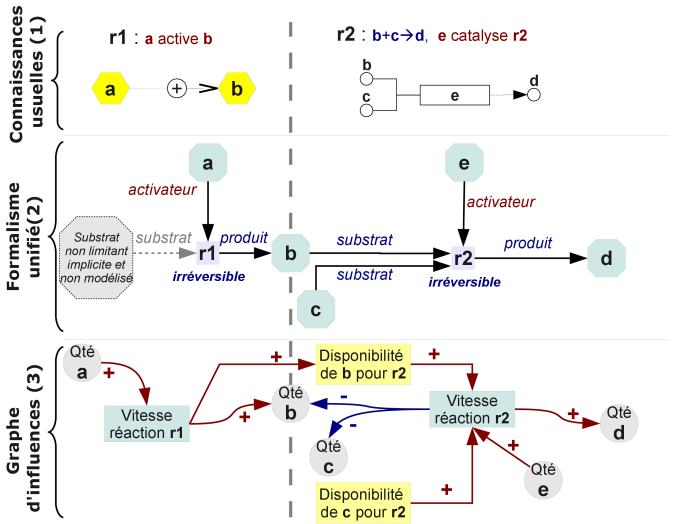


Figure 1. Formalisme unifié pour les réactions et les régulations : interprétation en graphe de causalité

(**1**) Les connaissances usuelles relatives au métabolisme et à sa régulation modélisent deux concepts différents : les *reactions* (b+c→d) qui consomment ou produisent des molécules et les *effets* (**a** active **b** et **e** catalyzes **r2**) qui décrivent les conséquences de la variation d'une quantité de molécule sans consommer cette dernière.

(**2**) Ces deux concepts sont unifiés dans un formalisme commun : les *réactions régulées*, en supposant que les effets (comme **a** active **b**) sont des réactions dont la vitesse est régulée par le régulateur (**a**) qui consomment un substrat non limitant non modélisé et qui produisent l'élément régulé (**b**). Ce formalisme fait clairement la distinction entre les flux ($\overset{substrats}{\longmapsto}$ et $\overset{produit}{\longmapsto}$) et les régulations de la vitesse de réaction par des molécules non consommées ($\overset{activateur}{\longmapsto}$, $\overset{inhibiteur}{\longmapsto}$ et $\overset{modulateur}{\longmapsto}$).

(**3**) Sous l'hypothèse que la vitesse de réaction est une fonction croissante de la disponibilité de chacun des substrats, une fonction croissante de la quantité d'activateur, une fonction décroissante de la quantité d'inhibiteur et une fonction monotone de la quantité de modulateur, les *réactions régulées* peuvent être interprétées sous forme de graphe de causalités. Ce dernier modélise les conséquences ($\overset{+}{\longmapsto}$,$\overset{-}{\longmapsto}$,$\overset{?}{\longmapsto}$) des variations des flux (rectangles : vitesses de réactions et disponibilité) et des quantités de molécules (cercles).

Table 1. Modélisation de Transpath en réaction régulées et en causalités.

(a) Transpath est une base de données dévrivant des réactions et leurs régulations (b) Les réactions et les effets sont unifiés sous forme de "réactions régulées" (c.f. 2.1) (c) Les "réactions régulées" sont ensuite converties en graphe de causalité (c.f. 2.2) afin d'exprimer les conséquences des variations de quantité de molécules, de flux et de vitesse de réactions. Au final, l'introduction de nouveau noeuds crée un graphe contenant 1.5 fois plus de noeuds et 4.9 fois plus d'arêtes que le graphe d'origine. Ceci s'explique notamment par les liens créés lorsque une même molécule intervient dans de nombreuses réactions.

(a) Transpath

Molécules	**158 545**
métabolites	122 951
genes	35 594
Relations	**224 080**
effect +	2 493
effect -	854
effect ?	10 251
reaction réversible	41 278
reaction irréversible	110 838
gene→protein	59 956

(b) Reactions régulées

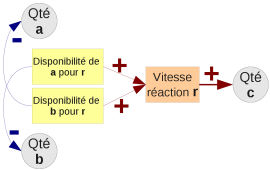
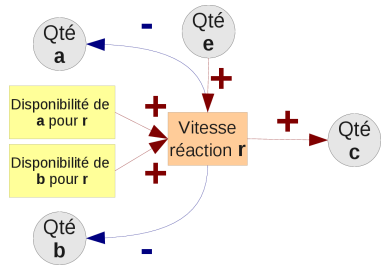
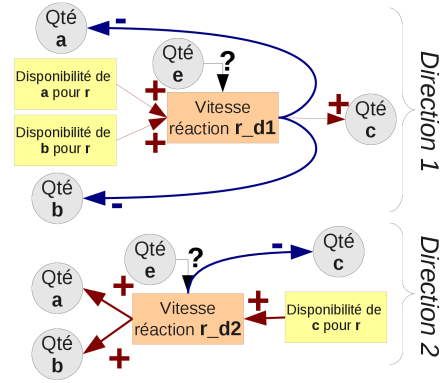
Noeuds	**291 306**
mol ou gènes	132 752
reactions rév	40 541
reactions irr	118 013
Arêtes	**383 732**
substrats	27 418
produits	115 630
activateurs	18 813
inhibiteurs	498
modulateurs	9 307

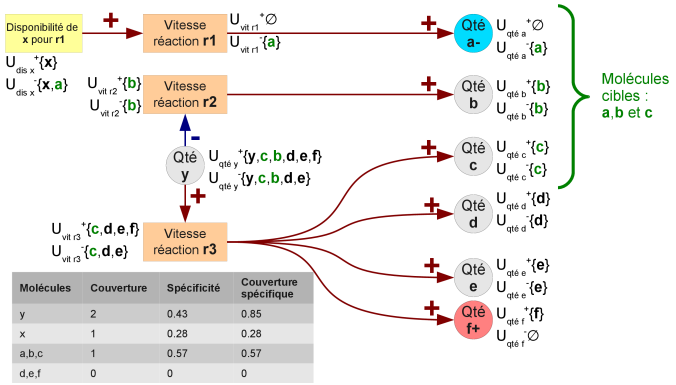
(c) Graphe de causalité

Noeuds	**422 652**
quantité	131 065
vitesse reac.	109 408
disponibilité	181 179
Arêtes	**1 891 492**
$\overset{+}{\longmapsto}$	1 656 275
$\overset{-}{\longmapsto}$	217 279
$\overset{?}{\longmapsto}$	17 938

Table 2. Règles de création d'un graphe de causalités à partir de réactions régulées

Chaque *reaction régulée* est modélisée sous la forme d'un graphe de causalités à l'aide des règles du tableau. Ensuite, chaque vitesse de réaction produisant une molécule **a** est liée par $\overset{+}{\longmapsto}$ à la disponibilité en **a** pour les autres réactions. Les lettres SPAIM représentent respectivent les ensembles de substrats, produits, activateurs, inhibiteurs et modulateurs, décrivant les *réactions régulées* ; $\emptyset$ désigne l'ensemble vide. Une réaction est considérée comme sans régulateur si elle n'a ni activateur ni inhibiteur ni modulateur. Ces règles décrivent les effets locaux des variations de flux et de quantités sous l'hypothèse quasistationnaire sans prendre en compte les effets globaux de la dynamique du système. On obtient ainsi une sur-approximation des comportements du système.

Règles		
Irréversible sans régulateur	Irréversible avec régulateur	Réversible
$\forall \mathbf{s} \in S, \forall \mathbf{p} \in P$, disponibilité de s pour $r \overset{+}{\longmapsto}$ vitesse de la réaction r. vitesse de la réaction $r \overset{+}{\longmapsto}$ quantité de p $\forall \mathbf{s1}, \mathbf{s2} \in S^2, \mathbf{s1} \neq \mathbf{s2}$, disponibilité de $s1$ pour $r \overset{-}{\longmapsto}$ quantité de $s2$.	$\forall \mathbf{s} \in S, \forall \mathbf{p} \in P$ disponibilité de s pour $r \overset{+}{\longmapsto}$ vitesse de la réaction r, vitesse de la réaction $r \overset{-}{\longmapsto}$ quantité de s, vitesse de la réaction $r \overset{+}{\longmapsto}$ quantité de p. $\forall \mathbf{a} \in A, \forall \mathbf{i} \in I, \forall \mathbf{m} \in M$, quantité de $a \overset{+}{\longmapsto} r$ vitesse de la réaction r, quantité de $i \overset{-}{\longmapsto} r$ vitesse de la réaction r, quantité de $m \overset{?}{\longmapsto} r$ vitesse de la réaction rx.	$\forall \mathbf{s} \in S, \forall \mathbf{p} \in P$, disponibilité de s pour $r \overset{+}{\longmapsto}$ vitesse de la réaction r_{d1}, disponibilité de p pour $r \overset{+}{\longmapsto}$ vitesse de la réaction r_{d2}, vitesse de la réaction $r_{d1} \overset{-}{\longmapsto}$ quantité de s, vitesse de la réaction $r_{d2} \overset{-}{\longmapsto}$ quantité de p, vitesse de la réaction $r_{d1} \overset{+}{\longmapsto}$ quantité de p, vitesse de la réaction $r_{d2} \overset{+}{\longmapsto}$ quantité de s. $\forall \mathbf{o} \in A \bigcup I \bigcup M$, quantité de $o \overset{?}{\longmapsto}$ vitesse de la réaction r_{d1}, quantité de $o \overset{?}{\longmapsto}$ vitesse de la réaction r_{d2}.
Examples		
a+b⟶c S={**a**,**b**},P={**c**},A=I=M= ∅,irreversible	**a+b⟶c**, enzyme **e** S={**a**,**b**},P={**c**}, A={**e**},I=M=∅,irreversible	**a+b↔c**, enzyme **e** S={**a**,**b**},P={**c**}, A={**e**},I=M=∅,reversible
		



Molécules	Couverture	Spécificité	Couverture spécifique
y	2	0.43	0.85
x	1	0.28	0.28
a,b,c	1	0.57	0.57
d,e,f	0	0	0

Figure 2. Exemple de recherche de régulateurs clefs dans le graphe de causalités à partir d'un ensemble de molécules cibles

Dans cet exemple, nous souhaitons identifier les régulateurs clefs de trois molécules cibles : **a**, **b** et **c** à l'aide de la procédure décrite dans la partie 2.3. Nous observons deux molécules : **a** et **f** qui augmentent. Nous supposons que la partie (1) de la procédure nous a conduit à travailler sur l'ensemble de réactions régulées suivant : la réaction irréversible **r1** de substrat **x** et de produit **a**, la réaction irréversible **r2** de produit **b** et d'inhibiteur **y** et la réaction irréversible **r3** de produits **c**, **d**, **e** et **f** et d'activateur **y**.

Dans un premier temps l'ensemble des réactions régulées est converti en graphe de causalité tel que représenté sur la figure. Les noeuds de *quantité* (Qté) et de *disponibilité en substrats* sont mis en relations avec la molécule qu'ils décrivent. Ainsi, le noeud : *disponibilité de **x** pour **r1*** est mis en relations avec la molécule **x**, le noeud *quantité de **a*** avec la molécule **a** et ainsi de suite pour toutes les autres molécules. Seuls les noeuds de *vitesse de réaction* sont associés à aucune molécule. Le sens de variation des molécules observées est représenté sur les noeuds associés à ces molécules dans le graphe de causalités (*Qté de **a*** - et *Qté de **f*** +).

Pour chaque noeud N_1 associé à une molécule M_1, les ensembles $\mathcal{U}^+_{N_1}$ et $\mathcal{U}^-_{N_1}$ sont initialisés par la molécule associée M_1 en tenannt compte des observations. Ainsi, le noeud *Qté de **a*** associé à la molécule **a** qui décroît, va être initialisé par $\mathcal{U}^+_{Qté\ de\ a}=\emptyset$ et $\mathcal{U}^-_{Qté\ de\ a}=\{\mathbf{a}\}$. Le noeud *Qté de **b***, associé à une molécule non observé sera initialisé avec : $\mathcal{U}^+_{Qté\ de\ b}=\mathcal{U}^-_{Qté\ de\ b}=\{\mathbf{b}\}$.

Les éléments des ensembles prédécents sont propagés récursivement aux prédecesseurs de manière à identifier pour chaque sommet les molécules qu'il régule, de manière cohérente avec les observations et les signes des influences (i.e. $\overset{+}{\longmapsto}$, $\overset{-}{\longmapsto}$ et $\overset{?}{\longmapsto}$). Par exemple, lorsque la *vitesse de réaction **r1*** augmente, elle régule **a** de manière cohérente avec les influences (il y a un chemin positif entre ce noeud et le noeud *Qté de **a***), et avec les observations : **a** étant observé comme diminuant, le noeud *Qté de **a*** explique bien **a** quand il diminue ($\mathbf{a} \in \mathcal{U}^-_{vit.\ r1}$) mais pas quand il augmente ($\mathbf{a} \notin \mathcal{U}^+_{vit.\ r1}$). Le résultat final de cette étape est représenté sur la figure par les ensembles $\mathcal{U}^+_n$ et $\mathcal{U}^-_n$ associés à chaque noeud n.

Finalement, les ensembles relatifs aux noeuds $\mathcal{U}^+_n$ (et $\mathcal{U}^-_n$) qui sont en relations avec une même molécule sont fusionnés pour se rapporter aux molécules associées à ces noeuds. Ensuite les scores de *couverture* et de *specificité* sont calculés. Dans cet exemple, la molécule **y** a une bonne couverture, car elle régule de nombreuses cibles (**b** et **c**) et une mauvaise spécificité car elle régule aussi beaucoup de molécules non cibles (**y**, **d**, **e** et **f**). D'un autre coté, **a**, **b** et **c** ont un mauvais score de couverture (ils ne régulent qu'une molécule : eux-mêmes), mais une meilleure spécificité car ils ne régulent aucune molécule non cible. En faisant le produit du score de couverture et de spécificité, on obtient un compromis qui fait ressortir **y** comme la meilleure explication des cibles.

Références

[1] E.G. Cerami, B.E. Gross, E. Demir, I. Rodchenkov, Ö. Babur, N. Anwar, N. Schultz, G.D. Bader, and C. Sander. Pathway commons, a web resource for biological pathway data. *Nucleic acids research*, 39(suppl 1) :D685–D690, 2011.

[2] C. Choi, M. Krull, A. Kel, O. Kel-Margoulis, S. Pistor, A. Potapov, N. Voss, and E. Wingender. TRANSPATH®a high quality database focused on signal transduction. *Comparative and functional genomics*, 5(2) :163–168, 2004.

[3] H. De Jong. Modeling and simulation of genetic regulatory systems : a literature review. *Journal of computational biology*, 9(1) :67–103, 2002.

[4] C. Desert, M.J. Duclos, P. Blavy, F. Lecerf, F. Moreews, C. Klopp, M. Aubry, F. Herault, P. Le Roy, C. Berri, et al. Transcriptome profiling of the feeding-to-fasting transition in chicken liver. *BMC genomics*, 9(1) :611, 2008.

[5] A. Faure, A. Naldi, C. Chaouiya, and D. Thieffry. Dynamical analysis of a generic Boolean model for the control of the mammalian cell cycle. *Bioinformatics*, 22(14), 2006.

[6] David Fell. *Understanding the Control of Metabolism*. Portland Press, 1997.

[7] H. Ge, A.J.M. Walhout, and M. Vidal. Integrating [] omic'information : a bridge between genomics and systems biology. *TRENDS in Genetics*, 19(10) :551–560, 2003.

[8] J.L. Gouzé and T. Sari. A class of piecewise linear differential equations arising in biological models. *Dynamical Systems : An International Journal*, 17(4) :299–316, 2002.

[9] R.M. Gutiérrez-Ríos, D.A. Rosenblueth, J.A. Loza, A.M. Huerta, J.D. Glasner, F.R. Blattner, and J. Collado-Vides. Regulatory network of Escherichia coli : consistency between literature knowledge and microarray profiles. *Genome research*, 13(11) :2435, 2003.

[10] C. Guziolowski, A. Bourdé, F. Moreews, and A. Siegel. BioQuali Cytoscape plugin : analysing the global consistency of regulatory networks. *BMC genomics*, 10(1) :244, 2009.

[11] S. Imoto, T. Higuchi, T. Goto, K. Tashiro, S. Kuhara, and S. Miyano. Combining microarrays and biological knowledge for estimating gene networks via bayesian networks. *Journal of Bioinformatics and Computational Biology*, 2(1) :77, 2004.

[12] H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, and A.L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804) :651–654, 2000.

[13] C. Jiang, Z. Xuan, F. Zhao, and MQ Zhang. Tred : a transcriptional regulatory element database, new entries and other development. *Nucleic acids research*, 35(suppl 1) :D137–D140, 2007.

[14] Y.H. Jin, P.E. Dunlap, S.J. McBride, H. Al-Refai, P.R. Bushel, and J.H. Freedman. Global transcriptome and deletome profiles of yeast exposed to transition metals. *PLoS Genetics*, 4(4), 2008.

[15] G. Karlebach and R. Shamir. Modelling and analysis of gene regulatory networks. *Nat Rev Mol Cell Biol*, 9(10) :770–780, 2008.

[16] M. Krull, S. Pistor, N. Voss, A. Kel, I. Reuter, D. Kronenberg, H. Michael, K. Schwarzer, A. Potapov, C. Choi, et al. Transpath® : an information resource for storing and visualizing signaling pathways and their pathological aberrations. *Nucleic acids research*, 34(suppl 1) :D546–D551, 2006.

[17] J. Labaer. Mining the literature and large datasets. *Nature Biotechnology*, 21(9) :976–977, 2003.

[18] Kanehisa M. and S. Goto. KEGG : Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res. 28, 27-30*, 2000.

[19] S. Mathivanan, B. Periaswamy, TKB Gandhi, K. Kandasamy, S. Suresh, R. Mohmood, YL Ramachandra, and A. Pandey. An evaluation of human protein-protein interaction data in the public domain. *BMC bioinformatics*, 7(Suppl 5) :S19, 2006.

[20] EI Prokudina, RY Valeev, and RN Tchuraev. A new method for the analysis of the dynamics of the molecular genetic control systems. II : application of the method of generalized threshold models in the investigation of concrete genetic systems. *Journal of theoretical biology*, 151(1) :89–110, 1991.

[21] I. Shmulevich, E.R. Dougherty, and Wei Zhang. From Boolean to Probabilistic Boolean Networks as Models of Genetic Regulatory Networks. *Proceedings of the IEEE*, 90(11), 2002.

Using Mutual Information and Answer Set Programming to refine PWM based transcription regulation network

Andrés ARAVENA[1,2], Carito GUZIOLOWSKI[3], Anne SIEGEL[2] and Alejandro MAASS[1,4]

[1] Mathomics, Center for Mathematical Modeling (UMI 2807 CNRS), University of Chile, and Center for Genome Regulation (Fondap 15090007), Santiago, Chile
{andres.aravena,amaass}@dim.uchile.cl

[2] IRISA Team Dyliss, CNRS - Université de Rennes 1 (UMR 6074) INRIA, 35042 RENNES Cedex - France
anne.siegel@irisa.fr

[3] Hamamatsu TIGA Center, Institute for Medical Biometry and Informatics, University of Heidelberg, Germany
carito.guziolowski@bioquant.uni-heidelberg.de

[4] Departamento de Ingeniería Matemática, University of Chile, Santiago, Chile

Abstract *Transcriptional regulatory network models can be reconstructed* ab initio *from DNA sequence data by locating the binding sites, defined by position specific score matrices, and identifying transcription factors by homology with known ones in other organisms. In general the resulting network contains spurious elements, because the pattern matching methods for binding site location have low specificity, while homology to known transcription factors does not always identify correctly new ones.*

In the case of A. ferrooxidans*, one of the bacterias involved in industrial bioleaching processes, the sequence based network reconstruction results in 66 transcription factors and 182 binding site motifs represented in 27 435 sites. In this work we use differential expression experimental data, in the form of Mutual Information, as logical constraints to be satisfied by any valid regulatory network subgraph. These rules are expressed as an Answer Set Program, a logical programming paradigm, and used to determine the minimal sets of motif and transcription factors which constitute a genetic regulatory network compatible with the experimental evidence. The resulting network comprises 27 transcription factors and 14 motifs in 2 428 instances, satisfying all constraints.*

Keywords Transcriptional regulatory network reconstruction, mutual information, logical programming.

1 Introduction

Regulation plays a major role in the transition from cell genotypes to phenotypes. The transcription of each gene is controlled by a regulatory region upstream of the transcription start site, where transcription factors can bind [1]. The chain of interactions between genes, transcription factors (TFs) and binding sites (BS) is known as the transcription regulation network, usually represented as a signed oriented graph [2]. Knowledge of this graph is the foundation of the qualitative or quantitative modeling of cell behavior, [3] thus constituting one of the main tools in system biology modeling of development as well as organism maintenance.

Experimental methods are used to identify the regions in the chromosome to which a protein binds, and for purifying and identifying a DNA-bound protein [4]. These experiments have not been widely applied because they are expensive and time consuming [5]. In addition, most of the results of these methods are focused on human and other eukaryotic organisms. However, some results for bacteria are organized in databases such as Prodoric [6], RegulonDB, RegTransBase and CoryneRegNet. Besides the direct wet-lab experimentation, there are two main strategies for regulatory network reconstruction: (1) directly recognizing regulatory elements in the genomic sequence, and (2) using expression data to determine gene regulatory relationships based on statistical correlation.

The first strategy for *in silico* network reconstruction uses the experimental sequences to extrapolate to other taxonomically related organisms [5]. Usual computational approaches focus either on: (1) mapping TFs and their target genes from one genome to another, or (2) using the binding site motif described in one organism to look for binding sites in a new one [7]. Neither of these approaches is perfect since orthology by itself cannot ensure that DNA-binding properties are conserved, or that both proteins respond to the same signals. On the other hand, significant variability has been observed among the binding site motifs for the same TF (e.g. see LexA in [8]) in the different bacterial phylogenetic groups. Finally, TFs can act either as activators or repressors, depending on the precise placement of their binding sites respect to the transcription start site [1]. In practice, this strategy requires two steps: (1) orthologous detection for TFs and target genes; as well as (2) new binding site prediction using position weight matrices (PWM) or other motif model. The first step can be accomplished using bidirectional Blast. For the second step there are several alternative tools such as MEME/MAST or PoSSuMsearch. Both utilize existing motif descriptions characterized as PWM and a probabilistic model to determine plausible binding sites. Unfortunately all these models produce a large number of false positives [9]. One strategy to reduce the number of false positives is the use of comparative genomics to see if the binding site is conserved among phylogenetically related organisms [10]. These strategies are known to limit the number of false positives at the cost of increasing the number of false negatives [11].

The second strategy uses a set of differential expression results, from the same organism in several stress conditions, to statistically determine the relationship between genes, from the point of view of expression regulation. Mutual information is a statistical concept describing a probabilistic relationship between variables, which can be used to identify and characterize relationships that are not detected by linear correlation [12]. In this case these variables are gene expression levels measured along several stress conditions. If two genes have a related behavior —for example if one goes up every time another goes down— then knowing the expression of one gives some information about the other. This is evaluated by mutual information. These relationships are not necessarily causal, for example if genes A and B are regulated by a third one C, mutual information will be significative between A and B as well as between A and C [13]. High mutual information is a necessary but not sufficient condition for causality. Once mutual information is evaluated, there are also several strategies to determine which are the mutual information values that correspond to significant gene relationships, and to prune the resulting graph. The general idea is to keep only the direct causal links and to drop the edges which can be better explained by a third gene. Among these strategies are Relevance Networks, ARACNe, CLR, MRNET and C3NET [14], which is the method used in this work.

In this paper, we propose an hybrid strategy to reduce the size of the set of putative TFs and binding sites using mutual information relationships derived from differential expression experimental data coded as rules in a logical program. The result is an unoriented graph which may have a small intersection with the transcriptional regulation network, because transcription factors can act at very low concentrations, and thus may be below the detection threshold of differential expression experiments. To handle this we consider operons instead of genes as the network nodes. Thus, if a TF regulates a given gene, all the operon where the TF belongs appears to control all the operon of the target gene. This makes sense because, in bacteria, operons are the minimal transcriptional unit [15].

This logical program is coded in Answer Set Programming (ASP), a declarative problem solving paradigm in logic programming and knowledge representation, which offers a rich modeling language along with highly efficient inference engines based on Boolean constraint solving technology. In ASP, a problem encoding is a set of logic programming rules which are first transformed into an equivalent propositional logic program and then processed by an answer set solver, which searches for specific solutions to the rules, called Answer Sets. ASP allows solving search problems of high complexity [27].

In summary we have two sources of data for regulation network: (1) the putative TFs and their possible binding sites, and (2) behavioral relationships stated by mutual information. Complementing these with operon prediction, we ask two questions:

- Can each of the mutual information relationships be explained by a shared regulatory element controlling both genes?

– If so, which is the minimal core of TFs and binding sites required to explain all mutual information relationships?

The plan of this work is the following. First, we reconstructed a genetic regulatory network for *Acidithiobacillus ferrooxidans* based in the publicly available sequences. Second, we estimate mutual information from microarray experiments in several stress conditions. Finally, we encode both information (based on regulation and experiments) in Answer Set Programming (ASP) to find the minimal set of TFs and binding site families that explains the observed correlations. This procedure optimizes over the complete space of possible configurations and defines a core regulatory network whose properties are further analyzed.

2 Materials

Genomic DNA sequences were downloaded from NCBI with accession number NC_011761. The functional annotation corresponds to [16]. TFs where determined as the best Blast hit with e-value under 1E-5 against Prodoric database of proteins [6]. Binding sites where located using FIMO from the MEME/MAST suite [17,18] and motifs from Prodoric database, represented as position weight matrices. Gene grouping in operons is given by the prediction of the recent database ProOpDB [19].

Microarray slides where printed with DNA segments from a native *Acidithiobacillus ferrooxidans* strain (Wenelen DSM16786). Genomic DNA was shotgun by sonication and 5 568 segments of 2Kbp nominal size where sequenced by both ends and printed in duplicate [20]. *Acidithiobacillus ferrooxidans* grows naturally in ferric or sulfuric medium. The set of experiments evaluated expression in ferric medium in the green channel and compared it versus: (i) sulfur medium, (ii) shift to sulfur, that is, ferric medium with last minut addition of sulfur, (iii) shift to Chalcopyrite ($CuFeS_2$), (iv) shift to Pyrite (FeS_2), (v) shift to coveline (CuS), (vi) shift to raw mine ore, and (vii) shift to quartz (SiO_2) in the red channel. Cell culture, RNA extraction, hybridization and scanning were performed by BioSigma S.A. (Colina, Chile) as described in [20].

Spot quality assessment was performed using several indices summarized in a *Qcom* value following [21], which is further used as spot weight. Expression was normalized intra-slide using *Lowess* regression and inter-slides using *Gquantile*, since green channel was the same condition for all 7 experiments considered [22]. Each DNA segment expression value was calculated as the weighted average of the spots representing it. Since each spot contains DNA segments which can include several genes, we developed a protocol to determine each gene contribution to total spot luminescence. Using Blast each spot sequence was mapped to ATCC23270 genome. To recover the individual gene expression we solve the minimization

$$\vec{g} = \arg\min ||\vec{c} - M\vec{g}||^2 \qquad \text{s.t.} \quad \vec{g} \geq 0. \tag{1}$$

where $\vec{c}$ is the vector of observed spots luminescence, $\vec{g}$ has components representing the equivalent luminescence (and thus concentration) of each gene, and M is a matrix whose element $M_{i,j}$ is the length of the intersection between spot i sequence and gene j in the Blast alignment.

This transformation is applied to each channel in each slide. The resulting experiment set was analyzed using the standard procedure implemented in *Limma* framework [23] for the R statistical package [24]. A linear model was fitted to each gene using *lmFit* method and *eBayes* was used to estimate statistical significance of observed values [25].

Differential expression data from 56 microarrays corresponding to 7 stress conditions was used to estimate the mutual information between genes using `c3net` library in R [14]. ASP code was executed using the Potassco implementation [26].

3 Results

In this section we detail the results of network reconstruction based on sequence data and its further pruning using logical programming to constraint the network to experimental evidence from differential expression results. Our main result is a reduced network, consistent with the experimental data, with a lower number of spurious regulatory elements and with a size more amenable to be analyzed and simulated using classical tools.

Element	**Network 1** (genes)	**Network 2** (operons)	**Network 3** (ASP optimization)
Nodes_1	3143	1542	1542
Nodes_2	182	182	14
$\text{Edges}_{1\to2}$	66	66	**27**
$\text{Edges}_{2\to1}$	27435	22011	**2428**
MI Restrictions	1245	1206	1206

Table 1. Summary of results for each stage of network reconstruction. Regulatory networks are bipartite graphs, connecting Nodes_1 (genes or operons) to Nodes_2 (binding site motifs). The first network is the full reconstruction linking transcription factor (determined by homology to known transcription factor sequences in Prodoric database) to motifs (found with MEME/FIMO using PWM profiles in Prodoric database) and back to genes (from official annotation in NCBI). $\text{Edges}_{1\to2}$ derive from literature relating transcription factor and motifs, while $\text{Edges}_{2\to1}$ exists when a motif has a instance in the promoter region of a gene. Restrictions were determined from mutual information calculated on experimental data as described in the text. Network 2 consolidates genes into operons (as predicted in ProOpDB), reducing network size based on biological hypothesis. The final network is the result of the method we present in the text. The size of the final network, supporting the same experimental evidence, is much smaller than the original one, TF are reduced by one third, BS are reduced to 8% of the initial number, as marked in boldface.

Transcription regulation networks can be represented as bipartite graphs, with genes and BS motifs as nodes. Our first result was obtained using MAST and motifs from Prodoric database to find 182 motifs represented in 27 435 putative binding sites upstream of 3 143 genes in *A. ferrooxidans*. Using reference protein sequences from the same database, we found by homology 66 genes whose product could encode TFs which can bind in 36 of the 182 motifs. This first network is summarized in the first column of TABLE 1.

Mutual information was calculated from 56 microarrays representing 7 stress conditions as described in Materials, and then filtered using the Conserved Causal Core Network protocol (C3NET, [14]). This protocol discards the weakest mutual information links and keeps a spanning-tree with 1 245 edges between genes whose behavior is experimentally related and constitute restrictions that any network simplification must satisfy.

A second network, also summarized in TABLE 1, was obtained simplifying the first one by grouping the 3 143 genes into 1 542 predicted operons, as predicted in ProOpDB. The 27 435 edges connecting motifs to genes are replaced by 22 011 edges, while the 1 245 mutual information relations are reduced to 1 206.

Our third result uses our method to express the network and minimize the number of relations subject to the mutual information restrictions. We encode biological constraints as disjunctive rules that can be processed by ASP, that is as a finite set of rules of the form

$$a_1; \ldots; a_l \leftarrow a_{l+1}, \ldots, a_m, not\ a_{m+1}, \ldots, not\ a_n$$

where a_n are atoms. Intuitively, atoms can be viewed as facts and rules as deductions to determine new facts. Rules shall be read from right to left: at least one fact in the left part $a_1; \ldots; a_l$ (called "head") shall be true whenever all facts in the right part (called "body") $a_{l+1}, \ldots, a_m, not\ a_{m+1}, \ldots, not\ a_n$ are satisfied. Consequently, the rule with empty head $\leftarrow a.$ means that the fact a is always false. The answers set of a logical program is a set of atoms that satisfy all the logical rules, together with minimality and stability properties, ensuring that every atom appears in at least one rule.

A strong feature that we use in the following is that atoms can be stated as predicates describing relationships between variables. For instance, $upstream(M, G)$ means that the motif M is present in the regulatory region of the gene G; $canBind(G, M)$ means that gene G codes for a transcription factor which can bind to motif M; and $miRel(G, G')$ represent that genes G and G' are related by mutual information, that is, this edge is conserved by C3NET protocol. The predicate $inOperon(G, O)$ is used to group a set of genes G into an operon O. Several genes can be part of the same operon, but no gene can be in two operons.

The first ASP program takes each gene pair related by mutual information and looks for shared motifs either in their upstream region or in the respective promoting region for operons which can regulate indirectly. Since

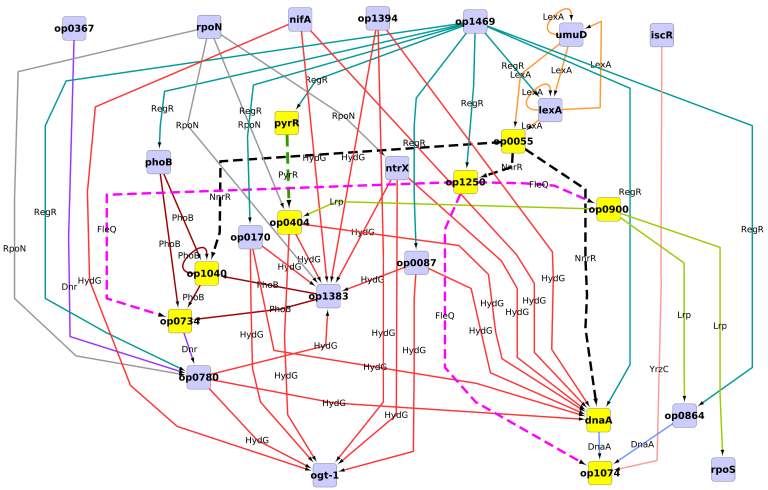


Figure 1. Core regulatory network of *A. ferrooxidans*. Nodes stand for operons, while colored edges stand for different transcription factors transcribed by the operons. This graph represents the connected component of the union of all optimal solutions, being its intersection represented by yellow nodes and dashed edges. Different operons having a common predecessor in the graph contain genes which are correlated in mutual information outputs. To facilitate visualization we show only the operons involved in at least one of the optimal solutions and omit "leaf" operons, that is, those that not control other operons.

mutual information is an index of a related behavior, we expect that each mutual-information-related gene pair should share at least one motif in the sense previously described. We found only one case: AFE_0560 and AFE_2588 are two hypothetical proteins, each in a monocystronic operon, related by mutual information but no common motif was found. This will be further examined, but in a first instance we ignore them.

Leaving this particular case aside we can restrict the model to forbid mutual information relationships which can not be explained. Since several TFs are represented by many genes, we explore the combinatorial space looking for the smallest subsets of TFs satisfying the constraints given by mutual information. This optimization procedure, made feasible by ASP code, also reduces the number of binding site motif used. For each gene G that can bind we generate an eventual bound instance

$$\{bound(G, M)\} \leftarrow canBind(G, M).$$

In ASP syntax this means that for each gene G which can bind to motif M, there will be at most one instance of $bound(G, M)$, or none. This is the key part of the program, as this generates all the combinatorial possibilities.

Then a motif M controls an operon O either if it is directly upstream it or if there is a transcript factor G mediating

$$controles(M, O) \leftarrow upstream(M, G), inOperon(O, G).$$
$$controles(M, O) \leftarrow controles(M, O'), inOperon(O', G'), bound(G', M'), upstream(M', G), inOperon(O, G).$$

Now a mutual information link is explained when there is a motif M that can control both

$$explained(O, O') \leftarrow controles(M, O), controles(M, O'), inOperon(O, A), inOperon(O', B), miRel(A, B).$$

So the main restriction is that all mutual information links must be explained, or equivalently, no mutual information link must remain unexplained

$$\leftarrow miRel(A, B), inOperon(O, A), inOperon(O', B), not\ explained(O, O').$$

Under this ruleset we minimize the number of $bound(G, M)$ predicates, that is, we look for minimization of the number of TFs used, which can explain the observed mutual information relationships, in particular when

Motif	Name	Protein Description	Count
MX000042	RegR	two-component response regulator	27
MX000061	**PyrR**	**transcriptional attenuator and uracil phosphoribosyltransferase activity**	**97**
MX000070	SigE (38-mer)	RNA polymerase sporulation mother cell-specific (early) sigma factor	5
MX000098	DnaA	DNA biosynthesis; initiation of chromosome replication; can be transcription regulator	41
MX000099	PhoB	positive response regulator for pho regulon, sensor is PhoR (or CreC)	21
MX000100	RpoN	RNA polymerase, sigma(54 or 60) factor; nitrogen and fermentation regul.	2
MX000102	Dnr	transcriptional regulator Dnr	77
MX000104	**FleQ**	**transcriptional regulator FleQ**	**97**
MX000114	Ada	O6-methylguanine-DNA methyltransf.; transcription activator/repressor	16
MX000140	**LexA**	**regulator for SOS(lexA) regulon**	**97**
MX000164	Lrp + Leucin	regulator for leucine (or lrp) regulon and amino acid transport system	5
MX000180	HydG	response regulator of hydrogenase 3 activity (sensor HydH)	77
MX000192	**NnrR**	**Crp-Fnr regulatory protein**	**97**
MX000198	YrzC	similar to hypothetical proteins	20

Table 2. Binding Site Motifs used in any of the optimal solutions. All mutual information relationships can be explained using just 7 TFs at a time, in what we call an *optimal solution*. There are 97 different optimal solutions, which use only these motifs. The four motifs marked in boldface are used in all solutions, suggesting they play a key role in the transcriptional regulation. Motif and protein identifiers were taken from Prodoric Database.

several genes could encode TFs binding to the same motif. The minimal number of TFs achieved is 7, which can be realized by 97 different configurations. In the union of all 97 optimal solutions there are 14 motifs (see TABLE 2) and 27 TFs (see TABLE 3), forming a connected subgraph of the initial regulation network, shown in FIG. 1. All operons can be regulated by one of these TFs through binding sites matching one of these motifs. This intersection of all optimal solutions is represented by dashed edges.

As seen in TABLE 3, in all 97 solutions the TF PyrR binds to motif MX000061 and regulates the operon $op0404$, while FleQ binds always to motif MX000104 regulating three operons. Finally, NnrR binds always to motif MX000192 which regulates gene AFE_2696 ($op1250$) and operons $op1040$ and $op1543$ (which contains gene dnaA). At the same time, the motif MX000140 (LexA) was reported in 55 of the 97 instances to regulate gene *lexA* (AFE_1868), and in the other 42 instances regulates gene *umuD* (AFE_1750). Both genes encode putative LexA TFs.

The presence in the solution set of several sigma dependent TFs is consistent with the literature, which describes them as part of the basic transcription modification mechanisms. It is worth noticing that the housekeeping sigma-70 appears in only one solution, meaning that it is not involved in the regulations triggered by the experimental conditions. On the other side sigma-54, associated with nitrogen regulation, appears twice, as encoded by gene AFE_2696 in all solutions, and alternatively by genes AFE_0447, AFE_0957, AFE_174, AFE_2955 and AFE_3025. There are also other TFs related to nitrogen assimilation, as AFE_0024 and AFE_1527. This suggests that sigma-54 plays an important role in the adaptation from an pure iron energy source to a medium including sulfur as an energy source.

An enrichment analysis (data not shown) of GO categories associated to all genes in the operons of FIG. 1 shows significative presence of transcription factors, iron-sulfur cluster assembly and lypopolysaccharide transport, which has been related to the bioleaching activity of *A. ferrooxidans*.

4 Conclusions

We have presented an hybrid method for regulatory network reconstruction, which combines probabilistic tools for identification of regulatory elements and experimental relationships, with ASP logical programming to identify the core regulation elements consistent with experimental evidence considering the whole space of possible configurations between regulators and correlated genes. This last point is essential when comparing our approach with other integrative approaches, such as DISTILLER [28] which cannot afford exploring the whole space of configurations and therefore introduces additional criterias of minimum number of correlated genes and regulators in the motifs that could generate biased results.

Gene	Operon	Name	Description	Count
AFE_0024	op0008	ntrX	nitrogen assimilation regulatory protein NtrX, putative	11
AFE_0119	**op0055**	-	**transcriptional regulator, Crp/Fnr family**	**97**
AFE_0185	op0087	-	type IV fimbriae expression regulatory protein PilR, putative	14
AFE_0447	op0170	-	sigma-54 dependent DNA-binding response regulator	11
AFE_0470	op0180	rpoS	RNA polymerase sigma-38 factor	1
AFE_0672	op0282	iscR	iron-sulfur cluster assembly transcription factor IscR	20
AFE_0857	op0367	-	transcriptional regulator, Crp/Fnr family	12
AFE_0957	op0404	-	sigma-54 dependent transcriptional regulator	11
AFE_1434	op0640	phoB	DNA-binding response regulator PhoB	9
AFE_1527	op0668	nifA	Nif-specific regulatory protein	11
AFE_1658	op0734	-	transcriptional regulator, Crp/Fnr family	65
AFE_1741	op0780	-	sigma-54 dependent transcriptional regulator	11
AFE_1750	op0787	umuD	UmuD protein	42
AFE_1868	op0848	lexA	LexA repressor (EC:3.4.21.88)	55
AFE_1901	op0864	-	hypothetical protein	12
AFE_1990	op0900	-	transcriptional regulator, AsnC family	5
AFE_2271	op1040	-	transcriptional regulator, putative	9
AFE_2342	op1074	-	RNA polymerase sigm-70 factor family	1
AFE_2696	**op1250**	-	**sigma-54 dependent transcriptional regulator**	**97**
AFE_2750	op1276	rpoH	RNA polymerase sigma-32 factor	3
AFE_2798	op1307	ogt-1	methylated-DNA–protein-cysteine methyltransferase (EC:2.1.1.63)	16
AFE_2934	op1383	-	transcriptional regulator, putative	3
AFE_2955	op1394	-	sigma-54 dependent DNA-binding response regulator, putative	8
AFE_3025	op1420	rpoN	RNA polymerase sigma-54 factor	2
AFE_3060	**op1427**	**pyrR**	**PyrR bifunctional protein (EC:2.4.2.9)**	**97**
AFE_3137	op1469	-	DNA-binding response regulator	27
AFE_3309	op1543	dnaA	chromosomal replication initiator protein DnaA	29

Table 3. Genes encoding the transcription factors used in any of the optimal solutions. These are the elements which can explain all mutual information correlations in experimental data. A *Gene* putatively encoding a TF is contained in an *Operon* (as predicted in ProOpDB), its *Name* and *Description* are taken from NCBI's annotation. Column *Count* shows the number of optimal solutions in which each transcription factor is used, the ones marked in boldface are used in all solutions. Note that several sigma-54 dependent transcriptional regulators are included, which suggest that this TF plays an important role in adaptation to different energy sources.

We applied this method to *A. ferrooxidans*, an industrial relevant bacteria whose characteristics difficult classical experimental methods. The proposed network drastically reduces the number of nodes necessary to explain co-expressions in differential expression results, and thus can be a significant contribution to the understanding of the biology of this microorganism.

Our perspective is to improve this network by incorporating signs to the edges, either from experimental data or by homology to other organisms, to get a model which can be used to predict behavior under new experimental conditions. Transcriptional units other than operons can also be considered in an improved version of this model. Those predictions will be eventually validated experimentally.

Acknowledgements

This work was funded by Center for Genome Regulation (Fondap 15090007), Universidad de Chile; Laboratory of Bioinformatics and Mathematics of Genome, Center for Mathematical Modeling, UMI 2807 CNRS-Universidad de Chile; INRIA-U. de Chile IntegrativeBioChile Associate Team; a mobility grant from the International College Doctoral (IDC) of Université Européenne de Bretagne (UEB), INRIA-Conicyt 2010 mobility grant 2010-55 and "Estadías cortas de Investigación para Estudiantes de Doctorado de la Universidad de Chile" grant.

References

[1] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell, 4th edition*. Garland Science, 2002.

[2] E. Davidson and M. Levin, "Gene regulatory networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, p. 4935, Apr 2005.
[3] A.-L. Barabási and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nature Reviews Genetics*, vol. 5, pp. 101–13, Feb 2004.
[4] E. Bouveret and C. Brun, "Bacterial interactomes: from interactions to networks," *Methods Mol. Biol.*, vol. 804, pp. 15–33, 2012.
[5] J. Baumbach, S. Rahmann, and A. Tauch, "Reliable transfer of transcriptional gene regulatory networks between taxonomically related organisms," *BMC Systems Biology*, vol. 3, p. 8, Jan 2009.
[6] A. Grote, J. Klein, I. Retter, I. Haddad, S. Behling, B. Bunk, I. Biegler, S. Yarmolinetz, D. Jahn, and R. Münch, "PRODORIC (release 2009): a database and tool platform for the analysis of gene regulation in prokaryotes," *Nucleic Acids Res*, vol. 37, pp. D61–5, Jan 2009.
[7] T. M. Venancio and L. Aravind, "Reconstructing prokaryotic transcriptional regulatory networks: lessons from actinobacteria," *Journal of Biology*, vol. 8, p. 29, Jan 2009.
[8] G. Mazón, J. M. Lucena, S. Campoy, A. R. F. de Henestrosa, P. Candau, and J. Barbé, "LexA-binding sequences in Gram-positive and cyanobacteria are closely related," *Mol Genet Genomics*, vol. 271, pp. 40–9, Feb 2004.
[9] D. Guhathakurta, "Computational identification of transcriptional regulatory elements in DNA sequence," *Nucleic Acids Research*, vol. 34, pp. 3585–98, Jan 2006.
[10] X. Xie, J. Lu, E. J. Kulbokas, T. R. Golub, V. Mootha, K. Lindblad-Toh, E. S. Lander, and M. Kellis, "Systematic discovery of regulatory motifs in human promoters and 3' utrs by comparison of several mammals," *Nature*, vol. 434, pp. 338–45, Mar 2005.
[11] C. Lepoivre, A. Bergon, F. Lopez, N. B. Perumal, C. Nguyen, J. Imbert, and D. Puthier, "TranscriptomeBrowser 3.0 : introducing a new compendium of molecular interactions and a new visualization tool for the study of gene regulatory networks," *BMC Bioinformatics*, vol. 13, p. 19, Jan 2012.
[12] C. J. Cellucci, A. M. Albano, and P. E. Rapp, "Statistical validation of mutual information calculations: comparison of alternative numerical algorithms," *Phys Rev E Stat Nonlin Soft Matter Phys*, vol. 71, p. 066208, Jun 2005.
[13] A. J. Hartemink, "Reverse engineering gene regulatory networks," *Nat Biotechnol*, vol. 23, pp. 554–5, May 2005.
[14] G. Altay and F. Emmert-Streib, "Inferring the conservative causal core of gene regulatory networks," *BMC Systems Biology 2010 4:132*, vol. 4, p. 132, Jan 2010.
[15] T. A. Brown, *Genomes, 2nd edition*. Wiley-Liss, 2002.
[16] J. Valdés, I. Pedroso, R. Quatrini, R. J. Dodson, H. Tettelin, R. Blake, J. A. Eisen, and D. S. Holmes, "Acidithiobacillus ferrooxidans metabolism: from genome sequence to industrial applications," *BMC Genomics*, vol. 9, p. 597, Jan 2008.
[17] T. L. Bailey, M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble, "MEME suite: tools for motif discovery and searching," *Nucleic Acids Research*, vol. 37, p. W202, Jul 2009.
[18] C. E. Grant, T. L. Bailey, and W. S. Noble, "FIMO: scanning for occurrences of a given motif," *Bioinformatics*, vol. 27, pp. 1017–1018, Apr 2011.
[19] B. Taboada, R. Ciria, C. E. Martinez-Guerrero, and E. Merino, "ProOpDB: Prokaryotic operon database," *Nucleic Acids Res*, vol. 40, pp. D627–31, Jan 2012.
[20] G. Levicán, J. A. Ugalde, N. Ehrenfeld, A. Maass, and P. Parada, "Comparative genomic analysis of carbon and nitrogen assimilation mechanisms in three indigenous bioleaching bacteria: predictions and validations," *BMC Genomics*, vol. 9, p. 581, Jan 2008.
[21] X. Wang, S. Ghosh, and S. W. Guo, "Quantitative quality control in microarray image processing and data acquisition," *Nucleic Acids Research*, vol. 29, pp. E75–5, Aug 2001.
[22] G. Smyth, N. Thorne, and J. Wettenhall, "limma: Linear Models for Microarray Data User's Guide," *Software manual available from http://www. bioconductor. org*, Jan 2003.
[23] G. K. Smyth, "Limma: linear models for microarray data," in *Bioinformatics and Computational Biology Solutions using R and Bioconductor* (R. Gentleman, V. Carey, S. Dudoit, and W. H. R. Irizarry, eds.), pp. 397–420, New York: Springer, 2005.
[24] R Development Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. ISBN 3-900051-07-0.
[25] G. K. Smyth, "Linear models and empirical bayes methods for assessing differential expression in microarray experiments," *Statistical Applications in Genetics and Molecular Biology*, vol. 3, p. Article3, Jan 2004.
[26] M. Gebser, R. Kaminski, B. Kaufmann, M. Ostrowski, T. Schaub, and M. Schneider, "Potassco: The Potsdam answer set solving collection," *AI Communications*, vol. 24, no. 2, pp. 105–124, 2011.
[27] C. Baral, *Knowledge Representation, Reasoning and Declarative Problem Solving*. Cambridge University Press, 2003.
[28] K. Lemmens, T. De Bie, T. Dhollander, S. C. De Keersmaecker, I. M. Thijs, G. Schoofs, A. De Weerdt, B. De Moor, J. Vanderleyden, J. Collado-Vides, K. Engelen, and K. Marchal, "DISTILLER: a data integration framework to reveal condition dependency of complex regulons in Escherichia coli," *Genome Biol.*, vol. 10, no. 3, p. R27, 2009.

Session 8 : Algorithms for Genomics

Conférence invitée

Bertil SCHMIDT

Institut für Informatik, Johannes Gutenberg University Mainz, Germany

Scalable Algorithms and Tools for Biological Sequence Analysis

High-throughput techniques for DNA sequencing have led to a rapid growth in the amount of digital biological data. The current state-of-the-art technology produces 600 billion nucleotides per machine run. Furthermore, the speed and yield of NGS (Next-generation sequencing) instruments continue to increase at a rate beyond Moore's Law, with updates in 2012 enabling 1 trillion nucleotides per run. Correspondingly, sequencing costs (per sequenced nucleotide) continue to fall rapidly, from several billion dollars for the first human genome in 2000 to a forecast US$1000 per genome by the end of 2012. However, to be effective, the usage of NGS for medical treatment will require algorithms and tools for sequence analysis that can scale to billions of short reads. In this talk I will demonstrate how parallel computing platforms based on CUDA-enabled GPUs, multi-core CPUs, and heterogeneous CPU/GPU clusters can be used as efficient computational platforms to design and implement scalable tools for sequence analysis. I will present solutions for classical sequence alignment problems (such as pairwise sequence alignment, BLAST, multiple sequence analysis, motif finding) as well as for NGS algorithms (such as short-read error correction, short-read mapping, short-read assembly, short-read clustering).

Keywords: Sequence alignment, Next-generation sequencing, parallel computing, GPUs.

References

1. Y. Liu, B. Schmidt, D. Maskell. CUSHAW: a CUDA compatible short read aligner to large genomes based on the Burrows-Wheeler transform, Bioinformatics, doi:10.1093/bioinformatics/bts276, 2012
2. Y. Liu, B. Schmidt, D. Maskell. DecGPU: distributed error correction on massively parallel graphics processing units using CUDA and MPI, BMC Bioinformatics, 12:85, 2011
3. W. Liu, B. Schmidt, W. Müller-Wittig, CUDA-BLASTP: Accelerating BLASTP on CUDA-enabled Graphics Hardware, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 8:6, 2011
4. H. Shi, B. Schmidt, W. Liu, W. Mueller-Wittig, A Parallel Algorithm for Error Correction in High-Throughput Short-Read Data on CUDA-enabled Graphics Hardware, Journal of Computational Biology, Vol. 17, No. 4, pp. 603-615, 2010
5. Y. Liu, D. Maskell, B. Schmidt: CUDASW++: optimizing Smith-Waterman sequence database searches for CUDA-enabled graphics processing units, BMC Research Notes, 2:73, 2009

Mapping Reads on a Genomic Sequence: a Practical Comparative Analysis

Sophie SCHBATH[1], Véronique MARTIN[1], Matthias ZYTNICKI[2], Julien FAYOLLE[1], Valentin LOUX[1] and Jean-François GIBRAT[1]

[1] INRA, UR1077 Unité Mathématique Informatique et Génome, F-78350 Jouy-en-Josas
{sophie.schbath, veronique.martin, valentin.loux, jean-francois.gibrat}@jouy.inra.fr
[2] INRA, Unité de Recherche Génomique Info, F-78026 Versailles
matthias.zytnicki@versailles.inra.fr

Abstract *Mapping short reads against a reference genome is classically the first step of many next-generation sequencing data analyses and it should be as accurate as possible. Because of the large number of reads to handle, numerous sophisticated algorithms have been developped in the last 3 years to tackle this problem. In this paper, we compare the performance of 9 mapping tools on a well controled benchmark built for this purpose. We built a set of reads that exist in single or multiple copies in a reference genome and for which there is no mismatch, and a set of reads with 3 mismatches. We considered as reference genome both the human genome and a concatenation of all complete bacterial genomes. On each dataset, we quantified the capacity of the different tools to retrieve all the occurrences of the reads in the reference genome. Special attention was paid to reads uniquely reported and to reads with multiple hits.*

Keywords NGS, mapping reads, benchmark

1 Introduction

Next-generation sequencing data are now the standard to produce genomic and transcriptomic knowledge about an organism, and they are massively produced due to an affordable cost. Mapping short reads against a reference genome is typically the first step to analyze such next-generation sequencing data and it should be as accurate as possible. Because of the high number of reads to handle, numerous sophisticated algorithms have been developped in the last 3 years to tackle this problem and many mapping tools exist now. These tools usually have their own specificities and the question of which one to use for a given application is a very vexing question. A very recent paper (Ruffalo *et al.* (2011)) presents a comparative analysis of 6 mapping tools run on the human genome. Their criteria to compare the performances of the mapping tools are based on the quality score, computed by the different tools, of the retrieved mappings. For a given set of reads, they define the accuracy of a mapping tool as the proportion of "good" mapping at the correct location. They globally analyze the accurracy with respect to various parameters such as the error rate, the size and the frequency of the indels in the reads. Our aim in this paper is also to compare some mapping tools (4 tools in common with the previous reference and 5 additional ones) but we choose to build a more controlled benchmark in order to study quantitatively detailed aspects of the mapping task. By more controlled benchmark, we mean that we know where the reads really map to the reference genome and that all reads have the same amount of errors. We thus address the following questions: Are the tools capable to systematically map a read occurring exactly (with no mismatch) in the reference genome? Can they always do it for a read having as many errors as the maximum number of mismatches allowed in the alignments? For reads occurring at several positions, do they retrieve all the occurrences or only a subset? Do the reads reported as unique really occur only once along the genome? As we will see, the answer will not always be positive, so it is important to know the limitation of each tools.

We have evaluated the performance of the 9 following mapping tools [1]: BWA (Li and Durbin (2009)), Novoalign (Novocraft (2010)), Bowtie (Langmead *et al.* (2009)), SOAP2 (Li *et al.* (2009)), BFAST (Homer

1. The 4 first tools were also evaluated in the study by Ruffalo *et al.*.

et al. (2009)), SSAHA2 (Ning *et al.* (2001)), MPscan (Rivals *et al.* (2009)), GASSST (Rizk and Lavenier (2010)) and PerM (Chen *et al.* (2009)). These tools have been selected either because they are widely used in the bioinformatics community, or because they represent a different algorithmic approach.

They can be divided into two main categories according to the type of algorithm they are based on: hash table based algorithms (indexing either the reads or the reference genome) and Burrows-Wheeler Transform based algorithms (see Table 2). MPscan uses a different approach based on suffix trees. A description of these algorithms is presented in the full paper Schbath *et al.* (2012). Section 2 will then describe the benchmarks we have built to compare these 9 mapping tools. Results are given in section 3 and are discussed in section 4.

2 Benchmark description

2.1 Datasets

The mapping tools are evaluated by two similar experiments. The first experiment (named $\mathcal{H}$ in the following) is run on the human genome (25 chromosomes for 2.7 Gbp). The second experiment (named $\mathcal{B}$) is run on bacterial genomes (904 genomic sequences for 1.7 Gbp).

In the experiment run on the human genome ($\mathcal{H}$), the reference genome $\mathcal{H}_{\text{ref}}$ is taken from the assembly 37.1 made by the NCBI. We have built two sets of reads, all of length 40. The first set of reads ($\mathcal{H}_0$) is composed of 10 millions reads drawn uniformly from the reference genome $\mathcal{H}_{\text{ref}}$. The drawing is done with `wgsim`[2]. Human chromosomes sometime contain a large proportion, as high as 30%, of the letter `N`. Mapping reads with long runs of `N` is of little information to assess the efficiency of mapping tools because these reads should map in numerous locations. We thus decided, beforehand, to remove runs longer than 10 `N`s from the reference genome[3]. The majority of the reads (8,877,107) from $\mathcal{H}_0$ occurs only once[4], but some reads can be repeated many times along the reference genome. For reads occurring more than once, the mean number of occurrences is 722.81 with a standard deviation of 2424.86. Moreover, the most frequent read occurs 53,162 times. The second set of reads ($\mathcal{H}_3$) is built from $\mathcal{H}_0$ by adding exactly 3 mismatches to each read. Therefore $\mathcal{H}_3$ contains also 10 millions reads. We are aware that modern sequencers are very unlikely to produce reads with such an error rate, but the rationale of this dataset is that many projects now produce resequencing and metagenomics data, which may diverge greatly from already sequenced genomes. The positions for the 3 mismatches are drawn uniformly within the 40 positions[5]. A nucleotide `A`, `C`, `G` or `T` is mutated to any of the three other nucleotide with equal probability $1/3$, whereas an `N` is mutated into `A`, `C`, `G` or `T` with probability $1/4$. Among the 10 millions reads from $\mathcal{H}_0$ and $\mathcal{H}_3$, only 49 reads contain some `N`s; the number of `N`s per read is given in Table 1.

	1	2	3	4	5	6	8	9	10	37	40	Total
$\mathcal{H}_0$	5	2	3	2	5	1	0	3	27	0	1	49
$\mathcal{H}_3$	5	2	6	1	4	0	4	15	11	1	0	49

Table 1. Number of reads with a given number of `N`s, from both data set $\mathcal{H}_0$, $\mathcal{H}_3$.

The second experiment ($\mathcal{B}$) is described in the full paper Schbath *et al.* (2012).

2.2 Mapping tools

We evaluated the performance of the following 9 mapping tools: BWA_v0.5.8, Novoalign_v2.06.09, Bowtie_v SOAP_v2.20, BFAST_v0.6.5a, SSAHA2_v2.5.2, MPscan, GASSST_v1.28 and PerM_v0.3.9.

2. Available from https://github.com/lh3/wgsim or SAMtools and developed by Heng Li.
3. One run of 40 `N`s is nevertheless kept in the experiment and this read is included in the $\mathcal{H}_0$ set.
4. The true number of occurrences of each read from $\mathcal{H}_0$ have been computed thanks to a dedicated naive algorithm.
5. We want each read to contain *exactly* 3 mimatches, hence a mismatch position cannot be drawn twice. Uniform drawing means that the first position is drawn uniformly within the 40 position, the second position is drawn uniformly within the 39 remaining positions, and the third position is drawn uniformly within the 38 remaining positions.

Table 2 gathers global characteristics of the tools, namely the type of algorithms they are based on, their output format, their ability to allow mismatches and/or indels in the alignments, and if they can use multiple threads.

Tool	Format	Algorithm	Threads	Gaps	Mismatches
BWA	SAM$^{(*)}$	BWT	yes	yes	yes
Novoalign	SAM	hash the ref.	yes	yes	yes
Bowtie	SAM	BWT	yes	no	yes
SOAP2	in-house	BWT	yes	no	at most 2
BFAST	SAM	hash the ref.	yes	yes	yes
SSAHA2	SAM	hash the ref.	no	no	yes
MPscan	in-house	suffix tree	no	no	no
GASSST	SAM	hash the ref.	yes	yes	yes
PerM	SAM	hash the ref.	no	no	yes

Table 2. Global characteristics of the mapping tools. $^{(*)}$ Sequence Alignments Map.

Additional information for each tool can be found in the full paper Schbath *et al.* (2012), in particular on the options we used to perform the comparisons. For each tool, our aim is indeed to retrieve all the alignments (so-called hits hereafter), either with no mismatch or with at most 3 mismatches, of our read datasets (see section 2.1). Moreover, datasets and command lines used are available on the web page `http://genome.jouy.inra.fr/ngs/mapping/`.

3 Results

We present here the results obtained when mapping the human datasets $\mathcal{H}_0$ and $\mathcal{H}_3$ to the human reference genome $\mathcal{H}_{\text{ref}}$. Results for the bacterial datasets are presented in the full paper Schbath *et al.* (2012) but they are globally similar to those obtained on the human datasets and genome even though their performances are slightly worse.

3.1 Exact mapping from $\mathcal{H}_0$ reads

We firstly performed an exact mapping, i.e. no mismatch is allowed, of the read set $\mathcal{H}_0$ onto the human genome with each of the 9 mapping tools presented in Section 2.2. Note that, by construction of the read set $\mathcal{H}_0$, all reads from $\mathcal{H}_0$ do exactly occur, at least once, in the reference genome. All the tools should then map all the reads with no mismatch, or at least the reads with no `N`s because alignments involving some `N`s account for mismatches for many tools.

3.1.1 Computation time and memory usage Computation times have been obtained by running the mapping tools in single-thread mode, on the same computer [6] and without any competition. However, as we can see in Table 3, SOAP2 and GASSST require more than 50 GB of memory, so they have been run on a more powerful computer [7] but without avoiding competition with other jobs (computation times could then be slightly over-estimated for both tools). For tools which are run through several command lines, we have reported the maximum of the memory required by each step. The most memory consuming step for Bowtie and BFAST is the indexing step, whereas it is the mapping step for SOAP2. Indexing times and mapping times have been reported separately whenever these steps corresponded to distinct command lines (see Table 3). The general trend is as follows: BWA, Bowtie, SOAP2, MPscan and GASSST take several hours to index the reference and map the 10 millions reads (MPscan is the fastest tool), Novoalign and PerM use half a day, and one day or more is required for SSAHA2 and BFAST.

6. Intel Quad Core 2.33 GHz 16 GB RAM.
7. Four Intel Six Core 2.40 GHz 132 GB RAM

Software	Memory usage (GB)	Indexing time	Mapping time	Unmapped reads	Mapped reads	Original position	
						retrieved	not ret.
BWA	2.18	1h 36mn	1h 13mn	49	9,999,951	9,999,951	0
Novoalign	8.12	8mn	13h 24mn	632	9,999,368	9,999,368	0
Bowtie	7.36	3h 25mn	2h 42mn	49	9,999,951	9,999,951	0
SOAP2	51.87	*1h 56mn*$^{(\ddagger)}$	*56mn*$^{(\ddagger)}$	49	9,999,951	9,996,385	3,566
BFAST	9.68	18h 01mn$^{(*)}$	15h 02mn	726,332	9,273,668	9,253,642	20,026
SSAHA2	9.60	24mn	1d 1h	35,875	9,964,125	9,770,914	193,211
MPscan	2.67	1h 20mn		26	9,999,974	9,999,974	0
GASSST	57.93	*8h 45*$^{(\dagger\ddagger)}$		49	9,999,951	9,999,897	54
PerM	13.77	13h 05mn		115,871	9,884,129	9,884,125	4

Table 3. Global characteristics of the run of each software on the exact human dataset ($\mathcal{H}_0$): memory usage, computation times, number of reads which have not been mapped among the 10 millions reads, number of mapped reads and number of mapped reads whose original position has been retrieved in the complete list of hits. $^{(*)}$ Average indexing time per spaced seed computed on 10 seeds. $^{(\dagger)}$ This time does not include the running time of the `gassst_to_sam` command. $^{(\ddagger)}$ This time is slightly over-estimated.

3.1.2 Unmapped reads None of the 9 mapping tools we have analyzed maps all the reads (see Table 3) but for some tools, it is just a question of `Ns`. Indeed, BWA, Bowtie, SOAP2 and GASSST correctly map all the reads except the 49 reads with some `Ns`. MPscan only fails to map 26 reads having some `Ns` and occurring only on the reverse strand. The other tools do not map some of the regular reads. Novoalign could not map 632 reads but succeeded in mapping 22 reads with `Ns` at their original [8] position. More dramatically, SSAHA2, PerM and BFAST fail to map many reads (35,875, 115,871 and 726,332 respectively). BFAST did not map any of the 49 reads with `Ns` whereas SSAHA2 mapped 36 reads with `Ns` but, most of the time, not at their original location. Despite the option `-k 54000`, PerM seems to discard reads with more than 2000 occurrences: the highest number of hits reported by PerM is 1999 and there are exactly 115,822 reads in $\mathcal{H}_0$ which occur 2000 or more times in $\mathcal{H}_{\text{ref}}$.

3.1.3 Is the original position retrieved? For all the mapped reads, we have checked if their original position, i.e. the one randomly drawn to generate the read, belongs to the complete list of returned hits. Indeed, we did not impose constraints on the number of hits per reads. BWA, Novoalign, Bowtie and MPscan retrieve the original position of all their mapped reads (see Table 3). PerM misses the original position for only 4 reads and GASSST for 54 reads. However, SOAP2, BFAST and SSAHA2 fail for many reads; we will point out some possible explanations in the next paragraph.

3.1.4 Unique and multiple reads We can now get in more details and distinguish between *unique* and *non unique* reads. By *unique read* we mean that the read has a unique occurrence into the reference genome whereas a *non unique* read will have more than one occurrence (repeats). This distinction is important when one knows that many post-analyses only consider reads mapping at a unique location. We know that $\mathcal{H}_0$ is composed of 8,877,107 unique reads (including 46 reads with `Ns`) and 1,122,893 multiple reads; the latter have on average 722.81 occurrences. These reference values are indicated at the bottom of Table 4 ("Reference" row).

We can see in Table 4 that BWA, Bowtie, SOAP2 and GASSST report 8,877,061 reads as unique and at their original position which is correct if one discards the 46 unique reads with `Ns`. MPscan also correctly reports the unique reads. Novoalign reports the correct number of unique reads, but this is a coincidence: indeed, Novoalign only maps 22 reads with `Ns`, meaning that some uniquely reported reads are in fact multiple (some hits have been missed). PerM seems to be quite reliable regarding the uniquely reported reads even if it misses some occurrences for 3 of them. SSAHA2 reports more unique reads than in reality, meaning that it misses some hits for some of these reads; This explains also why 9,847 of these reads are not mapped at their original location.

8. The original position of a read is defined as the position randomly drawn to generate the read.

Regarding the non unique reads, BWA, Bowtie and MPscan seem to provide the correct numbers of hits (722.81 hits on average with a standard deviation of 2424). GASSST, Novoalign and SOAP2 are also quite good. Note that SSAHA2 returns much less hits per read (around 80 on average), leading to original positions not retrieved, but this could be explained by the fact that SSAHA2 limits the number of hits per read to 500. This phenomenon is more drastic for BFAST which only returns 3 hits on average per read. It looks like these tools do not explore all the possible occurrences leading to many unmapped reads, to many reads not mapped at their original location and to many reads wrongly reported as unique. The case is different for PerM: its small mean number of hits (126) per mapped read can be explained by the fact that PerM does not map very frequent reads (more than 2000 occurrences).

		Reads uniquely retrieved		Reads with multiple hits		
Software	Non-mapped reads	Nb	Orig. pos. not retr.	Nb	Nb hits mean [sd]	Orig. pos. not retr.
BWA	49	8,877,061	0	1,122,890	722.81 [2424.86]	0
Novoalign	632	8,877,107	0	1,122,261	698.63 [2171.49]	0
Bowtie	49	8,877,061	0	1,122,890	722.81 [2424.86]	0
SOAP2	49	8,877,061	0	1,122,890	653.26 [1804.95]	3566
BFAST	726,332	8,840,305	9,193	433,363	2.96 [1.47]	10,833
SSAHA2	35,875	8,886,204	9,847	1,077,921	79.52 [151.74]	183,364
MPscan	26	8,877,081	0	1,122,893	722.81 [2424.86]	0
GASSST	49	8,877,061	0	1,122,890	722.47 [2422.11]	54
PerM	115,871	8,877,068	3	1,007,061	126.42 [333.29]	1
Reference		8,877,107		1,122,893	722.81 [2424.86]	

Table 4. For each software run on the exact human dataset ($\mathcal{H}_0$), the number of unmapped reads, the number of reads mapped to a unique location and the number of reads mapped to several locations are displayed. Moreover, for the two latter categories, the number of reads whose original position has not been retrieved is reported. The mean number of hits per reads mapped more than once is also presented.

3.2 Mapping reads from $\mathcal{H}_3$ with 3 mismatches

We then performed a mapping, allowing up to 3 mismatches, of the read set $\mathcal{H}_3$ onto the human genome. Since SOAP2 and MPscan do not allow 3 mismatches (SOAP2 is limited to 2 mismatches whereas MPscan only performs an exact mapping), we only used the 7 other tools. By construction of the read set $\mathcal{H}_3$, all the reads do occur at least once with exactly 3 mismatches, meaning that (i) all reads should be mapped, (ii) each read is expected to have more hits than in the previous experiment without mismatch (section 3.1), which leads to (iii) less reads uniquely retrieved.

3.2.1 Computation time and memory usage The amount of memory required by each tool is similar to the one required in the previous experiment for Novoalign, Bowtie, BFAST, SSAHA2 and PerM, it increases for BWA (due to the mapping step) while it decreases for GASSST. GASSST still needs more than 16 GB of memory, so it has been run on a different computer (see section 3.1.1).

Since the indexes are built on the reference genome, indexing times are identical to the previous experiment, only mapping times may change due to allowed mismatches. The running times of BFAST and Perm decrease, respectively, by 30% and 5%. The mapping times of other methods increase globally by a factor 4, except BWA

which becomes 10 times slower (cf. Table 5). To our surprise, BFAST is significantly faster when run on the $\mathcal{H}_3$ instance. The reason for this behavior is not clear to us.

Software	Memory usage (GB)	Indexing time	Mapping time	Unmapped reads	Mapped reads	Original position	
						retrieved	not ret
BWA	10.01	1h 38mn	17h 04mn	49	9,999,951	9,999,951	(
Novoalign	8.07	8mn	2d 6h	47	9,999,953	9,334,497	665,45(
Bowtie	7.36	3h 25mn	9h 57	49	9,999,951	9,999,951	(
BFAST	9.68	18h 01mn$^{(*)}$	10h 45mn	199,451	9,800,549	8,997,831	802,71
SSAHA2	9.60	24 mn	3d 11h	213	9,999,787	5,537,652	4,462,13
GASSST	27.14	*1d 12h*$^{(\dagger\ddagger)}$		326,598	9,673,402	9,631,509	41,89
PerM	13.75	12h 25mn		186,752	9,813,248	9,813,241	7

Table 5. Global characteristics of the run of each software on the 3 mismatches human dataset ($\mathcal{H}_3$): memory usage, computation times, number of reads which have not been mapped among the 10 millions reads, number of mapped reads and number of mapped reads whose original position has been retrieved in the complete list of hits. $^{(*)}$ Average indexing time per spaced seed computed on 10 seeds. $^{(\dagger)}$ This time does not include the running time of the `gassst_to_sam` command. $^{(\ddagger)}$ This time has been obtained on a computer with more memory.

3.2.2 Unmapped reads and reads mapped at their original position Again, none of the tools maps all the reads even when 3 mismatches are allowed. The read set $\mathcal{H}_3$ still contains 49 reads with `N`s (see Table 1). Bowtie and BWA [9] map all the reads with no `N`s and their original position is always retrieved. Novoalign maps more reads (47 unmapped reads) but globally the returned list of hits is incomplete for many reads (666,456 reads). SSAHA2 maps a reasonable number of reads but half of them are not mapped at their original position indicating a very partial list of reported hits; this is not really due to the `-best` option set to 1 because we still obtain 4,325,039 reads not mapped at their orginal position when `-best` is set at 0. BFAST, PerM and GASSST do not map many reads but the diagnosis is different: PerM does not map frequent reads (2000 occurrences or more) but maps the other reads at their original location, whereas GASSST and BFAST seem to produce incomplete lists of hits. PerM seems then to be the most reliable tool after Bowtie and BWA.

3.2.3 Unique and multiple reads The mean number of hits per read, for non unique reads, should be greater than 722.81 (the value for the exact mapping case) because one knows that all exact hits of $\mathcal{H}_0$ are hits with 3 mismatches of $\mathcal{H}_3$. Unlike the results obtained for the exact mapping case, Novoalign retrieves much less hits than expected (15 on average) indicating that many hits are missed. BFAST and SSAHA2 [10] report also very few hits per read leading to a high number of reads not mapped at their original location. Bowtie, BWA and GASSST provides a mean number of hits greater than 1100 which is more satisfactory. As already mentionned, PerM discards reads occurring 2000 or more times, which leads to less reads with multiple hits and to a smaller mean number of hits per mapped reads.

Regarding the reads uniquely retrieved by the different tools, they are less frequent than in the exact mapping case, as expected. However, Novoalign (which performed well in the exact case) returns too many unique hits, many of them at a position different from the orginal one. Unfortunately, this behavior could have a very negative impact on further analyses that only consider uniquely mapped reads. Concerning this point, BFAST is not good, SSAHA2 [11] is catastrophic whereas Bowtie, BWA and PerM are excellent and GASSST provides intermediate results.

9. If we omit the `-N` option, BWA has catastrophic results because it would only map half of the reads and 500,000 of them would not be mapped at their original positions.

10. With `-best 0`, the mean number of hits increases to 198.16 but is still low. Moreover, surprinsingly, 9,689,113 reads are reported with multiple hits.

11. With `-best 0`, SSAHA2's results are worst because it reports only 310,674 reads as unique and half of them are not mapped at their original position

		Reads uniquely retrieved		Reads with multiple hits		
Software	Non-mapped reads	Nb	Orig. pos. not retr.	Nb	Nb hits mean [sd]	Orig. pos. not retr.
BWA	49	8,496,649	0	1,503,302	1161.98 [4634.98]	0
Novoalign	47	8,699,117	202,440	1,300,836	15.12 [36.42]	463,016
Bowtie	49	8,496,649	0	1,503,302	1161.98 [4634.98]	0
BFAST	199,451	8,476,476	84,019	1,324,073	6.17 [4.92]	718,699
SSAHA2	213	8,286,416	3,085,913	1,713,371	6.81 [14.88]	1,376,222
GASSST	326,598	8,193,650	5,703	1,479,752	1139.25 [4554.11]	36,190
PerM	186,752	8,496,655	2	1,316,593	147.25 [342.7]	5

Table 6. For each software run on the 3 mismatches human dataset ($\mathcal{H}_3$), the number of unmapped reads, the number of reads mapped to a unique location and the number of reads mapped to several locations are displayed. Moreover, for the two latter categories, the number of reads whose original position has not been retrieved is reported. The mean number of hits per reads mapped more than once is also presented.

4 Conclusion

We performed a quantitative comparison of 9 mapping tools on well controled benchmarks built for this purpose. We designed a first setting in which all reads do occur exactly in the reference, in one or several copies, and a second setting in which all reads occur with 3 or less mismatches. The reads were uniformly drawn from the reference genome and the original position of these reads was expected to belong to the list of hits returned by each mapping tool. A special attention has been given to reads uniquely reported because many further analyses only consider reads with a single hit.

Regarding the mapping of reads with no mismatch, we can discern 3 groups. BWA, Bowtie, MPscan and GASSST belong to the first group. Programs in this group map all the reads without `N`s and the original position is returned in the list (with few exceptions for GASSST). MPscan is faster than BWA, Bowtie and GASSST but the latter programs can be threaded. The second group contains PerM, Novoalign and SOAP2. The first two programs do not align many reads but they retrieve the original position of the mapped reads. SOAP2 maps all the reads without `N`s but the original position of the reads, in case of multiple copies in the genome, is not always returned. The last group consists of BFAST and SSAHA2. They fail to align a large number of reads and they do not return the original position for a number of reads (the list of returned reads is incomplete). As a consequence, reads reported uniquely by these tools are not necessarily unique and could map elsewhere.

Regarding the mapping of reads with three mismatches, we can cluster the programs in four groups. The first group consists of Bowtie and BWA. These two programs map all the reads without `N`s and, for the human dataset, return their original position, even for unique reads. However, we note that both tools miss the original position of a few reads for the bacterial dataset, with a slight advantage for Bowtie. For comparable results, BWA is about 40% slower than Bowtie. The next group contains only PerM that, although it does not map a large number of reads, usually finds the original position of the reads. The third group consists of GASSST and BFAST that do not map a large number of reads and that also do not return the original position of many reads (BFAST being much worst than GASSST in this respect). The last group consists of Novoalign and SSAHA2 that align most reads without `N`s but that miss the original position of many reads (SSAHA2 being much worst than Novoalign).

It should be emphasized that the good results of BWA and, to a lesser extent, GASSST in our study rely on the fact that we did not use the default parameters, generally set to favor a rapid but incomplete search. We indeed forced both tools to perform an exhaustive search (options `-N -k 3` for BWA and options `-l 0 -s`

5 for GASSST) despite a higher computation time. This trade-off between sensitivity and speed remains one of the difficulties for tuning the numerous parameters of such tools. In this respect, Bowtie appears to be the tool that is the most finely tuned to keep an excellent sensivity and a reasonable execution time with the default parameters.

Reads with Ns are discarded by most of the tools unless the number of allowed mismatches is at least equal to the number of Ns. Although this may be a concern, reads with Ns usually have a poor quality and most would be unalignable anyway.

We reported the computation times using a single thread in order to compare the different tools in similar settings. Of course, several tools can use multi-threads and this advantage should be taken into account for real usage. Moreover, some tools, like GASSST and Bowtie, allow to reduce the computation time by decreasing the sensitivity, ie. by reporting only a subset of hits.

In this work, we did not consider indels mainly because only 4 tools can cope with indels (BWA, Novoalign, BFAST and GASSST), but clearly we will now investigate this aspect. We are clearly aware that our benchmarks favored tools that do not handle gaps. Finally, we left the evaluation of pair-end or mate-pair sequencing data —a work worth a paper *per se*— as future direction for another study.

Acknowledgements

We are grateful to the INRA MIGALE bioinformatics platform http://migale.jouy.inra.fr for providing computational ressources. This work is supported by the French "Agence Nationale de la Recherche" project CBME (ANR-08-GENM-028-01). We also thank Pierre Nicolas for hepful comments on these results.

References

Chen, Y., Souaiaia, T. and Chen, T. 2009. PerM: efficient mapping of short sequencing reads with periodic full sensitive spaced seeds. *Bioinformatics* 5, 2514–21.

Homer, N., Merriman B. and Nelson SF. 2009. BFAST: An Alignment Tool for Large Scale Genome Resequencing *PLoS ONE* 4, e7767.

Langmead, B., Trapnell, C., Pop, M. and Salzberg, SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10:R25.

Li, H. and Durbin, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics* 25, 1754–60.

Li, R., Yu, C., Li, Y., Lam, TW., Yiu, SM., Kristiansen, K. and Wang, J. 2009. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* 25, 1966–67.

Ning, Z., Cox, AJ. and Mullikin, JC. 2001. SSAHA: a fast search method for large DNA databases *Genome research* 11, 1725–29.

Novocraft 2010. novocraft.com

Rivals, E., Salmela, L., Kiiskinen, P., Kalsi, P. and Tarhio, J. 2009. MPscan: Fast Localisation of Multiple Reads in Genomes *Lecture Notes in Computer Science* 5724, 246–260.

Rizk, G. and Lavenier, D. 2010. GASSST: global alignment short sequence search tool *Bioinformatics* 26, 2534–2540.

Ruffalo, M., LaFramboise, T. and Koyutürk, M. 2011. Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics* 27, 2790–2796.

Schbath, S., Martin, V., Zytnicki, M., Fayolle, J., Loux, V. and Gibrta J.-F. 2012. Mapping Reads on a Genomic Sequence: an algorithmic overview and a practical comparative analysis *Journal of Computational Biology*. In press

SortMeRNA: a new software to filter total RNA for metatranscriptomic or RNA analysis

Evguenia KOPYLOVA[1], Laurent NOÉ[1] and Hélène TOUZET[1]

LIFL, UMR8022 CNRS Université Lille 1, and INRIA Lille Nord Europe, 59655 Villeneuve d'Ascq, France
{evguenia.kopylova,laurent.noe,helene.touzet}@lifl.fr

Session 9_A : Protein and Genome Structure

CSA : Comparaison compréhensible d'alignement de paires de structures de protéines

Inken WOHLERS[1], Noël MALOD-DOGNIN[2], Rumen ANDONOV[3] and Gunnar W. KLAU[1]

[1] Life Sciences, Centrum Wiskunde & Informatica, Science Park 123, 1098 XG Amsterdam, the Netherlands
{Inken.Wohlers, Gunnar.Klau}@cwi.nl
[2] INRIA Sophia Antipolis - Méditerranée, 2004 route des Lucioles, 06902 Sophia Antipolis Cedex, France
Noel.Malod-Dognin@inria.fr
[3] INRIA Rennes - Bretagne Atlantique and University of Rennes 1, Campus de Beaulieu, 35042 Rennes Cedex, France
randonov@irisa.fr

Abstract *CSA est un serveur web pour la comparaison compréhensible d'alignements de paires de protéines. Son moteur d'alignement exact calcule, pour des scores de similarité basés sur les distances inter-résidus CMO (Contact Map Overlap maximization), PAUL, DALI et MATRAS, soit des alignements optimaux (ayant le plus haut score), soit des alignements heuristiques avec une garantie de qualité. Ces alignements, plus ceux transférés par l'utilisateur, sont comparés en utilisant de nombreuses mesures de qualité et des visualisations intuitives. CSA apporte un nouveau regard sur les relations structurales entre paires de protéines, et constitue un outil prévu pour l'étude des similarités structurales. CSA est disponible sur* http://csa.project.cwi.nl.

Le ou les auteur(s) ne souhaite(nt) pas que ce document soit diffusé en ligne

CSA: Comprehensive Comparison of Pairwise Protein Structure Alignments

Résumé *CSA is a web server for the comprehensive comparison of pairwise protein structure alignments. Its exact alignment engine computes either optimal, top-scoring alignments or heuristic alignments with quality guarantee for the inter-residue distance based scorings of contact map overlap, PAUL, DALI and MATRAS. These and additional, uploaded alignments are compared using a number of quality measures and intuitive visualizations. CSA brings new insight into the structural relationship of the protein pairs under investigation and is a valuable tool for studying structural similarities. It is available at* http://csa.project.cwi.nl.

Mots-clés *Protein structure alignment, alignment comparison, web server, exact algorithm, scoring functions*

1 Introduction

L'alignement de structures protéiques est une méthode clé pour répondre aux questions biologiques impliquant le transfert d'information de protéines bien étudiées vers des protéines inconnues. Les structures étant mieux conservées durant l'évolution que les séquences, les alignements structuraux permettent des appariements plus précis entre résidus équivalents. Ceci est particulièrement important pour (i) détecter et étudier les motifs structuraux, les sites fonctionnels et les coeurs structuraux, et (ii) mesurer la similarité entre deux protéines et déterminer leur relation d'homologie, via la classification. De nombreux serveurs web sont disponibles et proposent des méthodes individuelles pour produire des alignements structuraux (par exemple [1,2,3,4]).

De nombreux scores structuraux ont été proposés, et il n'y a pas de consensus sur lequel est le meilleur [5]. Les études comparatives montrent que chaque score a ses propres forces et faiblesses, et que les alignements maximisant différents scores peuvent varier considérablement [6].

Dans cet article, nous présentons *CSA* (Comparative Structural Alignment), le premier serveur web pour la comparaison compréhensible des alignements structuraux au niveau des résidus. CSA facilite l'évaluation

A Novel Approach of Spatial Motif Extraction to Classify Protein Structures

Rabie SAIDI[1,2], Wajdi DHIFLI[1,2], Mondher MADDOURI[3,4] and Engelbert MEPHU NGUIFO[1,2]

[1] LIMOS - Blaise Pascal University - Clermont University, BP 10448, Clermont-Ferrand 63000, France
[2] LIMOS - CNRS UMR 6158, Aubière 63173, France
{saidi,dhifli,mephu}@isima.fr
[3] LIPAH - University of Gafsa, Tunisia
[4] King Fahd University, Saudi Arabia
mondher.maddouri@fst.rnu.tn

On the Complexity of two Problems on Orientations of Mixed Graphs

Guillaume FERTIN, Hafedh MOHAMED-BABOU and Irena RUSU

LINA, UMR 6241 CNRS, 2 rue de la Houssinière BP 92208, Nantes, 44322, Cedex 03, France

prenom.nom@univ-nantes.fr

Abstract *Interactions between biomolecules within the cell can be modeled by biological networks, i.e. graphs whose vertices are the biomolecules (proteins, genes, metabolites etc.) and whose edges represent their functional relationships. Depending on their nature, the interactions can be undirected (e.g. protein-protein interactions, PPIs) or directed (e.g. protein-DNA interactions, PDIs). A physical network is a network formed by both PPIs and PDIs, and is thus modeled by a mixed graph. External cellular events are transmitted into the nucleus via cascades of activation/deactivation of proteins, that correspond to paths (called* signaling pathways*) in the physical network from a source protein (cause) to a target protein (effect). There exists experimental methods to identify the cause-effect pairs, but such methods do not provide the signaling pathways. A key challenge is to infer such pathways based on the cause-effect informations. In terms of graph theory, this problem, called* MAXIMUM GRAPH ORIENTATION *(*MGO*), is defined as follows: given a mixed graph G and a set $\mathcal{P}$ of source-target pairs, find an orientation of G that replaces each (undirected) edge by a single (directed) arc in such a way that there exists a directed path, from s to t, for a maximum number of pairs $(s,t) \in \mathcal{P}$. In this work, we consider a variant of* MGO*, called* S-GO*, in which we ask whether* all *the pairs in $\mathcal{P}$ can be connected by a directed path. We also introduce a minimization problem, called* MIN-DB-GO*, in which* all *the pairs in $\mathcal{P}$ must be connected by a directed path, while we allow some edges of G to be doubly oriented (i.e. replaced by two arcs in opposite directions). We investigate the complexity of* S-GO *and* MIN-DB-GO *by considering some restrictions on the input instances (such as the maximum degree of G or the cardinality of $\mathcal{P}$). We provide several polynomial-time algorithms, hardness and inapproximability results that together give an extensive picture of tractable and intractable instances for both problems.*

Keywords Biological networks, computational biology, graph orientation, NP-completeness, APX-hardness.

1 Introduction

A *physical network* [16] is a biological network formed by protein-protein interactions (PPIs) and protein-DNA interactions (PDIs). While PPIs are undirected [5], PDIs are directed from the transcription factors to their target genes [8]. Thus, a physical network can be modeled by a mixed graph whose vertices are proteins, and edges (resp. arcs) are PPIs (resp. PDIs). Such a network is helpful to understand the processes that occur between a cell and its external environment, notably when an external stimulus is to be propagated into the nucleus. Indeed, this propagation is realized via cascades of activation/deactivation of proteins, and these cascades correspond to paths - in the physical network - from a source protein (cause) to a target protein (effect) [16]. The cause-effect pairs can be identified experimentally, for instance by the measure of transcription changes in response to a gene knock-out [16]. However, experimental methods do not provide the paths going from the source to the target proteins. A key problem in biology is to infer these paths by combining causal information on cellular events [10]. This question leads to the well-studied MAXIMUM GRAPH ORIENTATION problem (MGO) [3,4,6,10,14]: given a mixed graph G and a set $\mathcal{P}$ of source-target (cause-effect) pairs of vertices, replace each (undirected) edge (u,v) by a single (directed) arc (either uv or vu), so that in the new graph, there exists a directed path for a maximum number of source-target pairs. In this work, we focus on a variant of MGO, called S-GO, in which we ask whether *all* the pairs in $\mathcal{P}$ can be connected by a directed path [1,7]. We also introduce a minimization problem, called MIN-DB-GO, in which we allow some edges of G to be *doubly*

oriented (i.e. replaced by two arcs in opposite directions), in such a way that all pairs of $\mathcal{P}$ are satisfied (i.e., can be connected by a directed path). In the context of biology, a doubly oriented edge reflects the presence of a reversible reaction. Furthermore, in a dynamic biological system, most reactions tend to be irreversible [9]. For this reason, MIN-DB-GO asks that the number of doubly oriented edges be minimized.

2 Problem Formulation

All along this paper, $G = (V, E, A)$ denotes a mixed graph without loops and with simple edges and arcs, where $V(G)$ (resp. $E(G), A(G)$) is the vertex set (resp. edge set, arc set) of G. The *underlying* graph of G, denoted G^*, is defined as follows: $V(G^*) = V(G)$ and $E(G^*) = E(G) \cup \{(u, v) : uv \in A(G)\}$. Finally, $\Delta(G^*)$ is the maximum degree over all vertices in G^*.

A *path* P in $G = (V, E, A)$ from vertex v_1 to vertex v_m is a sequence $P = v_1, v_2, \dots v_{m-1}, v_m$ of vertices $v_i \in V$ such that for all $1 \leq i \leq m-1$, $(v_i, v_{i+1}) \in E$ or $v_iv_{i+1} \in A$. A *cycle* C in G is a path $v_1, v_2, \dots v_{m-1}, v_m$ such that $v_1 = v_m$. A *circuit* in G is a special case of cycle $v_1, v_2, \dots v_{m-1}, v_1$ where $v_iv_{i+1} \in A$ for all $1 \leq i \leq m-1$. A *Mixed Acyclic Graph* [4] (or *MAG*) is a mixed graph that contains no cycle (and therefore no circuit).

An *orientation* G' of G is a directed graph G' obtained from G by replacing each edge $(u, v) \in E$ by an arc uv, or an arc vu, or by uv and vu simultaneously. An edge (u, v) replaced by both arcs uv and vu is called a *doubly oriented* edge. Any orientation G' of G that contains no doubly oriented edge will be called a *simple orientation*. A pair of vertices $(u, v) \in V \times V$ is said to be *satisfied* by the orientation G' of G if there is a (directed) path from u to v in G'. Let $P = v_1, v_2, \dots v_{m-1}, v_m$ be a path in G. In the following, we will often write *the orientation of P from v_1 towards v_m* to refer to the orientation that replaces every edge of the form (v_i, v_{i+1}), $1 \leq i \leq m-1$, by the arc v_iv_{i+1}.

DEFINITION 2.1. *[1] Let $G = (V, E, A)$ be a mixed graph and let $\mathcal{P} \subseteq V \times V$ be a set of source-target pairs of vertices. The graph G is said to be* $\mathcal{P}$-connected *if for all $(u, v) \in \mathcal{P}$, there is a path in G from u to v.*

DEFINITION 2.2. *[1] Let $G = (V, E, A)$ be a mixed graph and let $\mathcal{P} \subseteq V \times V$ s.t. G is $\mathcal{P}$-connected. A* $\mathcal{P}$-orientation *G' of G is a simple orientation of G that satisfies all pairs in $\mathcal{P}$.*

We call S-GO the problem of deciding whether a graph G admits a $\mathcal{P}$-orientation.

S-GO [1,7]
Instance : A mixed graph $G = (V, E, A)$ and $\mathcal{P} \subseteq V \times V$ s.t. G is $\mathcal{P}$-connected.
Question: Does G admit a $\mathcal{P}$-orientation ?

Analogously to a $\mathcal{P}$-orientation, we define a $(\mathcal{P}, k)$-*DB-orientation* as follows.

DEFINITION 2.3. *Let $G = (V, E, A)$ be a mixed graph and let $\mathcal{P} \subseteq V \times V$ s.t. G is $\mathcal{P}$-connected. Let $k \geq 0$ be an integer. A $(\mathcal{P}, k)$-DB-orientation G' of G satisfies the two following conditions: (i) G' is an orientation of G satisfying all the pairs in $\mathcal{P}$ and (ii) G' contains exactly k doubly oriented edges.*

We are now able to formulate the problem MIN-DB-GO (an illustration of such a problem is provided in Fig. 1).

MIN-DB-GO
Instance : A mixed graph $G = (V, E, A)$, $\mathcal{P} \subseteq V \times V$ s.t. G is $\mathcal{P}$-connected.
Question: Find a $(\mathcal{P}, k)$-DB-orientation of G that minimizes k.

In Section 3, we show that for both problems, we can always assume, without loss of generality, that G is a Mixed Acyclic Graph (MAG). In Section 4, we give some complexity results for S-GO. We study the complexity of MIN-DB-GO in Section 5. Section 6 is the conclusion, together with several open questions.

3 Reduction to Mixed Acyclic Graphs

It has been shown in [14] that starting with any instance $(G_1, \mathcal{P}_1)$ of problem MGO (defined in Section 1), one can construct an equivalent instance $(G_2, \mathcal{P}_2)$ s.t. G_2 is a MAG.

PROPERTY 1. *[14] Let $G_1 = (V_1, E_1, A_1)$ be a mixed graph and let $\mathcal{P}_1 \subseteq V_1 \times V_1$. One can construct a MAG $G_2 = (V_2, E_2, A_2)$ and a set $\mathcal{P}_2 \subseteq V_2 \times V_2$ with $|\mathcal{P}_2| = |\mathcal{P}_1|$ s.t. for every integer $k \geq 0$, there exists a simple orientation of G_1 satisfying k pairs in $\mathcal{P}_1$ if and only if there exists a simple orientation of G_2 satisfying k pairs in $\mathcal{P}_2$.*

Applying Property 1 with $k = |\mathcal{P}_1|$, we obtain that G_1 admits a $\mathcal{P}_1$-orientation if and only if G_2 admits a $\mathcal{P}_2$-orientation. Hence, in the S-GO problem, without loss of generality we can always consider the input mixed graph to be a MAG. We will show that this property is also valid for the MIN-DB-GO problem. For this, we first show in Property 2 that, for any mixed graph G, we can find a orientation G' such that the edges of any cycle C in G become arcs of a circuit C' in G'. We then show in Property 4 that the vertices of any circuit in G' can be contracted in a single vertex without changing the nature of the problem. This proof closely follows the one in [14].

PROPERTY 2 (ORIENTATION OF CYCLES). *Let $G = (V, E, A)$ be a mixed graph and let $\mathcal{P} \subseteq V \times V$. Let C be a cycle in G. There exists an optimal solution G'' for* MIN-DB-GO*, with inputs G and $\mathcal{P}$, in which C becomes a circuit in G''.*

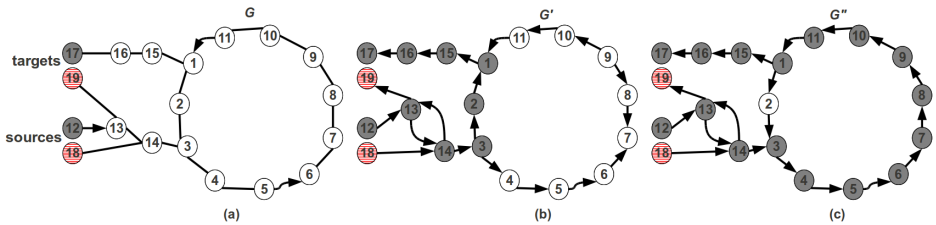


Figure 1. (a) $(G, \mathcal{P})$ is an instance of MIN-DB-GO with $\mathcal{P} = \{(12, 17), (18, 19)\}$, C is a cycle in G with vertex set $\{1, 2, \ldots, 11\}$ (b) G' is an optimal $(\mathcal{P}, 1)$-DB-orientation of G in which the orientation C' of C is not a circuit (c) G'' is an optimal $(\mathcal{P}, 1)$-DB-orientation of G in which the orientation C' of C is a circuit. Gray vertices in G' (resp. G'') induce a path satisfying the pair $(12, 17)$.

Proof. First, since G is $\mathcal{P}$-connected, there must exist an optimal solution G' for MIN-DB-GO with inputs G and $\mathcal{P}$. Let G'' be an orientation of G s.t. (i) C becomes a circuit C'' and (ii) the edges in $E(G) \backslash E(C)$ are oriented similarly as in G' (see Fig. 1 for an illustration). We now show that G'' is also an optimal solution for MIN-DB-GO with inputs G and $\mathcal{P}$. Let $(u, v) \in \mathcal{P}$. If $u, v \in V(C)$ then obviously the pair (u, v) is satisfied in G''. If $u \notin V(C)$ or $v \notin V(C)$, then let us consider a path $P' = a_1, a_2, \ldots a_m$ in G', from $u = a_1$ to $v = a_m$, that satisfies the pair (u, v). Let $x = min\ \{i : a_i \in V(C)\}$ and let $y = max\ \{i : a_i \in V(C)\}$. Then the pair (u, v) is satisfied in G'' by the path formed by (1) the path in G'' induced by the vertex set of the subpath of P' going from a_1 to a_x, (2) the subpath of C'' going from a_x to a_y and (3) the path of G'' induced by the vertex set of the subpath of P' going from a_y to a_m (see for example Fig. 1, in which $a_x = 3$ and $a_y = 1$). Let E_{DB} denote the set of doubly oriented edges in G'. The set of doubly oriented edges in G'' is $E_{DB} \backslash E(C)$. However, by minimality of $|E_{DB}|$, we know that $E_{DB} \cap E(C) = \emptyset$, and consequently G'' is also an optimal solution for MIN-DB-GO. □

In the following, we say that two instances $(G_1, \mathcal{P}_1)$ and $(G_2, \mathcal{P}_2)$ of MIN-DB-GO are *e*quivalent if and only if for every integer $k \geq 0$, there exists a $(\mathcal{P}_1, k)$-DB-orientation of G_1 if and only if there exists a $(\mathcal{P}_2, k)$-DB-orientation of G_2.

Let $G = (V, E, A)$ be a mixed graph and let $\mathcal{P} \subseteq V \times V$. Let $G_1 = (V_1, E_1, A_1)$ be the mixed graph obtained from G by orienting, iteratively, each cycle into a circuit. According to Property 2, $(G, \mathcal{P})$ and $(G_1, \mathcal{P})$ are equivalent. Let also $\mathcal{P}_1$ denote the set obtained from $\mathcal{P}$ by removing each pair $(u, v) \in \mathcal{P}$ s.t. there is a

directed path in G_1 from u to v. In that case, the instance $(G_1, \mathcal{P}_1)$ of MIN-DB-GO obtained from $(G, \mathcal{P})$ will be called a *reduced instance*. Clearly, $(G_1, \mathcal{P}_1)$ and $(G_1, \mathcal{P})$ are equivalent, and thus the following property holds.

PROPERTY 3 (REDUCED INSTANCES). *Let $(G_1, \mathcal{P}_1)$ be a reduced instance of* MIN-DB-GO *obtained from instance $(G, \mathcal{P})$. Then $(G, \mathcal{P})$ and $(G_1, \mathcal{P}_1)$ are equivalent.*

PROPERTY 4 (CONTRACTION OF CIRCUITS). *Let $(G_1, \mathcal{P}_1)$ be a reduced instance of* MIN-DB-GO*, and let C' be a circuit in G_1. Let $(G_2, \mathcal{P}_2)$ be the instance of* MIN-DB-GO *defined as follows: (i) $\mathcal{P}_2 = \mathcal{P}_1$ and (ii) G_2 is the graph obtained from G_1 by contracting the vertices of C' in a single vertex. Then, $(G_1, \mathcal{P}_1)$ and $(G_2, \mathcal{P}_2)$ are equivalent.*

Proof. The graph $G_2 = (V_2, E_2, A_2)$ is defined as follows: $V_2 = (V_1 \backslash V(C')) \cup \{x_0\}$, $E_2 = (E_1 \backslash \{(u,v) : u \in V(C') \text{ or } v \in V(C')\}) \cup \{(u, x_0) : u \notin V(C') \text{ and } \exists v \in V(C') \text{ s.t. } (u,v) \in E_1\}$, $A_2 = (A_1 \backslash \{uv : u \in V(C') \text{ or } v \in V(C')\}) \cup \{ux_0 : u \notin V(C') \text{ and } \exists v \in V(C') \text{ s.t. } uv \in A_1\} \cup \{x_0 v : v \notin V(C') \text{ and } \exists u \in V(C') \text{ s.t. } uv \in A_1\}$. In other words, G_2 is obtained from G_1 by contracting the circuit C' in a single vertex x_0. Obviously, $|E_1| = |E_2|$.

Let $\mathcal{P}_2 = \mathcal{P}_1$. Now, let us show that the two instances $(G_1, \mathcal{P}_1)$ and $(G_2, \mathcal{P}_2)$ are equivalent. Let $G'_1 = (V_1, A'_1)$ be a $(\mathcal{P}_1, k)$-DB-orientation of G_1. We construct an orientation G'_2 of G_2 as follows. Let $(u,v) \in E_2$. If $u \neq x_0$ and $v \neq x_0$, then (u,v) is oriented in G'_2 similarly as in G'_1. If $u = x_0$, then there is a vertex $w \in V(C')$ s.t. $(w,v) \in E_1$. If $wv \in A'_1$ (resp. $vw \in A'_1$) we replace in G_2 the edge (x_0, v) by the arc $x_0 v$ (resp. $v x_0$). The case $v = x_0$ is similar. Let $(u,v) \in \mathcal{P}_1$ and let $P'_1 = a_1, a_2, \ldots, a_m$ be a directed path in G'_1 from $u = a_1$ to $v = a_m$ satisfying the pair (u,v). Let $x = min\ \{i : a_i \in V(C')\}$ and let $y = max\ \{i : a_i \in V(C')\}$. Then the pair (u,v) is satisfied in G'_2 by the path formed by (1) the path in G'_2 induced by the vertex set of the subpath of P'_1 going from a_1 to a_{x-1}, (2) the vertex x_0 (3) the path of G'_2 induced by the vertex set of the subpath of P'_1 going from a_{y+1} to a_m. Obviously, G'_2 is a $(\mathcal{P}_2, k)$-DB-orientation of G_2. Reciprocally, starting with a $(\mathcal{P}_2, k)$-DB-orientation of G_2, by the same way, one can construct a $(\mathcal{P}_1, k)$-DB-orientation of G_1. Hence, the property follows. □

Now, using the previous properties, we are able to show that in the MIN-DB-GO we may, without loss of generality, assume that the input mixed graph is a MAG.

PROPERTY 5 (REDUCTION TO MAG). *Let $(G, \mathcal{P})$ be an instance of the* MIN-DB-GO *problem. Then there exists an equivalent instance $(G_M, \mathcal{P}_M)$ of* MIN-DB-GO *s.t. G_M is a MAG.*

Proof. We construct the graph G_M and the set $\mathcal{P}_M$ by applying the following process:

1. Construct the reduced instance $(G_1, \mathcal{P}_1)$ obtained from $(G, \mathcal{P})$
2. Construct the graph G_2, obtained by contracting, in G_1, every circuit into a single vertex, and let $\mathcal{P}_2 = \mathcal{P}_1$
3. If G_2 is a MAG then set $G_M = G_2$ and $\mathcal{P}_M = \mathcal{P}_2$. Otherwise, set $G = G_2$ and $\mathcal{P} = \mathcal{P}_2$, and return to step 1.

Properties 3 and 4 ensure that $(G_M, \mathcal{P}_M)$ is equivalent to $(G, \mathcal{P})$, which proves the property. □

Therefore, we will always assume, in the remaining of the paper, that for any instance $(G, \mathcal{P})$ of MIN-DB-GO (resp. S-GO), $G = (V, E, A)$ is MAG and G is $\mathcal{P}$-connected. Note that we can also assume that G^* to be connected. Otherwise, we can consider separately each graph $G_1, G_2, \ldots, G_r$ induced, in G, by the vertices of the connected components of G^*.

Let $G = (V, E, A)$ be a MAG and $\mathcal{P} = \{(s_i, t_i) \in V \times V : 1 \leq i \leq m\}$ be a set of pairs of vertices. For each i, $1 \leq i \leq m$, we note by n_i the number of distinct paths in G from s_i to t_i. All along the paper, the integer B is the following value: $B = \max\{n_i : 1 \leq i \leq m\}$.

In the next sections, we study the complexity of S-GO (Section 4) and MIN-DB-GO (Section 5), by considering different constraints on the three following parameters: $\Delta(G^*)$, B and $|\mathcal{P}|$.

Remark that when $B = 1$ we can easily solve the S-GO and the MIN-DB-GO problems. Indeed, G is $\mathcal{P}$-connected and G^* is connected, thus if in addition we have $B = 1$, then for each pair $(s,t) \in \mathcal{P}$ there is a unique path P_i in G from s_i to t_i. Consequently, in order to satisfy the pair (s_i, t_i) we must orient P_i from s_i towards t_i. Then, we orient each remaining edge (u,v), in G, in a unique arbitrarily direction. Obviously, in this orientation we create a minimum number of doubly oriented edges, and thus MIN-DB-GO is optimally solved. If there is no doubly oriented edge at the end of the process, then we obtain a $\mathcal{P}$-orientation of G. Otherwise, G has no $\mathcal{P}$-orientation. The case $\Delta(G^*) = 1$ is obvious.

	$\Delta(G^*) = 2$	$\Delta(G^*) = 3$		
		$B = 2$	$B = 3$	B is unbounded
S-GO	P [Cor. 1]	P [Th. 1]	NPC [Thm. 3]	NPC [Th. 3]
MIN-DB-GO	Open	APX-h [Th. 9]	NPC [Th. 6], Non-approx. [Th. 7], APX-h [Th. 9]	NPC [Th. 6], Non-approx. [Th. 7], APX-h and W[1]-h [Th. 8]

Table 1. Complexity of S-GO and MIN-DB-GO when G is a MAG and G^* is a bounded degree graph. Recall that $B = \max\{n_i, 1 \leq i \leq |\mathcal{P}|\}$, where n_i is the number of distinct paths in G from s_i to t_i. Note also that the result provided in Theorem 1 remains valid even when $\Delta(G^*)$ is unbounded.

	$\lvert\mathcal{P}\rvert \leq 2$	$\lvert\mathcal{P}\rvert \geq 3$ (and $\lvert\mathcal{P}\rvert = \mathcal{O}(1)$)	
		$B = \mathcal{O}(1)$	B is unbounded
S-GO	P [Arkin et al. [1]]	P [Th. 2]	Open
MIN-DB-GO	P [Th. 4]	P [Th. 5]	Open

Table 2. Complexity of S-GO and MIN-DB-GO when G is a MAG and $|\mathcal{P}|$ is a constant.

The complexity results, when $B \geq 2$ and $\Delta(G^*) \geq 2$, are summarized in Table 1 and Table 2. Interestingly, Table 1 shows that parameter B defines the border between easy ($B = 2$) and difficult ($B = 3$) instances of S-GO, even when G^* is of small degree ($\Delta(G^*) = 3$). Unlike S-GO, the problem MIN-DB-GO is difficult even when $B = 2$. In Table 2, we show the complexity of both problems when $|\mathcal{P}|$ is a constant. Due to space constraints, some proofs are omitted in this paper.

4 Complexity of the S-GO problem

It has been shown that the S-GO problem is polynomial-time solvable on undirected graphs [7] and NP-complete on general MAGs [1]. In this section, we investigate the complexity of the S-GO problem for MAGs with bounded $\Delta(G^*)$ and/or bounded B (see Table 1), and for bounded $|\mathcal{P}|$ (see Table 2).

4.1 Easy cases

THEOREM 1. *The* S-GO *problem is polynomial-time solvable when G is a MAG and $B = 2$.*

Proof. For each pair $(s_i, t_i) \in \mathcal{P}$ there are in G at most two paths from s_i to t_i and such paths can be computed in polynomial-time. If for a pair $(s_i, t_i) \in P$, there is only one path from s_i to t_i, then we orient it from s_i towards t_i and we remove the pair (s_i, t_i) from the set $\mathcal{P}$. We continue this process until (1) G is no longer P-connected or (2) $\mathcal{P} = \emptyset$ or (3) for each pair $(s_i, t_i) \in P$ there are exactly two paths from s_i to t_i. The first case implies that G has no $\mathcal{P}$-orientation. In the second case we arbitrarily orient each edge, in the resulting graph, in a unique direction to obtain a $\mathcal{P}$-orientation. Finally, in the last case we reduce to an instance of the S-GO problem in which there are, in G, exactly two paths from s_i to t_i for all $(s_i, t_i) \in \mathcal{P}$.

We note by X_{i1} and X_{i2} the two paths, in G, from s_i to t_i. Given $i, j \in \{1, 2, \ldots |\mathcal{P}|\}$, $i \neq j$, and $a, b \in \{1, 2\}$, we say that the two paths X_{ia} and X_{jb} are in *conflict* if orienting X_{ia} from s_i towards t_i and X_{jb} from s_j towards t_j, creates a doubly oriented edge. Now, we construct an instance $(\mathcal{X}, \mathcal{C})$ of the problem 2-SAT as follows. Let $\mathcal{X} = \{x_{i1}, x_{i2} \ : \ 1 \leq i \leq |\mathcal{P}|\}$ be the variable set. For all $i \in \{1, 2, \ldots, |\mathcal{P}|\}$, we add the clause $c_i = (x_{i1} \vee x_{i2})$. For all $i, j \in \{1, 2, \ldots, |\mathcal{P}|\}$, $i \neq j$, and $a, b \in \{1, 2\}$, we add the clause

$(\overline{x_{ia}} \vee \overline{x_{jb}})$, if paths X_{ia} and X_{jb} are in conflict. Let us show that there is an assignment of the variables in $\mathcal{X}$ that satisfies all clauses in $\mathcal{C}$ if and only if G has a $\mathcal{P}$-orientation. Indeed, consider a truth assignment of clauses in $\mathcal{C}$ and let x_{ih_i}, $1 \leq h_i \leq 2$, be a true literal of clause c_i, $1 \leq i \leq |\mathcal{P}|$. We orient in G the path X_{ih_i}, from s_i towards t_i, for all i, $1 \leq i \leq |\mathcal{P}|$. This orientation cannot create any doubly oriented edges. Otherwise, there are $i, j \in \{1, 2, \dots |\mathcal{P}|\}$, $i \neq j$ such that the paths X_{ih_i} and X_{jh_j} are in conflict, implying that the clause $(\overline{x_{ih_i}} \vee \overline{x_{jh_j}})$ is unsatisfied. To finish the orientation of G, we orient arbitrarily the remaining edges in G without creating any doubly oriented edge.

Now, let us show the reverse implication. We consider the set $\{Y_{1h_1}, Y_{2h_2}, \dots, Y_{|\mathcal{P}|h_{|\mathcal{P}|}}\}$ s.t. Y_{ih_i} is a directed path from s_i to t_i, in a $\mathcal{P}$-orientation of G. Each path Y_{ih_i} is the orientation (from the source towards the target vertex) of a mixed path $X_{ih_i} = G[V(Y_{ih_i})]$, $h_i \in \{1, 2\}$, for all $1 \leq i \leq |\mathcal{P}|$. We set to $true$ the variable set $\{x_{ih_i} : 1 \leq i \leq |\mathcal{P}|\}$ and we set to $false$ the remaining variables. Obviously, this assignment satisfies the clause set $\{c_i : 1 \leq i \leq |\mathcal{P}|\}$. By contradiction, assume now that some clause $(\overline{x_{ia}} \vee \overline{x_{jb}})$ is not satisfied. Then $x_{ia} = true$ and $x_{jb} = true$. Consequently, in the resulting $\mathcal{P}$-orientation of G, the path X_{ia} (resp. X_{jb}) is oriented from s_i towards t_i (resp. from s_j towards t_j). A contradiction, because a $\mathcal{P}$-orientation cannot use, simultaneously, two paths that are in conflict. As the problem 2-SAT is polynomial-time solvable [2], we deduce that one can solve in polynomial-time the S-GO problem when graph G is a MAG s.t. there are in G at most two paths from s_i to t_i, for all $(s_i, t_i) \in \mathcal{P}$. □

COROLLARY 1. *The* S-GO *problem is polynomial-time solvable when G is a MAG and $\Delta(G^*) = 2$.*

Proof. The graph G^* is connected. Thus when $\Delta(G^*) = 2$, the graph G^* must be a path or a cycle, and consequently $B \leq 2$. If $B = 1$ the S-GO problem is trivial. If $B = 2$, we deduce from the previous result (Theorem 1) that the S-GO problem is polynomial-time solvable. □

Now, we show that the S-GO problem is polynomial-time solvable when both parameters B and $|\mathcal{P}|$ are bounded.

THEOREM 2. *The* S-GO *problem is polynomial-time solvable when G is a MAG, $|\mathcal{P}| = \mathcal{O}(1)$ and $B = \mathcal{O}(1)$.*

4.2 Difficult cases

We showed in Theorem 1 that the S-GO problem is easy when $B \leq 2$. However, in the following theorem, we show that when $B = 3$ the problem S-GO becomes difficult.

THEOREM 3. *The* S-GO *problem is* NP*-complete even when the graph G is a MAG, $\Delta(G^*) = 3$ and $B = 3$.*

Proof. Arkin et al. [1] provided an NP-completeness proof of the S-GO problem on general MAGs. Their proof is based on a reduction for the Satisfiability problem (SAT). Here, we modify the mixed graph G constructed from their reduction to ensure that $\Delta(G^*) = 3$. For this, we reduce to our problem, the problem 3-SAT in which each variable appears at most in four clauses [15]. Let $(\mathcal{C}_m, \mathcal{V}_n)$ be an instance of 3-SAT s.t. $\mathcal{C}_m = \{c_1, \dots c_m\}$ is a set of clauses and $\mathcal{V}_n = \{x_1, \dots, x_n\}$ is a variable set, and for all j, $1 \leq j \leq n$, the variable x_j satisfies the following condition: (1) x_j and $\overline{x_j}$ appear at most in four clauses. In addition, one can require second condition: (2) for each variable x_j, there is at least one clause that contains x_j and at least one clause that contains $\overline{x_j}$. Otherwise, w.l.o.g the variable x_j can be fixed to *true* or *false*. Now, let us construct an instance $(G, \mathcal{P})$ of the S-GO problem. For each clause c_i, we create two vertices s_i and t_i, $1 \leq i \leq m$. For each variable x_j, we create these 15 vertices: $\{u_j, v_j\} \cup \{a_{jk}, b_{jk}, a'_{jk}, b'_{jk}\}_{1 \leq k \leq 3}$. Then, we add an edge (u_j, v_j) and the four following directed paths: $a_{j1}a_{j2}a_{j3}u_j$, $v_j b_{j3} b_{j2} b_{j1}$, $a'_{j1}a'_{j2}a'_{j3}v_j$ and finally $u_j b'_{j3} b'_{j2} b'_{j1}$, for all $1 \leq j \leq n$. For each variable x_j, there are k_j clauses containing x_j and k'_j clauses containing $\overline{x_j}$ s.t. $1 \leq k_j \leq 3$, $1 \leq k'_j \leq 3$ and $k_j + k'_j \leq 4$. Let $\{c_{i_1}, c_{i_2}, \dots, c_{i_{k_j}}\}$ (resp. $\{c_{i'_1}, c_{i'_2}, \dots, c_{i'_{k'_j}}\}$) be the set of clauses that contain x_j (resp. $\overline{x_j}$). We add an arc $s_{i_\alpha} a_{j\alpha}$ and an arc $b_{j\alpha} t_{i_\alpha}$, for all $\alpha \in \{1, 2, \dots k_j\}$. Also, we add an arc $s_{i'_\beta} a'_{j\beta}$ and an arc $b'_{j\beta} t_{i'_\beta}$, for all $\beta \in \{1, 2, \dots k'_j\}$. To finish our construction, we set $\mathcal{P} = \{(s_i, t_i), 1 \leq i \leq m\}$. An example of construction is illustrated in Fig. 2. According to conditions (1) and (2), one can easily show that $\Delta(G^*) = 3$. In addition, for each pair (s_i, t_i) there are exactly three paths in G fro s_i to t_i, because each clause in $\mathcal{C}_m$ contains exactly three literals. Thus $B = 3$.

We claim that there is an assignment satisfying all the clauses in $\mathcal{C}_m$ if and only if there exists a $\mathcal{P}$-orientation of G. Indeed, consider an assignment satisfying all the clauses in $\mathcal{C}_m$, similarly to the proof presented in [1], if $x_j = true$ (resp. $x_j = false$) then we orient the edge (u_j, v_j) from u_j to v_j (resp. from v_j to u_j). Let l_i be a true literal of clause c_i. Then, there is a variable x_j s.t. $l_i = x_j$ or $l_i = \overline{x_j}$. If $l_i = x_j$ (resp. $l_i = \overline{x_j}$) then there is an integer k_i, $1 \leq k_i \leq 3$, such that $s_i a_{jk_i}, b_{jk_i} t_i \in A(G)$ (resp. $s_i a'_{jk_i}, b'_{jk_i} t_i \in A(G)$). Thus, the pair (s_i, t_i) is satisfied by the path $s_i a_{jk_i} a_{j(k_i+1)} \dots u_j v_j b_{j3} \dots b_{jk_i} t_i$ (resp. $s_i a'_{jk_i} a'_{j(k_i+1)} \dots v_j u_j b'_{j3} \dots b_{jk_i} t_i$).

Now, let us prove the reverse implication. Given a $\mathcal{P}$-orientation G' of G, we set the variable x_j to $true$ (resp. to $false$) if the edge $u_j v_j \in A(G')$ (resp. $v_j u_j \in A(G')$). Let c_i be a clause in $\mathcal{C}_m$. Then the pair (s_i, t_i) is satisfied in by a directed path P in G', from s_i to t_i, going through an arc $u_j v_j$ or an arc $v_j u_j$. If P contains the arc $u_j v_j$ then the clause c_i must contain the literal x_j and thus c_i is satisfied. If P contains the arc $v_j u_j$ (consequently $x_j = false$) thus the clause c_i must contain the literal $\overline{x_j}$ and thus c_i is also satisfied. □

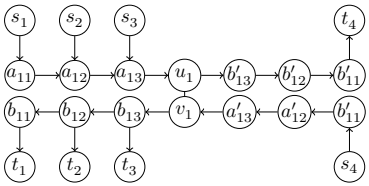


Figure 2. Construction of an instance $(G, \mathcal{P})$ of the S-GO problem, from an instance of 3-SAT in which each variable appears at most in four clauses. Here, the variable set is $\mathcal{V} = \{x_j, 1 \leq j \leq 6\}$ and the clause set is $\mathcal{C} = \{c_i, 1 \leq i \leq 4\}$ s.t. $c_1 = (x_1 \vee \overline{x_2} \vee x_3)$, $c_2 = (x_1 \vee x_4 \vee x_5)$, $c_3 = (x_1 \vee \overline{x_4} \vee \overline{x_6})$ and $c_4 = (\overline{x_1} \vee \overline{x_5} \vee x_6)$. The set of pairs of vertices is $\mathcal{P} = \{(s_i, t_i), 1 \leq i \leq 4\}$. In this figure, we show only the subgraph corresponding to variable x_1.

5 Complexity of MIN-DB-GO

Let DB-GRAPHORIENTATION denote the decision version of the minimization problem MIN-DB-GO. The S-GO problem, investigated in the previous section (Section 4), is a particular case of the problem DB-GRAPHORIENTATION when no doubly oriented edge is allowed. Hence, each $\mathcal{P}$-orientation of G is a solution of MIN-DB-GO. However, if there is no $\mathcal{P}$-orientation of G, then we conclude just that at least one edge must be doubly oriented in a solution of MIN-DB-GO, but in general that gives no information about the number of edges to be doubly oriented to solve the MIN-DB-GO problem.

In this section, we study the complexity of MIN-DB-GO when the input graph is a MAG (see Table 1 and Table 2). As in the previous section, we suppose that G is a $\mathcal{P}$-connected MAG.

5.1 Easy cases

We first show that, similarly to the S-GO problem, the MIN-DB-GO problem is also polynomial-time solvable for general MAGs when $|\mathcal{P}| \leq 2$.

THEOREM 4. *The* MIN-DB-GO *problem is polynomial-time solvable when G is a MAG and $|\mathcal{P}| \leq 2$.*

Proof. The case $|\mathcal{P}| = 1$ is obvious. Let $G = (V, E, A)$ be a MAG and $\mathcal{P} = \{(s_1, t_1), (s_2, t_2) \in V \times V\}$.

A *$\mathcal{P}$-essential* edge is an edge $e = (u, v) \in E$, s.t. if we orient e from u to v or from v to u, the graph G is no longer $\mathcal{P}$-connected. One can compute the $\mathcal{P}$-essential edges in polynomial-time [1].

Let E_{ess} (resp. E_{min}) be the set of $\mathcal{P}$-essential edges (resp. the set of doubly oriented edges in a solution of the MIN-DB-GO problem).

We show that $E_{min} = E_{ess}$. Let $e = (u, v) \in E_{ess}$. If we orient e in a unique direction (from u to v or from v to u) then, by definition of $\mathcal{P}$-essential edges, there is an integer i, $1 \leq i \leq 2$, s.t. there is no path in G from s_i to t_i. Thus, whatever the orientation of edges in $E - \{e\}$, if e is not doubly oriented, then the pair (s_i, t_i) would not be satisfied. Hence, we must doubly orient each edge $e \in E_{ess}$, which implies that $E_{ess} \subseteq E_{min}$.

Conversely, let $G' = (V, E', A')$ denote the mixed graph obtained from G after replacing each $\mathcal{P}$-essential edge (u, v) by the arcs uv and vu, i.e., $V(G') = V(G)$, $E(G') = E(G) \backslash E_{ess}$ and $A' = A \cup \{uv, vu : (u, v) \in E_{ess}\}$. The graph G' does not contain any $\mathcal{P}$-essential edge, then G' has a $\mathcal{P}$-orientation G'' [1]. Thus, G'' is an orientation of G that satisfies all the pairs in $\mathcal{P}$ and creates $|E_{ess}|$ doubly oriented edges, which implies that $|E_{ess}| \geq |E_{min}|$. Since we have already shown that $E_{ess} \subseteq E_{min}$, we conclude that $E_{min} = E_{ess}$.

Now, to solve the MIN-DB-GO problem when $|\mathcal{P}| = 2$, we apply the following process.

1. Compute the $\mathcal{P}$-essential edges of G using the polynomial-time algorithm presented in [1];
2. Construct a mixed graph G' by replacing each $\mathcal{P}$-essential edge (u, v) by two arcs uv and vu;
3. Apply the polynomial-time algorithm presented in [1] to compute a $\mathcal{P}$-orientation of G'.

□

As in the S-GO problem, the MIN-DB-GO problem is polynomial-time solvable when both parameters $|\mathcal{P}|$ and $|B|$ are bounded.

THEOREM 5. *The* MIN-DB-GO *problem is polynomial-time solvable when G is a MAG, $|\mathcal{P}| = \mathcal{O}(1)$ and $B = \mathcal{O}(1)$.*

5.2 Difficult cases

Let DB-GO denote the decision version of the MIN-DB-GO problem. We show in the next result that the DB-GO problem is NP-complete even when $\Delta(G^*) = 3$.

THEOREM 6. *The problem* DB-GO *is* NP*-complete when G is a MAG and $|\mathcal{P}|$ is unbounded even when $\Delta(G^*) = 3$ and $B = 3$.*

In the remaining of this section, we study the approximability and the parameterized complexity of the MIN-DB-GO problem.

THEOREM 7. *Unless* P = NP*, the* MIN-DB-GO *problem is non-approximable when the graph G is a MAG and $|\mathcal{P}|$ is unbounded even when $\Delta(G^*) = 3$ and $B = 3$.*

The MIN-DB-GO problem is also APX-hard and W[1]-hard when $|\mathcal{P}|$ and B are unbounded even when $\Delta(G^*) = 3$.

THEOREM 8. *The* MIN-DB-GO *problem is* APX*-hard and* W[1]*-hard (parametrized by the number of doubly oriented edges) when G is a MAG and $|\mathcal{P}|$ and B are unbounded, even when $\Delta(G^*) = 3$.*

Proof. We propose an L-reduction from the APX-hard problem MINIMUM SET COVER [12,13]: given a ground set $\mathcal{X} = \{X_1, \ldots, X_n\}$, and a collection of sets $\mathcal{C} = \{S_1, \ldots, S_m\}$ s.t. $S_i \in 2^{\mathcal{X}}$, for all $1 \leq i \leq m$, the goal is to find a minimum set cover $\mathcal{C}'$, i.e., a set $\mathcal{C}' \subseteq \mathcal{C}$ s.t. $\mathcal{C} = \bigcup_{S_i \in \mathcal{C}'} S_i$ and $|\mathcal{C}'|$ is minimum.

We note by α_j, $1 \leq j \leq n$, the number of the sets containing X_j. Let us construct an instance $(G, \mathcal{P})$ of the MIN-DB-GO problem. For each $X_j \in \mathcal{X}$, we create the vertex set $\{x_j\} \cup \{x_j^k, 1 \leq k \leq \alpha_j\}$, then we create the directed path $x_j^1 x_j^2 \ldots x_j^{\alpha_j} x_j$. For each S_i, we add the vertex set $\{s_i, s'_i\} \cup \{s_i^j, 1 \leq j \leq |S_i|\}$, and we add an edge (s'_i, s_i) and a directed path $s_i s_i^1 s_i^2 \ldots s_i^{|S_i|}$.

Let $X_l \in \mathcal{X}$ be the $j-$th element in a set S_i, for each j, $1 \leq j \leq |S_i|$, we add an arc from s_i^j towards *only one* vertex from the set $\{x_l^k, 1 \leq k \leq \alpha_l\}$. Such a vertex is chosen in such a way that the indegree of each vertex x_l^k is at most two, for all $1 \leq k \leq \alpha_l$. To finish the construction of G, we add a vertex r_1 connected, by two edges, to the vertices s'_1 and s'_2. Then, we add a new vertex r_2 connected, by two edges, to the vertices r_1 and v'_3. We continue the creation of vertices r_i connected, by edges, to r_{i-1} and s'_{i+1}, for all $3 \leq i \leq m-1$. The set of pairs to satisfy is $\mathcal{P} = \{(r_{m-1}, x_j), 1 \leq j \leq n\} \cup \{(s_i, s'_i), 1 \leq i \leq m\}$ with $m = |\mathcal{C}|$. An example of construction is illustrated in Fig. 3. The degree of each vertex r_i in G is at most three and also each vertex s_i^j is connected to exactly one vertex $x_{i'}^{j'}$. Thus, one can easily check that $\Delta(G^*) = 3$.

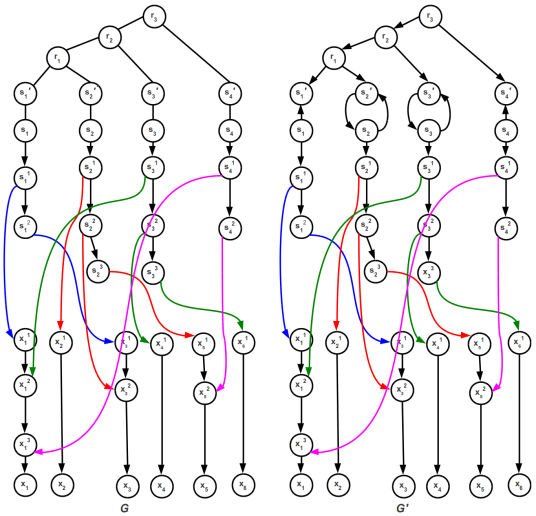


Figure 3. (a) Construction of an instance $(G, \mathcal{P})$ of the S-GO problem from an instance of MINIMUM SET COVER problem. Here, $\mathcal{X} = \{X_1, X_2, X_3, X_4, X_5, X_6\}$ and $\mathcal{C} = \{S_1, S_2, S_3, S_4\}$ s.t. $S_1 = \{X_1, X_3\}$, $S_2 = \{X_2, X_3, X_5\}$, $S_3 = \{X_1, X_4, X_6\}$ and $S_4 = \{X_1, X_5\}$. The set of pairs is $\mathcal{P} = \{(s_1, s'_1), (s_2, s'_2), (s_3, s'_3), (s_4, s'_4), (r_3, x_1), (r_3, x_2), (r_3, x_3), (r_3, x_4), (r_3, x_5), (r_3, x_6)\}$. (b) The graph G' is an orientation of G obtained from the set cover $\mathcal{C}' = \{S_2, S_3\}$ and satisfying all the pairs in $\mathcal{P}$.

We claim that, for every integer $k \geq 0$, there is a set cover of $\mathcal{C}$ of cardinality k if and only if there is a $(\mathcal{P}, k)$-DB-orientation of G.

$\Rightarrow$: Given a set cover $\{S_{i_1}, S_{i_2}, \ldots, S_{i_k}\}$, we doubly orient the edge (s_{i_j}, s'_{i_j}), for all $1 \leq j \leq k$. Then, we replace each edge (v_i, v'_i) by the arc $s_i s'_i$, for all $i \in \{1, 2, \ldots, m\} \backslash \{i_1, i_2, \ldots, i_k\}$. Finally, we orient the tree induced by the vertex set $\{r_i, 1 \leq i < m\} \cup \{s'_i, 1 \leq i \leq m\}$, to create a directed tree of root r_{m-1}.

$\Leftarrow$: Let E_{DB} be the set of doubly oriented edges in a $(\mathcal{P}, k)$-DB-orientation. Let $\mathcal{C}' = \{S_i : (s_i, s'_i) \in E_{DB}\}$. We will show that the set $\mathcal{C}'$ is a set cover of $\mathcal{C}$. Suppose that there is $X_j \in \mathcal{X}$ s.t. $X_j \notin S_i$ for all $S_i \in \mathcal{C}'$. Let $\mathcal{C}_j$ denote the collection of the sets containing X_j, i.e., $\mathcal{C}_j = \{S \in \mathcal{C} : X_j \in S\}$. The graph G is constructed in such a way that, to satisfy any pair (r_{m-1}, x_j), we must replace an edge (s'_i, s_i) by the arc $s'_i s_i$ s.t. $S_i \in \mathcal{C}_j$. On the other hand, we have to orient each edge (s'_i, s_i), from s_i towards s'_i, to satisfy the pair $(s_i, s'_i) \in \mathcal{P}$. Then the edge (s'_i, s_i) must be doubly oriented, which implies that $S_i \in \mathcal{C}'$. This is a contradiction, because $\mathcal{C}' \cap \mathcal{C}_j = \emptyset$.

The above reduction is an L-reduction that preserves the parameter k (the cardinality of the set cover and the number of doubly oriented edges). Since the problem MINIMUM SET COVER is APX-hard and W[1]-hard when parametrized by k [12,13], then the MIN-DB-GO problem is also APX-hard and W[1]-hard when parametrized by the number of doubly oriented edges. □

Now let us show that, unlike the S-GO problem, the MIN-DB-GO problem remains difficult even when $B = 2$.

THEOREM 9. *The problem* MIN-DB-GO *is* APX*-hard when G is a MAG and $|\mathcal{P}|$ is unbounded, even when $B \in \{2, 3\}$ and $\Delta(G^*) = 3$.*

Proof. Again, we use the above reduction (proof of Theorem 8), but we consider the variant MINIMUM SET COVER-K of the MINIMUM SET COVER problem in which each $X_j \in \mathcal{X}$ appears in exactly K sets in $\mathcal{C}$ s.t. K is a constant ≥ 2. For each pair of vertices $(s_i, s'_i) \in \mathcal{P}$ there is a unique path in G, from s_i to s'_i (that

is the edge (s_i, s'_i)), for all $1 \leq i \leq m$. In addition, the fact that each $X_j \in \mathcal{X}$ appears in exactly K sets in $\mathcal{C}$, implies that for each pair (r_{m-1}, x_i) there are, in G, K paths from r_{m-1} to s_i, for all $1 \leq i \leq n$. Thus $B = K$ and also the graph G is constructed so that we have $\Delta(G^*) = 3$.

As the problem MINIMUM SET COVER-K is APX-hard [11], then we conclude that MIN-DB-GO is APX-hard when G is a MAG s.t. $\Delta(G^*) = 3$ and $|\mathcal{P}|$ is unbounded, even when, B is a constant ≥ 2 (and thus when $B \in \{2, 3\}$) . □

6 Conclusion

In this work, we considered two problems that are concerned with the orientation of mixed graphs, both motivated, among others, by biological applications. We studied the complexity of both problems, and in particular we provided polynomial-time algorithms for some restricted instances, and several hardness and inapproximability results. However, some interesting problems remain open so far, such as the following ones:

- Explore the possibility of obtaining fixed-parameterized tractable (FPT) algorithms for MIN-DB-GO.
- Study the approximability of MIN-DB-GO on specific graph classes.
- Study the complexity of S-GO and MIN-DB-GO when $|\mathcal{P}| \geq 3$ is a constant.

References

[1] E. M. Arkin and R. Hassin. A note on orientations of mixed graphs. *Discrete Applied Mathematics*, 116:271–278, 2002.

[2] B. Aspvall, M. F. Plass, and R. E. Tarjan. A linear-time algorithm for testing the truth of certain quantified boolean formulas. *Inf. Process. Lett.*, 8:121–123, 1979.

[3] B. Dorn, F. Huffner, D. Kruger, R. Niedermeier, and J. Uhlmann. Exploiting bounded signal flow for graph orientation based on cause–effect pairs. *Algorithms for Molecular Biology*, 6:21, 2011.

[4] M. Elberfeld, D. Segev, C. R. Davidson, D. Silverbush, and R. Sharan. Approximation algorithms for orienting mixed graphs. *In Proc. CPM*, 6661:416–428, 2011.

[5] S. Fields and O. Song. A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230):245–246, July 1989.

[6] I. Gamzu, D. Segev, and R. Sharan. Improved orientations of physical networks. *In Proc. WABI*, 6293:215–225, 2010.

[7] R. Hassin and N. Megiddo. On orientations and shortest paths. *Linear Algebra and its Applications*, 114:589–602, 1989.

[8] T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J.-B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford, and R. A. Young. Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, 298(5594):799–804, Oct. 2002.

[9] U. Lucia. Probability, ergodicity, irreversibility and dynamical systems. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 464:1089–1104, 2008.

[10] A. Medvedovsky, V. Bafna, U. Zwick, and R. Sharan. An algorithm for orienting graphs based on cause-effect pairs and its applications to orienting protein networks. *In Proc. WABI*, 5251:222–232, 2008.

[11] C. Papadimitriou and M. Yannakakis. Optimization, approximation, and complexity classes. *In Proc. STOC*, 43:229–234, 1988.

[12] A. Paz and S. Moran. Non deterministic polynomial optimization problems and their approximations. *Theor. Comput. Sci.*, 15:251–277, 1981.

[13] R. Raz and S. Safra. A sub-constant error-probability low-degree test, and a sub-constant error-probability PCP characterization of NP. *In Proc. STOC*, pages 475–484, 1997.

[14] D. Silverbush, M. Elberfeld, and R. Sharan. Optimally orienting physical networks. *Journal of Computational Biology*, 18:1437–1448, 2011.

[15] C. A. Tovey. A simplified NP-complete satisfiability problem. *Discrete Applied Mathematics*, 8:85–89, 1984.

[16] C.-H. Yeang, T. Ideker, and T. Jaakkola. Physical network models. *Journal of Computational Biology*, 11:243–262, 2004.

Session 9_B : Sequence Analysis

Fast Homology Search Using Domain-Architecture Alignment

Nicolas TERRAPON[1], Sonja GRATH[1], January WEINER[2], Andrew D. MOORE[1] and Erich BORNBERG-BAUER[1]

[1] Westfalian Wilhelms University, Institute of Evolution and Biodiversity, Huefferstr. 1, 48149 Muenster, Germany
{n.terrapon, s.grath, radmoore, ebb}@uni-muenster.de
[2] Max Planck Institute for Infection Biology, Charitéplatz 1, 10117 Berlin, Germany
january.weiner@mpiib-berlin.mpg.de

Abstract. Homology detection, i.e. the identification of genes that share common ancestry, is key to understanding the evolutionary history of gene families. However, homology detection using local sequence similarity alone (such as employed by BLAST) can be problematic when applied to multi-domain proteins, as true homology can be masked by local similarity of domain sequences. Here, we present a new approach for detecting homology that is based on the global alignment of domain arrangements. It combines the ability of global alignments to capture homology across the entire sequence with the efficiency and sensitivity of local alignments.
A web-based interface is available under [illegible]

1 Introduction

Assessing homology between sequences has become a cornerstone of modern biology. In general, homology detection involves pairwise alignments between the DNA or protein sequence of a gene of interest (query) and a large collection of other sequences (database). Methods to align two sequences to each other are based on either global or local alignment strategies. For a global alignment, full sequences are aligned, whereas for a local alignment only the best matching fragments of the sequences are taken into account. In practice, heuristics of local alignment such as BLAST [1] are frequently used to determine sequence homology across large datasets. Local alignment strategies allow for detection of structural or functional similarity (e.g. a protein domain, the units of protein structure, function and evolution) even if the context within the two proteins is different. However, homology detection due to local sequence similarity alone can bear drawbacks. On the one hand, distant homologs can be missed by local approaches if the sequences have strongly diverged. On the other hand, strong similarity in only parts of the sequences is not necessarily an indicator of homology or functional similarity as different parts of proteins may have different evolutionary histories, in particular in multi-domain proteins [2].

Domains form the structural, functional and evolutionary units of the underlying protein sequences. A protein can be further described by the ordered sequence of the domains it consists of, the so-called domain arrangement or domain architecture. Domain architecture in comparison with pure domain composition has been found to provide additional insights into functional and structural constraints within proteins [3].

Hence, the problems described above can be addressed by taking the domain composition or architecture of the considered proteins into account. Different domain-based approaches for homology detection have been developed over the last years [4, 5, 6] and it could be shown that several of these algorithms perform very well on curated data sets when compared to complete sequence based approaches.

Here, we present a new method that is based on global alignments of domain arrangements and thus combines the ability of global protein alignments to capture the homology information of the complete sequence with the efficiency of local alignments. All domain-based homology approaches are able to detect homology only for sequences sharing at least one domain with the query [7]. We show that the error caused by this characteristic trait of domain-based methods for homology search is negligible. Further, we illustrate that our approach reasonably accomplishes results achieved by sequence based techniques while providing additional relevant results. This work, still in progress, can therefore become an important tool for future analysis of protein function and evolution.

Fast estimation of posterior probabilities in change-point models through a constrained hidden Markov model

The Minh LUONG[1], Yves ROZENHOLC[1] and GREGORY NUEL[1]

MAP 5, Université Paris Descartes, 45 rue des Saints Pères, 75006 Paris, France
the.minh.luong@parisdescartes.fr

Abstract *The detection of change-points in heterogeneous sequences is a statistical challenge with applications across a wide variety of fields. In bioinformatics, a vast amount of methodology and numerous efficient algorithms have been developed to identify an ideal set of change-points for detecting Copy Number Variation (CNV). However, relatively few approaches consider the important problem of assessing the uncertainty of the change-point location. Having quadratic complexity, these approaches typically are intractable for large datasets of tens of thousands points or more. In this paper, we assess uncertainty through a constrained hidden Markov model with a fixed number of segments of contiguous observations with the same distribution, separated by each change-point. Forward-backward algorithms from this model estimate posterior probabilities of interest with linear complexity. The methods are implemented in the R package postCP, which uses the results of a given change-point detection algorithm to estimate the probability that each observation is a change-point. Due to its frequentist framework, postCP obtains less conservative confidence intervals than previously published Bayesian methods. On another data set of high-resolution data* ($n = 14,241$)*, the implementation processed high-resolution data in less than one second on a mid-range laptop computer.*

Keywords Change-point estimation, Posterior distribution of change-points, Segmentation, Constrained hidden Markov model, Forward-backward algorithm, Fast computation.

1 Introduction

The detection of *change-points* in heterogeneous sequences is a statistical challenge with many applications in fields such as finance, reliability, signal analysis, neurosciences and biology [22]. In bioinformatics in particular, a vast amount of methodology [16,7,11] has been developed for identifying an ideal set of change-points in data from array Comparative Genomic Hybridization (aCGH) techniques, in order to identify Copy Number Variation (CNV).

Given the numerous efficient algorithms currently available for finding the best segmentation of the data, this paper focuses on characterizing the uncertainty of the estimated change-point locations using the classical segmentation model. Moreover, this paper addresses the problem of complexity for large datasets of tens of thousands points or more. Change-point estimation is increasingly challenging from a computational point of view given the need to handle large amounts of data from emerging high-throughput technologies.

A typical expression of the change-point problem is as follows: given a dataset $X = (X_1, X_2, \ldots, X_n)$ of real-valued observations, and $K \geqslant 2$ number of segments, find the best partitioning $S \in \mathscr{M}_K$ of the data into K non-overlapping intervals, assuming that the distribution is homogeneous within each of these intervals. For K segments of contiguous observations, the *segment-based model* expresses the distribution of X given a segmentation $S \in \mathscr{M}_K$ as:

$$\mathbb{P}(X|S;\theta) = \prod_{i=1}^{n} g_{\theta_{S_i}}(X_i) = \prod_{k=1}^{K} \prod_{i,S_i=k} g_{\theta_k}(X_i) \tag{1}$$

where $g_{\theta_k}(\cdot)$ is the parametric distribution (typically: Poisson or Gaussian) with parameter θ_k, $\theta = (\theta_1, \ldots, \theta_K)$ is the global parameter, and S_i is the segment index at position i. For example, if $n = 5$, $K = 2$, and if the change-point is located between positions 2 and 3, then $S = (1,1,2,2,2)$.

Define $\mathscr{M}_K$ as the set of all possible combinations of hidden states $S_1,\ldots,S_n$ for fixed K number of segments. Introducing a prior distribution $\mathbb{P}(S)$ on any $S \in \mathscr{M}_K$ obtains a posterior distribution of the segmentation:

$$\mathbb{P}(S|X;\theta) = \frac{\mathbb{P}(X|S;\theta)\mathbb{P}(S)}{\sum_R \mathbb{P}(X|R;\theta)\mathbb{P}(R)}. \tag{2}$$

For a uniform prior, set $\frac{1}{\mathbb{P}(S)} = \binom{n-1}{K-1} = |\mathscr{M}_K|$.

A common alternative to the above segmentation procedure is to consider an unsupervised hidden Markov Model (HMM). Assuming that S is a Markov chain, this approach [19] can be thought of as being *level-based*, as the parameter of the k^{th} segment takes its value in the finite set of $L \geqslant 1$ levels [1]: $\{\theta_1,\theta_2,\ldots,\theta_L\}$. This simply is equivalent to the model defined by Equation (1), with the noticeable difference that $S \in \{1,2,\ldots,L\}^n$. With this level-based approach $K \geqslant L$ in general, and the HMM is unconstrained in the sense that transitions are possible between any pair of states. This is an appropriate model when the conditional distribution within contiguous observations can be shared among other segments.

A convenient feature of these HMM approaches is that the posterior distribution $\mathbb{P}(S|X;\theta)$ can be computed efficiently in $O(L^2 n)$ using classical forward-backward recursions [6], making them suitable for handling large datasets. One such algorithm uses a discrete HMM [7] to map the number of states and the most likely state at each position, with Various extensions to this HMM approach have been proposed [26,14] to improve the results. HMM-based implementations for dealing with higher-resolution data have also been developed for current array technologies [3,25]. A wide amount of segment-based approaches for identifying CNV in DNA data have also been explored, which typically find contiguous segments differing according to a pre-specified threshold such as a p-value. They include a non-parametric approach that uses permutation [16,24], a likelihood-based approach to estimate parameters in a Gaussian model after adaptive weights smoothing [11] and a wavelet-based change-point detection procedure with data smoothing [10]. To estimate the number of segments, K, methods include a modified Bayes Information Criterion to adjust for the number of change-points [27], a least squares minimization procedure with a penalization criterion for the number of contrasts [4], and an adaptive method for estimating the location and number of change-points, with penalization terms for likelihoods [17].

In this paper, we exploit the effectiveness of the level-based HMM approach through a constrained HMM corresponding *exactly* to the above segment-based model, providing a fast algorithm for computing $\mathbb{P}(S|X;\theta)$. While the unconstrained HMM is preferable in many practical situations, the segment-based model requires less assumptions and is thus a more general model.

While most of the current procedures are designed to detect the ideal number or set of change-point locations, the problem of characterizing the uncertainty in these locations has been considered in relatively few articles. Previous estimates have generally focused on asymptotic behaviour whose conditions are delicate due to the discreteness of the problem [1,15], on bootstrap techniques [12] and on stochastic methods [13]. For finding the exact posterior distribution of change-points, algorithms were introduced to compute $\mathbb{P}(X|S;\theta)$ in theoretical $O(Kn^2)$ [8,20] in a Bayesian framework. However, the complexity of these approaches provides for very slow processing in current implementations for large datasets with sequences of tens of thousands or more.

2 Methods

2.1 Constrained HMM

Let us assume that S is a heterogeneous Markov chain over $\{1,2,\ldots,K,K+1\}$ (where $K+1$ is a "junk" state only considered for consistency reasons) such as $\mathbb{P}(S_1 = 1) = 1$ and with the following transitions: for all $2 \leqslant i \leqslant n$, and $1 \leqslant k \leqslant K$ we have $\mathbb{P}(S_i = k|S_{i-1} = k) = 1 - \eta_k(i)$ and $\mathbb{P}(S_i = k+1|S_{i-1} = k) = \eta_k(i)$. For consistency, choose $\mathbb{P}(S_i = K+1|S_{i-1} = K+1) = 1$. For example, if $n = 5$, $K = 2$, and $S = (1,1,2,2,2)$ then $\mathbb{P}(S) = (1-\eta_1(2))\eta_2(3)(1-\eta_2(4))(1-\eta_2(5))$. With this Markov chain, it is clear that $\{S \in \mathscr{M}_K\} = \{S_n = K\}$.

1. Similar to the segment-based model, the choice of L is critical and is usually addressed through penalized criteria.

In the particular case where the Markov chain is homogeneous with $\eta_k(i) = \eta \in]0,1[$ for all k and i, $\mathbb{P}(S) = (1-\eta)^{n-K}\eta^{K-1}$ for all $S \in \mathscr{M}_K$ [2]. In other words, only positive state jumps of $+1$ are possible. Therefore $\mathbb{P}(S|S \in \mathscr{M}_K) = 1/|\mathscr{M}_K|$, which corresponds to the canonical choice of a uniform prior on $\mathscr{M}_K$. Note that a choice of different transition coefficients $\eta_k(i)$ allows us to specify informative priors.

2.2 Forward-backward procedure and posterior probabilities

The constrained HMM provides a framework for additional inference on the uncertainty in the estimated change-point model; in particular, after obtaining the segmentation from any previous procedure, we can obtain confidence intervals around each of the identified change-points. In terms of practical applications, this approach is helpful when dealing with situations where both very short and very long segments may be present, and the exact location of change-points may not be identifiable.

The forward-backward algorithm [19], also known as posterior encoding, is a recursive algorithm that can be used to estimate the posterior probabilities of each observation i being in a particular hidden state S_i, and being a change-point such that $S_i = S_{i-1} + 1$. The algorithm is of complexity $O(Kn)$ for K segments and n observations; the $O(K^2 n)$ complexity of the classical forward-backward algorithm is reduced due to the sparse transition matrix between states.

We define the forward and backward quantities as follows, for observation i and state k:

For $1 \leqslant i \leqslant n-1$:

$$F_i(k) = \mathbb{P}(X_{1:i} = x_{1:i}, S_i = k) \tag{3}$$
$$B_i(k) = \mathbb{P}(X_{i+1:n} = x_{i+1:n}, S_n = K | S_i = k) \tag{4}$$

We obtain the forward quantities by recursion through the following formulae:

Forward:

$$F_1(k) = \begin{cases} g_{\theta_1}(x_1) & \text{if } k = 1 \\ 0 & \text{else} \end{cases} \tag{5}$$
$$\begin{aligned} F_i(k) &= \sum_\ell F_{i-1}(\ell)\mathbb{P}(S_i = k | S_{i-1} = \ell, S \in \mathscr{M}_K) g_{\theta_k}(x_i) \\ &= [F_{i-1}(k)(1-\eta_k(i)) + \mathbf{1}_{k>1} F_{i-1}(k-1)\eta_k(i)] g_{\theta_k}(x_i) \end{aligned} \tag{6}$$

where $g_{\theta_k}(x_i)$ is the density function of the chosen emission distribution G with parameter θ_k, when k is the underlying segment for observation i.

We use a similar recursive procedure to obtain the backward quantities:

Backward:

$$B_{n-1}(k) = \begin{cases} \eta_K(x_n) g_{\theta_k}(x_n) & \text{if } k = K-1 \\ (1-\eta_K(x_n)) g_{\theta_k}(x_n) & \text{if } k = K \\ 0 & \text{else} \end{cases} \tag{7}$$
$$\begin{aligned} B_{i-1}(k) &= \sum_\ell \mathbb{P}(S_i = \ell | S_{i-1} = k, S \in \mathscr{M}_K) g_{\theta_\ell}(x_i) B_i(\ell) \\ &= (1-\eta_k(i)) g_{\theta_k}(x_i) B_i(k) + \mathbf{1}_{k<K} \eta_{k+1}(i) g_{\theta_{k+1}}(x_i) B_i(k+1) \end{aligned} \tag{8}$$

To obtain the posterior probabilities of the state $S_i = k$ at position i, we note that :

$$\mathbb{P}(X_{1:n} = x_{1:n}, S \in \mathscr{M}_K) = F_1(1) B_1(1) \tag{9}$$
$$\mathbb{P}(S_i = k | X_{1:n} = x_{1:n}, S \in \mathscr{M}_K) = \frac{F_i(k) B_i(k)}{F_1(1) B_1(1)}. \tag{10}$$

2. Note that the particular value of η does affect $\mathbb{P}(S)$ but has no effect whatsoever on $\mathbb{P}(S|S \in \mathscr{M}_K)$ of the chosen segmentation S. We can therefore safely make an arbitrary choice like $\eta = 0.5$ for practical computations.

The constrained HMM estimates the probability of changing state while being at state $S_i = k$ at observation i as:

$$\mathbb{P}(S_i = k | X_{1:n} = x_{1:n}, S_{i-1} = k-1, S \in \mathscr{M}_K) = \frac{\eta_{k-1}(i-1) g_{\theta_k}(x_i) B_i(k)}{B_{i-1}(k-1)}. \tag{11}$$

The posterior probability of the k^{th} change-point occurring after observation i, or in other words $i+1$ being the first observation in the $k+1^{th}$ segment, is:

$$\begin{aligned} \mathbb{P}(CP_k = i | X_{1:n} = x_{1:n}, S \in \mathscr{M}_K) &= \mathbb{P}(S_i = k, S_{i+1} = k+1 | X_{1:n} = x_{1:n}, S \in \mathscr{M}_K) \\ &= \frac{F_i(k) \eta_k(i) g_{\theta_{k+1}}(x_{k+1}) B_{i+1}(k+1)}{F_1(1) B_1(1)} \end{aligned} \tag{12}$$

2.3 Statistics package postCP

We apply the preceding methods in the statistics package postCP, available on the CRAN website http://cran.r-project.org/web/packages/postCP.
The forward and backward recursive algorithms are programmed in C++ to optimize the speed of the computatior intensive process.

The following is a typical *R* command line for segmenting a sequence *LRR*, with a vector of length $K-1$ change-points (or last index of segments) *initseg*, and 95% confidence intervals. The options also save forward-backward and posterior change-point probabilities in the output (using `keep=TRUE` option) in addition to the set of change-points with the highest posterior probability (by Viterbi algorithm), and generate 100 different change-point vectors (each of length $K-1$):

```
postCP(data=LRR,seg=initseg,ci=0.95,keep=TRUE,nsamples=100)
```

3 Detecting copy number variation in array CGH breast cancer data

We apply the methods to a widely referenced data set from a breast cancer cell line BT474 [22]. The data consist of log-reference ratios signifying the ratio of genomic copies of test samples compared to normal. The goal is to segment the data into segments with similar copy numbers, with change-points specifying a copy number variation that may point to genetic mutations of interest [18]. We compare the results of postCP to the Bayesian confidence intervals previously published on the same data [20], consisting of 120 observations from Chromosome 10. The observations are sorted according to their relative position along Chromosome 10.

We use a modification of the greedy K-means algorithm [9] to obtain an initial segmentation for 3 and 4 segments (Table 1). The locations refer to the indices of the last observations of the first $K-1$ segments. Afterwards, we used postCP to obtain estimates of the forward and backward matrices in Section 2.2, and afterwards, estimates of the posterior probabilities of the change-points being at each observation. We assumed a homoscedastic Gaussian model for the observations.

The estimated posterior change-point probabilities of the aCGH data are displayed in Fig. 1 for three and four segments. For both segmentations, the probability of the last change-point had a sharp peak close to 1.0. There is an additional change-point found at position 80 by the four segment model with a relatively high uncertainty. In both the 3 and 4 segment instances, the distributions of the change-points are not regular due to the discreteness of the segmentation procedure.

At these predefined observations, the corresponding posterior change-point probabilities were higher than reported by Rigaill, and confidence intervals were slightly narrower. This is expected, as Rigaill's method uses a Bayesian framework that accounts for the uncertainty in the parameter estimates. In particular, the rightmost change-point estimation for both segmentations of 3 and 4 had posterior probabilities above 0.95, while the corresponding posterior probabilities from the Bayesian method were closer to 0.5.

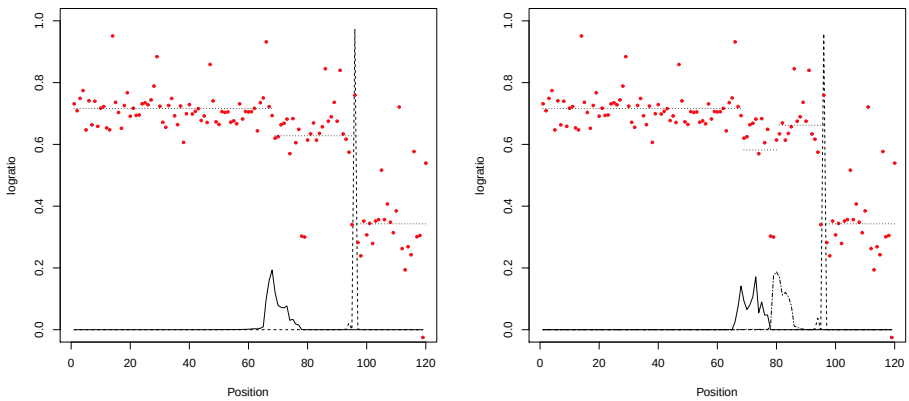


Figure 1. Plots of estimated posterior change-point probabilities for Chromosome 10 data [22]. LRR data and horizontal lines representing the estimated means within segments are scaled to the probability axes. (left) Posterior probabilities for three segments, first change-point initialized at $i = 60$ is a solid line, 2nd initialized at $i = 96$ is a dashed line. (right) Posterior probabilities for four segments, additional change-point after $i = 80$ is a dotted-dashed line.

Table 1. Change-point confidence intervals aCGH Chromosome 10

CP #	Δ mean	Est location	95 % CI (postCP)	95% CI (Bayesian)
		Three segments		
1	-0.22	68	66-76	64-78
2	-0.71	96	96-96	92-97
		Four segments		
1	-0.34	68	66-76	66-78
2	-0.20	80	79-85	78-97
3	-0.80	96	96-96	91-112

Change-point estimates of aCGH data for Chromosome 10 [22], with (95% confidence interval), by postCP and Bayesian confidence intervals [20]. Narrower confidence intervals were found by postCP.

4 Detecting copy number variation in SNP array colorectal cancer data

To illustrate the efficiency of the constrained HMM model, we apply the methods to a copy number variation profile obtained through current SNP (single nucleotide polymorphism) microarray technology [23]. The data contain 261,563 SNPs across the genome obtained from an Affymetrix chip from a colorectal cancer tumour cell line with intensities providing information regarding mutations, specifically duplications and deletions [5]. After normalization, the log of the intensities in the tumour sample were compared to normal values, obtained from a reference sample, to obtain the log-reference ratios (LRR) across each of 23 chromosomes.

We use the circular binary segmentation (CBS) procedure implemented in a widely used segmentation package, DNAcopy [16,24] for the statistics software *R*. This package obtains an initial estimate of change-points in LRR, for each of the 23 chromosomes in tumour sample 103. We ran DNAcopy, after smoothing, at the default parameters of $\alpha = 0.01$.

We ran *postCP* using the set of change-points identified by DNAcopy as initial segmentation to obtain estimates of the posterior probabilities of the change-points being at these ten locations. We assumed a homoscedastic Gaussian model for the observations. The forward-backward algorithm was practically instantaneous (less than 0.1 seconds) for this sequence of over 14,000 observations on a mid-range dual-core 2.5

GHz, 4GB RAM laptop PC. Not surprisingly, the most narrow confidence intervals, and most accurate change-point estimates, were found for larger differences (Table 2). Of the ten estimated change-points, six separated segments whose means differed by greater than one standard deviation; these points all were found to have posterior change-point probabilities greater than 0.5. Eight of the ten change-points from DNAcopy had the highest posterior change-point probabilities for their positions, with the exception of the fifth and tenth change-points whose probabilities were slightly lower than the respective maximum of their position.

Table 2.

Change	CP	Post	Δ	Size of confidence interval				
point	Est	Prob	Mean	ci:0.5	0.6	0.7	0.8	0.9
1	211	0.973	-0.582	1	1	1	1	1
2	215	0.918	0.523	1	1	1	1	1
3	273	0.556	-0.293	1	2	2	2	3
4	383	0.580	0.381	1	2	2	2	3
5	736	0.028[a]	-0.081	31	37	46	53	61
6	3091	0.860	-0.456	1	1	1	1	2
7	3102	0.880	0.466	1	1	1	1	2
8	8308	0.050	0.075	21	30	50	91	132
9	8761	0.064	-0.075	15	21	48	63	78
10	12383	0.006[b]	0.042	233	336	412	501	522

Information for 10 change-points identified within Chromosome 10. [a]: point 721 had slightly higher change-point probability (0.031),[b]: point 11943 had slightly higher change-point probability (0.008).

Fig. 2 (left) displays the posterior change-point probabilities for the first five change-points. Given the much narrower segment lengths, and greater segment differences, the probabilities are much higher for the first four change-points than the fifth change-point, whose 95% confidence interval was greater than 60 SNPs wide. Fig. 2 (right) displays the posterior change-point probabilities for the tenth change-points, whose 95% confidence interval was greater than 500 SNPs wide. Note that the shapes of the distributions of the change-points are also highly irregular.

5 Discussion

A common point of interest in current genomics studies is to find genetic mutations pointing to phenotypes susceptible to diseases such as cancers or Type II diabetes. The detection of change points in copy number variation (CNV) is a critical step in the characterization of DNA, including tumoral DNA in cancer. A CNV may locate a genetic mutation such as a duplication or deletion in a cancerous cell that is a target for treatment. However, due to the often large nature of DNA data sequences it is often difficult to reliably identify CNVs. The latest technologies for detecting CNV, such as single nucleotide polymorphism (SNP) arrays, are able to produce sequences of tens or hundreds of thousands of observations.

This paper describes a procedure that extracts further useful information from segmenting large sequences such as those found in CNV estimates from SNP array data. The unsupervised HMM, which is a level-based approach used in many current implementations, includes forward-backward calculations in linear time. However they are less flexible than segment-based methods, in part because they require prior transition probabilities between states that assume a geometric prior on segment lengths, which may be inappropriate in some practical situations. On the other hand, most segment-based methods can allow for a wider variety of specified priors, but in many situations are unfeasible due to their quadratic complexity for large data sets such as those generated by SNP arrays.

The described constrained HMM model's worst case complexity of $O(Kn)$ enables it to handle very large data sets of tens of thousands of observations in reasonable time, unlike implementations with $O(Kn^2)$. The procedure allows the full distribution of change-points to be estimated conditional to the observations, allowing for practical estimation of confidence intervals. In high-resolution data sets where change-points are unlikely to be estimated at the exact correct position, the confidence intervals may yield important information. In

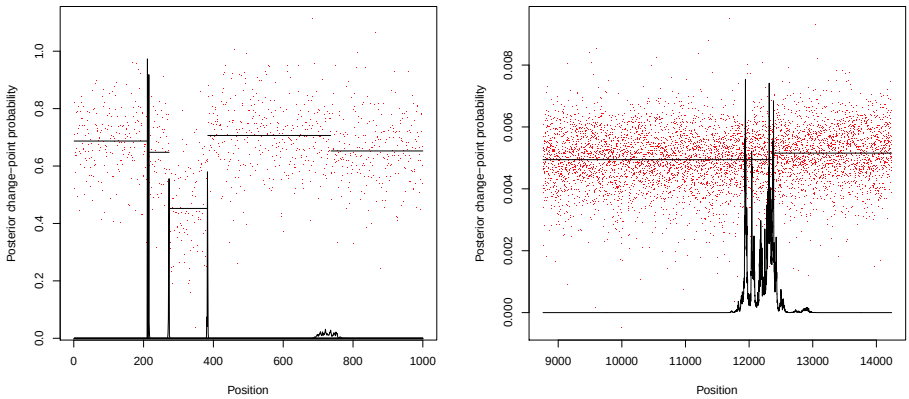


Figure 2. Plot of estimated posterior change-point probabilities. LRR data and horizontal lines representing the estimated means within 11 segments are scaled to the probability axes. (left) For first five change-points: The first four change-point estimates (SNPs at position: $211,215,273,383$) are precise, with a wide uncertainty around the fifth change-point (position: 736). (right) For tenth change-point: wide confidence interval. The change-point estimate from the CBS algorithm (position 12,383) is the fourth rightmost peak.

SNP array technology, the change-points need to be detected with high-precision and the differences between segments may not be very large. Additionally, overlapping confidence intervals for CNV across several different cell tumours or lines may identify associated copy number variations, and help in identifying similar disease phenotypes and treatments in patient subgroups.

The described methods are a useful tool in the segmentation of large-scale sequences such as those involving CNVs, specifically when combined with any current implementation designed for detecting an ideal set of change-points. The constrained HMM approach is sensitive to its starting point and tends to provide a local rather than global optimum. To this end, a fast dynamic programming algorithm of order $O(n\log n)$ [4] obtaining an initial segmentation of the points is to be bundled in the *postCP* package, or alternately a greedy approach with $O(Kn)$ for fixed K may be used. The posterior probabilities obtained by *postCP* may also be used to quickly estimate criteria for model selection, such as the Bayesian Information Criterion [21] and the Integrated Completed Likelihood [2].

Other planned refinements to the implementation include allowing the specification of informative prior transition probabilities $\eta_k(i)$, which can be estimated directly from an initial run of *postCP*, and an expectation-maximization algorithm to optimize the simultaneous estimation of parameters and posterior probabilities. Maximum *a posteriori* encoding is also possible through the Viterbi algorithm by modifying the forward quantities. Another approach is to combine the segment-based and level-based approaches by limiting possible values of segment parameters $(\theta_1,\ldots,\theta_K)$ to $L \leq K$ different values. Other applications include more complex models to simultaneously model multiple datasets, or accounting for values from multiple patients and samples as random effects in a mixed model. Another practical extension is to handle multi-dimensional output, such as current CNV technology [23] that simultaneously includes LRR and baseline allelic frequency (BAF) data.

References

[1] J. Bai and P. Perron. Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 18(1):1–22, 2003.

[2] C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(7):719–725, 2000.

[3] S. Colella, C. Yau, J.M. Taylor, G. Mirza, H. Butler, P. Clouston, A.S. Bassett, A. Seller, C.C. Holmes, and J. Ragoussis. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Research*, 35(6):2013, 2007.

[4] F. Comte and Y. Rozenholc. A new algorithm for fixed design regression and denoising. *Annals of the Institute of Statistical Mathematics*, 56(3):449–473, 2004.

[5] W. De Roock, B. Claes, D. Bernasconi, J. De Schutter, B. Biesmans, G. Fountzilas, K.T. Kalogeras, V. Kotoula, D. Papamichael, P. Laurent-Puig, et al. Effects of KRAS, BRAF, NRAS, and PIK3CA mutations on the efficacy of cetuximab plus chemotherapy in chemotherapy-refractory metastatic colorectal cancer: a retrospective consortium analysis. *The Lancet Oncology*, 11(8):753–762, 2010.

[6] R. Durbin. *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge Univ Pr, 1998.

[7] J. Fridlyand, A.M. Snijders, D. Pinkel, D.G. Albertson, and A.N. Jain. Hidden Markov models approach to the analysis of array CGH data. *Journal of Multivariate Analysis*, 90(1):132–153, 2004.

[8] Y. Guédon. Exploring the segmentation space for the assessment of multiple change-point models. Technical Report 6619, INRIA, 2008.

[9] J.A. Hartigan and M.A. Wong. Algorithm AS 136: A K-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[10] LI Hsu, S.G. Self, D. Grove, T. Randolph, K. Wang, J.J. Delrow, L. Loo, and P. Porter. Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics*, 6(2):211–226, 2005.

[11] P. Hupé, N. Stransky, J.P. Thiery, F. Radvanyi, and E. Barillot. Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, 20(18):3413–3422, 2004.

[12] M. Hušková and C. Kirch. Bootstrapping confidence intervals for the change-point of time series. *Journal of Time Series Analysis*, 29(6):947–972, 2008.

[13] T.L. Lai, H. Xing, and N. Zhang. Stochastic segmentation models for array-based comparative genomic hybridization data analysis. *Biostatistics*, 9(2):290–307, 2008.

[14] JC Marioni, NP Thorne, and S. Tavaré. BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. *Bioinformatics*, 22(9):1144–1146, 2006.

[15] V.M.R. Muggeo. Estimating regression models with unknown break-points. *Statistics in Medicine*, 22(19):3055–3071, 2003.

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

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

[18] D. Pinkel, R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W.L. Kuo, C. Chen, Y. Zhai, et al. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics*, 20:207–211, 1998.

[19] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, volume 77, pages 257–286. IEEE, 1989.

[20] G. Rigaill, E. Lebarbier, and S. Robin. Exact posterior distributions and model selection criteria for multiple change-point detection problems. *Statistics and Computing*, pages 1–13, 2011.

[21] G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[22] A.M. Snijders, N. Nowak, R. Segraves, S. Blackwood, N. Brown, J. Conroy, G. Hamilton, A.K. Hindle, B. Huey, K. Kimura, et al. Assembly of microarrays for genome-wide measurement of DNA copy number by CGH. *Nature Genetics*, 29:263–264, 2001.

[23] J. Staaf, D. Lindgren, J. Vallon-Christersson, A. Isaksson, H. Goransson, G. Juliusson, R. Rosenquist, M. Hoglund, A. Borg, and M. Ringnér. Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays. *Genome Biol*, 9(9):R136, 2008.

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

[25] K. Wang, M. Li, D. Hadley, R. Liu, J. Glessner, S.F.A. Grant, H. Hakonarson, and M. Bucan. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research*, 17(11):1665–1674, 2007.

[26] H. Willenbrock and J. Fridlyand. A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics*, 21(22):4084–4091, 2005.

[27] N.R. Zhang and D.O. Siegmund. A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, 63(1):22–32, 2007.

Detecting Outliers in HMM modeling through Relative Entropy with Applications to Change-Point Detection

Vittorio PERDUCA and Gregory NUEL

MAP5, UMR 8145 CNRS, 45 Rue des Saints-Pères, 75006, Paris, France
{vittorio.perduca, gregory.nuel}@parisdescartes.fr

Abstract *Hidden Markov models (HMMs) are a standard tool in many applications, including change-point (or segmentation) data analysis. Since HMMs are intrinsically heterogeneous, the detection of outliers in data modeled by HMMs is a challenging problem. This problem can be modeled by an* ad hoc *model which extends the HMM by explicitly taking into account variables for the outlier status of the observations. We suggest a novel and model free method based on relative entropy and show a dynamic programming algorithm to implement it in linear time. We validate the two methods on simulated data. We apply our method based on relative entropy on Copy Number Variation (CNV) data and show its effectiveness.*

Keywords Hidden Markov models, Bayesian networks, Kullback-Leibler distance, belief propagation, segmentation, cancer.

1 Hidden Markov Models

Hidden Markov models (HMMs) are a standard tool in many applications, including signal processing, speech recognition and biological sequence analysis [5]. A typical HMM is defined by the joint probability distribution

$$\mathbb{P}(X,S)=\mathbb{P}(S_1)\prod_{i=2}^{n}\mathbb{P}(S_i|S_{i-1})\prod_{i=1}^{n}\mathbb{P}(X_i|S_i)$$

where n is the total number of observations, $X=(X_1,\dots,X_n)$ is the vector of all the observation variables, $S=(S_1,\dots,S_n)$ is the vector of all the hidden variables. The dependencies among the variables are shown in Fig. 1.

For *homogeneous* HMMs we define $\mathbb{P}(S_i=s|S_{i-1}=r)=\alpha(r,s)$ and $\mathbb{P}(X_i=x|S_i=s)=\beta_s(x)$ for all $i=2,\dots,n$. Moreover $\mathbb{P}(S_1=s)=\gamma(s)$. The inference problem in HMMs is efficiently solved in linear time in n by the forward-backward algorithm [5]. We recall the recursive formula behind this algorithm in the standard case where the probability laws are conditioned on the evidence $\mathcal{E}=\{X=x\}=\{X_1=x_1,\dots,X_n=x_n\}$. In this particular case, the forward and backward quantities are defined as

$$F_i(s):=\mathbb{P}(S_i=s,X_{1:i}=x_{1:i}) \text{ for } i=2,\dots,n \text{ and } F_1(s)=\gamma(s)\beta_s(x_1) \tag{1}$$

and

$$B_i(s):=\mathbb{P}(X_{i+1:n}=x_{i+1:n}|S_i=s) \text{ for } i=1,\dots,n-1 \text{ and } B_n\equiv 1. \tag{2}$$

The two main results in this context are:

1. $\mathbb{P}(S_i=s,X=x)=F_i(s)B_i(s)$ for $i=1,\dots,n$
2. $\mathbb{P}(S_{i-1}=r,S_i=s,X=x)=F_{i-1}(r)\alpha_i(r,s)\beta_s(x_i)B_i(s)$ for $i=2,\dots,n$.

These two formulas allow deriving the forward/backward recursions and provide explicit EM update formulas for the maximum likelihood estimation (the EM algorithm for HMMs is also known as the Baum-Welch algorithm [1]). More precisely, the forward and backward quantities can be computed recursively with

$$F_i(s)=\sum_r F_{i-1}(r)\alpha(r,s)\beta_s(x_i), \text{ for } i=2,\dots,n \tag{3}$$

and

$$B_{i-1}(r)=\sum_s \alpha(r,s)\beta_s(x_i)B_i(s), \text{ for } i=n,\dots,2. \tag{4}$$

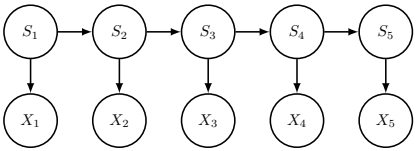


Figure 1. HMM topology with $n = 5$. For $i = 1 \dots 5$, S_i are the hidden variables and X_i the observed variables.

2 *Ad hoc* Model for Outliers in HMMs

Since HMMs are intrinsically heterogeneous, the detection of outliers in these models is a challenging problem. By definition, an observation is an outlier if it is not generated by the model described in the previous section. An *ad hoc* model for outliers in data modeled by HMMs was introduced in the Bayesian framework [11]. The authors of this paper suggest to consider explicit variables for outliers, thus extending the underlying HMM to a simple Bayesian network. We describe this model in the frequentist case. For each i, the model takes into account a Bernoulli variable O_i with $O_i = 1$ iff X_i is an outlier. We suppose that the outlier variables O_i are independent and identically distributed with $\mathbb{P}(O_i = 1) = \rho$. In the resulting Bayesian network, observation X_i depends not only on S_i but also on O_i as shown in Fig. 2. In the case of the homoscedastic Gaussian model we have

$$\mathbb{P}(X_i|S_i, O_i = 0) = \mathcal{N}(\mu_{S_i}, \sigma^2) \tag{5}$$

$$\mathbb{P}(X_i|S_i, O_i = 1) = \mathcal{N}(\mu_{S_i}, \sigma^2) + \mathcal{N}(0, \delta^2) \tag{6}$$

where the last term $\mathcal{N}(0, \delta^2)$ is the noise density that characterizes outliers. Standard inference algorithms for HMMs can be easily extended to this Bayesian network and allow the estimation of parameters and the detection of outliers: 1) the standard forward and backward quantities for HMMs described above can be extended to compute the posterior probabilities $\mathbb{P}(O_i|X = x)$, and 2) the EM algorithm for HMMs can be extended in order to estimate the parameters $\gamma, \alpha, \rho, \sigma^2, \delta^2, \mu_{S_i}$. More specifically, the forward and backward quantities are defined exactly as for standard HMMs - Equations (1) and (2) - and are computed recursively through the formulas

$$F_i(s) = \sum_r F_{i-1}(r)\alpha(r, s) \sum_{o=0,1} \mathbb{P}(X_i = x_i|S_i = s, O_i = o)\mathbb{P}(O_i = o)$$

with $F_1(s) = \gamma(s) \sum_{o=0,1} \mathbb{P}(X_1 = x_1|S_1 = s, O_1 = o)\mathbb{P}(O_1 = o)$, and

$$B_{i-1}(r) = \sum_s \alpha(r, s) \sum_{o=0,1} \mathbb{P}(X_i = x_i|S_i = s, O_i = o)\mathbb{P}(O_i = o)B_i(s)$$

with $B_n(r) \equiv 1$ by convention. The posterior probability distributions of outliers are $\mathbb{P}(O_i = 1|X = x) =$

$$= \sum_s \left(\frac{\rho\mathbb{P}(X_i = x_i|S_i = s, O_i = 1)}{\rho\mathbb{P}(X_i = x_i|S_i = s, O_i = 1) + (1 - \rho)\mathbb{P}(X_i = x_i|S_i = s, O_i = 0)} \cdot \frac{F_i(s)B_i(s)}{\mathbb{P}(X = x)} \right) \tag{7}$$

The probability of the evidence is $\mathbb{P}(X = x) = \sum_s F_n(s)$.

At last, we observe that this Bayesian network is easy to program and therefore provides a convenient way for simulating HMM data with outliers.

3 Model Free Outlier Detection in HMMs based on Relative Entropy

In the present work, we suggest a model free approach for outlier detection which is based on relative entropy measures. The relative entropy (or Kullback-Leibler distance) is a standard (asymmetric) dissimilarity

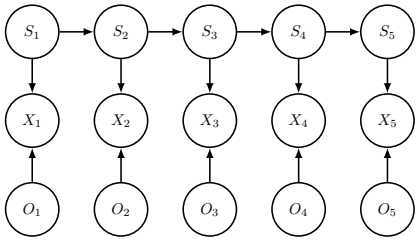


Figure 2. *Ad hoc* model for the HMM in Fig. 1 with outliers: S_i are the hidden variables, X_i the observed variables and O_i the variables encoding the outlier status: $O_i = 1$ iff X_i is an outlier.

measure between two probability density functions [2]. Given two probability distributions p and q their relative entropy is

$$K(p||q) = \int_z p(z) \log \frac{p(z)}{q(z)} dz.$$

Roughly speaking, the relative entropy between p and q measures the extra number of bits required for encoding events sampled from p using a code based on q.

Our method is based on the following assumption: suppose that we observe $X = x$, then if the i-th observation $X_i = x_i$ is an outlier, it must have a strong influence on the posterior distribution $\mathbb{P}(S|X = x)$ which, as a consequence, must differ significantly from the posterior distribution $\mathbb{P}(S|X_{-i} = x_{-i})$ of the hidden variables conditioned on all the observations but the i-th (X_{-i} denotes the set of all the observation variables but X_i).

More specifically, we suggest to use the relative entropy

$$K_i := \sum_S \mathbb{P}(S|X_{-i} = x_{-i}) \log \frac{\mathbb{P}(S|X_{-i} = x_{-i})}{\mathbb{P}(S|X = x)} \tag{8}$$

as a measure of the influence of the observation X_i: if $X_i = x_i$ is an outlier, then K_i is large.

By using standard recursions, it is possible to compute K_i in linear time in n for a fixed i. As a consequence, the complexity of computing K_i for all i is quadratic in n.

We suggest a new recursive formula that reduces the complexity to the linear case through dynamic programming based on the backward and forward quantities for HMMs. The key point is the following lemma:

LEMMA 3.1. *For an arbitrary fixed $i = 1, \ldots, n$, the relative entropy between the two posterior distributions $\mathbb{P}(S|X_{-i} = x_{-i})$ and $\mathbb{P}(S|X = x)$ is equal to the relative entropy between $\mathbb{P}(S_i|X_{-i} = x_{-i})$ and $\mathbb{P}(S_i|X = x)$*

$$K_i = \sum_s \mathbb{P}(S_i = s|X_{-i} = x_{-i}) \log \frac{\mathbb{P}(S_i = s|X_{-i} = x_{-i})}{\mathbb{P}(S_i = s|X = x)}.$$

The factors in the equation above can be calculated by using the following formulas:

$$\mathbb{P}(S_i = s|X = x) = \frac{B_i(s)F_i(s)}{\sum_r B_i(r)F_i(r)}$$

$$\mathbb{P}(S_i = s|X_{-i} = x_{-i}) = \frac{F_i^*(s)B_i(s)}{\sum_r F_i^*(r)B_i(r)}$$

where F_i, B_i are computed recursively as in equations (3) and (4), and

$$F_i^*(s) = \sum_r F_{i-1}(r)\alpha(r, s) \text{ for } i = 2, \ldots, n.$$

with $F_1^*(s) = \gamma(s)$.

δ	AUC S	AUC T
0.00	0.50 [0.47,0.54]	0.52 [0.48,0.55]
0.50	0.50 [0.46,0.53]	0.49 [0.46,0.53]
1.00	0.52 [0.49,0.56]	0.56 [0.52,0.59]
1.50	0.55 [0.52,0.59]	0.74 [0.71,0.78]
2.00	0.69 [0.66,0.73]	0.87 [0.85,0.89]
2.50	0.76 [0.73,0.79]	0.93 [0.91,0.95]
3.00	0.84 [0.81,0.87]	0.97 [0.95,0.98]
3.50	0.87 [0.85,0.90]	0.99 [0.98,0.99]
4.00	0.92 [0.91,0.94]	0.99 [0.99,1.00]
4.50	0.96 [0.95,0.97]	0.99 [0.99,1.00]

Table 1. Statistics S based on relative entropy vs statistics T based on the *ad hoc* model for a few values of the noise δ characterizing the outliers. The 95% confidence intervals for the AUC are shown.

4 Validation of the two Methods by Simulations

We compared our method based on relative entropy and the *ad hoc* model on simulated data. Simulations where performed using the *ad hoc* model described in Section 2 in the homoscedastic Gaussian case for Gaussian outliers. In both cases, we considered $n = 200$ observations with three hidden states $s \in \{1, 2, 3\}$; data were generated accordingly to the following parameter values:

- $\gamma(1) = 1$ and $\gamma(s) = 0$ for $s \in \{2, 3\}$
- $\alpha(r, s) = \begin{pmatrix} 1-\eta & \eta/2 & \eta/2 \\ \eta/2 & 1-\eta & \eta/2 \\ \eta/2 & \eta/2 & 1-\eta \end{pmatrix}$ with $\eta = 0.05$
- $\mu_1 = 1$, $\mu_2 = 2$, $\mu_3 = 3$
- $\sigma = 1$
- $\rho = 0.05$.

We considered several values of δ in Eq. (6) and for each of those we simulated 500 datasets. For each simulation and for each observation, we computed the relative entropy K_i - Eq. (8) - and the posterior probability that the ith-observation is an outlier - Eq. (7). All the computations were done with the true parameters, i.e. the parameters used for simulating the data.

We tested the hypothesis that the data contain outliers using the two alternative global statistics

$$S = \max_{i=1,\dots,n} K_i$$

and

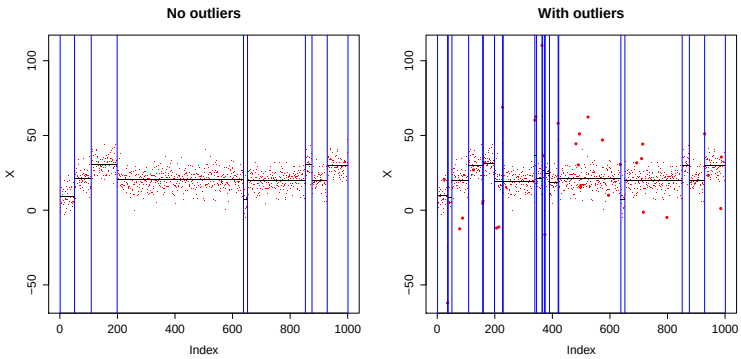
$$T = \max_{i=1,\dots,n} \mathbb{P}(O_i = 1 | X = x).$$

We compared the performance of the two statistics by means of the Area Under the Curve (AUC) which measures the surface under the Receiver Operating Characteristic (ROC) curve. The results are depicted in Table 1.

Results show that both the statistics S and T have good performance in detecting the global presence of outliers when the noise characterizing the outliers is $\delta \geq 2$ (AUC ≥ 0.7). Not surprisingly the *ad hoc* model gives better results than our entropy-based method.

5 Application to Change-Point Detection in CNV data

The segmentation of heterogeneous sequences is a statistical challenge with many applications in fields such as finance, signal analysis, neuro-sciences and biology [12]. A wide variety of literature exists for finding an ideal set of change-points that characterize the data, in terms of both the number of change-points in a given sequence and their corresponding locations. More precisely, the aim is to identify the intervals in which the signal being analyzed is homogeneous, i.e. the change points in which the signal is subject to abrupt variations. In order to do so, one possible approach is to model the signal with an HMM in which the hidden



(a) (b)

Figure 3. Simulated CNV data using the model described in Section 2 with $\sigma = 6, \delta = 35$ in Equations (5),(6). Change points are detected using the algorithm introduced in [4] and correspond to the vertical blue lines. Fig. (a): Original data. Fig. (b): Original data with outliers added in 5% of the positions (the outliers are depicted with larger filled dots).

state S_i pertains to the level (or underlying distribution) of observation X_i (*level-based model*). This is an appropriate model when underlying properties can be shared between observations in different, non-adjacent segments. Another very common approach consists in considering HMMs in which the hidden states are the homogeneous segments (*segment-based model*).

The detection of change-points in data from array Comparative Genomic Hybridization (aCGH) techniques, in order to identify Copy Number Variation (CNV), is crucial for the characterization of DNA, including tumoral DNA in cancer. A CNV may locate a genetic mutation such as a duplication or deletion in a cancerous cell. For CNV data, the segment-based approach is commonly adopted in order to detect segments in data differing more than a certain threshold [9]. The segment-based approach has been widely studied and many variations exist, see [7] for a review. The level-based approach was suggested to characterize mutations and was originally introduced in [6]. Various extensions to this approach have been proposed, including implementations for dealing with higher-resolution data [3,13]. Other research directions in this domain include estimating the number of change-points [10] and characterizing the uncertainty in the change-point locations [8].

While most of the current procedures are designed to detect the ideal number or set of change-point locations, here we address the problem of characterizing the outliers in data. Segmentation models are known to be particularly sensitive to the presence of outliers: for instance a single outlier can result in a segment consisting in just one point. As an example, we simulated CNV without outliers and in presence of outliers, see Fig. 3a and Fig. 3b respectively. Data were simulated using the aforementioned *ad hoc* model. Change-points were detected using the method described in [4] (consisting in a dynamic programming exhaustive search on thin grid with penalized likelihood model selection) and correspond to the vertical blue lines in the figures. As the figure clearly shows, outlier determines an over segmentation. This example shows that it is crucial to detect outliers in the data prior to perform change-point detection.

We applied our method based on relative entropy to a widely referenced data set from a breast cancer cell line BT474 [12]. The data consist of log-reference ratios signifying the ratio of genomic copies of test samples compared to normal. After estimating the parameters with the EM algorithm for HMMs with homescedastic Gaussian emissions, we computed the relative entropy K_i for each observation i ($n = 120$). Fig. 4a shows the original data whereas Fig. 4c shows the value of K_i for each i. Peaks in the plot correspond to observations with greater influence on the posterior distributions of the hidden levels.

Fig. 4b shows the same data in which we arbitrarily added three outliers in positions $i = 20, 40, 100$ (triangular red dots). After estimating the parameters for the new dataset, we computed the new values of the relative entropy. Fig. 4d shows that the relative entropy has peaks in the positions corresponding to the outliers which were previously arbitrarily added. The horizontal bar at height $h = 5.30$ corresponds to an empirical 5% threshold under H0; there are seven observations with K_i greater than this threshold including the three artificial outliers ($i = 19, 20, 40, 96, 100, 111, 119$).

6 Discussion and Conclusions

In this abstract we presented two alternative methods for detecting the presence of outliers in data modeled by HMMs. The first method relies upon a simple *ad hoc* Bayesian network which extends the HMM by including indicatrix variables for the outliers. The second method is our original contribution and consists in assessing the influence of each observation on the hidden distribution by means of relative entropy measures.

We validated the two methods on simulated datasets. The two methods proved to have good power in detecting the global presence of outliers. On simulated data, the *ad hoc* method showed better performance than our method. This is not surprising because data were simulated using the *ad hoc* model and computations were performed using the true parameters. It would be interesting to compare the performance of the two methods when the parameters are estimated. In order to do so, it is necessary to extend the EM algorithm for HMMs to the *ad hoc* Bayesian network described in Section 2; this problem is left for future work. It is worth pointing out that the *ad hoc* model requires estimating more parameters than in our method. For instance, in the specific homoscedastic Gaussian case that we have presented, the *ad hoc* models depends on seven parameters $(\eta, \mu_1, \mu_2, \mu_3, \sigma, \delta, \rho)$, whereas in order to apply our method, one has to estimate five parameters $(\eta, \mu_1, \mu_2, \mu_3, \sigma)$.

At last, we applied our method on real CNV data. After estimating the parameters, we were able to locate the outliers which we had previously added by looking at the relative entropy values.

Aknowledgments. We would like to thank the anonymous referees for their useful comments and remarks and The Minh Luong for many valuable discussions on change-point detection.

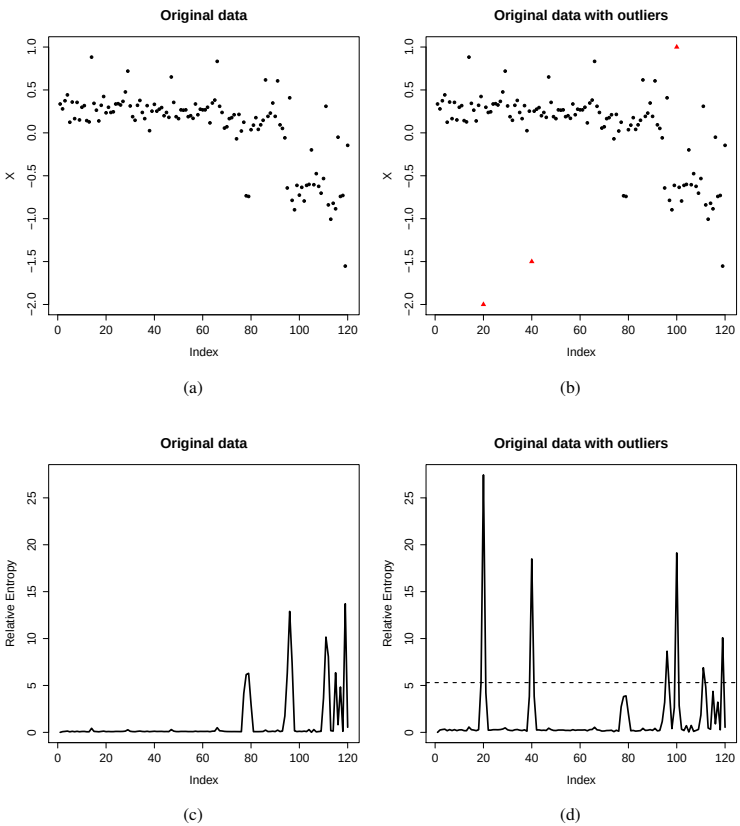


Figure 4. Application of the method based on relative entropy for the detection of outliers. Fig. (a): Original CNV data from [12]. Fig. (b): Original data with outliers added in positions $i = 20, 40, 100$. Fig. (c) and (d): Relative entropy K_i computed after estimating the parameters; the horizontal bar indicates the empirical 5% threshold under H0.

References

[1] J. Bilmes. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *International Computer Science Institute*, 4:126, 1998.

[2] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[3] S. Colella, C. Yau, J. Taylor, G. Mirza, H. Butler, P. Clouston, A. Bassett, A. Seller, C. Holmes, and J. Ragoussis. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Research*, 35(6):2013–2025, 2007.

[4] F. Comte and Y. Rozenholc. A new algorithm for fixed design regression and denoising. *Annals of the Institute of Statistical Mathematics*, 56(3):449–473, 2004.

[5] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge Univ Pr, 1998.

[6] J. Fridlyand, A. Snijders, D. Pinkel, D. Albertson, and A. Jain. Hidden Markov models approach to the analysis of array CGH data. *Journal of multivariate analysis*, 90(1):132–153, 2004.

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

[8] T. Luong, Y. Rozenholc, and G. Nuel. Fast estimation of posterior probabilities in change-point models through a constrained hidden Markov model. *Arxiv preprint arXiv:1203.4394*, 2012.

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

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

[11] S. Shah, X. Xuan, R. DeLeeuw, M. Khojasteh, W. Lam, R. Ng, and K. Murphy. Integrating copy number polymorphisms into array CGH analysis using a robust HMM. *Bioinformatics*, 22(14):e431–e439, 2006.

[12] A. Snijders, N. Nowak, R. Segraves, S. Blackwood, N. Brown, J. Conroy, G. Hamilton, A. Hindle, B. Huey, K. Kimura, et al. Assembly of microarrays for genome-wide measurement of DNA copy number by CGH. *Nature genetics*, 29:263–264, 2001.

[13] K. Wang, M. Li, D. Hadley, R. Liu, J. Glessner, S. Grant, H. Hakonarson, and M. Bucan. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome research*, 17(11):1665–1674, 2007.

Session 9_C : Distributed Data and Computation

Towards a Distributed Infrastructure for Bioinformatics: French GRISBI Perspective

Christophe BLANCHET[1], Clément GAUTHEY[1], Christophe CARON[2], Olivier COLLIN[3], Stéphane DELMOTTE[4], Tiphaine MARTIN[5], Simon PENEL[4], Aurélien ROULT[3], Franck SAMSON[6] and Bruno SPATARO[4]

[1] Université Lyon 1, Univ Lyon, France;CNRS, FR 3302;Institut de Biologie et Chimie des Protéines, IBCP, 7 passage du vercors, F-69367, France
{christophe.blanchet, clement.gauthey}@ibcp.fr

[2] ABiMS, FR2424 CNRS-UPMC, Station Biologique, Place Georges Teissier, 29680, Roscoff, France
christophe.caron@sb-roscoff.fr

[3] GenOuest, CNRS IRISA - UMR6074, Campus de Beaulieu, 35042 RENNES Cedex - France
{olivier.collin, aurelien.roult}@irisa.fr

[4] LBBE, UMR5558 CNRS, 43 bd du 11 novembre 1918, 69622 VILLEURBANNE cedex, France
{bruno.spataro, stephane.delmotte}@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

Abstract *Bioinformatics is now facing a deluge of data that is modifying drastically the practices. Now, Bioinformatics requires research infrastructures able to store very large biological data sets, and to make these data available for intensive scientific computing. The GRISBI infrastructure (Grid Support for Bioinformatics, www.grisbio.fr) has built such infrastructure over the own resources of French bioinformatics platforms. To make challenging biological applications possible, usual biological data and bioinformatics tools have been deployed on the distributed infrastructure. In terms of science, comparative genomics and analysis of next-generation sequencing data suit well the potential added-value of a grid. Structural biology analysis has also demonstrated a relevant usage of a distributed infrastructure, such as the automatic structure determination of proteins from experimental NMR data. Next steps will see the extension of the infrastructure with other French platforms, to insert it in the European landscape in collaboration with the ELIXIR infrastructure, and to open it to other biological applications and scientific fields linked to Life Science.*

Keywords Bioinformatics, national scientific community, big data, distributed infrastructure.

1 Introduction

Bioinformatics is now facing a deluge of data that is modifying drastically the landscape. Most common analyses have evolved to a larger scale, for example from the study of a single gene/protein to a whole genome or family of proteins, from a single metabolic pathway to systems biology. Now, Bioinformatics is requiring research infrastructures able to store very large biological data sets, and to make these data available for intensive scientific computing.

Grid computing concept [1] is defined as a set of information resources (computers, databases, networks, instruments) that are integrated to provide users with tools and applications that treat those resources as components within a virtual system. The goal of a grid infrastructure devoted to Bioinformatics is to gather computing resources at the national level to make possible challenging bioinformatics applications dealing with large scale systems.

Among the different fields of Bioinformatics some are very CPU- and data-demanding, for example comparative genomics and genome annotation, protein structure automatic determination, molecular interactions such as protein-protein or DNA-protein, etc. The requirements of such applications concern

storing large data in a large number of files and transferring them among different geographical sites, analyzing these large datasets with large-scale computations that can be parallelized by the data.

Bioinformatics applications such as macromolecular interactions [2], automated refinement of X-ray structure models [3] or proteins structure determination [4] have demonstrated a real improvement by being integrated in a distributed infrastructure such as a grid. The last application, the automatic structure determination of proteins from experimental nuclear magnetic resonance data was done with the ARIA software [5] that have been adapted and deployed on the production infrastructure of the GRISBI grid.

In this paper, we describe the work done to set up a national distributed infrastructure devoted to Bioinformatics. The gLite grid middleware was used to link the different computing resources, in collaboration with national and European infrastructure. We have extended the available services from gLite with components from other software frameworks. The infrastructure was validated with relevant bioinformatics applications requiring large-scale computing resources.

2 Materials and Methods

2.1 Grid middleware

From the web site of the gLite software, the distribution is described as «an integrated set of components designed to enable resource sharing: this is middleware for building a grid. The gLite middleware was produced by the EGEE project and it is currently being developed by the EMI project. [...] The distribution model is to construct different services ('node-types') from these components and then ensure easy installation and configuration on the chosen platforms».

The gLite middleware defines different components to build the grid: user interface (UI), computing element (CE), worker node (WN), storage element (SE), workload management service (WMS), logical file catalog (LFC), Berkeley database information index (BDII), Nagios box, virtual organization management service (VOMS), etc. (Figure 1).

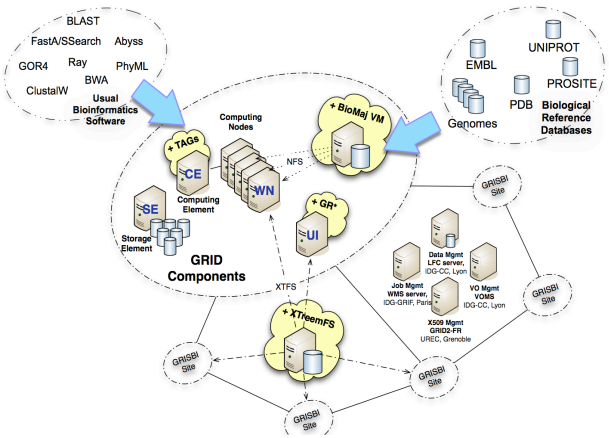


Figure 1. GRISBI infrastructure with the standard gLite components (in grey) and the supplementary components (in yellow): BioMAJ, XtreemFS and GR* commands.

The CE and WNs are in charge of the computation, the SEs store the files, the LFC organizes these files in a logical file tree with one root and one branch by virtual organisation (VO), for example «/grid/vo.renabi.fr», etc. Some of them are mandatory on each site joining its resources to the infrastructure, at least one SE, one CE and several WNs. One UI per site, while not mandatory, is recommended. Other components are mandatory but at the level of virtual community, i.e. per VO: one VOMS server, one LFC server and at least one WMS server. VOMS and LFC server are unique and can not be duplicated, while there could and should be several WMS per virtual organization. The service of VOMS, LFC and WMS for our VO

vo.renabi.fr are operated by 3 sites of the French Institute of Grid of CNRS: computing center of IN2P3 Lyon, institute LAL Orsay, institute IPHC Strasbourg.

2.2 Distributed filesystem

Availability of a personal volume for each user all over the grid is a an important requirement to help the adoption of the infrastructure by biologists and bioinformaticians. XtreemFS is one of the very few frameworks providing such a functionality. While distributed and including replication, XtreemFS is a POSIX filesystem in a sense that a volume can be mounted on a computer and used by legacy software without change of the code.

The web site of the project is describing XtreemFS as «clients and servers that can be distributed worldwide. XtreemFS supports installations across many data centers and is able to handle the failures that occur in wide-area installations. Clients can mount XtreemFS volumes from anywhere with an Internet connection. XtreemFS replicates files data across multiple storage servers, which can be distributed worldwide».

XtreemFS defines three different components: object storage device (OSD), directory service (DIR), metadata and replica catalog (MRC), DIR and OSD. The OSD servers are storing the data, while the DIR is the directory referencing the different services, and the MRC is managing the metadata, the logical tree of the files, their replicas and the authorizations. An OSD is a server hosting the physical disks. There have to be several OSDs, in fact as many OSDs than you have sites, and there could be also several OSDs per site. A MRC is mastering several OSDs, from the same site, from different geographical ones. And the DIR, that could be replicated for security reasons, is on the top of the pyramid, mastering all the MRCs and OSDs of the different sites.

The users manage their own volumes, even create new ones, and define the authorizations to access these volumes by themselves or by other users. The interaction with the volumes is done through a command line client, available for the different operating systems. For example, the command mount.xtreemfs mounts the volumes in a POSIX way, that means that users access them on their computer as if it was a local disk.

3 Results

The GRISBI infrastructure (Grid Support for Bioinformatics, www.grisbio.fr) is a joined initiative between several French bioinformatics sites, collaborating in the context of the national RENABI network (www.renabi.fr). The goal is to mutualize the computing and bioinformatics resources of these scientific datacenters in a distributed Research Infrastructure devoted to Bioinformatics.

The GRISBI infrastructure comprise 9 bioinformatics sites with 7 of them providing currently hardware resources to the infrastructure. These sites are sharing and mutualizing their storage and computing resources with grid software components of the gLite middleware. These sites are distributed over 8 cities in 5 different regions of the country (Table 1). Roughly 860 cores and 30 TB of storage have been integrated into the grid. The installation and the maintenance of all the GRISBI sites are managed with the QUATTOR system [6], installed on a centralized server located in Lyon. Each site administrator can define the profiles for his own grid servers, and deploy them in an automated way. To these initial bioinformatics hardware resources, three French «generic» computing centers have also authorized the RENABI's VO on their resources, extending the infrastructure. They are the IPHC institute in Strasbourg, the datacenter M3PEC in Bordeaux and the datacenter of University of Lille-1.

Bioinformatics regional center	Cities (nb sites)	CPU (core)	storage (TB)
APLIBIO	Jouy-en-Josas (1)	-	3
PRABI	Lyon (2)	716	20.5
RENABI-GO	Rennes (1), Roscoff (1)	88	4.05
RENABI-NE	Lille (1), Strasbourg (1)	-	-
RENABI-SO	Bordeaux (1), Toulouse (1)	52	2
TOTAL	9	856	29.55

Tab. 1. RENABI grid resources and the contributing bioinformatics regional centers (details on www.grisbio.fr/resources/hardware).

3.1 Integration of legacy batch farms

At the time of their involvement in the GRISBI project, the French bioinformatics platforms were already providing biologists and bioinformaticians with local computing farms through a web portal or a command line interface. These resources were running, for most of them, in production since several years. So, these clusters could not be interrupted nor disrupted by their integration in the grid, for evident reasons of quality of service for the scientific community. But the gLite middleware is recommended to be installed on new bare servers. Indeed, the only available installation guides and procedures are treating with OS- and data-free computers.

A new way of installing a computing element (CE) in front of an existing batch farm was required and has been developed to setup the GRISBI infrastructure. This adapted component was called «CE-gateway» (CEgw). This work is based on the native functionality of the gLite CE to be installed on 2 different servers, albeit most of time it is not used. Procedures and recommendations for the installation and the operations were defined, formalized, and used to run such CEs in production mode on the GRISBI infrastructure. The principles are to set up a complete CE connected to the existing master server of a running computing farm (Figure 2). A CEgw is registered in the infrastructure and received the jobs coming from the WMS. The CEgw then converts these grid jobs in local jobs, and forwarded them to the local master. Once computed, the CEgw send the status and the results of the computation back to the WMS and to the grid information system. Bulk of users need to be created on the local farm master for each VO authorized on this CE. A NFS shared filesystem is also needed between the master and the nodes to share the users home directories.

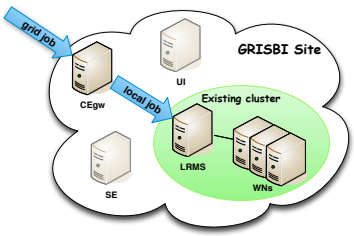


Figure 2. Principle of the integration of a existing production computing farm. A computer element is added as a gateway between the existing cluster and the component of the grid. This CE is relaying the jobs coming from the grid and convert them in local jobs.

This work was done for two major batch schedulers used in the community: Torque and Sun Grid Engine. The integration rules have been adapted from the installation of a classic gLite computing element for that specialized purpose. Today two CEgw are running on two sites of the GRISBI infrastructure, one with Torque in PRABI-IBCP site and one with SGE in ReNaBiGO-ABiMS site. An other one is planned to be deployed with Torque in PRABI-LBBE.

3.2 Storage of user data on an ubiquitous volume

To provide GRISBI users with a personal volume available all over the grid, even from their own personal computer, the XtreemFS framework was integrated to the environment of the GRISBI infrastructure. For each user a shared volume is created and made accessible only with its personal grid credentials, based on X509 certificate and proxy mechanisms. The user volume is labeled relatively to the denomination name (DN) attribute of the own user certificate.

Each user volume can be mounted at the same time from several points of the infrastructure, scaling to hundreds of nodes. The reliability and security of concurrent access on the files is ensured by the internal XtreemFS mechanisms. Two XtreemFS pools have already been deployed in PRABI-IBCP (Lyon) and RENABI-GO (Rennes) for a total space of 10 terabytes. Others storage nodes are planned to extend the storage space and the replication of the data.

Biological databases	Tags
pdb_derived50	VO-vo.renabi.fr-data-pdb_derived50
pdb_derived90	VO-vo.renabi.fr-data-pdb_derived90
pdb_seqres	VO-vo.renabi.fr-data-pdb_seqres
rfam	VO-vo.renabi.fr-data-rfam
uniprot_sprot	VO-vo.renabi.fr-data-uniprot_sprot
uniprot_sprot_varsplic	VO-vo.renabi.fr-data-uniprot_sprot_varsplic
uniprot_trembl	VO-vo.renabi.fr-data-uniprot_trembl
uniref100	VO-vo.renabi.fr-data-uniref100
uniref50	VO-vo.renabi.fr-data-uniref50
uniref90	VO-vo.renabi.fr-data-uniref90

Bioinformatics tools	Tags	Sites
ABYSS	VO-vo.renabi.fr-tool-ABYSS-1.2.6-bundle	3
R	VO-vo.renabi.fr-tool-R-2.13.0-bundle	6
Ray	VO-vo.renabi.fr-tool-Ray-1.3.0	1
bwa	VO-vo.renabi.fr-tool-bwa-0.5.8c	3
cap3	VO-vo.renabi.fr-tool-cap3-0.0	5
clustalw2	VO-vo.renabi.fr-tool-clustalw2-2.0.5	2
clustalw2	VO-vo.renabi.fr-tool-clustalw2-2.1	1
fasta35	VO-vo.renabi.fr-tool-fasta35-35.4.12	3
gor4	VO-vo.renabi.fr-tool-gor4-0.0	4
hmmer	VO-vo.renabi.fr-tool-hmmer-3.0-bundle	3
meme	VO-vo.renabi.fr-tool-meme-4.4.0	2
meme	VO-vo.renabi.fr-tool-meme-4.6.0_1-bundle	1
ncbi_blast	VO-vo.renabi.fr-tool-ncbi_blast-2.2.23-bundle	3
phyml	VO-vo.renabi.fr-tool-phyml-3.0	3
predator	VO-vo.renabi.fr-tool-predator-2.1.2	4
simpa96	VO-vo.renabi.fr-tool-simpa96-1.1	1
ssearch35	VO-vo.renabi.fr-tool-ssearch35-35.4.12	3

Tab. 2. Representative biological databases and bioinformatics tools that have been deployed on the GRISBI infrastructure (details on www.grisbio.fr/bioapps/votags).

3.3 Bioinformatics resources on the GRISBI grid.

In parallel to the deployment of the hardware resources, biological data and bioinformatics tools have also been replicated and installed on the nodes of the infrastructure. A list of representative data sources and bioinformatics software are installed in different places (See Table 2). For example the UNIPROT database and three genome analysis tools have been deployed on the GRISBI grid: Abyss, BWA and Ray.

Scientists running their jobs on GRISBI can ask to use the specific sites having these resources, data and tools (Figure 3). This is based on the gLite VO-tags that are defined by the user in the job description file (JDL file). Then the workload management system of the infrastructure schedules the job to run on the site comprising the requested bioinformatics resources.

Access to public reference biological databases from any compute node is a strong requirement. Such databases contain, for example, protein or gene sequences and associated data, protein structures or complete genomes. The GRISBI infrastructure has been extended with a new type of grid component, a biological data repository. This server acts as a proxy between Internet where all the reference databases are published and the infrastructure where the bioinformatics analyses are computed. The BioMAJ system [7] is used to import and maintain these databases. This repository node stores the data on a local disk, and then exports it through a NFS share, in read-only mode for security reasons, to all the computing nodes of the site (Figure 4).

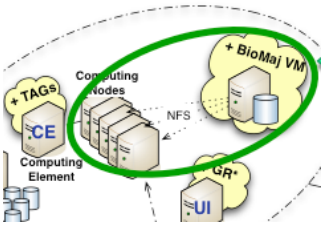


Figure 3. GRISBI infrastructure with the standard gLite components (in grey) and the supplementary components (in yellow): BioMAJ, XtreemFS and GR* commands.

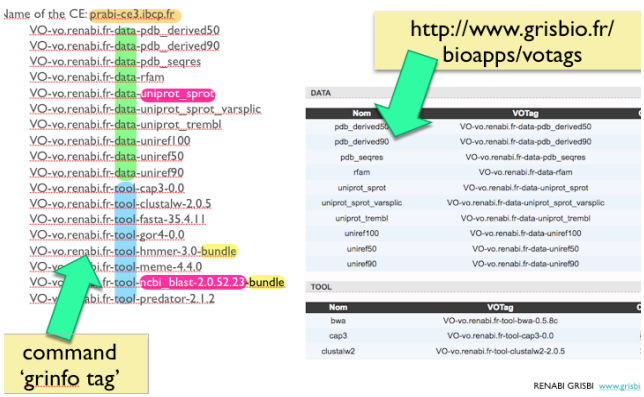


Figure 4. GRISBI infrastructure with the standard gLite components (in grey) and the supplementary components (in yellow): BioMAJ, XtreemFS and GR* commands.

4 Discussion

Regarding the deployment of the infrastructure, several issues were encountered that were not raised by previous site deployments of the gLite stack. For example, French bioinformatics platforms do not dispose most of time of a large number of public network addresses (IP). The platforms are running their local network with private IP addresses (like 192.168.0.0 or 172.16.0.0) and are using network address translation (NAT) for their public servers. This was a strong constraint in that sense that it was leading to troubles when deploying the grid middleware. Moreover, network could also be a bottleneck in terms of bandwidth. Sites aiming to join the infrastructure need to be connected to Internet (RENATER for the French academic sites) through at least a gigabit link. An other issue raised and solved was about the operating system. gLite comes for Scientific Linux only. That makes no difference to deploy it on the CentOS distribution that most of the GRISBI sites are running, but have raised some troubles for the sites using the Debian distribution.

Two applications were chosen as representative of the global requirements of the community and were run on the infrastructure. The first one was about large scale phylogenomics and comparative genomics that require complex computational methods (parsimony, maximum likelihood, bayesian methods, MCMC, etc.) associated with massively distributed calculation [8]. The partner PRABI-LBBE is studying the evolution of Life at the molecular level: evolution and dynamic of genomes, building of the Tree of Life, species history, etc. Regarding the comparative field, PRABI-LBBE developed the HOGENOM database whose maintenance demands large computing resources [9]. That work has requested pairwise comparisons of millions of protein sequences with the BLAST software. An other kind of analysis used to build this database was the computation of phylogenies for hundred of thousands of gene families. Regarding the BLAST comparisons used for HOGENOM, the update of the database has implied to maintain a list of the BLAST hits among all the protein sequences publicly available. That represents today roughly 12 millions sequences obtained from different sources: UNIPROT, CDS translated from ENSEMBL genomes, etc. The update of

the database was done with different computing resources, GRISBI was one of them and used for part of 8 millions of sequences. The results were stored in several hundreds of thousands of files representing more than one terabyte of data. During the computations run on the GRISBI infrastructure, the data were stored and exchanged between the different grid jobs through a XtreemFS distributed volume. The experiment computed on GRISBI was representing 47,713 job for a total of 71,362 'HEP-SPEC 2006' normalized hours. During the experiment, the higher number of jobs running in parallel was 830. This computation has used the pre-installed version of the BLAST tool. And the shared XtreemFS volume has supported the load of all these jobs running at the same time.

The second representative application was about comparative genomics and the analysis of next-generation sequencing (NGS) data. Albeit that such application is well suited to get benefits from the potential added-value of the grid -storing large data in a lot of files and transferring them among different geographical sites-, the usage of the GRISBI infrastructure during real experiments have proved its usefulness but also raised some issues. These issues were mainly about the data transfer between the data production sites and the computation sites of the grid. Massive data such as NGS ones need network link with high transfer rates that are not yet the general case in France as in other countries. An other issue was about memory consumption and computation error management in the grid. Some genome assembling tasks require a lot of memory and then need to be run on specific machines. We have put in the GRISBI grid such big-memory machines, but during some analyses they have not been big enough and the computation has crashed. The related computation nodes were in an undetermined state, leading to a node interruption on a site, but without a good error feedback to the scientist having submitted the jobs.

As a general perspective, such a distributed infrastructure is biologically relevant with the requirements of the community in term of storage and computation. The different components deployed for the storage, especially the XtreemFS one, fulfill the needs from scientists to transfer their data easily from their desktop or the instruments in their laboratory to the national big storage and computation facilities. The mutualized computation resources available in GRISBI provide also a large ready-to-use reservoir for temporary requirements in terms of large-scale (hundreds of cores) or specific computation (for example machine with 1 TB memory) from the biological laboratories and bioinformatics platforms. The distributed GRISBI infrastructure is a step in the direction of a national bioinformatics infrastructure relying on the mutualization of the own community resources of the regional platforms, and complementary with the national HPC centers.

5 Conclusion

The GRISBI infrastructure (Grid Support for Bioinformatics, www.grisbio.fr) was built as a distributed infrastructure over French bioinformatics platforms. Today 7 sites are mutualizing their computing resources in a distributed Research Infrastructure devoted to Bioinformatics. This infrastructure was used for scientific experiments dealing with challenging bioinformatics applications such as comparative genomics, genome annotation or protein structure automatic determination.

Perspective are to extend this model to other biological applications and other datacenters, such as the other RENABI platforms or multidisciplinary regional datacenters. Applying the distributed GRISBI organizational model to connected fields of Health and Environment communities will also been foreseen. First candidate is certainly the neighboring community of Environment, and in a second time the Health community that have more organizational constraints related to data privacy of the patients. Efforts have also to be done in terms of interfaces available to scientists. Having bioinformatics web portals able to manage grid jobs and data will help to run large bioinformatics studies and pipelines on such a distributed infrastructure.

Most of European Countries have begun to build a distributed computing infrastructure at a national scale: Spain & Portugal (IBERGRID), Italy (LIBI), United Kingdom (eScience), Switzerland (SWING), Sweden (SNIC), etc. These infrastructures have a sense of purpose to provide different scientific communities with computing grid infrastructures, and a part of their activities are devoted to Biology/Bioinformatics. The GRISBI platform has already several contacts and ongoing collaborations with these infrastructures in other European countries. GRISBI will enforce these links with the foreign national bioinformatics infrastructures to identify the common requirements, and to promote a model of interoperability between the national distributed bioinformatics infrastructures.

Acknowledgements

This work has been supported by the national framework IBISA «Infrastructures en Biologie Santé et Agronomie» (AO2008), the French «Institut des Grilles du CNRS», and by the national network of bioinformatics platforms RENABI. The authors also thank for their participation to the infrastructure: David Benaben, Christelle Eloto, Pierre Gay, Daniel Jacob, Didier Laborie, Nouredine Melab, Alexis Michon, Delphine Naquin and Frédéric Plewniak.

References

[1] I. Foster and C. Kesselman, Eds. The Grid 2, *Second Edition: Blueprint for a New Computing Infrastructure (The Elsevier Series in Grid Computing)*, 2nd ed. Morgan Kaufmann, 2003, p. 748.

[2] K. Zakrzewska, B. Bouvier, A. Michon, C. Blanchet, and R. Lavery, Protein-DNA binding specificity: a grid-enabled computational approach applied to single and multiple protein assemblies. *Phys Chem Chem Phys*, vol. 11, no. 45, pp. 10712–10721, Dec. 2009.

[3] R. P. Joosten, J. Salzemann, V. Bloch, H. Stockinger, A.-C. Berglund, C. Blanchet, E. Bongcam-Rudloff, C. Combet, A. L. Da Costa, G. Deléage, M. Diarena, R. Fabbretti, G. Fettahi, V. Flegel, A. Gisel, V. Kasam, T. Kervinen, E. Korpelainen, K. Mattila, M. Pagni, M. Reichstadt, V. Breton, I. J. Tickle, and G. Vriend, PDB_REDO: automated re-refinement of X-ray structure models in the PDB. *J Appl Crystallogr*, vol. 42, no. 3, pp. 376–384, Jun. 2009.

[4] F. Mareuil, C. Blanchet, T. E. Malliavin, and M. Nilges, Grid computing for improving conformational sampling in NMR structure calculation. *Bioinformatics*, vol. 27, no. 12, pp. 1713–1714, Jun. 2011.

[5] 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*, vol. 23, no. 3, pp. 381–382, Feb. 2007.

[6] S. Childs, M. E. Poleggi, C. Loomis, L. F. M. Mejías, M. Jouvin, R. Starink, S. De Weirdt, and G. C. Meliá, Devolved Management of Distributed Infrastructures with Quattor. *LISA*, pp. 175–189, 2008.

[7] 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*, vol. 24, no. 16, pp. 1823–1825, Aug. 2008.

[8] V. Miele, S. Penel, and L. Duret, Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics*, vol. 12, p. 116, 2011.

[9] S. Penel, A.-M. Arigon, J.-F. Dufayard, A.-S. Sertier, V. Daubin, L. Duret, M. Gouy, and G. Perrière, Databases of homologous gene families for comparative genomics. *BMC Bioinformatics*, vol. 10, p. S3, 2009.

Présentation Industrielle

Sven NEIRYNCK

Panasas

Parallel Storage: Addressing the Big Data Challenge

The exponentially growing volumes of data generated by technical applications compound the challenge of selecting a storage infrastructure capable of linearly scaling capacity and performance. Panasas will discuss how to address this big data storage challenge with high-performance parallel storage and how the emerging open standard parallel NFS (pNFS) protocol will further enable performance at scale.

Panasas, Sunnyvale, USA

Présentation Industrielle

Jean-Christophe Baratault

NVIDIA

Plus de simulations pour des résultats encore plus rapide avec les GPU NVIDIA

Plus de 200 papiers techniques ont été publiés ces 12 derniers mois sur des projets de recherche en génomique et bio-informatique, ce en utilisant la puissance des processeurs NVIDIA dans des stations de travail et des serveurs.

Découvrez comment le GPU, conçu initialement pour des applications graphiques, peut maintenant booster radicalement vos projets de recherche en décuplant le nombre de vos simulations. NVIDIA a développé un écosystème complet à destination des chercheurs en travaillant conjointement avec de nombreux développeurs d'applications scientifiques Open Source. Bénéficiez dès maintenant de la puissance des GPUs dans vos travaux de recherche !

NVIDIA - Tesla Sales EMEA

Session 10 : Evolution

Conférence invitée

Toni GABALDÓN

Comparative Genomics Group. Bioinformatics and Genomic Programme. Centre for Genomic Regulation (CRG). Barcelona. Spain

Phylogenomics in the light of ever-growing sequencing data

A pressing challenge in phylogenomics is the need to cope with the massive production of complete genomic sequences, especially after recent technological developments. Problems that are particularly affected by the increasing flow of genomic data and that require continuous update are: i) the establishment of evolutionary relationships between species (the so-called Tree Of Life (TOL)), ii) the inference of orthology and paralogy relationships across genomes, and iii) the study of the evolution of large, widespread super-families that evolved through complex patterns of duplications and losses. To face such challenges we have developed two sophisticated pipelines that allow high scalability and continuous update, while achieving highest levels of accuracy. The first such pipeline automatically reconstructs entire species-centric collections of gene phylogenies (the so-called phylome), and combines this with phylogenetic information from various other sources to derive unique orthology and paralogy predictions. The second pipeline, which we apply to the superfamily and the Tree of Life assembly problems, is able to reconstruct large phylogenies by means of an iterative strategy that provides scalable resolution and allows continuous update. In this talk, I will illustrate the use of such approaches in the context of the assessment of the evolution of important traits in fungi, and the reconstruction of a genome-based, eukaryotic tree of life.

Evolution of conjugation and type IV secretion systems

Julien GUGLIELMINI[1], Fernando DE LA CRUZ[2] and Eduardo P.C. ROCHA[1]
[1] MICROBIAL EVOLUTIONARY GENOMICS, Institut Pasteur, CNRS UMR355, F-75015, Paris, France
{jgugliel, erocha}@pasteur.fr
[2] DEPARTAMENTO DE BIOLOGÍA MOLECULAR E INSTITUTO DE BIOMEDICINA Y BIOTECNOLOGÍA DE CANTABRIA, Universidad de Cantabria-CSIC-SODERCAN, C. Herrera Oria s/n,
39011 Santander, Spain
delacruz@unican.es

Abstract *Conjugation is a key mechanism of genetic exchange responsible for the spread of resistance, virulence and social traits among prokaryotes. We analyzed the phylogeny of key conjugation proteins to unravel the evolutionary history of conjugation and the associated type IV secretion systems (T4SS). We show that ssDNA and dsDNA conjugation, while both based on a key AAA$^+$ ATPase, arose independently. The two key ATPases of ssDNA conjugation are monophyletic, having diverged at an early stage from dsDNA translocases. Our data suggests that ssDNA conjugation arose first in diderm bacteria, possibly Proteobacteria, and then spread to other bacterial phyla, including bacterial monoderms and Archaea. Identifiable T4SS fall within 8 monophyletic groups, determined by both taxonomy and the structure of the cell envelope. Transfer to monoderms might have occurred only once, but followed diverse adaptive paths. Remarkably, some Firmicutes developed a new conjugation system based on an atypical relaxase and an ATPase derived from a dsDNA translocase. Analyses of the evolutionary rates and the patterns of presence/absence of specific T4SS proteins show that conjugation systems are often and independently exapted for other functions. The possibility of classing all kinds of conjugative systems together provides a natural base for their classification, a problem that is growing at least as fast as genomic databases. Our analysis provides the first global picture of the evolution of conjugation and shows how a self-transferrable complex multi-protein system has adapted to different taxa and often been domesticated by the host. As conjugation systems became specific to certain clades and cell envelopes, they may have biased the rate and direction of gene transfer within prokaryotes.*

Keywords Bacterial conjugation, horizontal gene transfer, type IV protein secretion, exaptation, plasmid evolution.

1 Introduction

Prokaryotic genomes adapt quickly to new environmental conditions largely because they can acquire pre-evolved traits by horizontal gene transfer (HGT) [1]. Conjugation is thought to be the most frequent mechanism of HGT because it allows single-event transfer of large DNA fragments, up to entire chromosomes, is not limited by homology to the recipient genome and has a broader host range than transduction or transformation [2]. Conjugation is also involved in the establishment of social processes, promoting biofilm formation and spreading of cooperative traits. There are two known modes of conjugation that differ both in the type of translocated DNA, ssDNA versus dsDNA, and in the complexity of the transport system. Both types of conjugative systems are either encoded by autonomously replicating plasmids or inserted in chromosomes as integrative conjugative elements (ICE). We recently made a large-scale identification of conjugation systems both in plasmids and in ICEs and found them to be essentially short-term variants of otherwise identical backbone elements [3].

Two major protein complexes are involved in ssDNA conjugation: relaxosomes and type IV secretion systems (T4SS, Figure 1). The relaxosome is composed of the relaxase (MOB) and often includes auxiliary proteins. It nicks the dsDNA and binds the resulting ssDNA at the origin of transfer. The diversity and evolution of the different families of relaxases has been extensively studied [4]. The essential type IV coupling protein (T4CP) binds the DNA-relaxase substrate and couples it to the T4SS, possibly using ATP to translocate the complex across the inner membrane. The majority of T4CPs belong to the VirD4$_{Ti}$ family, but

some T4SS were recently found to use a distantly related ATPase as T4CP ($TcpA_{pCW3}$) [5]. Translocation through the membrane of the donor cell to the cytoplasm of the recipient cell is done by the T4SS. In proteobacteria, the T4SS is a large protein complex including a ubiquitous ATPase ($VirB4_{Ti}$ or the distant homolog $TraU_{R64}$), mating-pair formation (MPF) proteins that form the transport channel, and a pilus that attaches to the recipient cell [6]. Four MPF families have been described in proteobacteria: MPF_T (based on the T-DNA conjugation system of *A. tumefaciens* plasmid Ti), MPF_F (based on plasmid F), MPF_I (based on the IncI plasmid R64) and MPF_G (based on ICEHIN1056) [7]. These four models describe all functionally studied and nearly all T4SS identified by bioinformatic methods among proteobacteria, both in plasmids and in chromosomes [3]. T4SS systems from Cyanobacteria, Bacteroides, Firmicutes, Actinobacteria and Archaea have homologs to VirB4 [3].

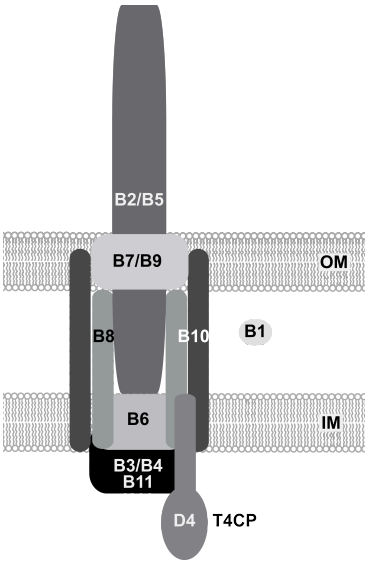


Figure 1. Scheme of the vir T4SS. The different vir proteins are depicted (VirB1 to VirB11) as well as the coupling protein (T4CP) VirD4 and the inner (IM) and outer membrane (OM).

Conjugation of dsDNA takes place in mycelia-producing Actinobacteria. It relies on a single protein: $TraB_{pSG5}$ that translocates dsDNA between neighboring cells in mycelia. This protein resembles, in sequence and function, the essential protein FtsK that segregates sister-chromosomes in the last stages of chromosomal replication. They are both members of the AAA^+ motor ATPase family, which also includes both types of T4CP (VirD4 and TcpA), and both types of essential T4SS ATPases (VirB4 and TraU). Hence, all key proteins of the dsDNA and ssDNA conjugation systems are evolutionarily related. This association has not yet been clarified from a phylogenetic point of view.

T4SS are also often recruited by bacterial pathogens to secrete protein effectors to eukaryotic cells [8]. These MOBless T4SS, so-called because they do not contain a relaxase gene, are closely related to conjugative systems. The extreme flexibility of T4SS has allowed at least two other types of exaptations, i.e. evolutionary events in which part of the pre-existing machinery of conjugation was co-opted for other functions. *Helicobacter pylori* genomes encode a MOBless T4SS that is used for natural transformation. It is essential to import environmental DNA. In *Neisseria gonorrhoeae*, a T4SS is responsible for DNA export to the extracellular space, an intermediate step in the process of natural transformation among these bacteria. A previous analysis of MPF_T systems suggests that exaptation events occurred several times in evolution [9]. Since we recently found that MOBless T4SS are significantly more abundant than previously thought [3], this point needs to be reassessed for MPF_T and developed for other MPF types.

While studies on conjugation are as old as molecular biology itself, several recent works have changed significantly our understanding of this process [3,5,6,10,11,12]. This succession of works opens the opportunity to infer a global scenario for the evolution of conjugative systems and T4SS, which is the goal of the present work.

2 Methods

2.1 Data

Data on complete chromosomes and plasmids of prokaryotes was taken from Genbank Refseq (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). This included 1,207 chromosomes, 891 plasmids that were sequenced along with these chromosomes and 1,391 plasmids that were sequenced independently. The information on T4SS was taken from [3].

2.2 Construction of protein profiles and genome searches.

Unless mentioned explicitly, the protein profiles used are those described in [3]. In order to study the presence/absence of the different component of the *vir* system, we made additional protein profiles, namely for VirB1, VirB2, VirB5, VirB7, VirB10, VirB11. We first used PSI-BLAST (e-value 0.1) to search for distant homologs using as query each of these genes from the VirB locus of the A. tumefaciens plasmid pTi SAKURA (Refseq entry NC_002147) and the abovementioned databank of completely sequenced replicons. Given the problems of convergence of PSI-BLAST when using complete genomes, and the extensive similarity of plasmid and chromosomal conjugative systems [3], we restricted homology searches to plasmid sequences when building protein profiles. We retrieved the proteins with hits for each protein family and built multiple alignments using MUSCLE. We used HMMER 3.0 to produce the HMM profiles and to perform searches within genomes. In the analysis of the evolution of the MPFT system we only considered the hits that co-localized with previously detected *vir* proteins (VirB3, VirB4, VirB6, VirB8, VirB9). FtsK proteins were retrieved directly by using the PFAM PF01580 profile. TraB proteins, being closely related to FtsK, were retrieved by BLASTP searches of TraB from Streptomyces plasmid pCQ3 (YP_003280879) on the Actinomycetales proteins from the Refseq database. We sampled the top results. We then built a protein profile for this protein and searched for its occurrences as for the other profiles. We built a web-server to allow running the protein profiles.

2.3 Phylogenetic analysis.

Unless explicitly stated, all phylogenetic analyses were performed with the following procedure. First, sequences were aligned using MUSCLE with default parameters and manual edition. Third, 100 replicate trees were built with RAxML 7.2.7 (Stamatakis 2006) using the model GTRGAMMA. We kept the one with the best likelihood. There were two exceptions to this method. We aligned the ATPases using MAFFT and did manual edition. We performed the phylogenetic inference as above and additionally with PhyML 3.0 under the LG model and with the bioNJ starting tree in order to get aLRT support values. The alignment of the set of VirB4 and VirD4 was built with MAFFT and then manually edited. MAFFT was used instead of MUSCLE because it provided better alignments in these cases. We used PhyML 3.0 to build the phylogenetic tree, under the LG model and with the bioNJ starting tree. aLRT support values were also calculated for each node.

3 Results and Discussion

3.1 Deep phylogeny of key conjugation ATPases

The two families of T4CPs (with prototypes given by the $VirD4_{pTi}$ and $TcpA_{pCW3}$), the two families of ATPases (based on $VirB4_{Ti}$ and $TraU_{R64}$), the dsDNA conjugation protein $TraB_{pSG5}$ and FtsK are all part of the super-family of AAA+ motor ATPases. Hence, we investigated the events at the onset of the natural history of conjugation from the analysis of the phylogeny linking homologs for all these protein profiles among 3,489 replicons (Figure 2). The tree was rooted using the distantly related protein family derived from $VirB11_{Ti}$. This phylogenetic reconstruction separated the VirD4/VirB4 clade from the others. This fits previous genomic and structural analysis showing the proximity between the dsDNA translocators FtsK and

TraB and between the ssDNA translocators VirD4 and VirB4. T4CPs and VirB4s show clear structural similarities underscoring a common functional mechanism. Overall, these analyses fit structural work suggesting that the common ancestor of the VirB4/VirD4 families consisted of a soluble protein engaged in polypeptide transport (as it's still the case in most studied VirB4 proteins).

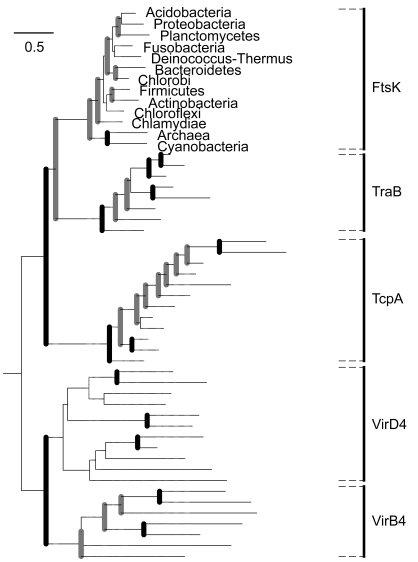


Figure 2. Phylogeny of AAA$^+$ ATPases associated with conjugation. Bold vertical black lines represent nodes with high support value (bootstrap > 66% and aLRT > 0.66). Bold grey lines represent nodes with high aLRT score (> 0.66) but a weaker bootstrap (=< 66%).

The other basal branch in the phylogeny includes TraB, TcpA and FtsK. The elements of the TraB family are found only in Actinobacteria and are closely related with FtsK, but they do not emerge from within the FtsK. Instead they derive independently from the ancestor of this protein. FtsK is an essential protein whose phylogeny follows approximately the one of bacteria and thus provides a guideline to the timing of the diversification of these protein families. This data does place the origin of ssDNA conjugation very early in the history of life. While TraB and TcpA seem to diversify after FtsK, in agreement with their presence only in Firmicutes and Actinobacteria, the diversification of the pair VirB4/VirD4 could be contemporaneous to that of FtsK.

3.2 T4SS phylogeny

We aligned the proteins matching the VirB4 and TraU profiles to infer the evolutionary history of all VirB4-homologs. We then used VirD4 to root this tree. This root is highly supported showing that these neighbor groups of proteins are monophyletic. The tree shows that all VirB4 and TraU-related proteins (over one thousand) can be classified in eight groups, which are represented by eight well-supported clades (Figure 3).

Four groups correspond to the different T4SS families of Proteobacteria obtained from the analysis of plasmids (MPF_F, MPF_G, MPF_I, MPF_T). These four groups are clearly separated because each contains a set of 4 to 9 genes that are specific, i.e. their protein profiles match loci of a given MPF but not those of the other MPF types [7]. The four remaining groups correspond to Cyanobacteria (MPF_C), Bacteroidetes (MPF_B) and two clades of monoderms (MPF_{FA} and MPF_{FATA}).

The trees of VirD4 and VirB4 are not congruent. Yet, they share many features. The proteins encoded by the *virD4* genes co-localizing in replicons with *virB4* tend to form similar clades. Notably, the VirD4

associated with each of six of the eight VirB4 clades also clustered in nearly monophyletic clades of T4CP (MPF_{FA}, MPF_{FATA}, MPF_B, MPF_G, MPF_I and MPF_C). VirD4 of the two remaining clades (MPF_T and MPF_F) are scattered in a small number of clades. Most of the MPF_{FA} use TcpA instead of a VirD4-like T4CP. The few VirD4 proteins found in MPF_{FA} are monophyletic and distant from any other group. It was previously shown that plasmid T4CP are sometimes scattered in different groups corresponding to given relaxases [7]. This result is still valid with the present much larger dataset. Hence, evolution of conjugation is driven by two main constraints, one acting mainly on the T4SS, represented by VirB4, and other on the relaxosome, represented by the relaxases. T4CP tend to co-evolve with both components.

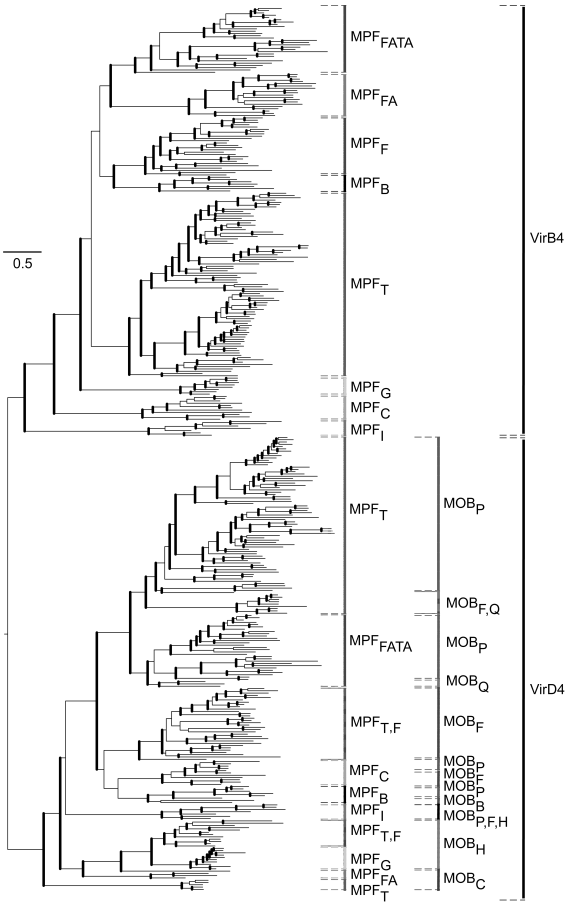


Figure 3. Joint phylogenetic reconstruction of the VirD4 and VirB4/TraU families of proteins from conjugative systems. Bold vertical lines represent nodes with a high support value (aLRT > 0.66).

3.3 Cell envelope-adaptation in monoderms

The most basal clades in both VirB4 and VirD4 phylogenies correspond to bacteria with both inner and outer membranes, i.e. diderms. As these nodes are well-supported, this strongly suggests that ssDNA conjugation was invented among diderms. In this scenario, ssDNA conjugation would have been acquired by monoderm prokaryotes, i.e. organisms devoid of an outer membrane, by horizontal gene transfer. This fits the observation of basal diderms but also the inclusion of all monoderm conjugation systems in two sister

clades: MPF_{FA} and MPF_{FATA}. The MPF_{FA} type includes Firmicutes and Actinobacteria, and the MPF_{FATA} also includes Tenericutes and Archaea.

Within the MPF_{FA} clade one of the two Firmicutes groups has TcpA as a putative T4CP. We found that 63% of the TcpA hits were co-localized with VirB4 in MPF_{FA} systems of Firmicutes and all 47 of these regions lacked a VirD4-like protein. This gives further credit to the hypothesis that TcpA is an alternative T4CP [5]. TcpA-associated systems are, with one single exception, also associated with MOB_T relaxase. Thus, although phylogenetically different, TcpA and VirD4 T4CPs seem to be both alternatives for ssDNA conjugation, suggesting the recruitment of a new dsDNA translocase to make ssDNA conjugation in this sub-clade of MPF_{FA}.

3.4 Evolution of MPF_T

Each MPF type contains a small number of proteins with homologues across other MPF types, e.g. VirB4 or VirD4. However, most protein profiles of a given MPF type match homologs only within the respective MPF systems. Several of these are nearly ubiquitous within a given MPF type and we have previously used them to class MPF types in plasmids and chromosomes [3,7]. To analyze in detail the patterns of presence and absence of MPF specific genes we analyzed the MPF_T system, the best-studied and most frequently found in sequenced genomes. Its prototype is the VirB system of the *A. tumefaciens* plasmid Ti, which encodes 11 genes: *virB1* to *virB11*. We built HMM profiles for each protein and used them to scan plasmids for homologs. Most of the systems include between 8 and 11 of the 11 genes, but not always the same genes are missing.

The names of the different *vir* genes correspond to their order within the prototype VirBTi system. This prototype gene order pattern (from 1 to 11 in ascending order) is conserved in a large fraction of the MPF_T. For almost all MPF_T loci, the order is strictly conserved for a core composed of virB2, virB3, virB4, virB8, virB9 and virB10. Importantly, the clusters of gene order in the tree reflect accurately the phylogeny of VirB4. This is further evidence that recombination of distant VirB4 variants rarely occurs, even within MPF types.

We made pairwise global alignments between the Ti plasmid VirB proteins and their homologs among MPF_T loci. This shows that some proteins evolve extremely fast, becoming undetectable by sequence analysis. Importantly, the proteins that are almost as frequently observed in MPF_T as VirB4 (VirB3, VirB9, VirB10, VirB11) are also the ones evolving more slowly. Inversely, the three proteins that were more often missed among MPF_T are also the ones evolving faster, i.e. VirB1, VirB5, VirB7.

These results suggest that some of the so-called MPF specific genes may have homologs in other systems, even if such homologs escape detection from our protein profiles because they evolve too rapidly. Since our protein profiles failed to uncover these similarities, we used BlastP hits of MPF_T proteins in regions flanking by 20 genes on each side of the VirB4 proteins of other MPF types. The proteins of the T4SS that are localized in the cytoplasmic and/or the inner membrane are among the most frequently hitting proteins in other systems: VirB3, VirB6 and VirB11. As expected, the monoderms show no hit to proteins locating at the outer membrane (VirB9, VirB7) or at the periplasm (VirB10). The substitution rates of VirB10 and VirB9 are almost as low as those of VirB4. Therefore these proteins are unlikely to be missed by sequence similarity and their absence from other systems is probably not caused by rapid loss of sequence similarity. Importantly, the two MPF types with fewer hits to MPF_T proteins are among the most distantly related to MPF_T (MPF_G and MPF_C). Their more distant ancestry may have resulted in significantly distinct structures (like for MPF_I), but we cannot exclude that homology may have simply become untraceable by sequence analysis.

3.5 T4SS domestication

We recently uncovered that a large fraction of T4SS lack neighboring relaxases [3]. These MOBless T4SS probably include some inactive elements. However, many of these elements lacked neighboring integrases, even though they were much more frequent in chromosomes than in plasmids. Furthermore, the T4SS that are known to secrete proteins to the eukaryotic cytoplasm fell in this group. These observations suggest that many MOBless T4SS are not undergoing degradation but that instead reflect frequent domestication of conjugation systems for other functions (exaptation). Since the T4CP is absent in some known protein export

systems, the only protein common to all experimentally verified T4SS is VirB4 [6]. The VirB4 phylogeny confirms that within MPF_T the loss of the relaxase occurred many times and that this pattern is also found among the other MPF types. Just like conjugative systems of ICEs and plasmids are interspersed in the phylogenetic trees [3], so MOBless T4SS are interspersed with conjugative systems. This shows that MOBless T4SS arose frequently and independently. The only exception concerns the Archaea and the Actinobacteria, for which the lack of known relaxases has been pointed out before [4]. In these clades it is likely that the abundance of MOBless T4SS reflects largely the presence of unknown relaxases. Importantly, when we mapped on the phylogenetic trees of MPF_T and MPF_F the T4SS that are experimentally known to have non-conjugation related functions, they are also interspersed in the tree. This suggests that conjugative T4SS have thus been "domesticated" and used to transport protein effectors or to export/import DNA to/from the extracellular milieu very frequently.

3.6 An evolution-based classification system for MPF

The lack of an all-encompassing classification scheme for conjugative systems and the extreme diverse gene nomenclature for homologous genes greatly and unnecessarily complicates the analysis of the literature of the domain. We suggest that the phylogeny of VirB4, the only ubiquitously recognizable protein of T4SS, could be used to class conjugative systems and other T4SS. This could be the foundation for much-needed gene name standardization in the literature and databases. The model systems of the VirB operon of *A. tumefaciens* Ti plasmid (MPF_T), of the F plasmid (MPF_F), of the R64 plasmid (MPF_I) and of ICEHin1056 (MPF_G) could be used for all Proteobacteria and possibly for other diderm clades such as Acidobacteria. Four other MPF types cover for now the diversity of all the other systems in so far as the VirB4 phylogeny is concerned. These would include a type that for the moment only includes Bacteroides (MPF_B) and another that includes only Cyanobacteria (MPF_C). The classification would also include the two types that are specific to monoderms, the MPF_{FA} and MPF_{FATA}. The MPF_{FA} type, given its heterogeneity in the use of T4CP might be split into two groups when more is known about the differences in the biochemistry of conjugation in the group. The advantage of this classification is that it is based on evolutionary biology, tends to reflect similarity between elements and can be done even when one knows yet relatively little of the biochemistry of the elements being classed.

This classification system can be applied to partial data, e.g. from metagenomics, because it requires the identification of a single gene.

4 Conclusions

Our work provides a scenario for the evolution of conjugation and T4SS from their origin to recent exaptations. These results suggest that conjugation is a very ancient process that arose in two independent ways for ssDNA and dsDNA mechanisms but starting from closely related AAA+ ATPases involved in DNA translocation. Conjugation of ssDNA is by far the best studied and also the mechanism most frequently found in prokaryotes. It probably appeared very early among bacteria with two cell envelopes, possibly ancient proteobacteria, and from there it spread to all clades of prokaryotes. The T4SS of monoderms seem simpler, in that they involve fewer genes, and could evolve by gene deletion from the larger T4SS of diderms. This evolutionary scenario links together all known ssDNA conjugative systems, and their T4SS, by the common ancestry of VirB4. Several observations show the validity of the use of this protein for the classification of T4SS. First, it is the only ubiquitous protein in T4SS. Second, its phylogeny closely matches those of other conserved proteins, notably the VirD4. Third, patterns of presence/absence of MPF specific genes match the VirB4 phylogeny. Fourth, the order of MPF-specific genes, at least in MPF_T, also matches the VirB4 phylogeny.

The structure of the VirB4 tree, with its robust separation in 8 large clades reflects in part an effect of the cell envelope. Indeed, once systems arose within a clade with a peculiar membrane structure, they tended to adapt to this cell structure and were not further passed on to other clades. This lead to large clades of VirB4 including monoderms - such as Archaea or Firmicutes - or diderms with peculiar membrane compositions such as Cyanobacteria or Bacteroides. Adaptation of the T4SS to such cell envelopes is likely to increase the efficiency of conjugation within taxa but at the cost of reducing its efficiency between taxa, effectively leading to T4SS specialization. This process has the potential to bias the rate and direction of genetic transfer

between prokaryotes and thus shape the networks of gene sharing. Notably, it might contribute to the observed coherence between high bacterial taxonomic ranks.

Surprisingly, our work shows that one group of ssDNA T4SS has radically changed into a system with a new T4CP (TcpA) and relaxase (MOB_T). While the cognate VirB4 protein fits clearly in our T4SS classification, and is presumably representative of the remaining T4SS, this suggests that the evolution of the coupling protein can in certain cases differ radically from the one of the T4SS. In several cases this leads to an evolution of the T4CP constrained by both the T4SS and the relaxases. However, in this case the T4CP change was dramatic since it involved the recruitment of a distantly related ATPase more closely related to dsDNA translocases. In line with these findings, our work also shows that exaptations of T4SS can occur frequently in the evolutionary history. Conjugation consists in the secretion of a nucleoprotein complex. Passing from this function to a purely protein secretion system is probably simple. Accordingly, several systems are known to transfer both proteins and relaxosomes [12]. Several other protein secretion systems are thought to be exaptations, e.g. T3SS are related with the bacterium flagellum and T6SS show structural homologies with phages. Yet, T4SS present an uncommon case in that exaptations occurred multiple times in the evolutionary history. Given the present results, it is not unlikely that novel exaptations, e.g. protein transfer among bacteria, are present among the poorly studied MOBless T4SS of free-living bacteria.

Acknowledgments

This work was supported by an ERC starting grant (EVOMOBILOME n°281605), the CNRS and the Institut Pasteur.

References

[1] H. Ochman, E. Lerat, V. Daubin, Examining bacterial species under the specter of gene transfer and exchange, *Proc. Natl. Acad. Sci. U.S.A.* 102 Suppl 1:6595-6599, 2005.

[2] C.F. Amabile-Cuevas, M.E. Chicurel. Bacterial plasmids and gene flux, *Cell* 70:189-199, 1992.

[3] J. Guglielmini, L. Quintais, M.P. Garcillàn-Barcia, F. de la Cruz, E.P.C. Rocha, The repertoire of ICE in prokaryotes undescores the unity, diversity, and ubiquity of conjugation, *PLoS Genet.*, 7(8): e1002222, 2011.

[4] M.P. Garcillàn-Barcia, M.V. Francia, F. de la Cruz. The diversity of conjugative relaxases and its application in plasmid classification. *FEMS Microbiol Rev* 33:657-687, 2009.

[5] J.A. Parsons, T.L. Bannam, R.J. Devenish, J.I. Rood, TcpA, an FtsK/SpoIIIE homolog, is essential for transfer of the conjugative plasmid pCW3 in Clostridium perfringens. *J Bacteriol* 189:7782-7790, 2007.

[6] C.E. Alvarez-Martinez, P.J. Christie, Biological diversity of prokaryotic type IV secretion systems. *Microbiol Mol Biol Rev* 73:775-808, 2009.

[7] C. Smillie, M.P. Garcillàn-Barcia, M.V. Francia, E.P.C. Rocha, F. de la Cruz. Mobility of plasmids. *Microbiol. Mol. Biol. Rev.* 74:434-452, 2010.

[8] E. Cascales, P.J. Christie, The versatile bacterial type IV secretion systems, *Nat. Rev. Microbiol.,* 1:137-149, 2003.

[9] A.C. Frank, C.M. Alsmark, M. Thollesson, S.G. Andersson, Functional divergence and horizontal transfer of type IV secretion systems. Mol Biol Evol 22:1325-1336, 2005.

[10] M. Juhas, D.W. Crook, I.D. Dimopoulou, G. Lunter, R.M. Harding, D.J. Ferguson, D.W. Hood, Novel type IV secretion system involved in propagation of genomic islands. *J Bacteriol* 189:761-771, 2007.

[11] R.A. Wozniak, M.K. Waldor, Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. *Nat Rev Microbiol* 8:552-563, 2010.

[12] E. Fernandez-Gonzalez, H.D. de Paz, A. Alperi, L. Agundez, M. Faustmann, F.J. Sangari, C. Dehio, M. Llosa, Transfer of R388 derivatives by a pathogenesis-associated type IV secretion system into both bacteria and human cells. *J Bacteriol*, 2011.

EvoluCode: an original view of Human Systems Evolution

Evolutionary Inference in Biological Networks

Benjamin LINARD[1], Olivier POCH[1] and Julie Dawn THOMPSON[1]

[1] LABORATOIRE DE BIOLOGIE ET GENOMIQUE INTEGRATIVE, Département de Biologie et Génomique Structurales CNRS/INSERM/UDS, Institut de Génétique et de Biologie Moléculaire et Cellulaire, 1 rue Laurent Fries, 67404, Illkirch, Cedex, France

{benjamin.linard, olivier.poch, julie.thompson, @igbmc.fr

Abstract *Evolutionary systems biology aims to uncover the general trends and principles governing the evolution of biological networks. An essential part of this process is the reconstruction and analysis of the evolutionary histories of the networks. Unfortunately, the formalisms for representing such complex evolutionary histories are currently limited and are not suitable for large scale, genome-wide studies. In this context, we have developed a new formalism, called EvoluCode (Evolutionary barCode), which allows the integration of different evolutionary parameters in a unifying format and facilitates the multilevel analysis and visualization of complex evolutionary histories at the genome scale. The applicability of the approach has been demonstrated by constructing barcodes representing the evolution of the complete human proteome. Various large-scale analyses have been performed: (i) the mapping and visualization of the barcodes on the human chromosomes (ii) clustering of the barcodes to highlight protein subsets sharing similar evolutionary histories and their functional analysis or (iii) the mapping of the EvoluCodes on human metabolic pathways, thus allowing the identification of specialized sub-networks. EvoluCode opens the way to the efficient application of other data mining and knowledge extraction techniques in evolutionary systems biology studies.*

Keywords Evolutionary Barcode, Evolutionary Analysis, Human Proteome, Biological Pathways.

EvoluCode: an original view of Human Systems Evolution

Evolutionary Inference In Biological Networks

Résumé *La biologie évolutive systémique vise à découvrir les tendances générales et les principes qui régissent l'évolution des réseaux biologiques. Une partie essentielle de ce processus est la reconstruction et l'analyse des histoires évolutives des réseaux. Malheureusement, les formalismes de représentation de ces histoires évolutives complexes sont actuellement limitées et ne conviennent pas à des analyses à l'échelle du génome. Dans ce contexte, nous avons développé un nouveau formalisme, appelé EvoluCode, qui permet l'intégration des différents paramètres de l'évolution dans un format fédérateur et facilite l'analyse multi-niveaux et la visualisation des complexes des histoires évolutives à l'échelle du génome. L'applicabilité de l'approche a été démontrée par la construction de codes à barres représentant l'évolution du protéome humain. Diverses analyses à grande échelle ont été réalisées: (i) la cartographie et de visualisation des codes à barres sur les chromosomes humains (ii) le clustering des codes à barres pour mettre en évidence des sous-ensembles de protéines partageant des histoires évolutives semblables et leur analyse fonctionnelle ou (iii) l'analyse des EvoluCodes sur les différents réseaux biologiques humains, permettant ainsi l'identification des comportements des gènes impliqués dans ces réseaux. Les EvoluCodes ouvrent la voie à l'exploration et l'application de nouvelles techniques d'extraction de connaissances pour l'étude des systèmes biologiques.*

Mots-clés *Code-barre évolutif, biologie évolutive systémique, protéome humain, réseaux biologiques*

1 Introduction

Systems biology aims to understand the structure and dynamic behavior of complex biological systems by modeling the components and their interactions at different functional levels [1]. Such a comprehensive understanding requires the integration of large-scale experimental data with computational analyses and mathematical modeling approaches [2]. In particular, successful systems biology will rely on our ability to integrate different types of multi-scale data across various levels of complexity [3] from individual molecules such as proteins, metabolites, etc. to cells, tissues, organisms or even ecosystems. These different levels are now being described by the large volumes of experimental data resulting from genomics technologies such as next-generation sequencing, transcriptomics, interactomics, etc.

In this context, the field of evolutionary systems biology aims to combine the modeling aspects of current systems biology with the long-standing quantitative experience in evolutionary genetics in order to uncover the general trends and principles underlying the evolution and function of complex biological networks [4]. Evolutionary based inference provides an incredibly powerful tool for comparing multiple sources of data, since features that are maintained in several organisms tend to be functionally important while variations or differences may indicate key innovations.

A number of groups have performed genome-scale studies aimed at investigating the potential correlations between variables characterizing different aspects of protein network functions and evolution [5,6,7]. While these studies were limited to the correlations observed between two variables, others have attempted to compile more diverse sets of evolutionary variables. Thus, principal component analysis was used to investigate the relationships between seven genome related variables, identifying three main axes reflecting a gene's "importance", "plasticity" and "adaptability" [8]. Although these studies have revealed several interesting trends, new standardized methodologies and tools are now needed that allow the integration of larger, more diverse sets of multilevel data and efficient, quantitative analyses at the genome scale. Similarly, despite some attempts to develop tools providing global overviews of complex evolutionary scenarios [9] original visualization tools will be required to facilitate rapid identification of specific behaviors.

2 EvoluCodes : evolutionary barcodes

We have developed a novel formalism, called EvoluCode [10], or the Evolutionary Barcode, which allows the integration of different data types in a unifying framework. Thus, a barcode is assigned to each component in a biological system and diverse evolutionary parameters from different biological levels can be incorporated, facilitating multi-scale evolutionary analyses. Visualization tools have also been developed to allow the human expert to view the barcodes and to identify interesting patterns in both low and high throughput studies.

In order to evaluate the pertinence of the evolutionary barcodes and to test their ability to represent complex evolutionary histories, we constructed evolutionary barcodes (figure 1) for the complete proteomes of 17 vertebrate species. In this context, we incorporated a number of different evolutionary variables, including primary sequence data, genome neighborhood and evolutionary conservation, but the barcode formalism can be easily extended to incorporate other variables representing different biological features.

. The columns of the matrix correspond to the studied organisms, which in this work consist of 17 vertebrates with almost complete genomes from the Ensembl database (version 51). The rows of the matrix correspond to different evolutionary parameters (Table 1) that were extracted from multiple alignments, synteny analysis and orthology data. For each vertebrate organism, the most closely related homolog to the human reference gene was identified (based on percent residue identity) and 10 parameters were calculated (table 1).

Parameter name	Description	Source
length	length of the vertebrate sequence	Multiple alignment
length_difference	difference in length between the human reference and vertebrate sequences	Multiple alignment
no_of_regions	number of conserved regions shared between the human reference and vertebrate sequences	Multiple alignment
sequence_identity	percent residue identity shared between the human reference and vertebrate sequences	Multiple alignment
no_of_domains	number of known protein domains (from the Pfam database) in the vertebrate sequence	Multiple alignment
domain_conservation	parameter indicating domain structure conservation between the human reference and vertebrate sequences: unchanged domain structure/domain gains/domain losses/domain shuffling	Multiple alignment
hydrophilicity	average hydrophilicity of the vertebrate sequence	Multiple alignment
inparalog	number of human inparalogs with respect to the vertebrate species, representing duplicability of a human gene compared to the other species	Ortholog/paralog database
co-ortholog	number of co-orthologs in the vertebrate species with respect to human, reflecting gene duplications in the non human lineage	Ortholog/paralog database
synteny	parameter indicating conservation of genome neighborhood: synteny on both sides of the gene / synteny either downstream or upstream of the gene / no synteny	Synteny database

Table 1. Evolutionary parameters included in EvoluCodes of Human Proteome.

To facilitate visualization of the EvoluCode, a color is assigned to each matrix cell representing typical or atypical parameter values. To do this, the distribution of each parameter in each organism is described and color gradients are assigned to three intervals:

- Interval 1 represents values that are lower than what is generally observed for a specific parameter in a specific organism and is assigned a blue-to-green gradient
- Interval 2 represents values that correspond to what is generally observed for a specific parameter in a specific organism and is assigned a green color
- Interval 3 represents values that are higher than what is generally observed for a specific parameter in a specific organism and is assigned a green-to-red gradient.

At this stage, the values of the barcode parameters are normalized to allow quantitative analyses and automatic comparisons, using standard data mining techniques such as clustering or classification. We show that, in addition to highlighting general evolutionary trends, the barcodes facilitate the identification of specific evolutionary histories, such as strict conservations or significant gene family expansions.

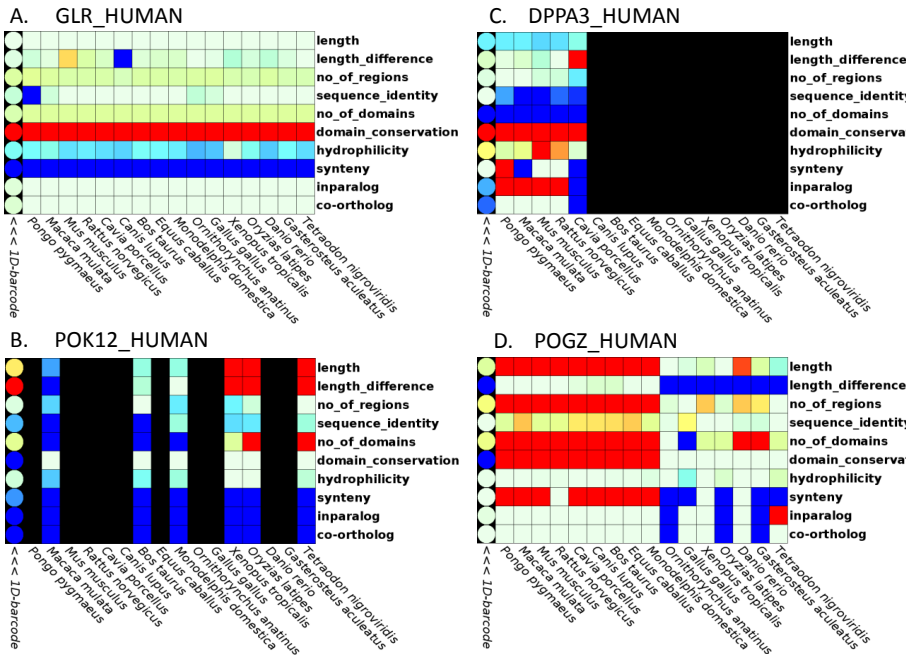


Figure 1. Several examples of evolutionary barcodes (EvoluCodes). To each human protein-coding gene is associated 1 barcode representing it's evolutionary history. Different barcodes profiles describe different histories.

3 High-throughput analysis of evolutionary histories

To estimate the reliability of our EvoluCodes and to test their ability to highlight evolutionary messages in high-throughput studies, we used several approaches. First, we mapped all barcodes to human chromosomes to confront evolutionary history and chromosomal positions of genes. Second, we clustered genes with similar evolutionary histories and designed a functional analysis of resulting clusters. Finally, to reach a systemic level in our analysis, we mapped our EvoluCodes to all human biological pathways.

3.1 Chromosomal mapping

We mapped the human proteome Evolucodes to the 24 human chromosomes, resulting in a barcode map of the complete genome. The visual inspection of this map allowed us to distinguish several previously published gene clusters. One example illustrated in figure 2 is the case of the keratin I gene clusters, separated by a keratin associated proteins (KRAP) cluster. Keratin clusters and KRAP cluster splitting the keratin cluster in two parts with very different EvoluCodes profiles.

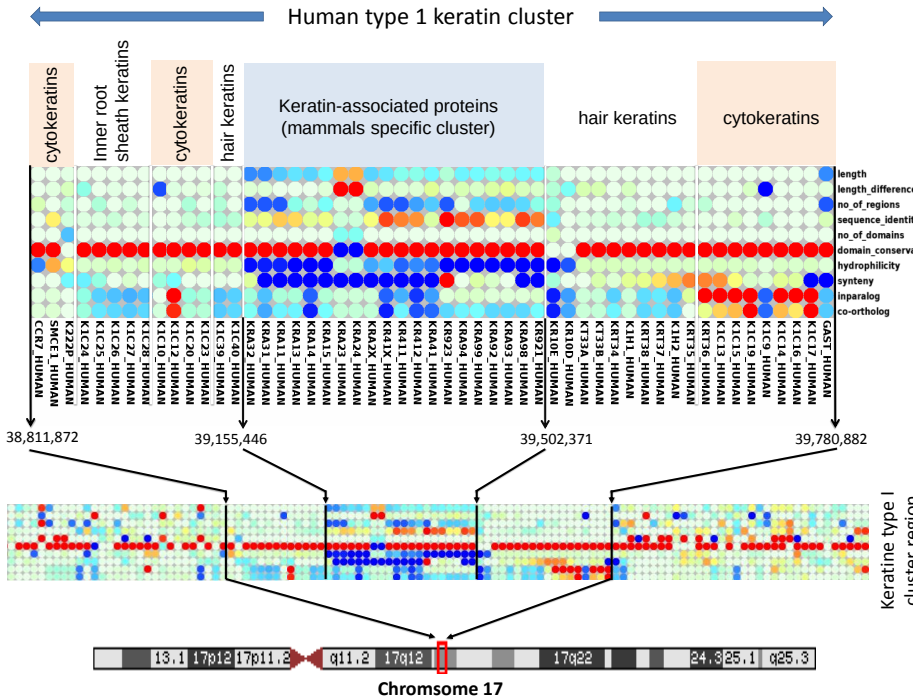


Figure 2. Evolutionary histories described by Evolucodes can be linked to genomic clusters of genes. Several keratin subfamilies are delimited by white vertical lines. The boundaries of the keratin cluster are delimited by black arrows.

The different Evolucode profiles of KRAP and Keratin clusters suggest original evolutionary histories for both. Indeed, barcode parameter analysis and several previously published studies confirmed this hypothesis. Similarly, many chromosomal gene clusters highlighted by their similar evolutionary histories were identified in all human chromosomes.

3.2 Clustering similar evolutionary histories

We identified subsets of genes that share similar barcodes, ie, similar evolutionary histories, among the whole set of 19,788 human genes. Using the Potts clustering model, 303 clusters were generated with a maximum cluster size of 380 proteins. To investigate the potential functional significance of these barcode clusters, we performed a GO enrichment analysis of the generated clusters. Surprisingly, most clusters were functionally enriched, indicating a clear linked between evolutionary histories and gene functions/localization.

3.3 Biological pathway analysis

To extend our analysis from gene sets to biological systems, we mapped our EvoluCodes to all human biological pathways defined in the Kegg database. For each pathway, we identified “outlier” EvoluCodes, i.e. barcodes corresponding to genes associated with peculiar evolutionary histories compared to other genes implicated in the pathway. As a preliminary analysis, we analysed the topological localization of these outliers in human metabolic pathways. We found that 10% of the genes implicated in such ancient pathways

have original evolutionary histories during vertebrate evolution (figure 3). Moreover, we observed that these genes are mostly located at interfaces between metabolic pathways or steps where several genes catalyse the same reaction.

		# of outlier reactions	% of reactions	% of outliers
A	Interface between pathways	18	5.31%	38%
B	Various enzymes for same reaction	17	5.01%	36%
C	Last reaction of a path	6	1.77%	13%
D	Parallel pathways	5	1.47%	11%
E	Atypical cascade	1	0.29%	2%
	Total	47	13.86%	100%

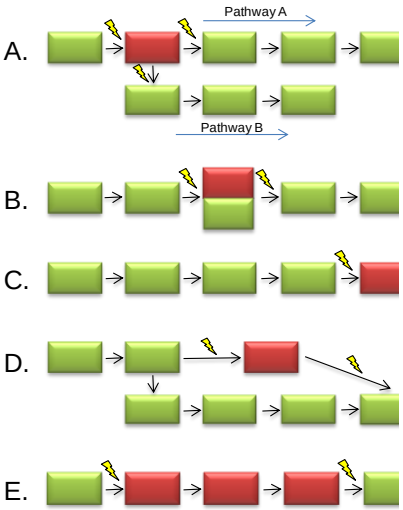


Figure 3. Outlier EvoluCodes for 23 human metabolic pathways have been classified into 5 categories. These categories correspond to different pathway branching profiles.

Our current work is aimed at extending this analysis to a cross pathway analysis. Indeed, a single gene can be considered as original (from an evolutionary point of view) in some biological systems but not others. Our goal is to observe general behaviors of genes over all biological pathways, in order to identify genes with unexpected evolutionary behavior in a systemic context.

Acknowledgements

We would like to thank Odile Lecompte for stimulating discussions, Huan Nguyen, Raymond Ripp and Laetitia Poidevin for help with database management and Nicolas Wicker and Alejandro Murua for help with the Potts Model clustering. The work was performed within the framework of the Decrypthon program, co-funded by Association Française contre les Myopathies (AFM), IBM and Centre National de la Recherche Scientifique (CNRS). We acknowledge financial support from the ANR (EvolHHuPro: BLAN07-1-198915 and Puzzle-Fit: 09-PIRI-0018-02) and Institute funds from the CNRS, INSERM, and the Université de Strasbourg.

References

[1] Kitano H. Systems biology: a brief overview. *Science*. Mar 1, 2002;295 (5560):1662–4.

[2] Kohl P, Noble D. Systems biology and the virtual physiological human. *Molecular Systems Biology*. Jul 2009;5.

[3] Hoehndorf R, Dumontier M, Gennari JH, et al. Integrating systems biology models and biomedical ontologies. *Bmc Systems Biology*. Aug 11, 2011;5.

[4] Loewe L. A framework for evolutionary systems biology. *Bmc Systems Biology*. 2009;3:27.

[5] Knight CG, Pinney JW. Making the right connections: biological networks in the light of evolution. *Bioessays*. Oct 2009;31(10):1080–90.

[6] Koonin EV, Wolf YI. Evolutionary systems biology: links between gene evolution and function. *Current Opinion in Biotechnology*. Oct 2006;17(5):481–7.

[7] Herbeck JT, Wall DP. Converging on a general model of protein evolution. *Trends Biotechnol*. Oct 2005;23(10):485–7.

[8] Wolf YI, Carmel L, Koonin EV. Unifying measures of gene function and evolution. *Proc Biol Sci*. Jun 22, 2006;273(1593):1507–15.

[9] Procter JB, Thompson J, Letunic I, Creevey C, Jossinet F, Barton GJ. Visualization of multiple alignments, phylogenies and gene family evolution. *Nat Methods*. Mar 2010;7(3 Suppl):S16–25.

[10] Linard, B., et al., EvoluCode: Evolutionary Barcodes as a Unifying Framework for Multilevel Evolutionary Data. Evol Bioinform Online, 2012. **8**: p. 61-77.

Haplotype-based method for detecting regions under selection in the domestic dog

Hillel Jean-Baptiste-Adolphe[1], Mathieu Emily[2,3], Amaury Vaysse[1], Catherine André[1] and Christophe Hitte[1]

[1] CNRS UMR6290, IGDR, Université Rennes1, 2 avenue du Professeur Léon Bernard, CS 34317, 35043 Rennes Cedex, France

{hillel.jba,amaury.vaysse}@gmail.com, {Catherine.andre,christophe.hitte}@univ-rennes1.fr

[2] Université Rennes II, Place du recteur Henri Le Moal, CS 24037, 35043 RENNES Cedex, France

mathieu.emily@univ-rennes2.fr

[3] IRMAR, UMR6625 CNRS, Campus de Beaulieu, 263 avenue du Général Leclerc, 35042 RENNES Cedex

Abstract *The domestic dog is a powerful genetic model for studying morphological traits and for its predisposing to genetics diseases comparable to human genetics diseases. Studying chromosomal regions under selection in the domestic dog is an important issue to better understand biological mechanisms underlying the emergence of phenotypes. Here, we propose a new approach to detect regions under selection based on the detection of specific haplotype profiles for breeds. Our method has been tested on a set of known regions in which the association between a trait and the causative gene has been identified for some breeds. In addition, the application of our method to discover regions with a genetic signature in a given breed (Weimaraner) at the haplotype level has confirmed the presence of selection at the haplotype level.*

Keywords Selection, Haplotype, Hierachical clustering, Dog.

Détection de régions sous sélection pour l'espèce canine par une méthode d'haplotypes

Résumé *L'espèce canine est un modèle génétique puissant pour l'étude de traits morphologiques et pour la forte prédisposition raciale aux maladies génétiques semblables à celles connues chez l'Homme. L'étude des régions sous sélection chez le chien domestique est donc un enjeu important pour comprendre certains mécanismes biologiques associés au développement de phénotypes. Nous proposons une nouvelle approche de détection de régions sous sélection qui s'appuie sur la mise en évidence de profils d'haplotypes spécifiques à une ou plusieurs races. Notre méthode a été testée sur des régions de référence pour lesquelles l'association entre un trait et le gène responsable a été démontrée expérimentalement. De plus l'application de notre méthode pour identifier les régions du génome comportant une signature génétique d'une race donnée (Braque de Weimar) a permis de confirmer le caractère spécifique de certaines régions.*

Mots-clés Sélection, Haplotypes, Classification hiérarchique, Chien.

1 Introduction

L'espèce canine est caractérisée par une formidable diversité phénotypique acquise depuis sa domestication il y a environ 15000 ans et par une intense sélection pratiquée par l'Homme depuis quelques siècles. A ce titre, elle constitue un modèle génétique particulièrement puissant pour l'étude de traits morphologiques (taille, poids, couleur et texture du pelage, etc.) et pour la forte prédisposition raciale aux maladies génétiques semblable aux maladies chez l'Homme (cancer, neuropathie, myopathie etc.).

Dans une récente étude, nous avons identifié près de 2000 régions du génome de différentiation allélique qui constituent autant de régions candidates associées aux différences phénotypiques qui existent entre races canines [1]. Dans cette étude, l'ensemble du génome canin a été analysé à l'aide d'un indice de sélection s'appuyant sur une mesure de différentiation génétique liée à la variance de la fréquence allélique entre populations : l'indice Fst. Cet indice a été calculé pour chacun des 170000 marqueurs de polymorphisme de

type SNP (Single Nucleotide Polymorphism) répartis le long du génome et analysés sur près de 500 chiens de 30 races. Pour chaque paire de race, nous avons dérivé de la valeur 'Fst' une métrique 'di' qui est une fonction des valeurs de Fst par paire entre une race 'i' et l'ensemble des autres races telle que : $di = \sum_j (F_{st}^{ij} - E(F_{st}^{ij}))/sd(F_{st}^{ij})$ où $E(F_{st}^{ij})$ et $sd(F_{st}^{ij})$ représentent respectivement la valeur attendue et l'écart-type du F_{st} entre les races i et j calculée à partir de la totalité des SNPs. La métrique 'di' mesure la valeur de la variance des fréquences alléliques de chaque SNP afin de différencier chaque race de l'ensemble des autres. Dans un second temps, les SNPs de fort indice 'di' et contigus ont été regroupés en fenêtres de ~200 kb et ont permis d'établir un répertoire de régions candidates à la sélection des races canines.

Les résultats obtenus par cette méthode ont permis d'établir une liste de régions candidates à la sélection, mais bien qu'efficace, cette méthode connaît certaines limitations. Tout d'abord, le phénomène de sélection a pour effet de réduire l'hétérogénéité de la région sélectionnée. Ainsi une région sous sélection a tendance à avoir un taux de conservation plus fort d'un individu à un autre au sein de la race d'intérêt. Etant donné le fort niveau d'homozygotie génétique au sein des races canines, il est attendu que les régions sous forte sélection aient fixé un nombre important de SNPs sur de grands segments, et donc produit peu de combinatoires alléliques par chromosome, appelées haplotypes. D'autre part, l'approche que nous avons utilisée dans Vaysse *et al.* [1] ne permet pas de détecter des régions pour lesquelles la sélection serait orientée. Dans le cas d'un trait continu, comme la taille, il est intéressant de pouvoir détecter des régions qui permettent de différencier les chiens de grandes tailles des chiens de petites tailles. Il est probable que pour ce type de régions de sélection, les races de petite et de grande taille se différencieront de l'ensemble des autres races sur les haplotypes les plus représentés au sein de chaque race.

Pour répondre aux limites posées par les méthodes existantes, nous proposons dans cette étude une nouvelle approche de détection de régions de différentiation allélique du génome en s'appuyant directement sur l'analyse des haplotypes. La méthode se décompose en trois étapes principales :

1. Inférer les haplotypes à partir des données de génotypes (phasage) issues de >10 individus non apparentés par races, pour une région donnée.
2. Regrouper les haplotypes en haplogroupes. Cette étape réunit les haplotypes caractérisant l'individu en une classe d'haplotypes donc d'individus qui ont une ascendance commune. Cette étape permet de réduire le poids des haplotypes rares pouvant être liés aux erreurs de génotypages et de phasage préalables.
3. Estimer les composantes principales de la matrice de contingence par une analyse en composantes factorielles (AFC). Une classification hiérarchique est ensuite appliquée sur les composantes principales pour extraire des groupes de races de chiens partageant des profils haplotypiques similaires.

La méthode proposée a été testée sur deux jeux de données caractéristiques. Le premier est composé de régions « benchmarks » pour lesquelles un trait phénotypique a été validé et associé expérimentalement à une mutation de type SNP et à une ou plusieurs races spécifiques. Le second jeu de données est constitué de régions précédemment présélectionnées [1] spécifiques de la race Braque de Weimar pour laquelle nous disposons du plus fort effectif d'individus (n=26).

Dans une première partie, nous décrirons les jeux de données utilisées puis nous présenterons les détails des trois étapes principales de notre méthode. Dans une seconde partie, nous présenterons les résultats obtenus sur les jeux de données traités. Puis nous discutons les résultats obtenus par la méthode que nous proposons.

2 Méthodes

2.1 Description des jeux de données

Le jeu de données utilisé, généré dans le cadre du consortium européen LUPA de génétique du chien, est constitué de 446 chiens répartis dans 30 races. Le génotype de chaque chien a été obtenu à l'aide d'une puce Illumina comportant 170K SNP répartis tous les 15 kb. Les données complètes ainsi que la nomenclature des races utilisées sont disponibles à http://dogs.genouest.org/SWEEP.dir/Supplemental.html.

Dans un premier temps, nous avons extrait 7 régions chromosomiques impliquées dans la variabilité phénotypique entre races canines dont la relation génotype/phénotype est connue. Parmi ces 7 régions, celle contenant le gène RSPO2 situé sur le chromosome 13, présente une insertion de 167 paires de bases associée au phénotype « pelage fourni » (sourcils et barbe fournis) [2], celle contenant le gène HAS2 associé aux phénotype « peau plissée » de la race Shar-pei. Ces régions (Table 1) nous servent de jeu de données de référence pour évaluer la méthode proposée.

Chromosome	Fin	Début	Trait	Races spécifiques
1	59080935	59280935	Brachycéphalie	EBD
13	11595558	11795558	Pelage Fourni	JRT_Bot_StP_IrW_TYo
13	23256842	23456842	Peau plissé	ShP
13	23654139	23854139	Syndrome fièvre périodique	ShP
16	61860252	62060252	Robe noire	NFd
18	23333636	23533636	Chondrodysplasie	Dac_TYo
27	5444094	5644094	Poil frisé	StP

Table 1. Liste des régions chromosomiques avec une relation génotype/phénotype avérée.

Dans un second temps, nous nous sommes intéressés aux régions décrites comme étant sous sélection sans la connaissance du phénotype associé. Nous avons extrait les 17 régions de différentiation allélique détectées dans la race Braque de Weimar (Weimar) situées sur le chromosome 1. Parmi ces 17 régions, 13 sont partagées avec d'autres races et 4 sont spécifiques du Weimar.

2.2 Analyse statistique

Considérons une région chromosomique définie par un numéro de chromosome, une position de départ et une position de fin. Une région correspond donc à une suite de bases nucléoniques adjacentes. Chaque région est analysée de façon indépendante. La première étape de l'analyse consiste à estimer les haplotypes pour chacune des régions. Pour cette étape nous considérons une région de 200 kb centrée sur le SNP associé ou responsable du phénotype ou définie par rapport au centre des régions détectées dans Vaysse et al 2011. L'estimation des haplotypes a été réalisée au sein de chaque race par le programme fastPHASE [3].

L'ensemble des haplotypes a ensuite été groupé en haplogroupes. Pour cette seconde étape, nous avons calculé le nombre de sites de ségrégation entre chaque paire d'haplotypes, générant ainsi une matrice de distance entre haplotypes. Une classification hiérarchique, utilisant la méthode de Ward, est ensuite appliquée à cette matrice de distance, permettant la construction d'un dendrogramme. Enfin, les haplogroupes sont déterminés en coupant horizontalement le dendrogramme pour obtenir le nombre de groupes fixés *a priori.* La génération d'haplogroupes permet d'éliminer certaines erreurs de phasage ou de génotypages. De plus, cette seconde étape permet de filtrer les haplotypes peu fréquents source de biais dans l'analyse.

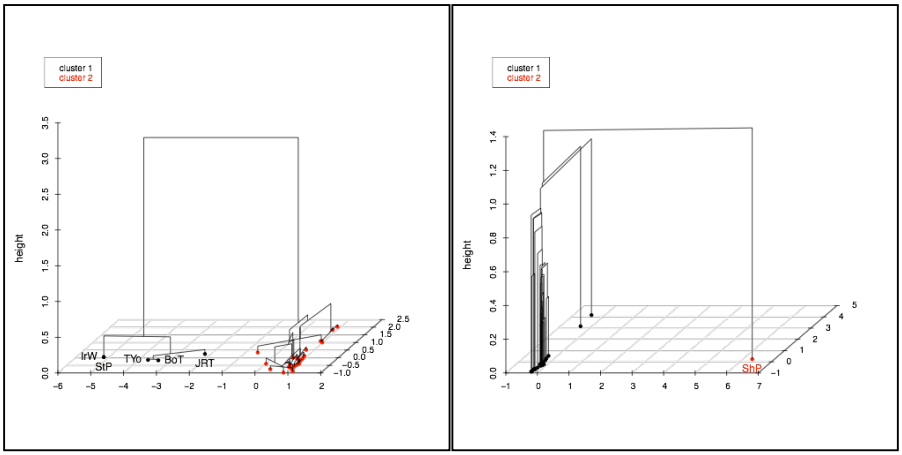
A l'issue du phasage des données en haplotypes, nous avons généré des matrices de contingence qui compte le nombre d'occurrences d'un haplotype donné pour une race donnée. A partir de la matrice de contingence, nous avons réalisé une analyse factorielle des correspondances (AFC) qui permet de d'extraire les composantes principales des données brutes. Puis une classification hiérarchique a été réalisée sur les composantes principales permettant ainsi de regrouper les races sur la base de profils haplotypiques similaires. A l'inverse, les races ayant des profils haplotypiques différents vont clusteriser dans des groupes distincts. Cette phase d'analyse de données a été réalisée à l'aide du package FactoMineR du logiciel R [4].

3 Résultats

3.1 Régions validées

Parmi les 7 régions analysées, 5 caractérisent un trait spécifique à une seule race, une caractérise un trait spécifique (la chondrodysplasie) à 2 races et une correspond à un trait partagé entre 5 races (le pelage

fourni). La méthode de classification hiérarchique des haplogroupes a permis de détecter les races présentant le phénotype d'intérêt pour les 7 régions témoins étudiées. La figure 1 représente les dendrogrammes de classifications hiérarchiques pour la région associée au trait 'pelage fourni' spécifique de cinq races et pour la région responsable de la peau plissée spécifique d'une race le Shar-pei. Ces résultats démontrent la séparation entre les profils haplotypiques associés au trait d'intérêt et les profils haplotypiques des autres races comme illustré dans la figure 1.

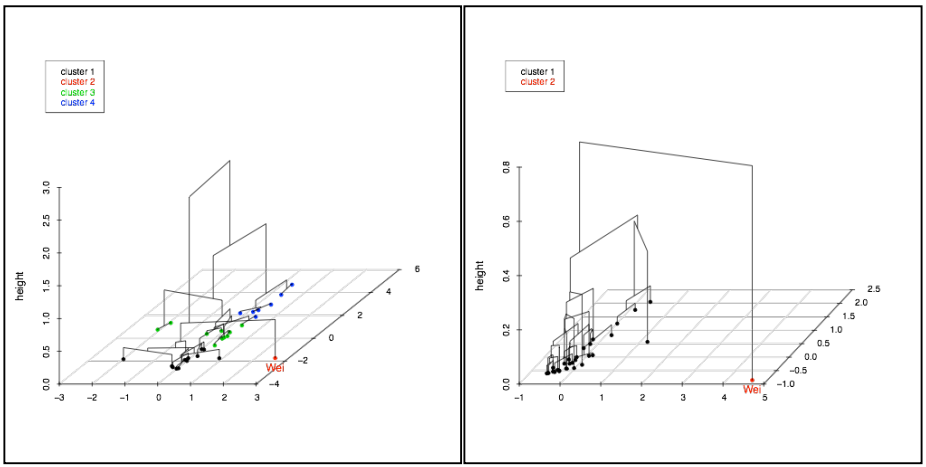


1.A. **1.B.**

Figure 1. Classifications hiérarchiques pour deux régions connues. Figure 1.A représente la région du chromosome 13:11595558-11795558 associé au trait « pelage fourni » et spécifique des races JRT, Bot, StP, IrW et TYo. Figure 1.B. représente la région du chromosome 13:23256842-23456842) spécifique de la race Shar-pei (ShP). Pour la clarté de la représentation, seuls les labels des races concernées sont reportés.

3.2 Régions spécifiques du Braque de Weimar

Parmi les 17 régions du chromosome 1 étudiées pour la race Weimar, la méthode proposée, classifiant les races à partir des haplogroupes, identifie 7 régions qui différencient le Weimar. Parmi ces 7 régions, nous retrouvons 3 régions spécifiques du Weimar, sur les 4 obtenues par Vaysse *et al.* Les 4 autres régions identifiées par notre méthode sont connues pour différencier d'autres races qui sont également identifiées. La spécificité de détection est améliorée en comparaison de l'approche menée dans Vaysse et al. La figure 2 reporte les arbres de classification obtenus par notre méthode pour deux régions spécifique du Weimar.



2.A. **2.B.**

Figure 2. Classifications hiérarchiques pour deux régions présélectionnées pour le Braque de Weimar. Figure 2.A représente la classification obtenue pour une région du chromosome 1:21308965-21567262. Figure 2.B. représente la classification de la région chr1:42320119-42424999.

4 Discussion

Dans cet article, nous décrivons une méthode originale de détection de régions génomiques de différentiation allélique appliquée à l'espèce canine. La méthode s'appuie sur la détection de profils haplotypiques spécifiques pour une ou plusieurs races, leur catégorisation en haplogroupe, et l'analyse factorielle des correspondances qui permet d'extraire les composantes principales des données brutes et de réaliser leur classification hiérarchique. Notre méthode testée sur un jeu de données de régions validées a permis de montrer sa sensibilité, à partir de la détection de régions tests, et son efficacité par la capacité de détection des régions spécifiques d'une seule race ou de plusieurs races. De plus, l'application de notre méthode sur des régions présélectionnées lors d'une étude pilote a confirmé le caractère spécifique de certaines régions chez le Weimar.

Afin d'explorer les limites de la méthode liées à la sensibilité des paramètres utilisées, notamment sur l'estimation des haplotypes [5], nous avons comparé les résultats obtenus par la méthode implémentée dans le programme fastPHASE avec ceux obtenus par une autre méthode de phasage implémentée dans le programme BEAGLE [6]. Bien que ces deux méthodes fournissent des spectres d'haplotypes sensiblement différents, nous obtenons des résultats très similaires en termes de profils haplotypiques et donc en termes de groupes de races. La méthode apparaît donc peu sensible à la méthode de phasage initiale. Nous avons également évalué la taille de la région sélectionnée en estimant des haplotypes de plus grandes régions de 500 kb centrées sur le SNP d'intérêt. Les résultats obtenus sont également très similaires à ceux obtenus pour des haplotypes de 200 kb.

Cette faible sensibilité à la méthode de phasage et à la taille des régions s'explique par la deuxième étape de formation d'haplogroupes. Cependant la seconde étape nécessite de définir la valeur k qui optimise le nombre de clusters ou haplogroupes. Ce choix dépend de la nature de chaque région et nécessite une formalisation. A l'inverse, la troisième étape de notre méthode, qui consiste à extraire les groupes de races similaires répond à un formalisme et est automatisée. Le nombre de groupes est choisi en fonction du gain en inertie dans la classification hiérarchique. L'utilisation de ce critère d'inertie peut engendrer un nombre de groupes de races relativement élevé. Les résultats de la figure 2 montrent que la méthode identifie 4 groupes pour la région chr1:21308965-21567262 (Fig 2.A) et 7 groupes pour la région chr1:42320119-42424999 (Fig 2.B). Parmi ces groupes, la race Weimar est un cluster constitué d'une race unique, cependant ce n'est pas la

seule race détectée qui est discriminée pour la région considérée. La détection de plusieurs clusters de races peut s'expliquer par l'utilisation initiale de régions détectées par l'approche Fst. L'indice Fst peut identifier des variances de fréquence allélique significativement différentes à la moyenne et détecter des régions de sélection diversifiante. Dans un souci de validation des résultats, il est donc nécessaire d'étudier l'association de chaque groupe à un trait spécifique.

Les résultats préliminaires de la méthode proposée permettent de détecter des régions candidates à la sélection et constitue en ce sens une alternative aux méthodes s'appuyant sur la mesure de Fst. L'application à l'échelle du génome nécessite une étape d'automatisation de la méthode afin de valider exhaustivement la méthode et d'identifier avec précision le catalogue des régions du génome candidates à la sélection chez le chien.

Remerciements

Nous remercions l'équipe de la plate-forme bio-informatique GenOuest (http://www.genouest.org) pour leur aide technique et leur assistance. Ce travail est en partie financé par la commission Européenne (LUPA - GA-201370). Nous remercions le CNRS et l'Université Rennes1 pour le support financier apporté à HJBA, CA et CH, et l'Université Rennes2 pour ME.

Réferences

[1] A.Vaysse, A. Ratnakumar, T. Derrien, E. Axelsson, P.G. Rosengren, S. Sigurdsson, T. Fall, E.H. Seppälä, M.S.T.Hansen, AND C.T. Lawley, E.K. Karlsson, D. Bannasch, C. Vilà, H. Lohi, F. Galibert, Francis, M. Fredholm, J. Hägström, A. Hedhammar, C. André, K. Lindblad-Toh, C. Hitte, M. T. Webster and The LUPA Selection Mapping. *PLoS Genetics*, 7(10): e1002316, 2011.

[2] E. Cadieu, M. W. Neff, P. Quignon, K. Walsh, K. Chase, H. G. Parker, B. M. VonHoldt, A. Rhue, A. Boyko, A. Byers, A. Wong, D.S. Mosher, A. G. Elkahloun, T. C. Spady, C. André, K. G. Lark, M. Cargill, C. D. Bustamante, R. K. Wayne, E. A. Ostrander, Coat Variation in the Domestic Dog Is Governed by Variants in Three Genes. *Science* 326:150-153, 2009.

[3] P. Scheet and M. Stephens. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics*, 78:629-644, 2006.

[4] S. Lê, J. Josse and F. Husson, F. FactoMineR : An R Package for Multivariate Analysis. *Journal of Statistical Software*. 25:1-18, 2008.

[5] O. Delaneau, J. Marchini and J.F. Zagury. A linear complexity phasing method for thousands of genomes. *Nature Methods,* 2:179-181, 2012.

[6] S.R. Browning and B.L. Browning. Rapid and accurate haplotype phasing and missing data inference for whole genome association studies using localized haplotype clustering. *American Journal of Human Genetics,* 81:1084-1097, 2007.

GECA: Gene Evolution/Conservation Analysis tool for Eukaryotic gene families.

Fawal Nizar[1, 2], Savelli Bruno[1, 2], Dunand Christophe[1, 2] and Mathé Catherine[1, 2],*
[1] Université de Toulouse, UPS, UMR 5546, Laboratoire de Recherche en Sciences Végétales, Castanet-Tolosan, France;
[2]CNRS, UMR 5546, Castanet-Tolosan, France;
{fawal, savelli, dunand, mathe}@lrsv.ups-tlse.fr

Abstract *GECA is a fast, user-friendly and freely-available web-based tool for representing genes exon/intron organization, comparing them and highlighting changes in gene structure among members of a gene family. It relies on a protein alignment, completed with the identification of common introns in the corresponding genes with CIWOG. Then, GECA produces as a main graphical representation the resulting aligned set of gene structures, where exons are up to scale. The important and original feature of GECA is that it overlays to these gene structures a symbolic display of the sequence similarity between subsequent genes. Noteworthy, the combination of a gene structure drawing with a similarity representation allows rapid identification of possible events of introns gains and losses, or points to erroneous structural annotations. The output image is generated in portable network graphics format which can be used for scientific publications. The web implemented version is freely available at: https://peroxibase.toulouse.inra.fr/geca_input_demo.php*

Keywords Evolution, Genome Analysis, Multiple sequence alignment, Comparative genomics, Visualization.

GECA: un outil pour l'analyse de l'évolution/conservation des gènes de familles multigéniques eucaryotes.

Résumé *GECA est un moyen rapide, convivial et disponible gratuitement servant à représenter les organisations exon/intron des gènes, en les comparant et mettant en évidence des changements dans la structure entre les membres d'une famille de gènes. Il s'appuie sur un alignement protéique, complété par l'identification des introns communs dans les gènes correspondants avec CIWOG. Ensuite, le graphique principal généré par GECA représente les structures des gènes après alignement, où seuls les exons sont à l'échelle. La caractéristique importante et originale de GECA est qu'au-delà de la structure des gènes, il schématise la similarité de séquence entre les gènes, deux à deux. Grâce à cette représentation qui superpose la similarité des séquences à la structure des gènes, GECA permet l'identification rapide des événements de gains et pertes d'introns, ainsi que des annotations structurales erronées. L'image de sortie est générée en format « png » utilisable pour les publications scientifiques. La version web est gratuitement disponible au lien suivant : https://peroxibase.toulouse.inra.fr/geca_input_demo.php*

Mots-clés Evolution, Analyse génomique, Alignement multiple, Génomique comparative, Visualisation.

1 Introduction

The Eukaryotic gene organization has a peculiar characteristic that the genomic sequences of the protein-coding genes are frequently interrupted by non-coding sequences, the introns. These quasi-random sequences are removed from RNA transcripts by the spliceosome prior to translation [1]. Spliceosomal introns' average numbers vary dramatically between eukaryotic species. For example, *Neurospora crassa* shows an average of 2 introns per gene, whereas the number in *Arabidopsis thaliana* is estimated to be greater than six [2].

These numbers vary even for the genes of the same class, where the Class III peroxidases of *Arabidopsis thaliana* show an average of three introns per gene, while *Oryza Sativa* and *Selaginella* show an average of 2 and 1 intron respectively [3].

From the fact that introns represent a large portion of eukaryotic pre-mRNA and that their numbers vary among organisms and among genes, emerges several evolutionary questions: how and when did they get to be situated within protein coding genes and is there a selective advantage of introns. Regarding their functional significance, introns are not selfish DNA with no distinct cellular functions, as initially stated by Cavalier-Smith in 1985 [4]. They have been shown to perform in various functions, for example as sources of non-coding RNA, and participate in gene evolution through alternative splicing and by containing regulatory elements [5]. However, this does not apply to all introns. Regarding the evolutionary questions, there is an ongoing debate concerning the behavior and dates of intron evolution [6-9]. The mechanisms and evolutionary dynamics of intron insertion and loss in eukaryotic genes remain poorly understood. A popular technique to study intron evolution is to compare homologous genes and try to detect common introns [10, 11]. The increasing pace of genome sequencing will allow performing global analysis of intron gain or loss through a large number of species. A significant bottleneck for large scale analysis was the lack of graphical tool to represent exon/intron conservation and/or evolution in eukaryotic families. To solve this technical problem, we have developed a new tool, namely GECA for Gene Evolution/Conservation Analysis [12].

2 State of the art

Currently, several tools are available that provide a representation of eukaryotic gene organization such as GSDS [13] and FancyGene [14]. The purpose of these programs is to represent the exon/intron structure of several genes in a single image in order to perform global gene structure comparison. However, these resources display the gene structures independently of each other, making them beyond comparison and without any notion of sequence conservation.

In order to accurately compare gene structures, we came to the conclusion that the following data is necessary. The position and level of similarity within a set of sequences is needed to highlight conserved regions. Once identified, a special focus to the conservation degree around introns is essential in order to determine whether they can be considered as conserved introns between paralogs/orthologs.

Since we failed to identify a tool that combines this information into a single output, we decided to develop our own tool, GECA. Its strategy relies on a simple observation: by aligning on the position of common introns shared by related sequences, we align the surrounding exons of the respective genes. Thus, GECA fully relies on the output of CIWOG [15], a freely available software that detects common introns (named cintrons) in a set of related protein-coding genes. In order to provide the protein alignment file required by CIWOG, GECA currently launches MAFFT [16] but any multiple alignment programs with a ClustalW-like output is suitable.

Once the protein sequences are aligned and the common introns identified, GECA is able to produce its main graphical representation: schematic gene structures, anchored on their first common intron position, overlaid with similarity content between pairs of genes.

GECA is a graphical tool for representing genes exon/intron organization, comparing them and highlighting changes in gene structure among members of a gene family. Therefore, it is mainly to analyze Eukaryotic gene families, but GECA can also be used on Prokaryotes where it will represent the similarities between the sequences.

3 Features

3.1 Input data format

Protein and genomic sequences in FASTA format together with gene structures in GenBank feature format are required to execute GECA. A specific FASTA header is needed and must be identical between the three data. It should be in the form of “>AccessionID | sequence name”.

3.2 GECA representation of aligned genes structures

Once the user-supplied data is uploaded, GECA uses PERL's GD and GD::Text::Align libraries to draw the gene structures. The intron/exon organizations of subsequent genes are aligned using their first common intron. The exons are represented up to scale while the introns are of fixed size with the color code used by CIWOG. To display the similarities between the sequences, GECA uses the protein alignment produced by MAFFT. The similarity between subsequent sequences in the alignment is represented at the amino acid level in the translated exons. Two amino acids are linked by a blue line if they are identical and a purple one if they stand for conservative substitutions (default matrix is BLOSUM62 but BLOSUM45 and 80 are also available). Gaps longer than 5 amino acids in the alignments are displayed in gray. Finally, the exon and intron sizes are calculated from the genomic coordinates and are displayed under the exons and introns in black and red respectively (Fig. 1). In this example, we analyze data of both orthologous and paralogous genes from the 2-Cysteine Peroxiredoxin family available at the Peroxibase database (http://peroxibase.toulouse.inra.fr/). Since alternative gene structures are provided and MAFFT [17]; is able to handle such data, this doesn't show any difficulty for GECA. We can see in Fig. 1 an example of alternative splicing, where the sequences "Pc2CysPrx01-a" and "Pc2CysPrx01-b" are two splicing variants.

Simultaneously to its main and original graphical output, GECA also produces two complementary outputs: 1) a classical, GSDS-like, gene structure display where exons and introns are scaled relatively to their length and cintrons are visible with gene structures aligned on the first common intron 2) a tab-delimited table containing for each gene its introns' IDs, lengths and phases.

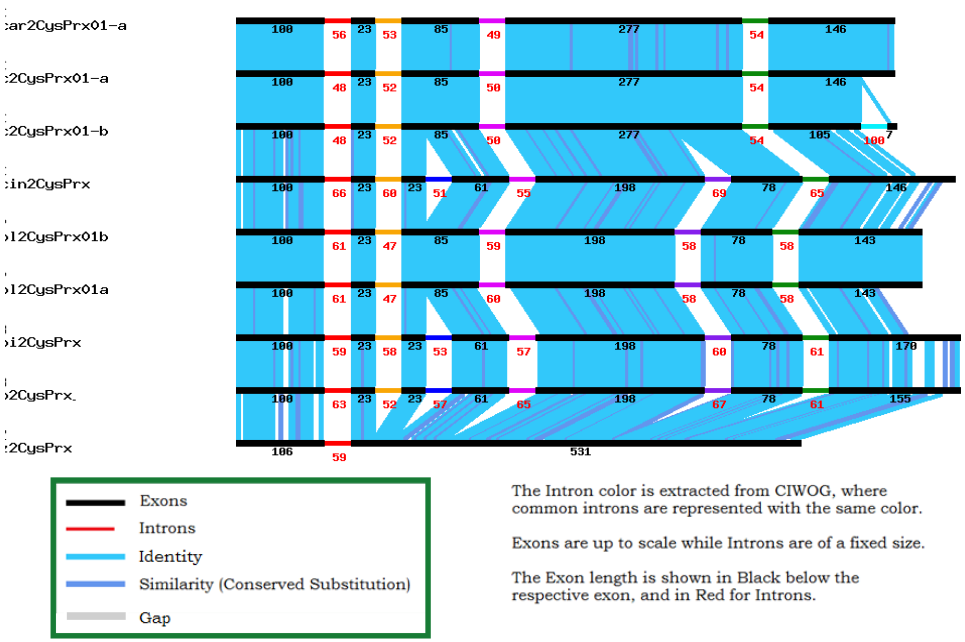


Figure 1. Screenshot of a result of GECA.

3.3 Web-server and standalone package

The idea of developing GECA came from our specific interest in large multigenic families, the peroxidases. The evolution of these families is not yet elucidated and new tools to understand their

evolutionary history was needed. Thus, GECA is available as a tool from our family database, the PeroxiBase [18]. The PERL CGI::LITE server is accessible in the PeroxiBase environment (http://peroxibase.toulouse.inra.fr/) as a further analysis after a BLAST or a multicriteria search.

An independent demo version, accessible from the PeroxiBase website, provides scientists with a user-friendly interface to GECA (https://peroxibase.toulouse.inra.fr/geca_input_demo.php). From this page, a package providing all PERL scripts for local installation and execution is available under open source license. GECA is controlled by a single program, customizable using a configuration file and requires pre-installation of PERL (minimum version tested 5.8.8) MAFFT and CIWOG.

Aligning exon/intron structures accompanied with the similarities between sequences and common introns information is very helpful while manually checking structural annotation. Moreover, it will provide major information about gene evolution, such as gain and loss of introns, and would bring new clues for the intron origin debate.

Acknowledgements

The Demo version of GECA is hosted by the Toulouse Midi-Pyrénées bioinformatics platform.

References

[1] Jurica, M. S. and Moore, M. J. "Capturing splicing complexes to study structure and mechanism" Methods 28 (3): 336 – 345, 2002.
[2] Roy SW and Gilbert W. The evolution of spliceosomal introns: patterns, puzzles and progress. Nat Rev Genet, 7:211–221, 2006.
[3] Mathé C, Barre A, Jourda C, Dunand C. Evolution and expression of class III peroxidases, Biochemistry and Biophysics, 500:58-65, 2010.
[4] Cavalier-Smith T. Eukaryotic gene numbers, noncoding DNA and genome size. In The evolution of genome size pp. 69-103, 1985.
[5] Fedorova L, Fedorov A. Introns in gene evolution. Genetica, 118(2-3):123-31, 2003.
[6] Coulombe-Huntington J, Majewski J. Characterization of intron loss events in mammals. Genome Res, 17:23–32, 2007.
[7] Roy SW, Irimia M. Mystery of intron gain: new data and new models. Trends Genet,25:67–73, 2009.
[8] Babenko VN, Rogozin IB, Mekhedov SL, et al. Prevalence of intron gain over intron loss in the evolution of paralogous gene families. Nucleic Acids Res, 32:3724–33, 2004.
[9] Roy SW, Penny D. On the incidence of intron loss and gain in paralogous gene families. Mol Biol Evol, 24:1579–81, 2007.
[10] Rogozin IB, Sverdlov AV, Babenko VN, et al. Analysis of evolution of exon-intron structure of eukaryotic genes. Brief Bioinform, 6:118–34, 2005.
[11] Irimia M, Roy SW. Spliceosomal introns as tools for genomic and evolutionary analysis. Nucleic Acids Res, 36:1703–12, 2008.
[12] Fawal N, Savelli B, Dunand C, Mathé C. GECA: a fast tool for Gene Evolution and Conservation Analysis in eukaryotic families. Accepted in Bioinformatics Application note, 2012.
[13] Chuan Y, Zhongguo H and Hui X GSDS: a gene structure display server. Bioinformatics, 29(8), 1023-1026, s2007.
[14] Rambaldi D and Ciccarelli F. FancyGene: dynamic visualization of gene structures and proteindomain architectures on genomic loci. Bioinformatics, 25,2281–2282, 2009.
[15] Wilkerson MD, Ru Y, Brendel VP. Common introns within orthologous genes: software and application to plants. Brief. Bioinform., 10(6),631-44, 2009.
[16] Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res., 30 (14),3059-3066, 2002.
[17] Golubchik T, Wise M.J, Easteal S and Jermiin L S. Mind the Gaps: Evidence of Bias in Estimates of Multiple Sequence Alignments. Mol Biol Evol 24 (11): 2433-2442, 2007.
[18] Koua D, Cerutti L, Falquet L, Sigrist CJA, Theiler G, Hulo N, Dunand C. PeroxiBase: a database with new tools for peroxidase family classifica-tion. Nucleic Acids Res, 37,D261-D266, 2009.

Posters, premier appel

Les résumés présentés dans cette session ont étés reçus lors du premier appel à communication. Ils ont fait l'objet d'un processus de relecture par le comité.

Djeen: A High Throughput Multi-Technological Research Information Management System for the Joomla! CMS

Hugo Duvergey[1], Olivier Stahl[2,3,4,5], Samuel Granjeaud[2,3,4,5], Oana Vigy[1] and Ghislain Bidaut[2,3,4,5]

[1] Plate-forme de Protéomique Fonctionnelle, c/o Institut de Génomique Fonctionnelle, F-34000 Montpellier, France.
[2] Inserm, U1068, Centre de Recherche en Cancérologie de Marseille, F-13009 Marseille, France.
[3] Institut Paoli-Calmettes, F-13009 Marseille, France
[4] Aix-Marseille Université, F-13284 Marseille, France.
[5] CNRS, UMR 7258, F-13009 Marseille, France
{hugo.duvergey; oana.vigy}@igf.cnrs.fr
{olivier.stahl; samuel.granjeaud; ghislain.bidaut}@inserm.fr,

Keywords RIMS, database, minimum information, Joomla! CMS.

1 Introduction

The advent of high-throughput data generation in biology implies great challenges to manage, store and share the generated data and their annotations [1]. Beyond that, manipulating and integrating heterogeneous data types in large scale collaborative projects involving several geographically separated laboratories remains a complex task for which proper management systems must be deployed, sometimes without the necessary bioinformatics expertise. 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.

To address these issues, we developed the Database for Joomla's Extensible Engine (Djeen). In contrast with previously proposed databases and Laboratory Information Management Systems (LIMS), Djeen is conceptually simple and adaptable to a large number of technologies and usages, thanks to a simple database scheme and organization of data within a file system. This makes it usable for any data types (no technology-dependent semantics was used in the core code) and adapted to the manipulation of very large numbers of data files and their annotations, generated in large scale assays..

Djeen has been designed as a Research Information Management System [4] which means that its structure is not modeled on data generated by a particular technology or instrument, but rather around laboratory experimentation workflows (figure 1). Longitudinal data integration concepts were implemented to address four previously identified fundamental issues in high throughput data management: data organization, data sharing, collaboration and publication [4].

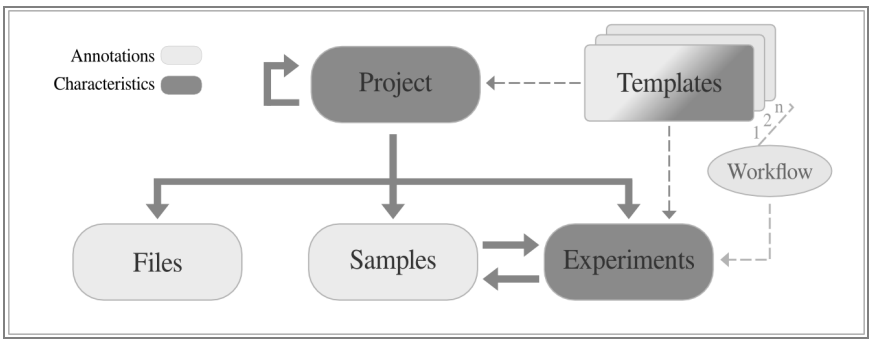


Figure 1. Data architecture and relationships.

2 Djeen Architecture and Features

Figure 1 highlights the internal Djeen structure which represents usual laboratory experimental processes. *Projects* gather coherent set of data (*samples*, *experiments* and *files*) and can be divided into sub-projects. *Samples* and e*xperiments* are connected structures: biological objects are described throughout *samples*, while *experiments* store the operations to transform these objects into other *samples*. *Templates* can be designed for *projects* and *experiments* with recurrent set of parameters to allow streamlining data annotation for high-throughput projects. In addition, it is possible to define ordered sets of experiment templates - known as *workflows* - to quickly create a sequence of experiments in a project.

Two types of parameters are used in Djeen: global information corresponds to *characteristics* (for example to describe experimental design) and specific data are stored as *annotations* (for example, patients clinical records). Both *characteristics* and *annotations* can be defined in *templates* and thus permit further data integration and minimum information gathering across collaborative work.

Each *project* has an owner who can *(i)* manage the user/group permissions and thus control the data sharing with designated collaborators, (*ii*) choose the template he or she wants a project or experiment to be built from. The *template* owner can define its structure and its required elements to enforce the capture of data by the experimenters. Moreover, authenticated users can backup a whole project through the *history* function.

3 Software and Availability

The 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 major developments include the conception of a generic data extraction engine for various uses (advanced search system, statistical queries execution, reports generation). Multiple databases management is also planned for a future release.

Acknowledgements

This project was initiated by the Cibi (CRCM[2] Integrative BioInformatics) team and has been funded by an Institut National du Cancer grant to GB. Servers running Djeen were funded by a Fondation pour la Recherche Médicale Grant to GB. Support for Olivier Stahl was obtained from Aix-Marseille Univ. Recently, the Functional Proteomics Platform[1] (IGF) has joined this project; support for Hugo Duvergey is given partly by the Région Languedoc-Roussillon (GEPETOS 2007 contract).

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.

Metagenomic gene clusters associate with diet-induced improvement of bioclinical factors in obesity

Aurélie COTILLARD[1], Edi PRIFTI[2], Ling Chun KONG[13], Nicolas PONS[2], Mathieu ALMEIDA[2], Salwa RIZKALLA[13], Sean KENNEDY[2], Joël DORÉ[2], Stanislav Dusko EHRLICH[2], Karine CLÉMENT[13] and Jean-Daniel ZUCKER[14]

[1] EQUIPE 7, U872 INSERM, 15 rue de l'école de médecine, Centre de Recherches des Cordeliers, 75006, Paris, France
aurelie.cotillard@crc.jussieu.fr

[2] MICALIS, UMR1319 INRA, Domaine de Vilvert, 78352, Jouy-en-Josas cedex, France

[3] ASSISTANCE PUBLIQUE-HOPITAUX DE PARIS, ICAN Institute of Cardiometabolism and Nutrition, CRNH-Ile de France, Pitié-Salpêtrière, 75013, Paris, France

[4]UMI 209 UMMISCO, IRD France Nord, F-93143, Bondy, France

Keywords Microbiota, metagenomic, obesity, gene clusters.

SAEMIX, an R version of the SAEM algorithm for parameter estimation in nonlinear mixed effect models

Audrey LAVENU[1], Emmanuelle COMETS[2] and Marc LAVIELLE[3]

[1] University Rennes-I, Rennes, France; INSERM CIC 0203, Rennes, France
audrey.lavenu@univ-rennes1.fr
[2] INSERM, UMR 738, F-75018 Paris, France; Univ Paris Diderot, Sorbonne Paris Cité, UMR 738, F-75018 Paris, France
emmanuelle.comets@inserm.fr
[3] INRIA, Saclay, France
Marc.Lavielle@inria.fr

Keywords Nonlinear mixed effect models, parameter estimation, SAEM algorithm, R, R package, pharmacokinetics, pharmacodynamics, longitudinal data.

SAEMIX, une librairie en R pour l'estimation des paramètres dans les modèles non-linéaires à effets mixtes par l'algorithme SAEM

Mots-clés Modèles non-linéaires à effets mixtes, estimation de paramètres, algorithme SAEM, R, librairie R, pharmacocinétique, pharmacodynamie, données longitudinales.

1 Introduction

The use of modelling and simulation in clinical drug development is now well established. Regardless of whether a single outcome is considered at the end of the study, clinical trials often collect longitudinal data, with each subject providing several measurements throughout the study. Longitudinal data is a staple in particular of pharmacokinetic (PK) and pharmacodynamic (PD) studies, which are a required part of a new drug application file. Nonlinear mixed effect models can help to characterise and to understand many complex nonlinear biological processes, such as biomarkers or surrogate endpoints, and are crucial in describing and quantifying the mechanisms of drug action and the different sources of variation, e.g., the interindividual variability.

Over the past decade, new and powerful estimation algorithms have been proposed to estimate the parameters of these models, which are in the process of superseding previous linearisation-based algorithms. The Stochastic Approximation Expectation Maximization (SAEM) algorithm has proven very efficient, quickly converging to the maximum likelihood estimators [1] and performing better than linearisation-based algorithms [2]. It has been implemented in the Monolix software [3] which has enjoyed increasingly widespread use over the last few years, more recently in the Statistics toolbox of Matlab (nlmefitsa.m), and is also available in NONMEM version 7 [4]. The objective of the present package was to implement SAEM in the R software [5].

2 Methods

Detailed and complete presentations of the nonlinear mixed effects model can be found in several reference textbooks, for instance [6]. We consider the following general nonlinear mixed effects model for continuous outputs:

$$y_{ij} = f(x_{ij}, \psi_i) + g(x_{ij}, \psi_i, \xi)\varepsilon_{ij} \quad , \; 1 \leq i \leq N \quad , \quad 1 \leq j \leq n_i \tag{1}$$

where y_{ij} is the jth observation of subject i, N is the number of subjects, n_i is the number of observations of subject i; the regression variables, or design variables, (x_{ij}) are assumed to be known, x_{ij}. We further assume that for subject i, the vector ψ_i is a vector of n_ψ individual parameters, and is a function of n unknown vector of fixed effects μ, an unknown vector of normally distributed random effects η_i, and possibly of individual

covariates c_i. The within-group errors (ε_{ij}) are supposed to be Gaussian random variables with mean zero and variance 1. Furthermore, we suppose that the ε_{ij} and the η_i are mutually independent. Different error models for g can be used in SAEMIX.

The SAEM algorithm is used to obtain maximum likelihood estimates of the parameters of nonlinear mixed effects models without any linearisation of the model. The log-likelihood for nonlinear mixed effect models is analytically intractable since it requires integration over the unknown individual parameters. The SAEM algorithm uses an EM algorithm [7], where the unknown individual parameters are treated as missing data, and replaces the usual E-step with a stochastic approximation step [8]. The missing parameters are simulated at each iteration via a MCMC procedure, which can be used after the algorithm has converged to obtain the conditional modes, the conditional means and the conditional standard deviations of the individual parameters.

3 Results

The library uses the S4 class system of R to provide a user-friendly input and output system; in particular functions like `summary` or `plot` have been developed and apply to fitted objects. The package provides summaries of the results, individual parameter estimates, standard errors (obtained using a linearised computation of the Fisher information matrix) Wald tests for fixed effects, and a number of diagnostic plots, including VPC plots and npde [9]. The log-likelihood can be computed by three methods: a linearisation of the model, an importance sampling procedure, or a Gaussian quadrature. The diagnostic graphs can be tailored to the user's individual preferences by setting a number of options, and are easily exported to a file.

We illustrate the use of the library with the well known PK dataset of theophylline. The dataset includes the concentration versus time data collected in 12 subjects given a single oral dose of theophylline, and for whom 11 blood samples were collected over a period of 24 h. We modelled this data using a one-compartment model with first-order absorption, parameterised as k_a, V, CL. The IIV was modelled using an exponential model with diagonal variance-covariance matrix, while the residual variability was modelled with a combined error model. Many diagnostic plots are available to evaluate convergence or model adequacy, such as individual plots. They can be accessed through options given to the `plot` function, and tailored to change specific features such as colours or axis scales.

4 Conclusion

The SAEMIX package provides the SAEM algorithm for R users. The current version handles models in analytical form, with continuous or binary covariates.

References

[1] Delyon B., Lavielle M., Moulines E. Convergence of a stochastic approximation version of the EM algorithm. *Annals of Statistics* 27:94–128, 1999.

[2] Girard P., Mentré F. A comparison of estimation methods in nonlinear mixed effects models using a blind analysis (oral presentation). *Meeting of the Population Approach Group in Europe (PAGE)*, Pamplona, 2005.

[3] Lavielle M. *MONOLIX (MOdèles NOn LInéaires à effets miXtes) User Guide*. MONOLIX group, Orsay, France, 2010. URL: http://software.monolix.org/

[4] Beal S., Sheiner L.B., Boeckmann A., Bauer R.J., *NONMEM User's Guides. (1989-2009)*, Icon Development Solutions, Ellicott City, MD, USA, 2009.

[5] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006.

[6] Davidian M., Giltinan D. *Nonlinear models for repeated measurement data*. Chapman & Hall, London, 1995.

[7] Dempster A. P., Laird N. M., Rubin D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B* 39:1–38, 1977.

[8] Kuhn E., Lavielle M. Maximum likelihood estimation in nonlinear mixed effects models. *Computational Statistics and Data Analysis* 49:1020–38, 2005.

[9] Brendel K., Comets E., Laffont C., Laveille C., Mentré M.. Metrics for external model evaluation with an application to the population pharmacokinetics of gliclazide. *Pharmaceutical Research*, 23:2036–49, 2006.

Relationship between protein surface exposure and antibody binding

Virginie LOLLIER[1], Sandra DENERY-PAPINI[1], Colette LARRÉ[1] and Dominique TESSIER[1]

[1] UR 1268 BIOPOLYMÈRES INTERACTION ASSEMBLAGES , INRA, rue de la Géraudière, 44316, Nantes, France
{virgine.lollier,sandra.denery,colette.larre,dominique.tessier}@nantes.inra.fr

Keywords Epitope prediction, 3D structures, Antigen.

1 Introduction

Immuno-informatics is an emerging field of research dedicated to the computational approaches applied to the heterogeneous immunological data integration [1]. It leads to the development of specific databases and prediction tools useful for the engineering of vaccines, immune therapeutics and diagnostics.

In a large part, experimental studies concern the localisation of epitopes. These small stretches of amino acids are the areas of the antigenic molecules which interact specifically with immune proteins. They are typed as T when the immune proteins are T cell receptors and as B when they bind to antibodies produced by B cells. In the case of food allergy, B cell epitopes are particularly searched to develop new industrial processes altering these sites and designing hypoallergenic food. Their identification is also useful for the construction of accurate diagnostic tools of allergic patients and for the detection of allergen traces in food products.

The B epitope prediction systems, coming from immuno-informatics [2], often include the basic assumption that these sites are exposed on the protein surface. However these tools display weak accuracy when they are applied either on antigen sequence or on antigen 3D structure [3,4,5]. The aim of this work is to measure this feature on antigenic proteins through a generic approach and to evaluate its relevance as a predictive criterion.

2 Method

Epitope sequences have been collected from the IEDB WEB site (Immune Epitope DataBase [6]) and divided according to their continuous/discontinuous feature. Briefly, continuous epitopes are identified by testing synthetic peptides and discontinuous epitopes are identified from the 3D structures of antigen-antibody complexes, by selecting side-chains which interact with antibody. The 3D protein structures related to these epitopes have been collected from the PDB WEB site.

Structural features have been computed on these structures and the resulting values of RSA (Relative Solvent Accessibility [7,8,9]) and PI (Protrusion Index [10]) have been compared between epitopic amino acids and non epitopic ones.

3 Results

Data relative to continuous epitopes are in a large majority (901 vs 56, on 61 3D structures). Actually, the resolution of complexed 3D structures is a longer process than scanning the binding ability of antibodies to a peptide library.

The values of RSA and PI display a similar distribution along the protein sequences. Considering the discontinuous epitopes, a statistically difference is observed between the distributions of the accessibility values of epitopic and non epitopic residues. However, no threshold distinguishes epitopes as surface exposed elements because their exposure varies from zero to some of the maximum values.

Considering only continuous epitopes, the distribution of epitopic residues is not different from non

epitopics. Moreover, in some cases, especially for the most studied antigens (like botulinum toxin or measles hemagglutinin), the continuous epitopes entirely cover the protein sequence.

Making the assumption that the sequence covering might reflect a lack of precision in the identification technique of continuous epitopes, then, the relevant residues belonging to epitopes should be more frequently identified. Considering 4 examples where antigens are covered by epitopes and are intensively studied (i.e. related to at least 4 bibliographic references), such residues are not systematically correlated with the surface exposure.

4 Conclusion

Our work does not affect the principle that, at any time, the epitope is necessarily presented on the surface of the molecule to make contact with the antibody, but it tries to evaluate if this feature in itself could be efficient for epitope predictions. Considering available data, it seems difficult to take benefit of surface exposure in any prediction system without knowledge of the structural form of the antigen which encounter the antibodies of the immune system [11]. Besides, the sequence covering by continuous epitopes calls out about its proper definition, notably to decipher B-cell epitope as an intrinsic feature of a protein.

Acknowledgements

This work was supported by the ANR (PREDEXPITOPE, ANR-08-ALIA-14).

References

[1] N. Tomar and R . K. De. Immunoinformatics: an integrated scenario. Immunology, 131, pp. 153-168, 2010.
[2] N. D. Rubinstein, I. Mayrose, E. Martz and T. Pupko. Epitopia: a web-server for predicting b-cell epitopes. BMC Bioinformatics, 10, p. 287, 2009.
[3] M. J. Blythe and D . R. Flower. Benchmarking b cell epitope prediction: underperformance of existing methods. Protein Sci., 14, pp. 246-248, 2005.
[4] J. V. Ponomarenko and P . E. Bourne. Antibody-protein interactions: benchmark datasets and prediction tools evaluation. BMC Struct. Biol., 7, p. 64, 2007.
[5] M. J. Sweredoski and P. Baldi. Cobepro: a novel system for predicting continuous b-cell epitopes. Protein Eng. Des. Sel., 22, pp. 113-120, 2009.
[6] R. Vita, L. Zarebski, J. A. Greenbaum, H. Emami, I. Hoof, N. Salimi, R. Damle, A. Sette and B. Peters. The immune epitope database 2.0. Nucleic Acids Res., 38, p. D854-62, 2010.
[7] B. Lee and F . M. Richards. The interpretation of protein structures: estimation of static accessibility. J. Mol. Biol., 55, pp. 379-400, 1971.
[8] The ccp4 suite: programs for protein crystallography. Acta Crystallogr. D Biol. Crystallogr., 50, pp. 760-763, 1994.
[9] S. Miller, J. Janin, A. M. Lesk and C. Chothia. Interior and surface of monomeric proteins. J. Mol. Biol., 196, pp. 641-656, 1987.
[10] A. Pintar, O. Carugo and S. Pongor. Cx, an algorithm that identifies protruding atoms in proteins. Bioinformatics, 18, pp. 980-984, 2002.
[11] V. Lollier, S. Denery-Papini, C. Larré and D. Tessier. A generic approach to evaluate how b-cell epitopes are surface-exposed on protein structures. Mol. Immunol., 48, pp. 577-585, 2011.

Full genome analysis for the prediction and prioritization of regulatory sequence variations

Virginie BERNARD[1], David J. ARENILLAS[1] and Wyeth W. WASSERMAN[1]

[1]Centre for Molecular Medicine and Therapeutics, Child and Family Research Institute
University of British Columbia Vancouver, BC, Canada

virginie@cmmt.ubc.ca

Abstract

Current software allow to predict variations within protein encoding genes that are likely to contribute to a disease phenotype, by focusing on missense or nonsense variants. With the emergence of full-genome analysis, the variations outside of exons are accessible and have to be prioritized. Our pipeline predict and rank the variation within transcription factor binding site that are likely to be involved in gene expression regulation failure, and so in the disease.

Keywords: Damaging variant prediction, regulatory variations, genome analysis, variant prioritization, NGS

The convergence of high-throughput technologies for sequencing individual exomes and full-genomes and rapid advances in genome annotation are driving a neo-revolution in human genetics. This wave of family-based genetics analysis is revealing causal variations. By mapping the reads to the human genome reference and by searching for variations relative to the reference, a list of small nucleotide variations, insertions and deletions is obtained. Selecting the ones shared by relatives having a same disorder and the ones unknown in polymorphism databases and former sequencing analysis removes a high number of false positive and polymorphism. In addition, analysis is required to reveal those variations most likely to contribute to a disease phenotype. Existing software score the severity of changes that lead to a missense or a nonsense variation. Nevertheless, no software focuses on scoring variations outside of exons while such variation may have a deleterious effect on the regulation of gene expression. Protein-encoding exons are only 2% of the genome and the remaining 98% of the genome control the developmental and physiological profile of gene activity - when and where a gene will be active. These regions are known to be linked to disease phenotype [1]. Functional contributions of cis-regulatory sequence variations to human genetic disease are numerous. The need to identify causal regulatory variations is becoming imperative.

With full genome sequencing becoming accessible to medical researchers, we can predict regulatory variations [2]. Our software system will enable genetics researchers to characterize such variations within individual full-genome sequences. As a first step, by analyzing all variations -inside as outside exons-, we focused on the prediction of the ones the more likely to have an impact on splice site recognition and on transcription binding to their specific DNA sequences. In order to score the impact of variations linked to transcription factor binding sites, we used position weight matrixes available in reference databases of regulatory elements or derived from experimental archives of protein-DNA interactions. We focused on the variations leading to the lost or the gain of a functional site prediction, both being damaging for the regulation of gene. We applied our approach on varitions putatively involved in Triple Negative Breast Cancer and prioritized 8 variations putatively damaging [3]. With our software, researchers will have greater capacity to identify variations potentially causal for disease.

References

[1] Vandermeer J.E. and Ahituv N. Cis-regulatory mutations are a genetic cause of human limb malformations. Dev Dyn (2011)

[2] Worsley-Hunt R., Bernard V., Wasserman W.W. Identification of cis-regulatory sequence variations in individual genome sequences. Genome Medicine,3(10):65 (2011).

[3] Shah S.P., Roth A., Goya R., Oloumi A., Ha G., Zhao Y., Turashvili G., Ding J., Tse K., Haffari G., Bashashati A., Prentice L., Khattra J., Bernard V., *et al.* Mutational landscapes of primary triple negative breast cancers reveal individual heterogeneity but distinct patterns of somatic mutation. Nature (in press)

Search for Three-dimensional Atom Motifs in Protein Structures

Efficient screening of the PDB for artificial protein design

Guillaume COLLET[1], and Philippe CUNIASSE[2]

[1] Commissariat à l'Énergie Atomique, DSV/iBiTec-S/SIMOPRO, CEA Saclay, 91191, Gif-sur-Yvette, France
{guillaume.collet, philippe.cuniasse}@cea.fr

Keywords Protein design, function transfer, structural motif graft, clique search.

1 Introduction

The design of novel protein functions by the transfer of a functional motif on an existing scaffold is a major goal of protein engineering. With the exponential growth of available protein structures in the Protein Data Bank (PDB), computational protein design approaches have led to many successful compounds. All these approaches involve a search for similar functional motifs in the available scaffolds. Some of these approaches try to find strictly identical motifs, i.e. the same residues in the same 3D topology, but if the goal is to graft a functional motif onto a suitable scaffold, such approaches are not relevant.

A few methods propose to search suitable sites to graft a functional motif: RosettaMatch [1], Scaffold Selection [2] and AutoMatch [3]. The two main limitations faced by these methods are: 1) the huge number of potential grafting sites to explore and 2) the evaluation of the functionality of the grafted motif.

To tackle the first limitation, the previous methods use efficient algorithms: geometric hashing, clique search and set reduction. Although these algorithms can reduce the complexity of the problem, its combinatorial essence forces to find a way to reduce the number of potential sites to explore. In consequence, these methods reduce the search to residues on the surface or into a pocket. Moreover, the functional motif is limited to 5 or 6 residues maximum.

We propose STAMPS, an approach based on geometrical features only, which can screen the entire PDB in a reasonable amount of time to find suitable grafting site of at least 3 residues. STAMPS uses the Cα-Cβ distances compatibility and a clique search algorithm to find suitable grafting sites. The quality of the identified sites is calculated using a RMSD optimization and a steric hindrance measure taking into account the protein target.

Although less flexible than the previous methods, we show that STAMPS retrieves functional motifs faster and with a better accuracy on the classical Zanghellini benchmark [1]. Moreover, we show that longer motifs (with more than 10 residues) can be searched with STAMPS without explosion of computing time.

Finally, STAMPS has been successfully applied for the design of Kv1.2 Potassium Channel Blockers [4], the analysis of facial two-histidine one-carboxylate binding motif through the entire PDB and the detection of potential scaffolds to design artificial inhibitors of metalloenzymes [6].

2 Motif search algorithm

In RosettaMatch and AutoMatch, the potential grafting sites are generated on the examined protein by grafting all the possible residues. This procedure leads to a huge amount of solutions to explore. In Scaffold Selection and STAMPS, the motif is directly search in the examined protein.

2.1 Motif definition

In STAMPS, the searched motif is called the *reference* motif and is defined as a set of residues $R=\{r_1,r_2,\ldots,r_k\}$. Its 3 letters name, its pdb number and the chain identifier it belongs to identify each residue. Then, the coordinates of its atoms are read in a PDB file.

2.2 Clique search

Given an examined protein chain $P=\{p_1,p_2,\ldots,p_n\}$, we define a graph G=(V,E) organized as a grid where each row i corresponds to residue r_i and each column x correspond to residue p_x. An edge between vertices v_{ix} and v_{jy} is created if the inter-atomic distances of Cα-Cβ between r_i and r_j are compatible with the inter-atomic distances of Cα-Cβ between p_x and p_y. Thus, finding all suitable sites corresponds to finding all the cliques of size k in G. The clique search is performed by a modified Bron-Kerbosh algorithm [5].

The number of cliques of size k in G is $O(n^k k^2)$ which is a huge number. In fact, the clique search is efficient for two reasons: i) the complexity of the algorithm is in $O(m^{3/2})$ with *m* the number of edges [7] and ii) the graph is sparse, less than 5% of edges are created (5% of distances are compatible) even for scaffolds with many suitable sites.

3 Zanghellini Benchmark

The Zanghellini benchmark [1] consists in ten enzyme catalytic sites. This benchmark has been designed to test the capacity of a method to detect catalytic sites in their native scaffold. Three catalytic sites were removed from the set (1H2J, 1OEX and 3VGC) because STAMPS searches on each chain independently. The seven remaining catalytic sites were searched in the SimpleScaffold Library defined by Zhang and Lai [3] (612 pdb files). The results show that STAMPS always find the native site at the best rank. Moreover, compared with AutoMatch, STAMPS is faster on difficult cases (AJCL and 1NEY).

	Native site ranking				Compute time (in seconds)	
PDB code	RosettaMatch	Scaffold Selection	AutoMatch	STAMPS	AutoMatch	STAMPS
1C2T	1	1	1	1	48	20
1DQX	37	830	1	1	432	78
1JCL	Not found	7	1	1	3180	22
1NEY	11	1	15	1	4236	19
1P6O	1	1	1	1	5	4
4FUA	1	5	1	1	6	17
6CPA	32	42	1	1	24	5

Table 1. Comparison of the native site rank and the computation time on the Zanghellini benchmark. The results for RosettaMatch, Scaffold Selection and AutoMatch are reported from [3].

References

[1] A. Zanghellini et al., New algorithms and an in silico benchmark for computational enzyme design, *Protein Science*, 15:2785-2794, 2006.

[2] C. Malisi, O. Kohlbacher and B. Höcker, Automated scaffold selection for enzyme design, *Proteins: Structure, Function, and Bioinformatics*, 77:74-83, 2009.

[3] C. Zhang and L. Lai, AutoMatch: Target-binding protein design and enzyme design by automatic pinpointing potential active sites in available protein scaffolds, *Proteins: Structure, Function, and Bioinformatics*, online publication, 2012.

[4] C. Magis, et al., Structure-based secondary structure-independent approach to design protein ligands: Application to the design of Kv1.2 potassium channel blockers. *J. Am. Chem. Soc.*, 128(50):16190-16205, 2006.

[5] F. Cazals and C. Karande, A note on the problem of reporting maximal cliques, *Theoretical Computer Science*, 407(1):564-568, 2008.

[6] B. Amrein, M. Schmid, G. Collet, P. Cuniasse, F. Gilardoni, F.P. Seebeck and T.R. Ward, Identification of two-histidine one-carboxylate binding motifs in proteins amenable to facial coordination to metals, Metallomics, online publication, 2012.

[7] N. Chiba and T. Nishizeki, Arboricity and Subgraph Listing Algorithms, SIAM Journal on Computing, 14:210-223, 1985.

Sequence, structure and function relationship of Baeyer-Villiger monooxygenases : insights for a better classification

Joseph Rebehmed[1], Véronique Alphand[2] , Véronique de Berardinis[3] and Alexandre G. de Brevern[1]

[1] DSIMB, INSERM, UMR_S 665, Université Paris Diderot, Sorbonne Paris Cité, INTS, Paris 75015, France
{joseph.rebehmed, alexandre.debrevern}@univ-paris-diderot.fr

[2] iSm2, UMR CNRS 7313, Université Aix-Marseille, Marseille, France
veronique.alphand@univ-amu.fr

[3] LCAB, Genoscope, CEA Evry, France
vberard@genoscope.cns.fr

Oxygenation reactions are key role reactions found in most of the living organisms [1]. One particular reaction, of great importance for organic chemistry, is the Baeyer-Villiger oxidation in which a ketone is oxidized into an ester. It can be performed by biocatalysis with a family of proteins called Baeyer-Villiger Monooxygenase (BVMO). This bio-reaction showed a number of aspects that are better than its chemical version; it does not induce the use of potentially harmful reagents (green chemistry) and displays better enantio- and regio- selectivity [2]. And because of all these features, pharmaceutical and perfumes companies start to show an interest.

We analyze the available x-ray structures of current BVMOs. Their flexibility will be examined by molecular dynamics simulations and Normal mode analyses. Binding site identifications and docking calculations will allow highlighting the existing interactions between the enzymes and their ligands and cofactors.

In parallel, many bacterial and fungal genomes were screened for new putative BVMOs based on their sequence identity and presence of a specific motif. Multiple sequence alignments will permit us to study the conservation of structurally and/or functionally important amino acids during evolution. Structural models for those new BVMOs, obtained by comparative modeling and threading techniques, alone or in complex with different ligands, will complement and confirm the previous results. At last, combining all those studies will allow us to discuss a better classification of the BVMO protein family, based on the sequence, structure and/or function.

Acknowledgements

This work is supported by a grant from the French National Research Agency (ANR).

References

[1] W.J.H. Van Berkel, N.M. Kamerbeek and M.W. Fraaije, Flavoprotein monooxygenase, a diverse class of oxidative biocatalysts. *Journal of Biotechnology*, 124:670-689, 2006.

[2] V. Alphand and R. Wohlgemuth, Applications of Baeyer-Villiger Monooxygenases in Organic Synthesis. *Current Organic Chemistry,* 14:1928-1965, 2010.

A pipeline for identification of mutations and correction of genome assemblies using the Illumina sequencing platform

Olivier ARNAIZ[1], Simone MARKER[2], Deepankar SINGH[2], Cyril DENBY WILKES[1], Quentin CARRADEC[2], Eric MEYER[2] and Linda SPERLING[1]

[1] Centre de Génétique Moléculaire, UPR3404 CNRS, Av de la Terrasse, 91198 Gif-sur-Yvette
[2] Institut de Biologie de l'Ecole Normale Supérieure, UMR8197 CNRS - INSERM U1024, 46, rue d'Ulm, 75005 Paris

Keywords SNP, Illumina sequencing, mutation, Paramecium

1 Introduction

The pioneering work of Tracy Sonneborn established *Paramecium tetraurelia* as an attractive genetic model organism over 50 years ago. The alternation of two modes of sexual reproduction facilitates genetic analysis. Conjugation, the reciprocal fertilization of two cells of complementary mating types, yields a pair of F1 cells that are genetically identical whereas autogamy, a self-fertilization process, results in entirely homozygous individuals. By the 1970s, numerous mutant strains had been isolated and characterized, affected in cellular morphogenesis, motility, swimming behavior, regulated secretion, mating type determination, the cell cycle etc [2]. However identification of the mutation responsible for the phenotype, when possible, was labor intensive and took several months.

We report here that the identification of point mutations is now feasible by whole genome sequencing using Illumina technology.

2 Results

2.1 Objectives and method

Sophisticated methods have been developed to identify SNPs and to increase the accuracy of SNP calling [1]. However SNP calling in *Paramecium* is straightforward. First, 100% homozygous cells are used. Second, good coverage can be obtained as the genome is relatively small (72 Mb [3]) : 30 million 75nt reads provide an average coverage of 30X. Thus only stringent read mapping and counting are required as in theory 100% of reads will differ from the reference at a SNP.

Our pipeline involves (Figure 1) :

- mapping reads to the reference genome assembly using the BWA short read aligner [4] allowing 2 mismatches
- indexing the mapped reads with samtools [5]
- use of custom Perl scripts and samtools (mpileup) to identify positions where the reads do not correspond to the reference

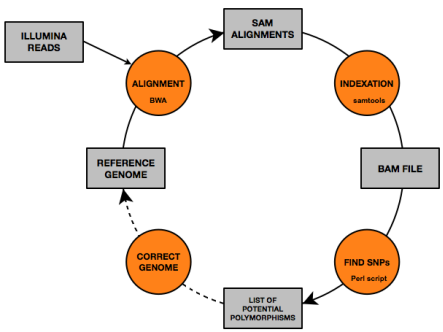


Figure 1. Pipeline for identification of SNPs

Positions are selected where at least 80% of the calls differ from the reference genome with a coverage between 10X and 300X.

2.2 Polishing the genome assembly

First we corrected the reference genome, assembled from 13X Sanger sequence reads [3]. This involved paired-end sequencing of exactly the same wild type DNA initially used to assemble the reference genome.

The distance between the paired-ends reads provides an important constraint for accurate alignment of the reads on the genome. The pipeline was used to correct assembly errors through successive rounds of mapping – SNP calling – correction. The iterative process is important, because the BWA aligner only tolerates a few mismatches. At the end, 13,758 substitutions were corrected, 929 deletions of 1-2 nt were filled and 10,339 insertions of 1-2 nt were removed.

2.3 Identification of mendelian mutations

DNA was isolated from each mutant strain, sequenced and candidate mutations were identified using the same pipeline. The results are presented in Table 1. The "putative SNPs" column corresponds to the output of our pipeline. These candidates were manually curated to avoid SNPs arising from differences in genetic background between mutant and reference and to see whether the SNP affects a translation product, to yield a list of candidate genes. The candidates were validated by PCR amplification and resequencing of DNA from mutant and wild type clonal cell lines, and by functional complementation.

	defect	million reads	% mapped reads	putative SNPs	candidate genes	validated gene	mutation
mtB[6]	mating type	30	92	6	1	transcription factor	ATG(START) to ACG(T)
mtC[6]	mating type	28	92	8	1	transcription factor	GTG(V) to ATG(M)
mtF^{E}[7]	mating type	62	95	9	1	novel	TTG(L) to TAG(Q)
RNAi 1.8	RNAi pathway	16	97	6	1	polyU polymerase	TCC(S) to TTC(F)
RNAi 3.1	RNAi pathway	24	95	3	1	RdRP1	CGA(R) to TGA(STOP)
RNAi 3.18	RNAi pathway	41	92	10	1	novel	TTA(L) to TTT(F)
cro1[8]	cell shape	39	95	4	1	NIMA-like kinase	GAG(E) to AAG(K)

Table 1. Identification of mutations

3 Conclusion

The identification of point mutations in *Paramecium* by whole genome sequencing is now straightforward and cost-effective and opens the way for more extensive use of genetic screens as a complement to the powerful reverse genetic techniques already widely employed in this organism. Our easy to implement pipeline may be useful for other organisms that are 100% homozygous or haploid.

Acknowledgements

This work was supported by the ANR and the CNRS.

References

[1] Nielsen R. et al. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet.* Jun;12(6):443-51, 2011.

[2] Sonneborn T.M. *Paramecium aurelia*, in R. King (ed.), *Handbook of Genetics* Plenum press, New York, pp. 469-594, 1974.

[3] Aury J.M. et al. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature* 444:171-8, 2006.

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

[5] Li H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078-2079, 2009.

[6] Byrne BC. Mutational Analysis of Mating Type Inheritance in Syngen 4 of *Paramecium aurelia*. *Genetics* 74: 63–80, 1973.

[7] Brygoo, Y. and Keller. A.-M. A mutation with pleiotropic effects on macronuclearly differentiated functions in *Paramecium tetraurelia*. Dev. Genet. 2: 23–34, 1981.

[8] Jerka-Dziadosz M. et al. Uncoupling of basal body duplication and cell division in crochu, a mutant of *Paramecium* hypersensitive to nocodazole. *Development* 125: 1305–1314, 1998.

Coarse-grained Simulations as a Bridge to Solve Atomic Structures of Dystrophin Essential Fragments from SAXS Envelopes

Emeline POLLET[1], Angélique CHERON[1], Mirjam CZJZEK[2], Jean-François HUBERT[1], Elisabeth LE RUMEUR[1] and Olivier DELALANDE[1]

[1] SIM Team, IGDR, UMR6290 CNRS, 2 av. du Pr. Léon Bernard, CS 34317, 35043 Rennes, France
emeline.pollet@etudiant.univ-rennes1.fr

[2] Station Biologique de Roscoff, UMR7139 CNRS-UPMC, Place Georges Teissier, 29680 Roscoff, France

Keywords Dystrophin, Coarse-grain, Molecular Dynamics, Small Angle X-ray Scattering.

1 Introduction

Dystrophin is a large (427kDa) filamentous protein whose absence or dysfunction are related to Duchenne or Becker muscular dystrophies. It contributes to the sarcolemma preservation during the muscle contraction-relaxation cycles [1]. Dystrophin interacts with many partners, such as F-actin of the muscular cell cytoskeleton. Its main and central domain is composed of 24 spectrin-like repeats (R1 to R24), associated in a three-helix coiled-coil fold. Dystrophin dynamics and repetitive structural pattern does not allow X-ray crystallography nor NMR studies and so make difficult its atomic structure determination [1,2].

Le ou les auteur(s) ne souhaite(nt) pas que ce document soit diffusé en ligne

Small Angle X-ray Scattering [illegible] information [illegible] size of biological macromolecules in solution but does not give precise atomic [illegible]. Molecular envelopes can be efficiently [illegible] following standard analysis (ATSAS program suite) and each [illegible] residue is represented by one single sphere [3]. By comparison, classical low-resolution coarse-grained (CG) models allocate two to four spheres per amino acid. Low-resolution CG molecular dynamics (MD) simulations are now largely considered as a well-suited approach to increase simulated timescales by keeping structural details of large biomolecular systems [4] and they could help in the processing of SAXS experimental data.

We here present a method coupling SAXS and CG MD applied to the atomic structure determination of two regions of dystrophin central domain close to its F-actin essential partner: the R1-3 fragment is linked to the first actin binding site (ABD1) and the R11-15 fragment composes the second actin-binding site (ABD2).

2 Methods

2.1 SAXS and Coarse-grained Models

SAXS acquisitions were performed at the SOLEIL synchrotron facility on highly pure samples of dystrophin expressed from *E. coli*. Data were analysed using the ATSAS package developed by Svergun lab [3]. Homology models were obtained from spectrin repeats using I-Tasser online server, and then converted in models at a CG resolution (Martini model). Molecular dynamics simulations (5xs trajectories) were computed and analysed with the Gromacs program.

2.2 Developing Protocol: Combining SAXS and Coarse-grained Modelling

From SAXS and CG MD data, we constituted a set of experimental or theoretical structures. While the SAXS signal results from the contribution of the multiple conformers present in the solution sample, the Oligomer program (ATSAS package) enabled to decompose the experimental SAXS record into individual contribution of scattering curves related to shapes generated by the GASBOR program (ATSAS). The set of theoretical models was obtained by a clustering analysis of the CG MD trajectories.

First, MD simulations allowed conformational space exploration to generate rational structures that could correspond to the best SAXS envelopes. Then, we used two empirical strategies to compare experimental and theoretical results: i) three dimensional particle distributions defining molecular volumes and ii) radial distribution functions from the two SAXS and CG MD structural sets.

Analysis of sexual differentiation in the brown alga *Ectocarpus* by RNA-seq

Reference genome and *de novo* approaches comparison

Alexandre CORMIER[1], Susana COELHO[1], Mark COCK[1] and Erwan CORRE[2]

[1] Génétique des Algues, UMR7139 CNRS, Place Georges Teissier, 29680, Roscoff, France
acormier@sb-roscoff.fr, coelho@sb-roscoff.fr, cock@sb-roscoff.fr

[2] ABIMS, FR2424 CNRS-UPMC, Place Georges Teissier, 29680, Roscoff, France
corre@sb-roscoff.fr

Abstract *We analyse in this study the differential expression of the gametophytic male and female transcriptome in the brown algae Ectocarpus siliculosus. Complementary assembly approaches (reference genome vs. de novo) have been used and will be discussed.*

Keywords RNAseq, differential expression, *de novo* and reference transcriptome assembly.

Le ou les auteur(s) ne souhaite(nt) pas que ce document soit diffusé en ligne

Résumé *Nous réalisons dans cette étude l'analyse de l'expression différentielle du transcriptome chez les gamétophytes mâles et femelles de l'algue brune modèle Ectocarpus siliculosus. Des approches d'assemblage avec génome de référence et des approches de novo ont été utilisées et sont discutées.*

Mots-clés RNAseq, analyse différentielle, assemblage de novo, assemblage avec génome de référence.

1 Introduction

Les bases moléculaires du déterminisme génétique du sexe et de la différenciation entre mâles et femelles ont beaucoup été étudiées chez les mammifères, les plantes et les champignons, mais rien n'est connu jusqu'à maintenant sur les mécanismes de détermination du sexe chez d'autres eucaryotes comme les algues brunes. L'identification et la caractérisation du locus contrôlant le caractère sexuel chez l'algue brune modèle *Ectocarpus* (dont le génome vient d'être entièrement séquencé [1]) est en cours dans l'équipe Génétique des Algues de l'UMR 7139 de la Station Biologique de Roscoff. Nous avons utilisé des techniques de séquençage haut-débit (Illumina HiSeq 2000) pour caractériser les transcriptomes de stades gamétophytiques mâles et femelles d'*Ectocarpus* avec l'objectif de identifier les gènes différentiellement exprimés chez les individus mâles versus femelles. En particulier, nous avons voulu comparer les approches d'analyse des données avec et sans génome de référence.

2 Matériel et Méthodes

Le séquençage a été réalisé par la société Fasteris (Suisse), sur des gamétophytes mâle et femelle matures, avec 2 réplicats biologiques et techniques pour chaque condition. Le séquençage a généré en moyenne 28 millions de reads par réplicat.

Un nettoyage des séquences a été fait à l'aide du logiciel « fastq », avec un filtrage des reads selon la qualité globale et un trimming des reads selon la qualité de chaque paire de base.

NEBULA

A web-server for advanced ChIP-seq data analysis.

Valentina BOEVA[1,2,3], Alban LERMINE[1,2,3], Camille BARETTE[1], Emmanuel BARILLOT[1,2,3]

[1]Institut Curie, 75005 Paris, France

[2]INSERM, U900, Bioinformatics and Computational Systems Biology of Cancer, 75248 Paris, France

[3]Mines ParisTech, 77300 Fontainebleau, France

Valentina.Boeva@curie.fr , Alban.Lermine@curie.fr , Camille.Barette@curie.fr , Emmanuel.Barillot@curie.fr

Abstract *We present a web service, Nebula, with which biologists can analyze their ChIP-seq data. ChIP-seq is chromatin immunoprecipitation followed by sequencing of the extracted DNA fragments. This technique allows accurate characterization of the binding sites of transcription factors and other DNA-associated proteins.*

Many existing tools for ChIP-seq data analysis are difficult to use by non-bioinformaticians. These tools map sequenced reads to the reference genome or predict binding site locations (ChIP-seq peaks). Several tools exist for peak filtering, motif discovery and genome feature association. Such tools are often command line applications or R packages.

Our web service, Nebula, was designed for biologists. It is based on the Galaxy open source framework. Galaxy already includes a large number of functionalities for mapping reads and peak calling. We added the following to Galaxy: (1) peak calling with FindPeaks and a module for immunoprecipitation quality control, (2) de novo motif discovery with ChIPmunk, (3) calculation of the density and the cumulative distribution of peak locations around gene TSSs, (4) annotation of peaks with genomic features, and (5) annotation of genes with peak information. Nebula generates the graphs and the enrichment statistics at each step of the process. During steps 3 to 5, Nebula optionally repeats the analysis on a control dataset and compares these results with those from the main dataset. Nebula can also incorporate gene expression (or gene modulation) data during these steps. In summary, Nebula is an innovative web service that provides an advanced ChIP-seq analysis pipeline, the output of which is directly publishable.

Availability: Nebula is available at http://nebula.curie.fr/.

Additional information: Nebula accepts mapped reads in SAM/BAM format. Each step of the pipeline produces several output files, which are mainly tab-delimited text files, .BED files or images. We used Perl and R to develop the tools used to perform the steps 3 to 5. The pipeline also includes several published tools (samtools, bedTools, MACS, FindPeaks, ChIPmunk). The most time consuming step in the pipeline is motif discovery (step 2), which takes several hours on a dataset containing ten thousand peaks. We provide a test dataset with which users can test the pipeline in less than 2 hours. The Nebula pipeline was tested by more than 30 scientists during an INSERM workshop in Marseille, France, in December 2011 (http://jacques.van-helden.perso.luminy.univmed.fr/INSERM_WS_211-212/).

Keywords ChIP-seq, Galaxy, peaks, motifs, genome feature association.

Computational modeling of FcεRI signaling during mast cell activation.

Anna Niarakis[1], Emrah Kamali[1], Yacine Bounab[2, 3], Marc Daëron[2, 3], Denis Thieffry[1]

[1]IBEns (ENS / CNRS UMR 8197 / INSERM U1024), Paris, France
niarakis@biologie.ens.fr, thieffry@ens.fr

[2]Institut Pasteur, Département d'Immunologie, Unité d'Allergologie Moléculaire et Cellulaire, Paris, France

[3]Inserm, Unité 760, Institut Pasteur, Paris, France
yacine.bounab@pasteur.fr, daeron@pasteur.fr

Keywords FcεRI signaling, mast cell activation, allergy, molecular map, dynamical modeling.

Summary

Mast cell activation is a pivotal event in the initiation of inflammatory reactions associated with allergic disorders. It is triggered by the aggregation of high-affinity IgE receptors (FcεRI), on the mast cell surface [1]. FcεRI aggregation is induced by the binding of a multivalent allergen to FcεRI-bound IgE antibodies. Mast cell activation is a complex process relying on multiple layers of tightly controlled intracellular signaling molecules, which form an intricate network [2, 3].

A global and rigorous understanding of the signaling and cross-regulatory processes involved in mast cell activation requires the integration of public and novel data into a comprehensive computational model.

Based on a survey of relevant data published in scientific journals or available in public databases, we are currently building and annotating a comprehensive regulatory map using the sofware CellDesigner [4]. This regulatory map currently encompasses 60 components, and more than 300 interactions, along with annotations and links to databases such as PubMED, EntrezGene and UniProt.

Our mast cell activation map will serve as a scaffold to generate a dynamical model of the underlying network, using a sophisticated logical modeling approach and the software GINsim [5]. Novel proteomic data will be used to delineate salient dynamical features of mast cell response under different conditions (e.g. how the FcεRI signaling network operates in the absence or in the presence of negative regulatory signals triggered by the FcγRIIB or by the transmembrane adaptor LAT2). To progressively improve the predictive power of the resulting model, computational results will be systematically confronted with experimental data.

Ultimately, our modeling analysis should contribute deepen our understanding of how the different functional outcomes of mast cell activation (degranulation, synthesis of lipidic mediators, induction of cytokine transcription) are articulated at the level of the underlying molecular network, and to delineate means to uncouple these functions and control them separately or collectively.

Acknowledgements

This work is funded by the ANR MI2 iSa project (2011-2014).

References

[1] Turner H, Kinet JP (1999). Signalling through the high-affinity IgE receptor Fc epsilonRI. *Nature Reviews* **402**: B24-30.

[2] Cao L, Yu K, Banh C, Nguyen V, Ritz A, Raphael BJ, Kawakami Y, Kawakami T, Salomon AR (2007). Quantitative time-resolved phosphoproteomic analysis of mast cell signaling. *Journal of Immunology* **179**: 5864-76.

[3] Gilfillan AM, Rivera J (2009). The tyrosine kinase network regulating mast cell activation. *Immunological Reviews* **228**: 149-69.

[4] Funahashi A, Matsuoka Y, Jouraku A.; Morohashi M, Kikuchi N, Kitano H (2008). CellDesigner 3.5: a versatile modeling tool for biochemical networks. *Proceedings IEEE* **96**: 1254-65.

[5] Chaouiya C, Naldi A, Thieffry D (2012). Logical modelling of gene regulatory networks with GINsim. *Methods in Molecular Biology* **804**: 463-79.

Logical modelling of MAPK network

Luca Grieco[1,2,3], Laurence Calzone[4], Andrei Zinovyev[4], Brigitte Kahn-Perlès[3], Denis Thieffry[1,3]
[1] IBENS (CNRS UMR 8197 / INSERM U1024), Paris, France
grieco@biologie.ens.fr, thieffry@ens.fr
[2] Université de la Méditerranée, Marseille, France
[3] TAGC (INSERM U1090), Marseille, France
kahn@marseille.inserm.fr
[4] INSERM U900, Institut Curie, Paris, France
laurence.calzone@curie.fr,andrei.zinovyev@curie.fr

Keywords: logical modelling, signalling network, MAPK pathway, cell fate decision, cancer.

Summary

Based on an extensive analysis of published data, we built a comprehensive molecular map for Mitogen-Activated Protein Kinase (MAPK) signalling network, using the CellDesigner software [1].

Mammalian MAPKs can be activated by a wide variety of stimuli, including growth factors and environmental stresses. Activation of MAPK pathways affects diverse cellular activities, including gene expression, cell cycle machinery, survival, apoptosis and differentiation. To date, three groups of MAPKs have been extensively studied: extracellular regulated kinases (ERK1/2), Jun NH2 terminal kinases (JNK1/2/3), and p38 kinases (p38 α/β/γ/δ).

A recurrent feature of MAPK pathways is a central three-tiered core signalling module, consisting of a set of sequentially acting kinases: a MAPK, a MAPK kinase (MAPKK) and a MAPK kinase kinase (MAPKKK). The MAPKKKs are activated by phosphorylation or by interaction with GTP proteins belonging to Ras/Rho family in response to extracellular stimuli. MAPKKKs activation leads to phosphorylation and activation of downstream MAPKKs, which in turn phosphorylate MAPK. Once activated, MAPKs phosphorylate the target substrates, which could be transcription factors, other kinases or proteins [2].

At each level of a MAPK cascade, protein phosphorylation is regulated by the opposing actions of phosphatases. Since the physiological outcome of MAPK signalling depends on the magnitude and duration of kinase activation, this regulation by phosphatases plays an important role.

Moreover, scaffold proteins bring together the components of a single pathway and insulate the module from activation by irrelevant stimuli and negative regulators (like phosphatases). However, the mechanism by which scaffold proteins act on MAPK cascades is not well known yet, so we provisionally ignored them in our model.

Finally, the three main MAPK cascades are strongly intertwined, making the prediction of cell fate decision an extremely difficult task.

Using the CellDesigner map as a reference, we derived a comprehensive logical model for the MAPK network, with the aim to reproduce the response of MAPK cascades to different stimuli and better understand their contributions to cell fate decision (between Apoptosis, Growth Arrest and Proliferation). The resulting logical model encompasses the three main MAPK cascades in response to four inputs: EGFR, FGFR3, TGFb, and DNA damage.

More precisely, our logical model includes the pathways starting from TGFb and DNA damage, leading to the activation of p38 and JNK cascades. Mainly based on the information integrated in the MAPK CellDesigner map, these pathways overlap in the MAPKs activation mechanisms, but their downstream targets of p38 and JNK are quite different.

In contrast, ERK cascade is mainly activated by growth factors. In particular, we consider EGFR and FGFR3 stimuli, whose different effects have been already studied in urothelial carcinoma (UC). A distinction between non-invasive (papillary) and invasive tumours has been documented [3, 4]. Non-invasive UC is characterised by a constitutive activation of the MAPK pathway, following mutations of FGFR3; the lower malignancy of this cancer may be due to the induction of cell cycle arrest by MAPK signalling. On the contrary, invasive UC shows alterations in the p53 and RB pathways, with the cells receiving growth signals by the MAPK pathway; epidermal growth factors seem to be over-expressed in this case.

The Boolean model, consisting of 54 variables, has been built using GINsim software [5, 6]. Exhaustive simulations of this model is beyond reach, given the huge dimension of its associated state transition graph (2^{54} nodes). We overcame this problem by taking advantage of a novel model reduction function implemented into GINsim [7], which preserves the attractors of the systems. Applying this reduction method to our MAK model, we obtained a 14-variable model, whose dynamics has been anlaysed in more details (determination of the attractors and of their reachability for different relevant initial states and for different genetic backgrounds). Our simulations show that the model dynamical behaviour is largely consistent with current biological knowledge.

Exploitation of our results led us to formulate novel hypotheses concerning the inter-play of the model components, suggesting wet experiments aimed at uncovering novel therapeutical targets. Such experiments will be finally performed in order to definitely validate our model.

Acknowledgements

This work was financed by the EU FP7 APO-SYS project (www.apo-sys.eu).

References

[1] A. Funahashi, Y. Matsuoka, A. Jouraku, M. Morohashi, N. Kikuchi, H. Kitano, CellDesigner 3.5: A Versatile Modeling Tool for Biochemical Networks. *Proceedings IEEE* 96(8):1254-65, 2008.

[2] M. Krishna, H. Narang H, The complexity of mitogen-activated protein kinases (MAPKs) made simple. *Cellular and Molecular Life Sciences*, 65:3525-3544, 2008.

[3] A.P. Mitra, R.H. Datar, R.J. Cote, Molecular pathways in invasive bladder cancer: new insights into mechanisms, progression, and target identification. *Journal of Clinical Oncology*, 24(35):5552-64, 2006.

[4] W.A. Schulz, Understanding urothelial carcinoma through cancer pathways. *International Journal of Cancer*, 119(7):1513-8, 2006.

[5] A. Naldi, D. Bérenguier, A. Fauré, F. Lopez, D. Thieffry, C. Chaouiya, Logical modelling of regulatory networks with GINsim 2.3. *BioSystems*, 97:134-9, 2009.

[6] C. Chaouiya, A. Naldi, D. Thieffry, Logical modelling of gene regulatory networks with GINsim. *Methods in Molecular Biology*, 804:463-79, 2012.

[7] A. Naldi, E. Remy, D. Thieffry, C. Chaouiya, Dynamically consistent reduction of logical regulatory graphs. *Theoretical Computer Science*, 412:2207-18, 2011.

Logical modelling of hematopoietic cell specification and reprogramming

Samuel Collombet[1], Cyrille Lepoivre[2], Denis Puthier[2], Thomas Graf[3], and Denis Thieffry[1,2]

[1] IBENS (CNRS UMR 8197 / INSERM U1024), Paris, France
{samuel.collombet, thieffry}@ens.fr
[2] TAGC-INSERM U1090, Marseille, France
{lepoivre, puthier}@tagc.univ-mrs.fr
[3] Center for Genomic Regulation, Barcelona, Spain
thomas.graf@crg.eu

Keywords: Hematopoiesis, transdifferentiation, regulatory network, logical modelling.

Hematopoiesis is a classic example of multi-lineage differentiation. Blood cells are derived from a common set of stem cells, which differentiate into more specific progenitors of erythroid, myeloid and lymphoid lineages, ultimately leading to functional cells such as erythrocytes, macrophages, B and T lymphocytes. This ontogenesis is controlled by a complex regulatory network involving environmental signals, as well as transcriptional and epigenetic factors.

Master regulators controls lineages differentiation and maintenance, forming the core regulatory network. For example, B cell development requires the expression of the cytokine receptor IL-7R, as well as the transcriptions factors E2F, E2A, Pax5, which activates the expression of a battery of cell-type specific genes involved in B cell functions. Macrophage development necessitates the CSF1 and IL3 receptors, as well as the PU1, CEBPa and CEBPb transcription factors. These factors regulate each other expression in a complex way, while some of them (e.g. PU1) are expressed and required in both cell types [1]. The ectopic expression of some of these factors, can induce the reprogramming of a cell type into another. For example B cells can be reprogrammed into macrophages by forcing the ectopic expression of CEBPa [2].

Using public data from molecular genetic experiments (qPCR, western blot, EMSA) or genome-wide essays (DNA-chip, ChIP-seq), we have built a regulatory network encompassing over 80 transcription factors and signaling components involved in myeloid (macrophages and neutrophils) and lymphoid (B and T cells) development.

Based on this network, we have developed a model using the logical modeling software GINsim [3], focusing on B cell and macrophage development. To date, this model recapitulates different experiments, including cytokine induced differentiation, pro-B cell reprogramming, and the effect of reported gene knock-downs.

We are currently extending this model to cover neutrophil and T-cell development, with the intention to predict the effect of various perturbations, including multiple gain- or loss-of-function. Interesting predictions will serve as a basis to design novel experiments.

Ultimately, this model analysis should provide useful insights into the molecular mechanisms controlling cell differentiation and reprogramming. Furthermore, simulations will be used to design promising experiments.

Acknowledgements

This work has been supported by a grant from the Belgian Federal Science Policy Office (PAI/IAP BioMaGNet).

References

[1] C.V. Laiosa, M. Stadtfeld and T. Graf, Determinants of lymphoid-myeloid lineage diversification. Annu. Rev. Immunol. 24: 705-738, 2006.

[2] LH. Bussmann, A. Schubert, T.P. Vu Manh, L. De Andres, S.C. Desbordes, M. Parra, T. Zimmermann, F. Rapino, J. Rodriguez-Ubreva, E. Ballestar and T. Graf, A robust and highly efficient immune cell reprogramming system. Cell Stem Cell 5: 554-566, 2009

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

From MOS1 to HsMAR-Ra, from C-ter to PEC by structure modelling

Jeanne CAMBEFORT[1, 2] and Corinne AUGE-GOUILLOU[1]

[1] Innovation Moléculaire et Thérapeutique, EA6306, Université de Tours,
UFR des Sciences Pharmaceutiques, Avenue Monge, 37200 Tours, France
{jeanne.cambefort, auge}@univ-tours.fr

[2] Génétique Immunothérapie Chimie et Cancer, UMR7292 CNRS

Abstract. *MOS1, SETMAR, and HsMAR-Ra are three closely related* mariner *transposases. The PDB contains two crystals of MOS1, the C-ter and the PEC, and two SETMAR C-ter structures. No HsMAR-Ra structure is available, thus preventing the improvement of specific inhibitors using docking approaches. We have made three C-ter structures of HsMAR-Ra: two by mutating C-ter crystals of SETMAR in Pymol, and the third in using Modeller. Our models were checked with the SAVES software. In future work, we would like to model the PEC of HsMAR-Ra thanks to the C-ter HsMAR-Ra obtained here.*

Keywords: Structure modelling, homology, MOS1, HsMAR-Ra.

1 Background

MOS1 est une transposase *mariner* appartenant à la super-famille des intégrases rétrovirales. C'est une transposase de drosophile, chez qui elle contrôle le cycle de transposition de l'élément génétique mobile, *Mos1*. Ce cycle comprend cinq étapes au cours desquelles la transposase adopte différentes conformations, dont le « PEC » (Paired-End Complex) dans lequel un dimère de transposase est associé aux extrémités ADN de l'élément qu'elle mobilise.

Nous avons découvert les premiers inhibiteurs de MOS1 [1], et montré que ces inhibiteurs étaient également actifs contre l'intégrase du VIH-1, une autre enzyme de la super-famille des intégrases rétrovirales. Toutes ces enzymes partagent un même domaine catalytique, appelé DDD/E. Dans le cadre du projet InhDDE (Région Centre, 2010-2012) nous proposions d'utiliser MOS1 et nos inhibiteurs comme modèles pour développer de nouvelles molécules actives contre le VIH-1. Pour cela, une stratégie impliquant l'amélioration des premiers inhibiteurs par docking moléculaire a été développée. MOS1 ne transposant pas en cellules humaines, nous utilisons HsMAR-Ra, une transposase *mariner* humaine [2] afin de valider *ex-vivo* nos modèles. HsMAR-Ra est la version ancestrale de la transposase codée par l'élément *Hsmar1*, présent en 200 copies - toutes inactives - dans le génome humain. Une copie fait exception : insérée et fixée en aval d'un gène *set* (codant une histone méthylase), elle fait partie d'un gène chimérique nommé *setmar* [3], codant une protéine (SETMAR) dont la partie MAR est homologue à la transposase HsMAR-Ra.

Toutes les transposases *mariner* sont organisées sur un même schéma : une partie C-terminale (C-ter) qui porte le domaine catalytique DDD et une partie N-terminale (N-ter), qui permet le fixation de la transposase aux extrémités de l'élément. Notre travail s'appuie sur plusieurs structures cristallographiques : la partie C-ter (2F7T [4]) et le PEC de MOS1 (3HOS et 3HOT [5]), ainsi que la partie C-ter de SETMAR (3K9J et 3K9K [6]). Aucune structure n'est disponible pour HsMAR-Ra, ce qui empêche tout docking. Nous voulons tout d'abord reconstituer la structure C-ter d'HsMAR-Ra par homologie avec les structures très similaires existantes (MOS1 et SETMAR). A moyen terme, nous espérons pouvoir reconstruire la structure du PEC d'HsMAR-Ra, plus complexe à réaliser.

2 C-terminal structure of HsMAR-Ra

Les séquences C-ter de MOS1 (résidus 119-345), HsMAR-Ra (résidus 118-343) et SETMAR (résidus 446-671) ont été alignées à l'aide de ClustalW (http://www.ebi.ac.uk/Tools/msa/clustalw2/). L'outil ProtScale (http://web.expasy.org/protscale/) a été utilisé pour analyser l'hydrophobicité des trois protéines.

Les parties C-ter de MOS1 et d'HsMAR-Ra partagent 40,09% d'identité de séquence, quant aux parties C-ter de SETMAR et de HsMAR-Ra, elles partagent 92,04 % d'identité de séquence (18 acides aminés différents sur 226).

Les alignements de structure (basé sur les Cα) entre MOS1 et SETMAR ont été réalisés avec CLICK [7]. Les domaines C-ter de SETMAR et de MOS1 s'alignent à 92,16% (RMSD 1.19), et le domaine C-ter de SETMAR s'aligne à 92,79% avec la partie correspondante du PEC de MOS1 (RMSD 1.27).

Puis nous avons utilisé Pymol (http://www.pymol.org/) pour muter dans les deux structures de SETMAR (3K9J/3K9K) les 18 acides aminés différents entre SETMAR et HsMAR-Ra, et ainsi obtenir deux structures d'HsMAR-Ra : HC-ter1 et HC-ter2. Pour finir, nous avons utilisé Modeller [8] pour créer une troisième structure (HC-ter3) d'HsMAR-Ra par homologie structurale avec SETMAR (3K9J/3K9K) et MOS1 (2F7T, 3HOS/3HOT). Nous avons affiné les boucles, lorsque ceci était justifié.

Pour valider nos modèles, nous avons utilisé le serveur SAVES (Structure Analysis and Verification Server, http://nihserver.mbi.ucla.edu/SAVES/) qui évalue la qualité stéréochimique et le repliement 3D. Parmi les trois modèles (HC-ter1, HC-ter2, HC-ter3), nous avons retenu le modèle de plus basse énergie.

3 Future works: Toward the PEC structure of HsMAR-Ra

L'intérêt scientifique de cette analyse réside dans la future prédiction par homologie du PEC d'HsMAR-Ra, en utilisant le PEC de MOS1. Etant donné que nous disposons maintenant de la structure C-ter d'HsMAR-Ra, il reste à (1) modéliser la région N-terminale d'HsMAR-Ra et (2) replacer l'ADN dans les structures obtenues.

Les séquences N-ter de MOS1 (résidus 1-118) et d'HsMAR-Ra (résidus 1-117) sont identiques à 28,21% seulement. Une approche par étape sera donc nécessaire. Nous nous baserons sur la phylogénie des régions N-ter des transposases *mariner* obtenue au laboratoire en 2000 [9], et nous comparerons les séquences candidates avec MOS1 et HsMAR-Ra. Par homologie avec la structure de MOS1, nous construirons un modèle intermédiaire (nommé XN-ter) de structure N-ter d'une transposase *mariner* située entre MOS1 et HsMAR-RA. Une fois le modèle XN-ter créé et validé par la méthode décrite précédemment, nous modéliserons la partie N-ter de HsMAR-Ra, par homologie avec le modèle XN-ter.

Les dernières étapes nécessaires pour construire la structure PEC d'HsMAR-Ra consistent à : utiliser Modeller pour associer les deux structures C-ter et N-ter précédemment modélisées ; finaliser le modèle de PEC en y introduisant l'ADN du transposon. A terme, ce modèle permettra d'améliorer nos inhibiteurs (dirigés contre MOS1 ou HsMAR-Ra) en utilisant une approche rationnelle basée sur le docking.

References

[1] N. Bouchet, et al., First Mariner Mos1 transposase inhibitors. *Mini Rev Med Chem*, 9(4):431-439, 2009.
[2] H. M. Robertson, K. L. Zumpano, Molecular evolution of an ancient mariner transposon, Hsmar1, in the human genome, *Gene*, 205:203-217, 1997.
[3] R. Cordaux, et al., Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. *Proc. Natl. Acad. Sci. U. S. A.*, 103:8101-8106, 2006.
[4] J. M. Richardson, et al,. Mechanism of Mos1 transposition: insights from structural analysis. *EMBO J.*, 25:1324-1334, 2006.
[5] J. M. Richardson, et al., Molecular architecture of the MOS1 paired-end complex, the structural basis of DNA transposition in a eukaryote. *Cell.*, 138(6):1096-1108, 2009.
[6] K. D. Goodwin, et al., Crystal structure of the human *Hsmar1*-derived transposase domain in the DNA repair enzyme Metnase. *Biochemistry*., 49(27):5705-5713, 2010.
[7] M. N. Nguyen, M. S. Madhusudhan, Biological insights from topology independent comparison of protein 3D structures. *Nucleic Acids Res.*, 39(14):e94, 2011.
[8] A. Sali, T. L. Blundell, Comparative protein structure modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, 234:779-815, 1993.
[9] C. Augé-Gouillou, et al., Phylogenetic analysis of the functional domains of *mariner*-like élément (MLE) transposases. *Mol. Gen. Genet.*, 264 :506-513, 2000.

A comparison of two statistical methods combining high-throughput data to predict the level of disease activity in patients with Rheumatoid Arthritis

Jonathan PLASSAIS[1,2], Julien CHIQUET[2], Alessandra C. L. CERVINO[1] and Christophe AMBROISE[2]
[1] TcLand Expression, 21 rue de la Noue Bras de Fer, 44200, Nantes, France
{jplassais, acervino}@tcland-expression.com
[2] Laboratoire Statistique et Génome, 23 boulevard de France, 91037 Evry, France
{christophe.ambroise, julien.chiquet}@genopole.cnrs.fr

Keywords Meta-analysis, multi-task, cooperative-Lasso, gene expression, biomarker, DAS28, rheumatoid arthritis.

Une comparaison de deux méthodes statistiques pour la combinaison de données en grande dimension afin de prédire le niveau d'activité de la maladie chez des patients atteints de polyarthrite rhumatoïde

Mots-clés méta-analyse, multi-tâches, cooperative-Lasso, expression génique, biomarqueur, DAS28, polyarthrite rhumatoïde.

1 Introduction

Disease activity for patients with Rheumatoid Arthritis (RA) is routinely measured in Europe using the Disease Activity Score (DAS28).The DAS28 is a continuous score based on multiple variables such as the count of tender joints and swollen joints, the general health score, and the level of C-reactive protein (CRP) or the level of erythrocyte sedimentation rate (ESR) [1]. Recently, a blood based biomarker has been developed by Crescendo Bioscience (Vectra™ DA) to estimate the DAS28(CRP) [2]. Vectra DA shows the feasibility of developing molecular biomarkers based on protein measurements to predict DAS28.

Although the DAS28 has been reported to be a reproducible score [1], it is known that there can be subjective differences based on who performs the clinical assessment and on how the patient feels at that moment. Accurate and objective measurements of disease activity are of key importance. One way to achieve this goal is through the use of biomarkers.

To this end, we identified a gene list, based on microarray data, by comparing two combinatorial methods, a meta-analysis approach by applying the inverse variance technique [3] and a penalized regression approach in a multi-task setup with the cooperative-Lasso [4].

2 Methods

With carefully selected inclusion criteria, we identified two microarray studies [5,6] from public repositories with whole blood samples. Following the recommendations from Ramasamy and collegues [7], we performed a six-step protocol: 1) identification of relevant datasets, 2) phenotype homogenization, 3) array processing, 4) quality controls for individual datasets, 5) probe to gene mapping (selecting only one probe for one gene by blasting all probes against the genome for each platform) and 6) choice of algorithm to combine all datasets. For the last step, we chose to focus only on two methods, the state-of-the-art inverse variance technique to perform a meta-analysis [3] and a new combined method, the cooperative Lasso to perform a multi-task analysis [4].

The meta-analysis method consists of two phases: first, the estimation of individual effects; and second, the aggregation phase. We estimated the Pearson's coefficient of correlation between gene expression and the DAS28 to determine the individual gene effects within each study. Z-scores were evaluated by applying a Fisher transformation. Under the null hypothesis that the gene expression levels do not correlate to the

DAS28, Z-scores follow a normal distribution. For the second phase, all Z-scores were aggregated between studies with the inverse variance technique and a Random Effect Model (REM).

Concerning the multi-task approach, we assumed that one task corresponds to one microarray study. The cooperative-Lasso is directly inspired from the Lasso procedure [8] with a group sign-coherent criterion to aggregate microarrays datasets. A group corresponds to an individual gene across all microarray experiments, the group size is the number of studies. The tuning parameter for penalization is chosen with 5-fold cross-validation and the final coefficients are aggregated with a Euclidean norm.

Finally, to compare the meta-analysis and multi-task approaches, we performed simulations from a multivariate linear model. To mimic a problem of response variability, we assumed different levels of Signal to Noise Ratio (SNR). Results were compared in term of Area Under the Receiver Operating Characteristic (AUROC).

3 Results

Through simulations, we showed that the meta-analysis and the multi-task are both more robust than independent approaches: we considered three simulated studies with different levels of SNR (signal to noise ratio), 10, 1 and 0.2. Performances of the combined approaches were evaluated by means of AUROC and compared with the independent approaches (meta-analysis with individual p-values and Lasso procedure on each study). Both meta-analysis and multi-task methods show AUROC values similar to the study with the highest level of SNR and thus show the robust character of these methods.

We then applied the algorithms to the two microarray data of peripheral blood sampled from 99 rheumatoid arthritis patients with an average DAS28 of 5.8. From 3931 selected genes, we identified a 43 gene signature with the meta-analysis and a 4 gene signature with the multi-task. The intersection between these two lists is weak. The two methods differ in various points. First the meta-analysis does not consider the correlation structure between genes on each dataset, unlike the multi-task. As such multi-task only identify genes that do not correlate to each other. Secondly, the multi-task tends to select a smaller subset of genes, contrary to the meta-analysis, despite of the multi-testing correction.

Considering the meta-analysis and the multi-task as a gene selection tool, we developed a prediction model based on these two gene lists. For each gene list, we applied a Lasso regression and we estimated the correlation between the prediction and the DAS28. We obtained a correlation of 0.82 and 0.74 for the meta-analysis gene list and a correlation of 0.84 and 0.38 for the multi-task gene list. These results are similar to those obtained with a protein biomarker but need to be validated in an external cohort.

References

[1] J. Fransen and P.L.C.M van Riel, The disease activity score and the EULAR response criteria, *Clin Exp Rheumatol.*, 23(suppl. 39):S93-S99, 2005.

[2] M. Bakker, Y. Shen, J.W.J. Bijlsma, J. Jacobs, F.P.J.G. Lafeber, and G. Cavet, Development of a multi-biomarker test for Rheumatoid Arthritis (RA) Disease Activity (Vectra DA) [abstract], *Arthritis Rheum.*, 68(suppl 10):1753, 2010.

[3] J.L. Fleiss, The statistical basis of meta-analysis, *Statistical Methods in Medical Research*, 2(2):121-145, 1993.

[4] J. Chiquet, Y. Grandvalet and C. Charbonnier, Sparsity with sign-coherent groups of variables via the cooperative-lasso, *http://imstat.org/aoas/next_issue.html*, 2012.

[5] J.R. Bienkowska, G.S. Dalgin, F. Batliwalla, N. Allaire, R. Roubenoff, P.K. Gregersen, and J.P. Carulli, Convergent Random Forest predictor: methodology for predicting drug response from genome-scale data applied to anti-TNF response, *Genomics*, 94(6):423-32, 2009.

[6] A. Julià, A. Erra, C. Palacio, C. Tomas, X. Sans, P. Barceló, and S. Marsal, An Eight-Gene Blood Expression Profile Predicts the Response to Infliximab in Rheumatoid Arthritis, *PLoS One*, 4(10):e7556, 2009.

[7] A. Ramasamy, A. Mondry, C.C. Holmes and D.G. Altman, Key issues in conducting a meta-analysis of gene expression microarray datasets, *PLoS Medicine*, 5(9):1320:1332, 2008.

[8] R. Tibshirani, Regression shrinkage and selection via the Lasso, *Journal of the Royal Statistical Society Series B*, 58(1):267-288, 1996.

TriAnnot: A High Performance Pipeline for the Automated Structural and Functional Annotation of Plant Genomes - New developments.

Philippe LEROY[1], Nicolas GUILHOT[1], Sébastien THEIL[1], Frédéric CHOULET[1], Hiroaki SAKAI[3], Michael ALAUX[2], 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, barley, rice, maize, oak.

In support of the international effort (IWGSC – *International Wheat Genomic Sequencing Consortium* - http://www.wheatgenome.org) to obtain a reference sequence of the bread wheat genome and provide plant communities dealing with large and complex genomes with a versatile, easy-to-use online automated tool for annotation, we have developed the TriAnnot pipeline [1]. Its modular architecture allows for the annotation and masking of transposable elements, the structural and functional annotation of protein-coding genes with an evidence-based quality indexing, and the identification of conserved non coding sequences and molecular markers. The performance of TriAnnot was evaluated in terms of sensitivity, specificity, and general fitness using curated reference sequence sets from rice (IRGSP/RAP build 5 - released on December 2009, last updated on August 2010) and wheat [2]. In less than 8 hours, TriAnnot was able to predict more than 83% of the 3,748 CDS from rice chromosome 1 with a fitness of 67.4%. On a set of 12 reference Mb-sized contigs from wheat chromosome 3B, TriAnnot predicted and annotated 93.3% of the genes among which 54% were perfectly identified in accordance with the reference annotation. It also allowed the curation of 12 genes based on new biological evidences, increasing the percentage of perfect gene prediction to 63%. TriAnnot systematically showed a higher fitness than other annotation pipelines that are not improved for wheat. The TriAnnot pipeline is parallelized on a 712 CPU computing cluster that can run a 1 Gb sequence annotation in less than five days. It is accessible through a web interface for small scale analyses or through a server for large scale annotations. For the later, the pipeline is launched automatically using FASTA files. After completion, the structural and functional annotation can be viewed through an online GBrowse and can be manually curated using Artemis (http://www.sanger.ac.uk/resources/software/artemis/) and GenomeView (http://genomeview.org/) graphical viewers which are complementary. As it is easily adaptable to the annotation of other plant genomes, TriAnnot should become a useful resource for the annotation of large and complex genomes in the future. Therefore, TriAnnot is currently improved for other plant genomes structural and functional annotations such as barley, rice, maize and oak species. Full description of the TriAnnot pipeline is available at http://www.clermont.inra.fr/triannot. Release 3.5 is on line in April 2012, and a new release 3.6 is underway.

References

[1] P. Leroy, N. Guilhot, H. Sakai, A. Bernard, F. Choulet, S. Theil, S. Reboux, N. Amano, T. Flutre, C. Pelegrin, H. Ohyanagi, M. Seidel, F. Giacomoni, M. Reichstadt, M. Alaux, E. Gicquello, F. Legeai, L. Cerutti, H. Numa, T. Tanaka, K. Mayer, T. Itoh, H. Quesneville, C. Feuillet, TriAnnot: a versatile and high performance pipeline for the automated annotation of plant genomes. *Frontiers in Plant Sciences,* 3:1-14, 2012

[2] 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.

Development of a Fully Flexible Protein-Protein Docking Method combining Normal Modes and Genetic Algorithm

Diego E. BARRETO GOMES[1], Luis Paulo B. SCOTT[2], Pedro G. PASCUTTI[1], Paulo M. BISCH[1], David PERAHIA[3]

[1] Instituto de Biofísica Carlos Chagas Filho, Universidade Federal do Rio de Janeiro, Avenida Carlos Chagas Filho, 373 Bloco G, Sala G1-019, Rio de Janeiro - RJ, Brazil, CEP 21941-902
{diego, pascutti, bisch}@biof.ufrj.br

[2] Centro de Matemática, Computação e Cognição, Universidade Federal do ABC, Rua Santa Adélia, 166. Bairro Bangu. Santo André - SP - Brasil. CEP 09.210-170, adresse, lieu, CodePostal, Ville, Cedex xx, France
luis.scott@ufabc.edu.br

[2] Laboratoire de Biologie e Pharmacologie Apliquée École Normale Supérieure de Cachan, UMR8113 CNRS, 61 Av du Président Wilson. Cachan cedex, 94235, France
david.perahia@ens-cachan.fr

Abstract *Protein-protein docking (PPD) is a multidimensional problem. In the binding processes conformational changes may occur and extend beyond local structure rearrangements. An efficient assessment of these changes remains a fundamental challenge. The linear combination (LC) of normal modes (NM) emerges as an effective approach to capture the structural transitions during binding. Here we propose a new PPD method that relies on: i. A genetic algorithm (GA) that evaluates the LC of NM space, for receptor and ligand, thru targeted molecular dynamics simulations and minimization, and produces populations of refined structures for docking. ii. A GA that finds the best receptor-ligand docking pairs within these populations. In every generation, the docking results are re-ranked by ZRANK, an improved scoring function (SF). The method was implemented as a highly parallel program, suited for clusters running the PBS scheduler, and examined on a subset of the PPD benchmark 4. The first GA rapidly evolves the candidate structures towards low energy conformations, however RMSD comparison shows that only some matched the known structural changes upon binding. Nevertheless, the docking algorithm should find the correct binding partners. The GA-conducted docking also quickly converged, though the near correct binding conformations, found in the first GA, were fairly sampled. In fact, those are poorly ranked, even with a sophisticated re-scoring function; as a result the false positives prevail. That is indeed a limitation of current SFs, addressed by the CAPRI competition. Combination of SFs was found to improve protein-ligand docking and could be considered as an alternative to enhance PPD.*

Keywords Docking, Genetic Algorithm, Normal Modes, Protein-Protein

The CycADS annotation database system to support the development and update of enriched BioCyc databases

Development of *ad hoc* BioCyc DB : AcypiCyc, ArthropodaCyc, …

Patrice Baa-Puyoulet[1,4], Augusto F. Vellozo[2,4], Jaime Huerta-Cepas[3], Gérard Febvay[1,4], Toni Gabaldon[3], Marie-France Sagot[2,4], Hubert Charles[1,4] and Stefano Colella[1,4]

[1] Biologie Fonctionnelle Insectes et Interactions, UMR203 INRA INSA Lyon BF2I, bat INSA Pasteur, 20 ave Albert Einstein, 69621, Villeurbanne Cedex, France
{patrice.baa-puyoulet, gerard.febvay, stefano.colella}@lyon.inra.fr, hubert.charles@insa-lyon.fr

[2] Laboratoire de Biométrie et Biologie Évolutive, UMR5558 CNRS Université Lyon 1, bat Grégor Mendel, 43 bd du 11 novembre 1918, 69622, Villeurbanne Cedex, France
augusto@vellozo.org, marie-france.sagot@inria.fr

[3] Centre for Genomic Regulation, Barcelona Biomedical Research Park, Barcelona, Spain
{jaime.huerta, toni.gabaldon}@crg.es

[4] BAMBOO, INRIA Rhône-Alpes, France

Keywords: metabolism, arthropods, gene annotation, metabolic pathways.

1 The CycADS Software Project

The Cyc Annotation Database System (CycADS) project started in 2008 during the genome annotation for the pea aphid, *Acyrthosiphon pisum*[1]. Since the early stages of the quest for all metabolism related genes/proteins in the genome, it was clear that an annotation data management system was needed to allow us (and others) to easily create and further update the BioCyc metabolism network reconstruction of the pea aphid. CycADS allows the collection of heterogeneous annotation information to create dedicated files that are processed with the PathwayTools Software (SRI International)[2] to produce BioCyc interfaces.

2 The CycADS Pipeline

CycADS[3] is centred on an *ad hoc* SQL database, complemented by a set of Java modules to import and export relevant information. Genomic sequence data from GenBank or from other genome community repositories, as well as data from different gene functional annotation tools (such as Blast2GO[4], KAAS[5], PRIAM[6], PhylomeDB[7], etc.), are collected into the database and later extracted, to generate an enriched input file in order to build or update a BioCyc-like database using PathwayTools. Thanks to its flexibility and text mining capabilities, CycADS can use many data sources in tabular flat file format. The CycADS pipeline, beyond allowing an easy update of a BioCyc database over time, can be applied for the metabolic reconstruction of organism for which the sequence of the genome becomes available. Using the CycADS system we can reconstruct, quickly and with the same set of criteria, different species metabolic networks, allowing comparative network analysis on different organisms.

The pipeline includes entry points for Java implemented functions grouped in two main modules, the Annotation Collector and the Annotation Generator. The **Annotation Collector** includes text parsers for commonly used file format such as GFF or Gb (GenBank), and generic text parsers for heterogeneous data file formats. Each of them can be configured to locate and extract the relevant information that will be processed by the program factory and stored in the SQL database. A score system can be used for later filtering. Thereby we decided to set an equivalent score to each inference of annotation method Blast2GO, PRIAM, KAAS-Gene and KAAS-Eukaryote (two different annotations in KEGG) for the EC numbers and PhylomeDB for the GO annotation. The **Annotation Generator** can be set for extracting the information from the SQL database. CycADS is highly configurable in data filtering, for example using the score system threshold it can produce files

containing more or less relevant information. Furthermore, CycADS allows the integration in the BioCyc database of specific links to custom information resources (in AcypiCyc: AphidBase, KEGG orthology and PhylomeDB) together with the already existing databases cross references included using PathwayTools (e.g. GenBank, BRENDA, Gene Ontology, ...). Filtering is possible on all objects managed by the system, for example the feature type, the annotations by scores or methods, the cross-references to external databases, etc. The output can be the Pathologic file format used by the PathwayTools software or a customized flat file. All the program function calls can be concatenated in a shell script to automate the workflow.

3 From AcypiCyc to ArthropodaCyc

CycADS has been successfully used to generate **AcypiCyc**[1], the pea aphid BioCyc database, and we decided to build a metabolic network database for other arthropods, for which the genome sequence is available. We kept the same workflow parameters for collecting data from all used annotation methods (Blast2GO, KAAS, PRIAM). The generated **ArthropodaCyc**[2] database includes, at present, metabolic reconstructions for 11 arthropods: *Acyrthosiphon pisum*, *Aedes aegypti*, *Anopheles gambiae*, *Apis mellifera*, *Culex quinquefasciatus*, *Daphnia pulex*, *Ixodes scapularis*, *Nasonia vitripennis*, *Pediculus humanus corporis*, *Tribolium castaneum* and *Drosophila melanogaster* (for this last species, both the CycADS version and the FlyCyc database manually curated by the FlyBase team are available). Collecting and organizing information into databases is useful for the researchers studying the metabolism of their newly sequenced model organisms (more arthropod genomes will be sequenced in the near future through the i5K Arthropod Sequencing Initiative), and it allows them to better understand different aspects of arthropod biology through comparative studies. Thanks to the CycADS software, we included, in each database, information on annotation sources and links to genomics databases (including AphidBase, BeetleBase, VectorBase, Hymenoptera Genome Database, FlyBase and wFleaBase). Future plans include adding other sequenced genomes to ArthropodaCyc and the generation of another BioCyc-like database centred on the arthropod endosymbiosis for which both host and symbiont genomes have been sequenced (a beta version of **ArtSymbioCyc**[3] is already available).

Acknowledgements

This work was supported by the ANR-BBSRC MetNet4SysBio project and INRA.

References

1. The International Aphid Genomics Consortium, *Genome sequence of the pea aphid Acyrthosiphon pisum*. PLoS Biol, 2010. 8(2): p. e1000313.
2. Karp, P., S. Paley, and P. Romero, *The Pathway Tools software*. Bioinformatics, 2002. 18(Suppl 1): p. S225.
3. Vellozo, A.F., A.S. Véron, P. Baa-Puyoulet, J. Huerta-Cepas, L. Cottret, G. Febvay, F. Calevro, Y. Rahbé, A.E. Douglas, T. Gabaldón, M.-F. Sagot, H. Charles, and S. Colella, *CycADS: an annotation database system to ease the development and update of BioCyc databases*. Database, 2011. 2011: p. bar008.
4. Conesa, A., S. Götz, J.M. García-Gómez, J. Terol, M. Talón, and M. Robles, *Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research*. Bioinformatics, 2005. 21(18): p. 3674-6.
5. Moriya, Y., M. Itoh, S. Okuda, A.C. Yoshizawa, and M. Kanehisa, *KAAS: an automatic genome annotation and pathway reconstruction server*. Nucleic Acids Res, 2007. 35(Web Server issue): p. W182-5.
6. Claudel-Renard, C., C. Chevalet, T. Faraut, and D. Kahn, *Enzyme-specific profiles for genome annotation: PRIAM*. Nucleic Acids Res, 2003. 31(22): p. 6633-9.
7. Huerta-Cepas, J., A. Bueno, J. Dopazo, and T. Gabaldón, *PhylomeDB: a database for genome-wide collections of gene phylogenies*. Nucleic Acids Res, 2008. 36(Database issue): p. D491-6.

[1] http://acypicyc.cycadsys.org/
[2] http://arthropodacyc.cycadsys.org/
[3] http://artsymbiocyc.cycadsys.org/

Combination of in silico and proteomic approaches to identify candidate genes responsible for the immunomodulatory properties of Propionibacterium freudenreichii

Caroline LE MARECHAL[12], Mahendra MARIADASSOU[3], Valentin LOUX[3], Amal HAMMANI[3], Julien BURATTI[3], Julien JARDIN[12], Valérie BION[12], Stéphanie-Marie DEUTSCH[12], Benoit FOLIGNE[4567], Gwenaël JAN[12], Hélène FALENTIN[12]

[1] INRA, UMR 1253, Science et Technologie du Lait et de l'Oeuf, 35000, Rennes, France

{caroline.lemarechal, julien.jardin, valerie.bion, stephanie-marie.deutsch, gwenael.jan, helene.falentin}@rennes.inra.fr

[2] AGROCAMPUS OUEST, UMR1253, Science et Technologie du Lait et de l'Oeuf, F-35000 Rennes, France UMR Science et Technologie du Lait et de l'Œuf, 35310, Rennes, France

[3] INRA, Mathématique Informatique et Génome, 78 352, Jouy en Josas , France

{mahendra.mariadassou, valentin.loux, amal.hammani, julien.buratti}@jouy.inra.fr

[4] Institut Pasteur de Lille, Lactic Acid Bacteria & Mucosal Immunity, Center for Infection and Immunity of Lille, 1, Rue du Pr Calmette,BP 245, F-59019 Lille, France

benoit.foligne@ibl.fr

[5] Université Lille Nord de France, 59000 Lille, France

[6] CNRS, UMR 8204, 59021 Lille, France

[7] Institut National de la Sante´ et de la Recherche Médicale, U1019, 59019 Lille, France

Abstract *Recent works have revealed highly strain-dependent immunomodulatory properties in Propionibacterium freudenreichii, a cheese starter bacterium. The underlying mechanisms are unknown but preliminary attempts have shown that surface components are involved in these immunomodulatory effects. The most promising strains/components exert promising anti-inflammatory effects through induction of regulatory cytokines.The aim of this work is to identify key surface components of Propionibacterium freudenreichii responsible for the immunomodulatory properties.*

de novo transcriptome assembly pipeline for *Scyliorhinus canicula* NGS data

Pierre PERICARD[1], Wilfrid CARRÉ[1], Christophe CARON[1] , Erwan CORRE[1] and Sylvie MAZAN[2]

[1] ABiMS, FR2424 CNRS-UPMC, Station Biologique, Place Georges Teissier, 29680, Roscoff, France
{pierre.pericard, wilfrid.carre, christophe.caron, erwan.corre}@sb-roscoff.fr
[2] UMR 7150 CNRS-UPMC, Station Biologique, Place Georges Teissier, 29680, Roscoff, France
sylvie.mazan@sb-roscoff.fr

Abstract. *Within the context of Dogfish (Scyliorhinus canicula) genome sequencing, mRNA libraries from different stages and tissues have been constructed and sequenced. We present here the pipeline used for the de novo transcriptome assembly of these sequencing data. A detailed workflow has been developed for the read cleaning and pre-processing steps, which emerged as critical in the final assemblies content and quality. Two complementary approaches have been used to obtain reliable and exhaustive assemblies and will be discussed.*

Keywords. *Transcriptome, Dogfish, Scyliorhinus canicula, chondrichtyans.*

Le ou les auteur(s) ne souhaite(nt) pas que ce document soit diffusé en ligne

Résumé. *Dans le cadre du séquençage du génome de la Roussette (Scyliorhinus canicula), des librairies d'ARN messagers provenant de différents stades et tissues ont été réalisées et séquencées. Nous présentons ici l'assemblage de transcriptome de novo à partir de ces séquences. Nous avons développé un pipeline détaillé pour le nettoyage et le prétraitement des données, qui se sont révélés des étapes cruciales pour le contenu et la qualité des assemblages finaux. Deux approches complémentaires ont été utilisées afin d'obtenir des assemblages fiables et exhaustifs, et elles seront discutées.*

Mots-clés. *Transcriptome, Roussette, Scyliorhinus canicula, chondrichtyens.*

1 Introduction

Despite an increasing number of vertebrate genomes available, most of them belong to mammals whereas other phylogenetic branches stay well under-represented. In chondrichtyans, a key group in vertebrates, only *Callorhinchus milii* (elephant shark) and *Leucoraja erinacea* (little skate) genomes have been released. Within the context of the *S. canicula* genome sequencing project and a project in collaboration with a pharmaceutical private partner, mRNA libraries have been constructed and sequenced in order to get transcriptomic data. Depending on the usage and purpose of these data, the assembly will have to fulfil different objectives. In one end an assembly with as few alternative transcripts as possible for annotation purpose and in the other end an exhaustive assembly. With these two objectives in mind, we present here the assembly pipeline we developed, emphasizing on the critical effects of cleaning and pre-processing steps to the final assemblies contents and quality.

2 Material and Methods

Two cDNA normalised libraries have been constructed from a pool of mRNA extracted from multiple adult tissues and embryonic stages of *S. canicula*. These libraries, one oriented, the other non-oriented, resulted respectively in 316,658,386 and 412,793,832 paired-end 101nt reads. In addition, 223,261 Sanger EST sequences (>700nt long) were also available.

Sequencing reads have been trimmed and filtered based on Ns, sequencing quality, length, low complexity and poly A/T tails. Remaining primers and adaptators have been removed, as well as

Molecular dynamics on truncated dystrophin in Becker Muscular Dystrophies

Aurélie NICOLAS [1], Emmanuel GIUDICE [1], Olivier DELALANDE [1], Frédérique BARLOY-HUBLER[1] and Elisabeth LE RUMEUR[1]

[1] UMR6290 CNRS, Avenue du Professeur Léon Bernard, CS 34317, 35043 Rennes Cedex
{aurelie.nicolas, emmanuel.giudice, olivier.delalande, frederique.hubler, elisabeth.lerumeur}@univ-rennes1.fr

Keywords Dystrophin, Becker Muscular Dystrophy, molecular dynamics.

1 Introduction

Dystrophin, encoded by the largest human gene *DMD*, is a 427 kDa sarcolemmal protein found predominantly in skeletal and cardiac muscles [1]. Dystrophin is composed of four structural domains with a major central rod domain composed of 24 repeats of about 100-110 residues and four flexible hinges. Each repeat has an identifiable structure in α-helical triple coiled-coil homologous to spectrin repeats while hinges are mostly unstructured regions. The succession of these repeats gives a filamentous structure to dystrophin, which plays a key role during contraction-stretch cycles in muscle by stabilizing sarcolemma against mechanical forces. However this specific structure is extremely hard to resolved experimentally and there is no atomic structure of dystrophin repeats available which is why our team has recently constructed structural homology models for the 24 repeats of dystrophin [2].

Hundreds of mutations, predominantly exon deletions, that conserve the open reading frame or not, have been observed in patients and lead respectively to, Becker (BMD) and Duchenne (DMD) Muscular Dystrophies. DMD is a severe disease that considerably reduces patient's life span and is therefore a target for gene therapy. In BMD, the reading frame is conserved and the deletions lead to internally truncated dystrophin molecules. These dystrophins are partially functional and have been used as patterns for the design of dystrophin to be expressed in gene therapy strategies. The pathology is generally less severe but present a large spectrum of severity depending on the mutation [3]. Therefore, a structural study of these pathological dystrophins, and in particular the deletion-flanking regions, is highly relevant to understand the molecular basis of the different phenotypes observed in BMD patients and to design better constructs for gene therapy.

Here we present the initial results of the modeling of the five most commonly observed mutations in BMD patients. We compare the structures and dynamics of the internally truncated dystrophin central domain with the corresponding region of the native dystrophin. Finally, we correlate our results (flexibility, stability) with experimental data to try to explain the differences in phenotypes observed.

2 Material and Methods

A collection of six proteins in the region of repeats 16 to 21 of central domain comprising the native and five internally truncated proteins were analyzed. Homology models were generated using I-Tasser online server and optimized with YASARA software. Molecular dynamics of these proteins were simulated with the program NAMD and the CHARMM force field on local and national computer facilities (CINES, IDRIS). Models were solvated with a 35x35x25Å thick layer of TIP3P water around the protein and neutralized with minimal salt conditions. After energy minimization systems were slowly heated and simulated for 30 ns under periodic boundary conditions, constant temperature (310K) and pressure (1 atm). The post-processing analysis of the trajectories was performed with GROMACS and VMD.

3 Results

The homology model of the native dystrophin R16-21 (Fig 1A) is composed of six repeats; each repeat is composed of 3 α-helices (A, B, C) joined with loops and folded in a triple coiled-coil as expected. Consecutive tandem repeats are joined by a common α-helix formed by the C helix of the first repeat and the A helix of the second repeat. But repeats 19 and 20 are connected by a mostly unstructured hinge. The homology models of the truncated proteins are divided in two distinct cases: i) truncated dystrophins that disrupt the spectrin repeat leading to very distorted dystrophin (exon deletion 45-49) ii) truncated dystrophins that reconstitute a hybrid repeat (exon deletions 45-51). Models containing these hybrid repeats conserve the same three-dimensional structure as in native dystrophin (fold in a triple coiled-coil) and maintain the filamentous structure.

Dynamic fluctuation of the native dystrophin R16-21 globally shows a higher flexibility in loops than in helices, and a well-maintained coiled-coil structure. However, repeats 19 and 21 fluctuate more than the other repeats (Fig 1A). Dynamic fluctuation of the dystrophin R16-21 del45-51 that reconstitutes a hybrid repeat shows the same flexibility profile than the native with the exception of truncated repeat 20, which increased rigidity, can be explained by hinge 3 deletion (Fig 1B). This may explain the overall mild symptoms observed in BMD patients presenting this deletion. On the contrary R16-21 del45-49 present a quite different flexibility profile. The initial conformation is very distorted when compared to coiled-coil, but remain stable along the simulation. However the angle between R16 and the rest of the molecule can jeopardize the interaction with nNOS and/or lipids, explaining the difference in phenotype.

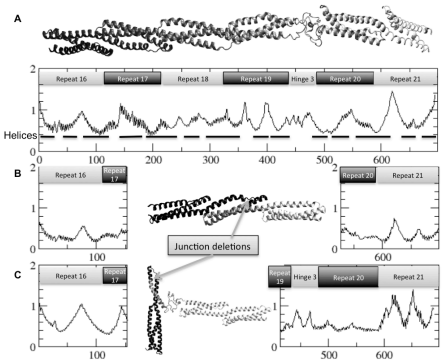


Figure 1. A) Wild-type dystrophin R16-21: initial three-dimensional structure model and flexibility (RMSF) of Cα measured during the 22 latest ns of simulation. Helix positions are indicated in the initial structure. B) R16-21 del45-51: initial three-dimensional structure model and flexibility (RMSF) of Cα measured during the 16 latest ns of simulation. C) R16-21 del45-49: initial three-dimensional structure model and flexibility (RMSF) of Cα measured during the 17 latest ns of simulation.

Acknowledgements

This project is supported by Rennes Metropole and Association Française contre les Myopathies and was performed using HPC resources from GENCI- [CCRT/CINES/IDRIS]. AN is supported by CNRS.

References

[1] E.P. Hoffman, R.H. Brown, Jr., L.M. Kunkel, Dystrophin: the protein product of the Duchenne muscular dystrophy locus. *Cell*, 51:919-928, 1987.

[2] B. Legrand, E. Giudice, A. Nicolas, O. Delalande, E. Le Rumeur, Computational study of the human dystrophin repeats: interaction properties and molecular dynamics. *PLoS One*, 6:e23819, 2011.

[3] A.P. Monaco, C.J. Bertelson, S. Liechti-Gallati, H. Moser, L.M. Kunkel, An explanation for the phenotypic differences between patients bearing partial deletions of the DMD locus. *Genomics*, 2:90-95, 1988.

A quantitative metagenomics analysis of the French cheese ecosystems

Anne-Laure ABRAHAM[1], Nicolas PONS[1], Sean KENNEDY[1], Antoine HERMET[2], Emmanuelle LE CHATELIER[1], Mathieu ALMEIDA[1], Jean-Michel BATTO[1], Benoît QUINQUIS[1], Nathalie GALLERON[1] and Pierre RENAULT[1]

[1] INSTITUT MICALIS, UMR1319 INRA, Domaine de Vilvert, 78352, Jouy-en-Josas, Cedex, France
anne-laure.abraham@jouy.inra.fr, pierre.renault@jouy.inra.fr
[2] LUBEM - EA3882, ESMISAB, Technopôle de Brest Iroise, 29280 Plouzané, France

Abstract *The diversity of the cheese ecosystem is not completely understood. Classical microbiological analysis are time consuming and can not give rapidly an accurate view of the flora. We propose a quantitative metagenomic approach to study the diversity of 40 french cheeses. We are developping a tool to rapidly determine the species present in a given cheese and we will provide an exhaustive catalog of the cheese ecosystem flora.*

Keywords Metagenomics, next generation sequencing, microbial genomes.

The manufacturing process of cheese, as for most fermented food, involves a complex flora, composed of bacteria but also yeast and filamentous fungi. The wide range of final products found on the dairy market is representative of the diversity of natural starters and ripening cultures used by dairy industries or coming from the food chain, from milk to the factory. However the cheese ecosystem is not completely understood. The natural starters are not constructed from pure strains and the knowledge of their exact composition remains incomplete, moreover number of species present in ripening cultures and in the food chain are little studied. Classical microbiological analysis or genetic methods (qPCR, MLST ...) can be used to identify and quantify to some extent these species. However these techniques are expensive and time consuming to provide a representative view of the flora. Therefore, in order to better understand cheese ecosystem, there is a need for developing new methods in order to achieve a more accurate view of the species playing a role in the cheese manufacturing.

To overcome these limitations, we have chosen to study the cheese ecosystem from a metagenomics point of view. This study has two main goals. The first one is to achieve a better understanding of the diversity of cheese flora by proposing an exhaustive and quantitative catalog of the present species. The second one is to provide a tool that gives a rapid and accurate vision of the species present in cheese in different production sites and over time in order to maintain a constant quality of the cheese product.

The metagenomics technique allows the sequencing of the species contained in a sample without isolating the species and culturing them. We have taken a sample of interior and/or surface of 40 traditional French cheeses and sequenced them with the SOLiD technology, by producing 15-20 millions of reads of 50 bases by run for each samples. Short reads were mapped to more than 3000 genomes of species available in microbial databases. Due to the similarity between genomes in the database, a high proportion of reads match present species, but also several closely relative genomes that are absent of the studied sample. We thus developed methodologies to detect and interpret false positives counts by taking into account several features such as the genome coverage, read distribution, annotations, the number of specific reads etc.

We will present some applications of this method on cheese ecosystems and the perspectives that these improvements will provide for metagenomic studies.

Acknowledgements

This work was supported by ANR FOOD-MICROBIOMES.

Dr Motifs: a web resource for pattern matching and discovery

Anthony Bretaudeau[1] and Olivier Collin[1]

[1] INRIA/Irisa, Campus de Beaulieu, 35042, RENNES, Cedex, France
{anthony.bretaudeau, olivier.collin}@irisa.fr

Abstract *Pattern discovery and pattern matching have become a very common task when analysing biological sequences. However, the diversity of algorithms and implementations to achieve these analyses make it difficult for users to choose the right tool. Dr Motifs is a web portal designed to ease analysis of motifs. It provides access to motif analysis tools via web interfaces and web-services. Tutorials, documentation and use-cases are also available to guide users in their experiments.*

Keywords Pattern matching, pattern discovery, web portal, web interface, web-service.

1 Context

Motifs can be found in biological sequences and are often associated to biological functions like transcription factor binding sites on DNA, or enzyme catalytic sites in proteins.

They can be represented in different forms: PROSITE-like patterns, profiles or HMM profiles. These formats differs in their expressivity and their use cases.

Many tools exist to describe new motifs from a set of related sequences (motif discovery), or to search motif occurrences in other sequences (motif matching). Each one is specific of a type of data (e.g. protein sequences or nucleic sequences) and uses different algorithms and different representations of the motifs. The results given by these tools can vary a lot depending on these specificities. For these reasons, it is often hard for users to choose the right tool for their use case.

2 Results

The Dr Motifs web portal is accessible at http://www.drmotifs.org. All the tools deployed on our servers are listed in a synthetic view on the Tools page. A link to web interfaces and web-services is displayed for each tool. Web interfaces were deployed on a Mobyle server [1], or developed using the Symfony PHP framework. In some cases, the web interfaces distributed with the tools source code were used directly with adaptations to our infrastructure. SOAP web-services were all deployed with the Opal toolkit [2]. These web-services can be used for example in workflow engines like Taverna [3]. Having both web interfaces and web-services allow to use the software with different quantities of data: from a few sequences to large scale analysis.

The most popular public tools are available for the common use-cases: MEME [4], Hmmer [5], InterProscan [6]. More specialized softwares are also available through this portal to test them and use them when the other tools show their limits. It is the case of Protomata [7], Stan [8] or Logol that are developed locally by the Dyliss and Genscale INRIA teams (Rennes, France). This gives access to other representations of motifs with better expressivity.

A blog is also available on this portal. It contains tutorials for using the tools in various use cases, general help on the study of motifs (what is a motif? how are they represented? when can they be used?), and news about the different tools (new versions, new tools).

Since its creation, the web portal is used by approximately 250 unique visitors per months, coming from various geographical regions. The portal has become the third official mirror for the access to MEME tools.

The portal will be updated in the future with new tools and tutorials to reflect the evolutions in the world of motifs.

Acknowledgements

This work was supported by an IBiSa grant 2009.

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*, 25:3005-3011, 2009.

[2] S. Krishnan, L. Clementi, J. Ren, P. Papadopoulos and W. Li, Design and Evaluation of Opal2: A Toolkit for Scientific Software as a Service, *The 2009 IEEE Congress on Services* (SERVICES-1 2009) 2009. July, 2009.

[3] D. Hull, K. Wolstencroft, R. Stevens, C. Goble, MR. Pocock, P. Li and T. Oinn, Taverna: a tool for building and running workflows of services. *Nucleic Acids Res.*, 34:W729-W732, 2006.

[4] T. L. Bailey, M. Bodén, F.A. Buske, M. Frith, C.E. Grant, L. Clementi, J. Ren, W. W. Li and W. S. Noble, MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research*, 37:W202-W208, 2009.

[5] R.D. Finn, J. Clements and S.R. Eddy, HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research, Web Server Issue* 39:W29-W37, 2011.

[6] E.M. Zdobnov and R. Apweiler, InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, 17(9): p. 847-8, 2001.

[7] F. Coste and G. Kerbellec, Learning Automata on Protein Sequences, *JOBIM*, 2006.

[8] J. Nicolas , P. Durand , G. Ranchy , S. Tempel and A.S. Valin, Suffix-tree analyser (STAN): looking for nucleotidic and peptidic patterns in chromosomes. *Bioinformatics*, 21(24):4408-4410, 2005.

Does a genome structure in divergent copies allow animal to evolve without sexual reproduction?

Martine Da Rocha[1], Lauriane Massardier[1], Laetitia Perfus-Barbeoch[1], Pierre Abad[1] and Etienne G.J. Danchin[1]

[1] Institut Sophia Agrobiotech, UMR INRA 1355 / Univ. Nice Sophia Antipolis / CNRS 7254, 400 rte des chappes , 06903, , Cedex , France
mdarocha@sophia.inra.fr

Keywords *comparative genomics, gene duplication, functional divergence, RNAseq.*

Un génome fait de multiples copies divergentes permet-il à un animal d'évoluer en l'absence de reproduction sexuée ?

Mots-clés génomique comparative, duplication génique, divergence fonctionnelle, RNAseq.

Les nématodes parasites de plantes sont des vers ronds microscopiques qui causent chaque année des pertes importantes au niveau agricole. Ils appartiennent à trois ordres différents, très éloignés au niveau phylogénétique : Tylenchida, Dorylaimida, Triplonchida. On trouve dans l'ordre Tylenchida les parasites de plantes les plus virulents, notamment ceux du genre Meloidogyne ou nématodes à galles. Dans ce genre, on retrouve trois modes de reproduction différents :

–reproduction exclusivement sexuée,

–reproduction sexuée facultative avec possibilité, dans certaines conditions, de reproduction asexuée avec méiose,

–reproduction exclusivement asexuée sans méiose.

Le tableau 1 met en parallèle le mode de reproduction de certaines espèces de Meloidogyne et leur spectre d'hôtes. De façon surprenante, on constate que chez les espèces où la reproduction est exclusivement asexuée sans méiose, le spectre d'hôtes est beaucoup plus important. Lorsqu'on compare les deux génomes actuellement séquencés de Meloidogyne, on constate que celui de *M. incognita* (reproduction asexuée sans méiose), est constitué d'une majorité de régions en deux copies présentant un taux de divergence moyen très élevé (plus de 7%)(1), alors que celui de *M. hapla* (reproduction sexuée)(2) ne présente pas cette caractéristique. Un de nos objectifs est de déterminer si la structure particulière du génome de *M. incognita* lui a permis d'aboutir à de la divergence fonctionnelle entre copies de gènes, lui assurant ainsi une plasticité en l'absence de reproduction sexuée. En effet, après une duplication, la redondance entre copies d'un même gène entraîne une relaxation de la pression de sélection et permet une accumulation de mutations. Cette accumulation de mutations peut, elle-même, conduire à une divergence fonctionnelle suivant un modèle de néo-fonctionnalisation ou sub-fonctionnalisation.

La première étape a consisté à inventorier les gènes présents en plusieurs copies dans le génome de *M. incognita* par rapport à celui de *M. hapla*. Pour comparer ces deux génomes nous sommes partis des protéines prédites et avons utilisé OrthoMCL(3). Nous avons obtenu 11 443 groupes d'orthologues dont 2 578 sont constitués d'un seul gène de *M. hapla* et d'au moins deux gènes de *M. incognita*. À partir de ces 2 578 groupes, nous avons voulu identifier les zones de synténies paralogues chez *M. incognita* contenant au minimum deux paires de gènes homologues en utilisant l'algorithme OrthoClusterDB(4). Nous avons obtenu au total 584 groupes de zones de synténie, dont 64 correspondent à des zones de synténies constituées uniquement de duplication en cis et 520 correspondent à des zones de synténies contenant des duplications

en bloc. Nous nous sommes concentrés sur les 520 zones de synténies en bloc, car elles reflètent la structure en deux copies du génome de *M. incognita*.

Pour identifier l'existence de divergence fonctionnelle entre les différentes copies de gènes nous avons utilisé des données de RNA-seq. Nous avons, au laboratoire, plusieurs banques de lectures Illumina correspondant aux différents stades de développement de *M. incognita*. L'alignement des lectures de ces banques sur le transcriptome, nous a permis d'obtenir le patron d'expression des gènes. Nous avons recherché les différences de patron d'expression entre copies d'un géne. Nous avons ainsi identifié 8 couples de copies de gènes présentant des différences d'expression significatives. Ces résultats constituent le premier signe indirect de divergence fonctionnelle entre ces copies. Ces divergences des profils d'expression entre copies seront confirmées par qPCR.

Parallèlement, nous allons mesurer le degré de divergence au niveau nucléotidique entre copies et, en particulier, nous nous intéresserons aux mutations non-synonymes, susceptibles d'impliquer une divergence fonctionnelle. Pour ce faire nous alignerons les paires de copies de gène de *M. incognita* avec leurs orthologues en copies uniques chez *M. hapla* et dans l'environnement datamonkey (5) nous serons en mesure d'évaluer les taux de mutations synonymes et non-synonymes.

Cette étude, nous permettra d'identifier des gènes présents en plusieurs copies chez *M. incognita* possédant des différences de patrons d'expression mais dont les gènes orthologues correspondant chez *M. hapla* sont en une seule copie. Une étude des fonctions de ces génes sera effectuée en vue d'étayer l' hypothése selon laquelle la présence de régions dupliquées dans le génome de *M. incognita* lui apporterait une plus grande diversité génétique et permettrait d'expliquer, en partie, le plus large spectre d'hôte et la répartition géographique plus importante de cette espèce à reproduction strictement parthénogénétique.

	M. incognita	*M. arenaria*	*M. javanica*	*M. hapla*	*M. subarctica*
Mode de reproduction	Parthéno-génétique sans méiose	Parthéno-génétique sans méiose	Parthéno-génétique sans méiose	Amphimictique ou Parthéno-génétique avec méiose	Amphimictique
Spectre hôtes *	+++	+++	+++	++	+

Tableau 1. Comparaison entre le mode de reproduction et le spectre d'hôtes de cinq espèces du genre Meloidogyne.
(*) : + une seule famille de plantes ; ++ plusieurs familles de plantes ; +++ majorité des familles de plantes.

References

[1]Abad P, Gouzy J, Aury J, Castagnone-Sereno P, Danchin EGJ, Deleury E & all, Genome sequence of the metazoan plant-parasitic nematode Meloidogyne incognita. *Nat Biotechnol.*, 26:909-915, 2008.

[2]Opperman CH, Bird DM, Williamson VM, Rokhsar DS, Burke M, Cohn J & all, Sequence and genetic map of Meloidogyne hapla: A compact nematode genome for plant parasitism. *Proc Natl Acad Sci USA*, 105:14802-14807, 2008.

[3]Li L, Stoeckert CJJ, Roos DS, OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.*, 13:2178-2189, 2003.

[4]Vergara IA, Chen N, Using OrthoCluster for the detection of synteny blocks among multiple genomes. *Curr Protoc Bioinformatics*, Chapter 6:Unit 6.10 6.10.1-18,2009.

[5] Delport W, Poon AF, Frost SD, Kosakovsky Pond SL, Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology, Bioinformatics, 26(19):2455-7, 2010.

Comparative Analysis of Next Generation Sequencing

Approaches for Exploring Microbiomes

Arnaud FELTEN, Charlotte JOUBERT, Pierre DEHOUX, and Catherine DAUGA

Institut Pasteur, Bio-informatique pour l'Analyse Génomique, 28 rue du Docteur Roux, 75015 Paris, France
{arnaud.felten, charlotte.joubert, pierre.dehoux, catherine.dauga}@pasteur.fr

Keywords Microbiome, MetaSim, GemSIM, *in silico* analysis.

1 Introduction

The advent of next generation sequencing has coincided with a growth of interest in using these approaches to better understand the role, the structure and function of the microbial communities in human, animal, and environmental health. Distinct microbial communities inhabit the body's surfaces: the skin, the oral cavity, the upper respiratory tract, the genital tract and the stomach of humans. Many past and present metagenomic studies generally do not address the effect of classification approaches on the accuracy of the results, even though it is of great significance for microbiome research. For instance, the bacterial component of the human gut microbiome consists of ≥ 1,800 genera and astonishing 15,000-36,000 species, depending on whether species are classified conservatively or liberally [1]. It is very interesting to search associations between microbiome structure and clinical phenotype in human diseases by showing significant levels of variation of some microbial components. However, this requires more precise processes and strategies of bioanalysis.

2 Simulations and analysis of metagenomic data

In this study, we have compared two technologies, pyrosequencing (454 Titanium) and Illumina, and have evaluated the classifications obtained by using 16S rRNA sequences based methods. First, we defined mock communities, which were composed of unique species and low complexity or high complexity populations of bacteria that reflect the diverse taxonomical composition of typical metagenome data sets. We generated collections of reads corresponding to complete 16S rRNA and its V4 and V5-V6 extended regions with the sequencing simulator called MetaSim [2]. In addition, we used GemSIM [3] with our own error models obtained from real data (mouse microbiota) [4] to generate reads as close as possible to one experiment. Blast search of the synthetic reads for each mock community was conducted against 6 reference databases currently used for metagenome studies.

3 Sequencing technologies and databases influence classification efficiencies

Classification efficiencies among sequencing technologies are significantly different and ultimately affect resolution. When comparing classification efficiencies (CE), defined as the proportion of reads confidently classified to a certain taxonomic level (genus-level in this case), we observed more acceptable values for the Titanium reads, as more unidentified and misidentified reads were generated by Illumina's technology. The CE can vary widely depending the database used. The Silva database (SSU Ref NR 104, February 2011) is the more reliable among the 6 databases tested. We observed the best taxonomic resolution (species level) by using V4 and the Greengenes database (May 2011). CE also depend on the taxonomic composition of the communities and Firmicutes appear to be the most unidentified taxons (Fig 1 shows an example on 2 databases). Futhermore, a majority species in the community can lead to over-identification of closely related species.

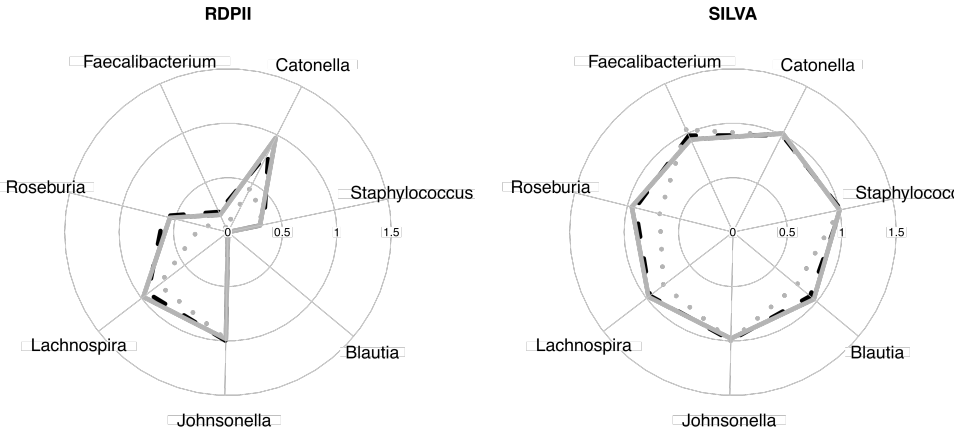


Figure 1. 454 Technology: proportions of identified reads from 7 species (Firmicutes) according to two databases. 16S rRNA : complete sequence (dashed line), V4 (grey circles), V5-V6 (plain line)

It is well established that the use of next-generation sequencing to perform 16S rRNA sequence surveys has resulted in important controversy surrounding the effect of sequencing errors on downstream analysis [5][6]. We now found that databases can have a significant impact on the result of metagenomic analyses.

One of the results of this evaluation suggest to adapt the sequencing method and the strategy of bioanalysis according to *a priori* knowledge of the taxonomic composition of the microbiota.

References

[1] M.J. Pallen, *Metagenomics of the Human Body*, Springer Science+Business Media LLC, pp 43-61, 2010.

[2] D.C. Richter, F. Ott, A.F. Auch, R. Schmid, and D.H Huson, A Sequencing Simulator for Genomics and Metagenomics. *PLoS ONE*, 3(10): e3373. doi:10.1371/journal.pone.0003373, 2008.

[3] K.E. McElroy, F. Luciani and T. Thomas, GemSIM: general, error-model based simulator of next-generation sequencing data. *BMC Genomics*, 13(74). doi:10.1186/1471-2164-13-74, 2012.

[4] T. Pedron, C. Mulet, C. Dauga, L. Frangeul, C. Chervaux, G. Grompone, and P. Sansonetti, A Crypt-Specific Core Microbiota resides in the mouse colon. *mBio*, In press, 2012.

[5] M.J. Claesson, O. O'Sullivan, Q. Wang, J. Nikkilä, J.R. Marchesi, H. Smidt, W.M de Vos, R.P. Ross and P.W. O'Tolle, Comparative Analysis of Pyrosequencing and a Phylogenetic Microarray for Exploring Microbial Community Structures in the Human Distal Intestine. *PLoS ONE*, 4(8): e6669. doi:10.1371/journal.pone.0006669, 2009.

[6] P.D. Schloss, D. Gevers, S.L. Westcott, Reducing the Effects of PCR Amplification and Sequencing Artifacts on 16S rRNA-Based Studies. *PLoS ONE*, 6(12): e27310. doi:10.1371/journal.pone.0027310, 2011.

Transcriptome-wide identification of CELF1 binding sites using PSSMs and *in silico* scoring selection

Stéphanie MOTTIER[1], Bernhard GSCHLOESSL[2], Olivier LE TONQUEZE[3], Luc PAILLARD[1] and Yann AUDIC[1]

[1] Institut Génétique et Développement, UMR6290 CNRS, Faculté de Médecine, 2 avenue du Prof Léon Bernard 35043, RENNES, Cedex, France
stephanie.mottier@univ-rennes1.fr, luc.paillard@univ-rennes1.fr, yann.audic@univ-rennes1.fr

[2] CBGP, Campus International de Baillarguet, CS30016, 34988, MONTERRIER-SUR-LEZ, Cedex, France
bernhard.gschloess@supagro.inra.fr

[3] Cancer Center, Massachusetts General Hospital, Harvard Medical School, MA 02114, BOSTON

Abstract *RNA-binding proteins (RBP) are involved in many regulatory processes. Experimental identification of their specific RNA targets is time consuming. Therefore it is of interest to take advantage of experimentally validated binding sites to build computational models that can contribute to predict potential target RNAs in other transcriptomes or cellular context. In this work, we used the experimentally RNA-binding site of a conserved regulatory RBP, called CELF1. A position-specific scoring matrix (PSSM) model which respects the specificities of the CELF1 binding site motif is then used to validated the prediction ability of the proposed in silico approach.*

Keywords Sequence analysis, Statistics, PSSM, Motif scoring, RNABP, SELEX, CLIPseq

1 Introduction

RBP bind specific RNAs to regulate post-transcriptional processes such as alternative splicing, mRNA stability and mRNA translation thereby controlling gene expression both qualitatively and quantitatively. The identification of the specific targets of RBP is crucial in understanding the extent and importance of post-transcriptional regulation in the dynamic of gene expression. To date, *in vitro* and *in vivo* experimentation like ”systematic evolution of ligand by exponential enrichment procedures” (SELEX) [1] and ”crosslinking and immunoprecipitation experiments and sequencing” (CLIPseq) [2] are the state of the art procedures for the identification of RBP binding sites on RNA molecules. However, these approaches are costly and time-consuming. The experimentally identified binding sites of a specific protein can be modeled as position-specific scoring matrices (PSSMs) that allow to score binding motifs in RNA sequences. Then, *in vivo* or *in vitro* derived motif prediction can greatly reduce the ”hands-on” experimental time if these information are used to other transcriptomes, other cell-types or other animal models.

2 Approach

The CELF1 protein [3,4] is a human RNA-binding protein highly conserved among vertebrates which is known to bind to single stranded UG-rich RNA motifs [5,6]. We use the MEME program [7] to identify enriched motifs in CELF1 interacting RNA regions on SELEX and CLIPseq data. Then we used Perl scripts to define PSSMs and calculated the motif-scores along every RNA strand. Motifs are then compared to each other for their ability to classify RNA sequences as CELF1 target or non CELF1 target.

3 Materials and methods

SELEX is an alternative approach in which synthetic RNA oligonucleotides are selected for their ability to bind to a recombinant CELF1 protein. After several rounds of selection, CELF1 interacting sequences [5] were selected and sequenced yielding 64 UG-rich sequences with an average length of 42 nucleotides. CLIPseq is a technique in which specific RNA/protein complexes are immunopurified from living cells in order to recover the fragment of RNA specifically associated with the RNABP *in vivo*. The co-purified

RNAs are then sequenced by high throughput next generation sequencing (NGS). About 20 millions 30nt-reads (SOLID) mapped uniquely onto the human genome. Regions of overlapping mapped RNA reads form 1,932 CELF1 interacting clusters (CICs) with an average length of 190 nt corresponding to 1,154 genes. Binding site motifs were identified in both datasets by applying the MEME program which is dedicated to pattern discovery. The first motif is defined from sequences identified *in vitro* by SELEX experiments [5] and the three other motifs are part of the *in vivo* binding sites identified by CLIPseq. We used the alignment sequences from MEME as input set to in-house Perl scripts. First to generate matrices for PSSM scoring and second to calculate scores. The scoring algorithm applies a sliding window along the RNA sequences and calculates a motif similarity score that takes into account the nucleotide frequency and the information content of each alignment position [6]. For scoring analysis, the 1,932 CICs represent the positive dataset and three negative control datasets were generated from the 1,932 CICs by shuffling individualy each sequence to loose motif constraint without altering nucleotidic composition.

4 Results

The SELEX sequences and three subsets of the CICs were subjected to motif identification using MEME. The CICs datasets were a) all the 1,932 CICs, b) only the first quartile (p-value-wise) of the CICs and c) only the top-20 scoring clusters. The top MEME motif identified from each dataset were all highly correlated and were UG-rich motif of different length. It is important to note that the motif identified on *in vitro* data and *in vivo* data are highly correlated suggesting that the intrinsic binding characteristic of CELF1 have been identified. This is further supported by the structural information available on the three RNA recognition motifs of the CELF1 protein.

We then compared the ability of these different motifs to specifically identify CELF1 interacting clusters versus the negative control datasets. At their best binary classification performance measures the different motif are compared. However, for the best performing motif from 1,932 CICs, about half of the CIC are retrieved. This indicates that it is likely that additional constraints modify the *in vivo* binding characteristic of CELF1, this may occur through protein-protein interaction between CELF1 and other RNA binding proteins that can allow CELF1 to have a slightly different repertoire for CELF1 *in vivo* depending on its interacting partners. Nevertheless, the combination of CLIPseq sequencing and identification of new binding motif without *a priori* knowledge will allow us to provide first an experimentally validated set of CELF1 interacting motifs, second a putative interaction map for CELF1 on all human transcript but also on transcript from closely related species considering the very high conservation of the CELF1 protein in particular at the level of its RNA recognition motifs.

Acknowledgements

This work was supported by ANR-07-JCJC-0097-01.

References

[1] Klug, S.J., Famulok, M.: All you wanted to know about selex. Molecular Biology Reports 20(2), 97–107 (1994)
[2] Ule J, Jensen KB, Ruggiu M, Mele A, Ule A, Darnell RB. CLIP identifies Nova-regulated RNA networks in the brain. Science 302(5648),1212-1215 (2003)
[3] Paillard, L. et al. EDEN and EDEN-BP, a cis element and an associated factor that mediate sequence-specific mRNA deadenylation in Xenopus embryos. EMBO J 17, 278-287 (1998)
[4] Timchenko, L.T. et al. Identification of a (CUG)n triplet repeat RNA-binding protein and its expression in myotonic dystrophy. Nucleic Acids Res 24, 4407-4414 (1996).
[5] Marquis, J., Paillard, L., Audic, Y., Cosson, B., Danos, O., Bec, C.L., Osborne, H.B.: Cugbp1/celf1 requires ugu-rich sequences for high-affinity binding. Biochemical Journal 400(2), 291–301 (2006)
[6] Tonqueze, O.L., Gschloessl, B., Namanda-Vanderbeken, A., Legagneux, V., Paillard, L., Audic, Y. Chromosome wide analysis of cugbp1 binding sites identifies the tetraspanin CD9 mRNA as a target for cugbp1-mediated down-regulation. Biochemical and Biophysical Research Communications 394(4), 884–889 (2010)
[7] Bailey, Williams, Misleh, Li MEME: discovering and analyzing DNA and protein sequence motifs, Nucleic Acids Research 34: W369-W373 (2006)

Automatized detection of duplicated copies from NGS data: application to 45S rDNA and coding genes in the hexaploid *Spartina maritima* (Poaceae)

Julien Boutte[1] , Benoît Aliaga[1] , Julie Ferreira[1], Oscar Lima[1], Sophie Coudouel[1], Delphine Naquin[1], Patrick Wincker[2], Julie Poulain[2], Corinne Da Silva[2], Malika Ainouche[1] and Armel Salmon[1]

[1] Ecobio UMR 6553 CNRS, University of Rennes 1. Bât 14A Campus Scientifique de Beaulieu, 35 042 Rennes Cedex, France

boutte.julien@gmail.com; armel.salmon@univ-rennes1.fr

[2] Genoscope, 2 rue Gaston Crémieux, 91 000 Evry, France

Spartina species (Poaceae, Chloridoideae) are perennial grasses exhibiting high ploidy levels (from 4x to 12x), making genomic analyses and gene orthology search particularly challenging.

We developed a program detecting duplicated variants from sequence alignments, an important step to understand gene and species history, using single nucleotide polymorphisms (SNPs). This program includes a *de novo* assembly process (using Genome Assembler 2.X), SNP detection (removing pyrosequencing-biased, false-positive SNPs), and a haplotype construction (with appropriate parameters such as considered SNP numbers for haplotype reconstruction, SNP weighting, read-depth, etc.).

We first applied this new bioinformatic tool to a contig obtained from a *de novo* assembly of Roche-454 reads coding the 45S ribosomal unit from *Spartina maritima*, a European-native hexaploid species. This gene family was chosen for its particular dynamic evolution in polyploids and its frequent use in phylogeny to reconstruct species history. Annotation of the different 45S rDNA domains was performed by aligning the *S. maritima* homologous consigns against phylogenetically related model species (rice, sorghum and maize). Sequence homogeneity was encountered in coding regions and internal transcribed spacers (ITS) while high intra-genomic variability was detected in the regions containing the intergenic spacer (IGS) and external transcribed spacers (ETS), agreeing with inter-individual or inter-specific variation levels reported in the literature. To validate the considered SNPs in the haplotype reconstruction, and the detected haplotypes, we resequenced the SNP-containing regions (Sanger sequencing) following appropriate primer design. As the sequence length of the detected haplotypes did not span the whole 45S rDNA contig (~9 kb), this resequencing strategy allowed the assembly of longer haplotypes, and confirmed the detection of two main 45S rDNA copies in *Spartina maritima*. This tool was also applied to assemble reads from coding regions using RNA-seq datasets from the hexaploid *S. maritima* and *S. alterniflora*, their homoploid hybrid derivative (*S. x townsendii*) and the allododecaploid (*S. anglica*) species.

These results are particularly interesting in the perspective of detecting duplicated copies at different ploidy levels in Spartina following the successive speciation events within this clade as well as in any other polyploid species.

Key-words: Duplication, Homoeology, SNPs, Sequence *de novo* Assembly, 45S rDNA.

Graph-based scaffolding for next-generation sequencing

Delphine NAQUIN[1] and Rayan CHIKHI[2]

[1] plateforme IMAGIF, FRC3115 CNRS, avenue de la Terrasse, 91198 Gif sur Yvette Cedex, France
delphine.naquin@cgm.cnrs-gif.fr
[2] ENS Cachan/IRISA, 35042 Rennes, France.
rayan.chikhi@ens-cachan.org

Abstract de novo *assembly of high-throughput sequencing reads is a fundamental task in current genome sequencing projects. The scaffolding step consists in ordering the assembled sequences using pairing information. This step has a great impact on the quality of assembly. However, current methods implement heuristics that often yield sub-optimal results. Better dedicated scaffolders are needed to fully take advantage of the pairing information. We introduce a novel scaffolding implementation, which improves the contiguity of current methods.*

Keywords Scaffolding, next-generation sequencing, de novo assembly, graph traversal

1 Context

With next-generation sequencing technologies, *de novo* assembly of a genome into one single sequence per chromosome is a difficult task. This is mainly because the sequenced reads are shorter than genomic repetitions. However, next-generation sequencers enable the production of pairs of reads separated by a distance ranging from 200 bp to 40 kbp (so-called paired-end or mate-pairs). This pairing information can, in principle, be used to remove the ambiguities due to repetitions. However, for algorithmic simplicity, pairing information is typically not used in initial assembly steps. Assemblers produce a set of contiguous genome sub-sequences (contigs), assembled without considering pairing information. Then, a dedicated assembly step (scaffolding) is used to improve the assembly by incorporating pairing information. The scaffolding problem consists in finding an optimal ordering of the contigs, supported by a maximal number of paired reads [3]. This step is essential in *de novo* assembly, as it enables to improve by typically an order of magnitude the N50 metric of assembled sequences.

Most genome assemblers implement a scaffolding step. There are also several recent, dedicated scaffolding implementations, e.g. SSPACE [2] and Opera [3]. These scaffolders have been shown to generally outperform integrated ones. SSPACE uses a greedy graph traversal strategy without explicitly constructing the contig graph, and aggressively discards edges with low pairs support. Opera builds a contig graph and focuses on finding an optimal graph traversal using only non-repeated contigs.

We propose a novel scaffolding approach, implemented in the Superscaffolder software. Our approach consists in explicitly constructing the classical contig graph and performing novel graph traversals. In essence, we model a class of graph structures that can be traversed unambiguously to create correct scaffolds. This class strictly extends structures (the so-called tips and bubbles) that are classically considered by genome assemblers and scaffolders. Our method detects and traverses structures that are a mixture of tips and bubbles; these are deemed too complex by other methods. Our implementation also features two practical improvements. First, it uses the software GASSST [4] to perform the most computationally intensive step (reads mapping) in parallel. Second, a novel step accurately detects the intrinsic parameters of the libraries (read pairs distance and deviation).

Superscaffolder was also used in a recent, international *de novo* assembly competition (Assemblathon 2, unpublished, http://www.assemblathon.org/).

2 Results

We have tested our method on two datasets: a bacterial genome (*E. coli*) and an insect genome (*M. persicae*). The bacterial genome was sequenced with the Illumina technology with 10 million paired reads of length 36 bp and insert size 200 bp (SRA: SRX000429). The insect genome was also sequenced with the Illumina technology, with three libraries: 162M paired-reads (insert size 450 bp), and respectively 72M and 81M mate-paired reads (insert size 2 kbp and 5 kbp). For both datasets, an initial assembly of contigs (resp. scaffolds for the *M. persicae* dataset) was built with the Monument assembler [5]. Our tool is compared with two other scaffolders: SSPACE (version 1.2) and Opera (version 1.0). For SSPACE and Opera, the pairing type of the reads (paired read or mate pair) and insert size must be specified. Superscaffolder is able to estimate the insert size from the data. The results of scaffolding the two datasets with all three scaffolders are shown in Table 1.

For the *E. coli* dataset, Superscaffolder produced the most contiguous and accurate scaffolding. However, SSPACE produced slightly less scaffolds, resulting in better average length. The scaffolding produced by Opera was inferior by all metrics. A better set of parameters may improve Opera results. However, this would require significant tuning compared to other tools, for which default parameters produce satisfying results. Note that all methods slightly deteriorate the original assembly accuracy. However, Superscaffolder produced the least amount of inaccurate chunks. For the *M. persicae* dataset, Superscaffolder outperforms SSPACE in all metrics by a wider margin. Given its results on the previous dataset, Opera was not evaluated on this dataset. Overall, the length of all assemblies is moderately increased by scaffolding.

Highly-contiguous assemblies are of fundamental importance in sequencing analysis. The Superscaffolder tool enables to significantly increase the contiguity of assemblies by using an improved scaffolding algorithm.

	Total bases (Mbp)	Nb contigs / scaffolds	Mean (bp)	N50 (bp)	Accuracy (%)
E.coli					
Original	4.51	754	5,994	12,078	100.0
Opera	4.52	414	10,951	27,473	95.6
SSPACE	4.53	**209**	**21,685**	54,361	98.6
Superscaffolder	4.51	216	20,906	**63,466**	**98.8**
M. persicae					
Original	386.3	14,711	26,263	348,563	-
SSPACE	390.3	11,133	35,061	497,117	-
Superscaffolder	388.2	**9,032**	**42,938**	**877,705**	-

Table 1. Results of scaffolding assemblies of *E. coli* and *M. persicae* using Opera, SSPACE and Superscaffolder (our contribution). The initial assembles are denoted as Original in the table. For *E. coli*, accuracy is measured from chunks of 10 kbp which align with more than 99% identity on the reference genome. For *M. persicae*, no close reference genome is available.

Acknowledgements

This work was supported by the ANR grant MAPPI. We wish to thank F. Legeai and T. Derrien from the Genouest platform for providing us with a quality-controled *M. persicae* dataset.

References

[1] S. Koren, T. J. Treangen, et M. Pop, Bambus 2: scaffolding metagenomes, Bioinformatics, vol. 27, n. 21, p. 2964–2971, 2011.

[2] M. Boetzer, C. V. Henkel, H. J. Jansen, D. Butler, et W. Pirovano, Scaffolding pre-assembled contigs using SSPACE, Bioinformatics, vol. 27, n. 4, p. 578, 2011.

[3] S. Gao, N. Nagarajan, et W. K. Sung, Opera: reconstructing optimal genomic scaffolds with high-throughput paired-end sequences, Research in Computational Molecular Biology, 2011, p. 437–451.

[4] G. Rizk et D. Lavenier, GASSST: global alignment short sequence search tool, Bioinformatics, vol. 26, n. 20, p. 2534, 2010.

[5] R. Chikhi et D. Lavenier, Localized genome assembly from reads to scaffolds: practical traversal of the paired string graph, Algorithms in Bioinformatics, p. 39–48, 2011.

Genomicus: Five Genome Browsers for Synteny and Ancestral Gene Content Information in Eukaryota

Alexandra LOUIS[1], Marlène GRATIGNY[1], Matthieu MUFFATO[2] and Hugues ROEST CROLLIUS[1]
[1] Institut de Biologie de l'Ecole Normale Supérieure (IBENS), Centre National de la Recherche Scientifique UMR8197, Paris, France
{alexandra.louis, marlene.gratigny, hrc}@biologie.ens.fr
[2] European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton Cambridge CB10 1SD, UK
muffato@ebi.ac.uk

Abstract *Visualisation interfaces are critical tools to explore complex multi-dimensional genomic data. Comparative genomic data in particular integrates spatial information related to gene organisation and temporal relations related to gene and genome evolution. For example, comparison between multiple orthologous and paralogous genomic loci in parallel in a single view is an efficient strategy to rapidly understand the evolutionary history of a locus from a common ancestor. Bioinformatics tools are available to visualise and compare genomes, but except for Genomicus, most are restricted to two or three genomes at a time. Here we present a major extension to the Genomicus database and web interface. Four new phylum including Plants, Fungi, non-vertebrate Metazoa and Protists are now represented in dedicated browsers, bringing to 148 the number of eukaryote genomes available for interactive comparisons.*

Keywords Genome browser, Synteny, Ancestral Gene Content, Evolution, Genomics.

1 Introduction

The Genomicus software is a dynamic visualisation interface that makes it possible to compare unlimited numbers of genomes [1]. It provides a comprehensive overview of genome organisation between modern species. In the case of vertebrates, Genomicus also provides extensive information on reconstructed ancestral genomic organisation at 48 nodes. Users can intuitively query and navigate within and between genomes in a phylogenetic context to examine the evolution and the conservation of a specific locus in a species of interest. Navigation in Genomicus can be performed in three dimensions: linearly along the axes and the position of chromosomal genes, transversally between different species and chronologically along an evolutionary axis.

2 Implementation and Access

Genomicus is composed of Perl scripts and modules, executed with mod_perl on an Apache2 server and querying a MySQL database. The pages embed inline-SVG drawings in XHTML while the JavaScript usage is limited to an information panel retrieved with AJAX calls. Genomicus sources and Mysql schema are available on request via a SVN server, and developers may contribute code to the main release.

Users may freely access Genomicus online at http://www.dyogen.ens.fr/genomicus. This main site provides an interface to genomic data derived from Ensembl [2], with additional information on reconstructed ancestral gene order [3]

Four additional sites are now available:

GenomicusPlants	http://www.dyogen.ens.fr/genomicus-plants
GenomicusProtists	http://www.dyogen.ens.fr/genomicus-protists
GenomicusMetazoa	http://www.dyogen.ens.fr/genomicus-metazoa
GenomicusFungi	http://www.dyogen.ens.fr/genomicus-fungi

Each derived from a different EnsemblGenome database. The same code is used for all versions, thus facilitating updates and maintenance.

3 Modules

The home page of each Genomicus server invites the user to enter its gene of interest, which will be defined as “reference gene” and belongs to a “reference species”.

Genomicus provides two main types of syntenic views:

1-The Phyloview module shows the gene order of the reference gene and its neighbouring genes, and the order of their respective orthologs and paralogs in different species that share the same ancestral "root" species with the reference species. It is the default view, after entering a gene name.

2-The AlignView module shows an alignment between the genes contained within the genomic region of the reference gene and all their respective orthologs in other species.

On both modules, a menu at the top of the page allows the user to switch between the two syntenic views, to focus on paralogs if the user is interested on a specific species, to hide “low coverage” species informations. The ancestral gene content can also be removed or shown. Each node of the gene tree or specie tree can be collapsed or hide to focus on specific species.

All the images can be exported in SVG format for subsequent editing and figure preparation.

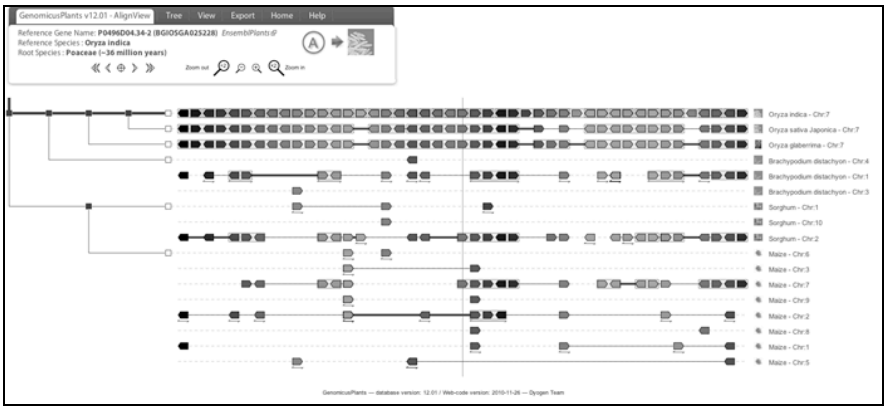


Figure 1. AlignView example of a gene in the rice genome (*Oryza indica*). The P0496D04.34-2 gene and its orthologs are centred on the thin vertical line. The genomic neighbourhood of P0496D04.34-2 is displayed on the top lane. When other species possess orthologous copies from this neighbourhood, additional lanes are added underneath, one per chromosome. For example the genomic neighbourhood of P0496D04.34-2 is contained on a single chromosome in *Oryza sativa*, (another rice species) but in 8 different chromosomes in maize.

4 Conclusion

Genomicus is a widely used web server, until now mainly addressing the needs of the community interested in vertebrate genomes. The aim of extending the server to a wider phylogenetic range answers a number of requests from collaborators and users, and will hopefully facilitate comparative genomics projects across a much wider phylogenetic range.

References

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

[2] Flicek, P, MR Amode, D Barrell, et al. 2012. Ensembl 2012. Nucleic Acids Res 40:D84-90.

[3] http://tel.archives-ouvertes.fr/docs/00/55/21/38/PDF/these.pdf

Identifying long range cis-regulation in the human genome using evolutionary co-segregation

Magali Naville, Alexandra Louis and Hugues Roest Crollius
IBENS, UMR8197 CNRS, 46 rue d'Ulm, 75005, Paris, France
{magali.naville, alexandra.louis, hrc}@ens.fr

Keywords Vertebrate genomics, transcriptional regulation, enhancers, genomes evolution

Identification de régulations longue distance dans le génome humain par co-ségrégation évolutive

Mots-clés Génomique des vertébrés, régulation transcriptionnelle, enhancers, évolution des génomes

The combined identification of regulatory sequences and their target genes is a major challenge in genomics. Due to the complex 3D structure of the vertebrate chromosome, enhancers often target genes at distances several hundred thousands of base pairs away. Simply hypothesizing that the nearest gene of a putative enhancer might be the regulatory target is thus often incorrect. Functional data such as histone modifications, ChIP-seq or DNAse hypersensitive sites now make it easier to find active regulatory regions, even if these methods are limited to specific tissues/cell types and specific differentiation stages. However, the identification of target genes still requires heavy experimental procedures, such as 3C/4C/5C or the generation of coordinated active chromatin signatures. Computational prediction of such functional association is thus of great interest. It represents a preliminary step for the better understanding of tight spatiotemporal controls of gene expression or of the coordination of complex developmental programs, and will greatly accelerate the identification of pathological mutations in non-coding loci.

Here we propose a bioinformatic method that allows the identification of evolutionary co-segregation between putative enhancers and specific genes, using the information on a large range of genomes. The method is currently developed on the human X chromosome.

Putative enhancers are, in a first approximation, assimilated to Conserved Non-coding Elements (CNEs). CNEs are defined based on their conservation in a range of vertebrate species, using an algorithm that scans the UCSC 46 species multiZ alignment and looks for regions with signatures of sequence conservation. This algorithm requires a minimal number of species in the multiple alignment and allows substitutions to occur, under a given threshold, in each column of the alignment. This approach led to the delimitation of 168.899 CNEs on the X chromosome, covering 5.840.833 bp (3.8% of its total length). One fifth (22%) of their positions overlap elements identified by the more conservative Siphy algorithm [1] (with a coverage corresponding to 61% of these elements), highlighting the ability of our method to detect less conserved regions. Both types of predictions were finally fused in a unified set containing 174.473 regions that cover 4.4% of the X chromosome.

Because sequence conservation may also identify other kinds of functional elements (such as promoters, insulators or matrix attachment regions), we further use different annotations to characterize CNEs: ENCODE data (on Transcription Factor Binding Sites, DNAse1 hypersensitive sites and histone modifications), and data from the literature (on co-activator p300 binding sites [2,3] or experimentally validated enhancers [4,5] for instance).

After the identification of putative enhancers, the first step of our target gene prediction method consists in

collecting genes immediately neighbouring (within 1.5Mb) each CNE. A scoring procedure is then applied on these genes to measure the frequency of co-retention of the CNE-gene pair during evolution. The score measures the presence/absence of the orthologous genes in the vicinity of the orthologous CNE in a large range of other genomes: it increases if an orthologous gene is found close to the CNE in any of the genomes that possess this CNE, and on the contrary decreases if such a gene is absent from the genome or located far from the CNE. The CNE/gene distance threshold depends on the genome considered. It is defined as 1Mb scaled to the size of the genome in comparison to human. The score also takes into account the rearrangement rate of the different genomes in comparison to human, as well as variations in their coverage rates. More than 95% of the targets have a score comprised between -20 and +20, and genes showing the highest values of this co-segregation score are considered as the most likely targets of the CNEs.

Strikingly, CNEs with high co-segregation scores are enriched in functional annotations. While 26% of all CNEs show H3K4me1 annotations, this proportion increases to 39% in CNEs comprised in the 30% top co-segregation scores. In the same way, the proportion of CNEs showing DNAse1 hypersensitivity increases from 13% in the whole set to 23% in CNEs with scores in the top 30%. This strongly supports the initial premises that elements segregating with genes are more likely to be true enhancers. The comparison of our predictions with a set of independently predicted PAX6 enhancer/target gene pairs on the X chromosome [6] is encouraging, with a coincidence rate of 11/15.

In collaboration with several human genetics groups, we are working on applying our procedure to genes involved in genetic disorders that could be explained by mutations in such non-coding regions. In addition, functional annotations, along with high co-segregation scores with a gene of interest, are used to prioritize the most interesting cases for independent validations using transgenic experiments in zebrafish, with a GFP reporter assay. More globally, this study will also allow the reconstruction of the evolution of chromosomal regulatory circuits, and a better understanding of the role of long range cis-regulation in negative selection of genomic rearrangements.

References

[1] K. Lindblad-Toh, M. Garber, O. Zuk, M.F. Lin, B.J. Parker, S. Washietl, P. Kheradpour, J. Ernst, G. Jordan, E. Mauceli & al. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, 478(7370):476-482, 2011.

[2] M.J. Blow, D.J McCulley, Z. Li, T. Zhang, J.A. Akiyama, A. Holt, I. Plajzer-Frick, M. Shoukry, C. Wright, F. Chen, V. Afzal, J. Bristow, B. Ren, B.L. Black, E.M. Rubin, A. Visel and L.A. Pennacchio. ChIP-Seq identification of weakly conserved heart enhancers. *Nat Genet*, 42(9):806-810, 2010.

[3] A. Visel, M.J. Blow, Z. Li, T. Zhang, J.A. Akiyama, A. Holt, I. Plajzer-Frick, M. Shoukry, C. Wright, F. Chen, V. Afzal, B. Ren, E.M. Rubin and L.A. Pennacchio. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, 457(7231):854-858, 2009.

[4] D. Lee, R. Karchin and M.A. Beer. Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Res.*, 21(12):2167-2180, 2011.

[5] A. Visel, S. Minovitsky, I. Dubchak and L.A. Pennacchio. VISTA Enhancer Browser - a database of tissue-specific human enhancers. *Nucleic Acids Res.*, 35:D88-92, 2007.

[6] P. Coutinho, S. Pavlou, S. Bhatia, K.J. Chalmers, D.A. Kleinjan and V. van Heyningen. Discovery and assessment of conserved Pax6 target genes and enhancers. *Genome Res.*, 21:1349-1359, 2011.

GnpSeq NGS: a new tool to manage NGS data in URGI's information system (GnpIS)

Célia MICHOTEY[1], Nacer MOHELLIBI[1], Hadi QUESNEVILLE[1], Delphine STEINBACH[1]

[1] URGI, INRA de Versailles, route de Saint-Cyr, 78000, Versailles, France
{celia.michotey, nacer.mohellibi, hadi.quesneville, delphine.steinbach
}@versailles.inra.fr

Keywords NGS, database, management.

GnpSeq NGS : un nouvel outil pour gérer les données NGS dans le système d'information de l'URGI (GnpIS)

Résumé *Le résumé du papier (optionnel pour les contributions courtes)*

Mots-clés NGS, base de données, management.

1 Contexte

Grâce aux avancées technologiques apportées par les technologies de séquençage à haut débit (NGS, Next Generation Sequencing), il est maintenant possible d'obtenir rapidement et à moindre coup un volume important de données. Pour cette raison, les projets de séquençage, génotypage et phénotypage à haut débit fleurissent et le volume des données à traiter augmente.

L'URGI (Unité de Recherche en Génomique-Info) est une unité de recherche de l'INRA dont le but est de développer des outils, intégrer des données et étudier la dynamique et l'évolution des génomes de plantes. Elle héberge une plate-forme bioinformatique (labellisée IBISA) dont l'une des missions est de développer et maintenir un système d'information, nommé GnpIS, dédié aux plantes d'intérêt agronomique et leurs bio-agresseurs. Accessible via une interface web (http://urgi.versailles.inra.fr/gnpis), GnpIS est composé de plusieurs bases de données relationnelles modulaires et interopérables. Le module GnpSeq est dédié aux séquences (technologie Sanger), GnpSNP à la gestion des polymorphismes, GnpMap au mapping (cartes génétiques, QTLs et MetaQTLs), GnpArray à la transcriptomique, GnpGenome à l'annotation génomique, SIReGal aux ressources génétiques et Ephesis aux phénotypes liés à l'environnement.

Afin de répondre à l'arrivée des technologies NGS, la plateforme s'est donné comme objectif d'étendre son système d'information pour permettre la gestion des expériences et des runs issus des différents projets dont elle a à gérer les données. Pour cela, elle a fait évoluer le schéma de ses bases de données (UML), les outils d'insertions (ETL Talend) et l'interface de consultation des données (JAVA J2E).

2 GnpSeq NGS, une base de données pour gérer les données NGS

L'outil qui a été développé permet par exemple de répondre à des questions telles que : 'combien de projets NGS sont pris en charge sur la plateforme et quel type de technologies utilisent-ils ?', 'quelle est la question biologique posée derrière ces projets ?', 'comment ont été obtenus les runs et quelles études sont réalisées avec ?'.

Ce nouveau module, « GnpSeq NGS », est complètement intégré dans le système d'informations GnpIS. C'est un outil qui permet de faire des tableaux de bord orientés 'plateforme', mais c'est également le point d'entrée pour naviguer plus précisément dans les données, les runs étant liés aux résultats des pipelines d'analyses qui ont été exécutés avec ces données. GnpSeq NGS permet de gérer des données de runs (technologies 454, Illumina, HiSeq...) obtenus à des fins de séquençage ou de reséquençage de génome, de détection de polymorphisme... pour différentes espèces, que ce soit des arbres ou des plantes comme le blé,

la vigne. Dans le cas de la détection de polymorphismes, ce module permet aussi d'associer aux runs les analyses faites par le pipeline URGI MAPHiTS (M. Bras et al.).

Un format de soumission dédié a été développé pour faciliter l'insertion automatique des données NGS dans notre base de données Oracle grâce à l'utilisation de le l'outil ETL (Extract Transform and Load) « Talend Open Studio ». Au même titre que les autres modules de GnpIS, GnpSeq NGS gère la confidentialité des données via une authentification apache et un système de gestion des droits d'accès (groupes utilisateurs et vues) mis en place au niveau de la base de données. L'outil ainsi que l'ensemble des modules de GnpIS est en cours de mise à jour de licence (passage de la licence OpenSource Cecill à la licence AGPL v3).

Des données de séquençage du blé et de génotypage de la vigne ont été insérées dans l'outil. Une partie de ces données est déjà accessible librement sur la version publique de l'application. D'autres données en attente de publication, sont uniquement mises à la disposition des collaborateurs en accès restreint.

3 GnpSeq NGS, une interface de visualisation

Ce nouveau module « séquences » a été développé en java J2E et ajouté au portail de GnpIS : http://urgi.versailles.inra.fr/sequence. Il est composé de différents formulaires de requêtes qui facilitent la recherche et la récupération d'informations dans la base de données. On peut ainsi effectuer des recherches à partir d'expériences de séquençage, d'analyses, de taxons ou de projets. La recherche peut se faire de manière exhaustive ou de manière restreinte grâce à l'utilisation de filtres, par exemple : sélection de taxons, de type de séquençage et/ou de séquenceur. Les résultats sont résumés dans un tableau dont les colonnes correspondent aux différents critères de recherche du formulaire. Des liens cliquables permettent l'affichage de fiches contenant des informations complémentaires, plus détaillées, sur une donnée précise.

Le point central de l'application est la fiche de run. Accessible à partir d'une expérience ou d'une analyse donnée, elle regroupe l'ensemble des informations caractéristiques d'un run (provenance, utilisation, fichiers liés…). Elle permet également l'interopérabilité avec le module GnpIS chargé de la gestion des données de polymorphisme : GnpSNP. En effet, la fiche run étant commune aux deux applications, elle possède un système d'onglets qui permet de basculer facilement et de façon transparente entre "GnpSeq NGS" et "GnpSNP", et inversement. Cet affichage de l'information en fonction du contexte permet d'avoir un autre regard sur une même information : savoir comment le run a été obtenu et à quelles analyses il a été soumis (GnpSeq NGS), puis voir les résultats obtenus par une analyse de détection des SNP effectuée sur ce même run (GnpSNP). Cette fiche permet également d'avoir accès aux séquences et pour certaines données facilite l'accès à des outils d'analyse comme BLAST ou des pipelines à disposition sur le serveur Galaxy de l'URGI.

Remerciements

Ce travail a été financé par le GIS IBISA dans le cadre de l'appel d'offre plateforme IBISA 2009.

Remerciements à toute l'équipe URGI GnpIS et en particulier Erik Kimmel, Cyril Pommier et Daphné Verdelet pour les développements, Nathalie Choisne pour les données de vigne des projets ANR Muscares et GrapeReSeq, Michael Alaux pour les données de blé des projets IWGSC et ANR 3BSeq, Aminah-Olivia Keliet pour l'administration des bases de données.

References

[1] D. Samson-Steinbach, F. Legeai, E. Karsenty, S. Reboux, JB. Veyrieras, J. Just, E. Barillot: GénoPlante-Info (GPI): a collection of databases and bioinformatics resssources for plant genomics. *Nucleic Acids Research* 2003 Jan 1;31(1):179-82 ; PMID: 12519976

[2] D. Steinbach, E. Kimmel, A.-O. Keliet, M. Alaux, D. Verdelet, N. Mohellibi, , N. Choisne, S. Durand, J. Amselem, C. Pommier, I. Luyten, S. Reboux,, H. Quesneville: GnpIS: an original information system to bridge genetic and genomic plant and fungi data. *in prep. 2012*

[3] N.Choisne, M. Bras, N. Mohellibi, S.Arnoux, O. Inizan, AF. Adam-Blondon, H. Quesneville : MAPHITS: an efficient workflow for SNP detection. computer demo at Plant and Animal Genome Conference January 14-18, San Diego, USA

Integration of experiments results coming from the meta-analysis of QTLs (meta-QTL) in the URGI information system GnpIS

Dorothée VALDENAIRE[1*], Erik KIMMEL[1*], Olivier SOSNOWSKY[2], Johann JOEST[2], Hadi QUESNEVILLE[1] and Delphine STEINBACH[1]

[1] Unité de Recherche en Génomique-Info, UR1164 INRA, Route de Saint Cyr, 78000, Versailles, France
{dorothee.valdenaire, erik.kimmel, hadi.quesneville, delphine.steinbach}@versailles.inra.fr

[2] UMR de Génétique Végétale, UMR INRA-UPS-CNRS-INA PG, Ferme du Moulon, 91190, Gif-sur-Yvette, France
{joets, sosnowsky}@moulon.inra.fr

[*]These authors contributed equally to this work.

Abstract *URGI is an INRA bioinformatics unit dedicated to plants and pest genomics. We develop and maintain an information system called GnpIs, for plants of agronomical interest. This information system includes data corresponding to different themes as transcriptomic, polymorphism, genetic resources, genotypic/environment interaction, physical map, genetic annotation and genetic mapping. We present here the work done in the frame of the bioinformatics project ANR MetaQTL whose aim was 1) to improve GnpIS genetic mapping module to include QTL meta-analysis and their results (meta-QTL), 2) to integrate on URGI Web site, the new version of the analysis and visualization map software, BioMercator, developed at INRA Moulon and 3) to promote the use of the two tools simultaneously.*

Keywords meta-QTL, meta-analysis, BioMercator, data.

Intégration de résultats d'expériences provenant de méta-analyses de QTL (méta-QTL), dans le système d'information GnpIS de l'URGI

Résumé *L'URGI est une unité de bioinformatique de l'INRA dédiée à la génomique des plantes et de leurs bioagresseurs. Nous développons et maintenons un système d'information appelé GnpIS, pour les plantes et en particulier les plantes d'intérêt agronomiques. Ce système d'information regroupe des données correspondant à différentes thématiques comme la transcriptomique, le polymorphisme, les ressources génétiques, les interactions génotype/environnement, les cartes physiques, les annotations génétiques et la cartographie génétique. Nous présenterons ici, le travail réalisé dans le cadre du projet bioinformatique ANR MetaQTL, dont le but était 1) de faire évoluer le module de cartographie génétique pour gérer les données de méta-analyses de QTL et leurs résultats (les méta-QTL) et 2) d'intégrer sur le site Web de l'URGI, la nouvelle version du logiciel d'analyse et de visualisation de cartes, BioMercator, développé à l'INRA du Moulon et de 3) favoriser l'utilisation simultanée des deux outils.*

Mots-clés méta-QTL, méta-analyse, BioMercator, données.

1 Introduction

Over several years, large efforts have been made to produce crop genetics and genomic data at genome scale. To extract knowledge of these data, integration is required at various levels of resolution. Integrative QTL meta-analysis [1] methods have been developed with the objectives to:

- identify key regions of the genome consistently involved in the variation of a trait over several mapping populations, or alternatively specific to a particular population or genetic background.
- refine the positions of QTL reducing the size of the confidence intervals in regions of the genome where significant effects were identified in independent experiments by taking advantage of all information available [2].

The first aim of the bioinformatics project, ANR MetaQTL, was to develop a new enhanced version of BioMercator [3] software offering 1) improved with state-of-the-art map compilation and QTL meta-analysis

algorithms 2) a wizard tool to assist a user in genetic/physical/sequence-based map integration and 3) a new module for projection of QTL/metaQTL onto physical/sequence map. This new version (BioMercator V.3 paper submitted 2012) will provide end-user with a unique user-friendly workbench for genetic/physical data integration.

The second aim of the project was to promote the sharing of results produced with BioMercator, by extending the URGI Information System GnpIS [4,5] and its genetic database module GnpMap to 1) store this new data and 2) make them available for query or display at the genetic map and genome levels.

2 Realization

Regarding the integration of this new type of data (i.e. meta-analysis experiments and their results, the meta-QTLs found), three new tools were developped: 1) a standardized exchange format (Excel file), 2) a processing tool (checking and conversion) and 3) a tool for loading data in GnpIS genetic mapping module.

To allow users to search, browse and retrieve data from meta-analysis of QTLs, we improved GnpIS genetic mapping module web interface to provide a user-friendly environment. With intuitive search query forms, it is now possible to restrict requests to find the most relevant information according to user biological questions. For example, it is possible to compile for wheat species, all meta-QTLs, associated to some defined meta-traits of agronomical interest (grain protein content, yield), filtering either by maps name or markers or linkage groups. It is also possible to get all meta-QTLs corresponding to a list of existing QTLs.

These results are available on the web in a synthetic result table, with links to more detailed forms (cards) that describe Meta-analysis experiments, meta-QTLs, meta-Traits and genetic maps.

XML was chosen to be the communication format used between GnpIS web interface and BioMercator. It is therefore possible to launch BioMercator directly from the genetic mapping interface. This feature is available now for project partners and will be available to public after publication.

The developments made in the MetaQTL project and presented here provide indeed a complete environment for data storage and results visualization of meta-QTL data.

Access for query : http://urgi.versailles.inra.fr/GnpMap

Acknowledgements

This work was supported by ANR MetaQTL.

We would like to thank all the GnpIS/URGI team and particularly Aminah Keliet for her work on databases administration.

References

[1] B. Goffinet, S. Gerber, Quantitative trait loci: a meta-analysis. *Genetics*, 155:463-73, 2000.

[2] F. Chardon, B. Virlon, L. Moreau, M. Falque, J. Joets, L. Decousset, A. Murigneux, A. Charcosset, Genetic architecture of flowering time in maize as inferred from quantitative trait loci meta-analysis and synteny conservation with the rice genome. *Genetics*, 168:2169-85, 2004.

[3] A. Arcade, A. Labourdette, M. Falque, B. Mangin, F. Chardon, A. Charcosset, J. Joets, BioMercator: integrating genetic maps and QTL towards discovery of candidate genes. *Bioinformatics*, 20:2324-6, 2004.

[4] D. Samson-Steinbach, F. Legeai, E. Karsenty, S. Reboux, J.-B. Veyrieras, J. Just, E. Barillot, GénoPlante-Info (GPI): a collection of databases and bioinformatics resssources for plant genomics. *Nucleic Acids Research*, 31:179-82, 2003, PMID:12519976.

[5] D. Steinbach, E. Kimmel, A.-O. Keliet, M. Alaux, D. Verdelet, N. Mohellibi, N. Choisne, S. Durand, J. Amselem, C. Pommier, I. Luyten, S. Reboux, H. Quesneville, GnpIS: an original information system to bridge genetic and genomic plant and fungi data. *in prep.* 2012.

Widgets integration in Mobyle to enable Bioinformatics data visualization and edition

Hervé MÉNAGER[1], Bertrand NÉRON[1], Olivier SALLOU[2], Pierre TUFFÉRY[3] and Bernard CAUDRON[1]

[1] Centre d'Informatique pour la Biologie, Institut Pasteur,
28, rue du Dr Roux, 75724 PARIS Cedex, France
{hmenager, bneron, caudron}@pasteur.fr
[2] IRISA / University of Rennes 1 Campus de Beaulieu, 35000 Rennes, FRANCE
[3] MTi, INSERM UMR-S 973, Université Paris Diderot (Paris 7), Paris, France
{pierre.tuffery}@univ-paris-diderot.fr

Keywords web, portal, integration, web-services, workflows, visualization.

Intégration de widgets dans Mobyle pour la visualisation et l'édition de données bioinformatiques

Keywords web, portail, intégration, services web, workflows, visualisation.

1 Introduction

Mobyle is an open source framework and web portal specifically designed for the integration of bioinformatics software and databanks. Its web interface allows scientists, without installing anything locally, to use command line-based bioinformatics tools to perform analyses on remote computing resources. The high level of integration between the different tools provided enables and guides users throughout the construction of potentially complex protocols, chaining interactively successive tasks in an exploratory mode, or automating their execution with workflows. However, it can be extremely tedious for users to understand and edit complex bioinformatics data (e.g., multiple sequence alignments, phylogenetic trees, structural data) in their native and often text-based formats, using traditional web interfaces. To solve this issue, Mobyle now offers the ability to use rich browser-embedded components, significantly reducing the previously mentioned usability problem.

2 Mobyle services outline

Mobyle is a system which provides an access to different software elements, in order to allow users to perform bioinformatics analyses. Such software elements are called services, and belong to three distinct categories:

- **programs** describe software that are executed on the server side (command line tools), and return results in the form of files,
- **workflows** describe compositions of successive and/or parallel calls to other server-side services (programs and/or subworkflows), and also return results as files,
- **viewers/widgets** describe browser-embedded visualization components, such as applets, that allow users to visualize data (results or bookmarks) stored in the Mobyle server.

3 Visualisation and Edition Widgets

Viewers are widgets which provide a way to embed type-dependent visualization components for the data displayed in the Mobyle portal. They provide a way to easily overcome the limitations of the basic previews offered by Mobyle for text-based-formatted data. Such custom interfaces can incorporate HTML-embeddable components such as Java applets. These widgets now also offer the possibility to bookmark data edited by the client back to the portal.

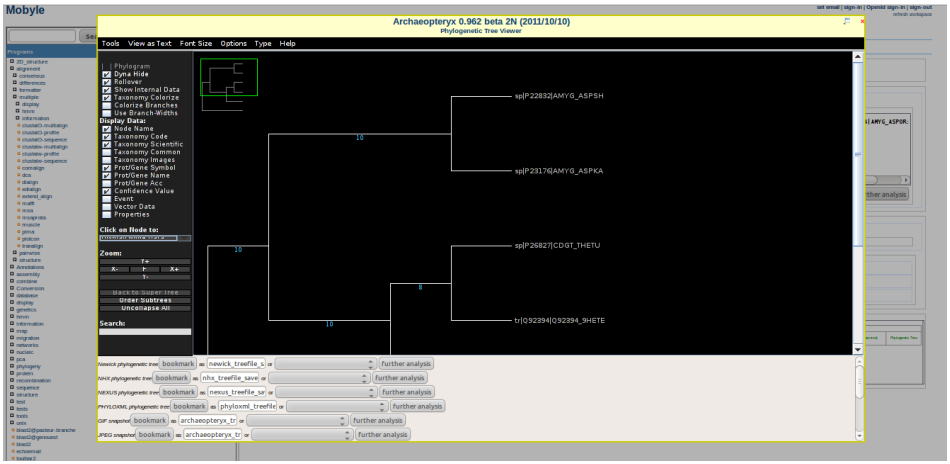


Figure 1. Archaeopteryx edition screen. This illustrates the edition of user data hosted in the web-hosted Mobyle workspace. Here users can manipulate phylogenetic trees using Archaeopteryx, and save the edited trees back on the server

These components are "plugged" to the Mobyle portal using XML description files similar to those used in other types of services, but replace the server-side command line generation and results detection code with web interface templates and javascript API calls. The description of each viewer/widget represents (1) each data item that has to be loaded as an input and a corresponding HTML template, and (2) each possible export provided by the component as an output and its corresponding javascript API call.

For instance, using viewers, the Archaeopteryx[2] component is automatically included wherever compatible data are displayed, so that users can immediately visualize the results of phylogenetic tree processing programs in the portal. As showed in Fig. 1, users can directly save the edited tree(s) on the server, in one of the many representation formats handled by the Applet (Newick, Nexus, PhyloXML, etc.), thus offering the possibility to reuse them in other Mobyle-hosted services.

4 Conclusion

The enhancement of viewer widgets capabilities to enable the edition of bioinformatics data completes the spectrum of possibilities offered by the Mobyle framework, by letting users perform operations such as manual adjustments or annotations, which are often required in Bioinformatics. Relying on the open and widely used protocols and formats of the web, the provided framework makes it easy to integrate additional existing "rich" components such as Java applets and Ajax-based or Flash-based components.

Acknowledgements

The MobyleNet project is funded by the IBISA (http://www.ibisa.net) initiative. The CIB/BCBB 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] M.V. Han and C.M. Zmasek, phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics*, 10:356, 2009.

Horizontally acquired adaptation and pathogenicity of *Mycobacterium tuberculosis*.

Ludovic MALLET, Sabine MENIGAUD and Patrick DESCHAVANNE

INSERM, U973, F-75205 Paris, France
Univ Paris Diderot, Sorbonne Paris Cité, UMRS 973, MTi, F-75205 Paris, France
ludovic.mallet@jouy.inra.fr
patrick.deschavanne@univ-paris-diderot.fr

Abstract

Keywords Comparative genomics, Evolutionary genomics, Horizontal transfers, *Mycobacterium tuberculosis*.

The *Mycobacterium tuberculosis* complex is the causative agent of tuberculosis, which still infects one people out of three worldwide and is responsible for millions of death each year. Emergence of highly virulent and drug-resistant strains raised therapeutic failures and fatal issues. Continuous species adaptation have to be fought by continuous drug development. Models of evolution of the *Mycobacterium tuberculosis* complex genomes proposed that harmless soil mycobacteria gained genes in a step-by-step manner through horizontal transfers that brought pathogenic lifestyle, host adaptation and virulence variability [1]. To detect those acquired elements, we adapted benchmarked parametric methods [2] known to assess and detect different features of transfers compared to phylogeny-based methods [3]. We applied a combination of two parametric methods based on the tetranucleotide frequencies in sliding windows and codon usage in 29 mycobacterial genomes. The genomes were selected to include non-pathogenic soil mycobacteria, pathogenic species of variable animals host range including Human and species of the *Mycobacterium tuberculosis* complex which are responsible for tuberculosis in mammals and unable to survive outside of their host. Multiple strains of *Mycobacterium tuberculosis* were included to assess their variability. Horizontally acquired orthologous genes were clustered with the Reciprocal Best Hit criterion. The groups were used to built a phylogenetic tree under a parsimonious model thus providing the gene content at each node of the tree. Each transfered group was then mapped on the tree to infer its most probable node of acquisition.

In the 29 genomes, an average of 11,7% of genomic DNA is horizontally acquired. The transferred regions include 4271 groups of orthologous genes acquired all along the evolution history of the species of this study. We built the phylogenetic tree and selected orthologous groups that were acquired around the nodes that segregate phenotypes of pathogenicity (P), adaptation to the Human as an host (H) and the formation of the *Mycobacterium tuberculosis* complex (C). Those three major evolution nodes present several acquisitions in which we found characterized virulence factors and essential products. Analysis of the origin of these transferred regions showed an evolution of the donor species spectrum, with an increase in proteobacteria and viruses compared to the transferred regions found in the soil species. Functional classes of the transferred genes significantly differs from the rest of the genome with an enrichment in virulence and adaptation factors, phages and insertion sequences and PE/PPE proteins family members.

We identified horizontally transferred regions acquired at major nodes of the species evolution containing factors of virulence and adaptation. The results match and support the proposed model of evolution of the *Mycobacterium tuberculosis* complex, in which the horizontal gene transfers provided features that changed life style and host range. This study implicates partners from CNRS and Pasteur Institute that are currently carrying functional characterization and assessing implication in strain variability and drugability of the transferred products.

Acknowledgements

This work was funded by the French ministry of reasearch (LM) and the French national agency for research (ANR MIE TB-Hits 2010) (SM). Thanks to Hélène Chiapello.

References

[1] Jichan Jang, Jennifer Becq, Brigitte Gicquel, Patrick Deschavanne, and Olivier Neyrolles. Horizontally acquired genomic islands in the tubercle bacilli. *Trends Microbiol*, 16(7):303–8, Jul 2008.

[2] Jennifer Becq, Cécile Churlaud, and Patrick Deschavanne. A benchmark of parametric methods for horizontal transfers detection. *PLoS One*, 5(4):e9989, 2010.

[3] Frédéric Veyrier, Daniel Pletzer, Christine Turenne, and Marcel A Behr. Phylogenetic detection of horizontal gene transfer during the step-wise genesis of mycobacterium tuberculosis. *BMC Evol Biol*, 9:196, 2009.

GOHTAM : A website for "Genomic Origin of Horizontal Transfers, Alignment and Metagenomics"

Ludovic MALLET, Sabine MENIGAUD, Géraldine PICORD, Cécile CHURLAUD, Alexandre BOREL and Patrick DESCHAVANNE

INSERM, U973, F-75205 Paris, France
Univ Paris Diderot, Sorbonne Paris Cité, UMRS 973, MTi, F-75205 Paris, France
ludovic.mallet@jouy.inra.fr
patrick.deschavanne@univ-paris-diderot.fr

Abstract

Keywords Website, horizontal transfers, phylogeny, genome alignment, databases, metagenomics

GOHTAM is a website dedicated to genomic and horizontal transfers (HT) analyses featuring several services such as HT detection, potential origin of HT assessment, inference of phylogenetic trees based on tetranucleotide frequencies, metagenomics analyses, and whole genome rearrangements studies [1].

The website provides access to a set of two efficient [2] parametric HT detection methods relying on tetranuclotide frequencies (genomic signature) and codons frequencies (Fig. 1A). These parametric methods have been reported to work on bacterial genomes [2,3,4] and tetranucleotide frequencies-based method has been used on lower eukaryotes that showed low intragenomic variability [5,6].

Detected transfered regions can be assessed for their potential origin by comparing their signatures to a database containing the signature of about 250,000 species present in GenBank (Fig. 1B).

The same database is also included in the phylogenetic tree inference tool relying on tetranuclotide frequencies distances [7]. Users can thus build a neighbor-joining tree with mixed personal sequences and database entries (Fig. 1C) available in png, svg and newick formats. Tetranucleotide frequencies allow one to build trees with entries sharing no sequence homology. Such feature is usefull for building trees of uncomplete genomic sequences or identifying taxonomics relatives of unknwon materials. The identification of potential donors of a fragment or an atypical region in a genome can provide insights on dna sharing between species. The potential origin of unknown DNA fragments can be assessed through a 'metagenomics' tool. The ten closest database entries in term of signature distance are returned and displayed along with confidence estimators (Fig. 1B).

Basic computation and display of tetranuclotide frequencies are available for illustrative purpose (Fig. 1D) as well as contingency and frequencies matrix. The website also embed the MUMMER software [8] to align genomes (Fig. 1E) and draw rearrangements events longer than 1 Kb (not shown).

Examples and help are available online. The results produced online can be fully downloaded in multiple formats on the result page or through a personal space. Further development will be done on quantitative estimation of HT donor species for a genome HT set.

Acknowledgements

This work was funded by the French ministry of research (LM) and the French national agency for research (ANR MIE TB-Hits 2010) (SM).

References

[1] Sabine Ménigaud, Ludovic Mallet, Géraldine Picord, Cécile Churlaud, Alexandre Borrel, and Patrick Deschavanne. A website for "Genomic origin of horizontal transfers, alignment and metagenomics". *Bioinformatics*, March 2012.

[2] Jennifer Becq, Cécile Churlaud, and Patrick Deschavanne. A benchmark of parametric methods for horizontal transfers detection. *PLoS One*, 5(4):e9989, 2010.

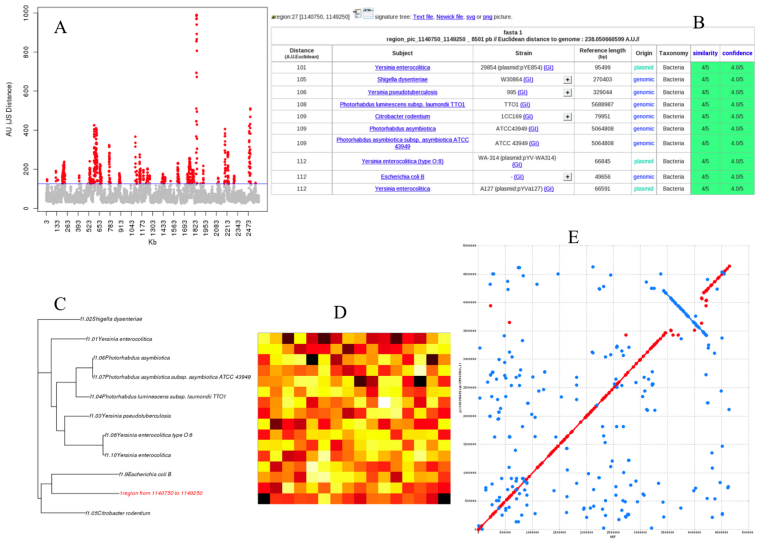


Figure 1. Different key capture from the website. **A** HT detection. **B** HT origin result. **C** Phylogenetic tree relying on tetranuclotide frequencies. **D** tetranuclotide frequencies representation. **E** MUMMER alignment.

[3] Christine Dufraigne, Bernard Fertil, Sylvain Lespinats, Alain Giron, and Patrick Deschavanne. Detection and characterization of horizontal transfers in prokaryotes using genomic signature. *Nucleic Acids Res*, 33(1):e6, 2005.

[4] Jennifer Becq, Maria Cristina Gutierrez, Vania Rosas-Magallanes, Jean Rauzier, Brigitte Gicquel, Olivier Neyrolles, and Patrick Deschavanne. Contribution of horizontally acquired genomic islands to the evolution of the tubercle bacilli. *Mol Biol Evol*, 24(8):1861–71, Aug 2007.

[5] P J Deschavanne, A Giron, J Vilain, G Fagot, and B Fertil. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol Biol Evol*, 16(10):1391–9, Oct 1999.

[6] Ludovic V Mallet, Jennifer Becq, and Patrick Deschavanne. Whole genome evaluation of horizontal transfers in the pathogenic fungus aspergillus fumigatus. *BMC Genomics*, 11:171, 2010.

[7] Charles Chapus, Christine Dufraigne, Scott Edwards, Alain Giron, Bernard Fertil, and Patrick Deschavanne. Exploration of phylogenetic data using a global sequence analysis method. *BMC Evol Biol*, 5:63, 2005.

[8] A L Delcher, S Kasif, R D Fleischmann, J Peterson, O White, and S L Salzberg. Alignment of whole genomes. *Nucleic Acids Res*, 27(11):2369–76, Jun 1999.

Optimizations to compute large correlation matrix onto GPU system of hybrid HPC clusters

Dany TELLO[1], Fouad BOUMEZBEUR[2], Victor ARSLAN[1], Vincent DUCROT[1], Pierre LÉONARD[2], Bouziane MOUMEN[2], Nicolas PONS[2], Tarik SAIDANI[1], Pierre RENAULT[2], Sean KENNEDY[2], Mathieu ALMEIDA[2], S.Dusko EHRLICH[2], Sébastien MONOT[1] and J.M. BATTO[2]

[1] AS Plus, 2 rue René Coche, 92170, Vanves, France

s.monot@asplus.fr

[2] UMR1319 INRA, Domaine de Vilvert, 78352, Jouy-en-Josas, France

jean-michel.batto@jouy.inra.fr

Keywords GPGPU, GPU, Supercomputer, Big Data, Correlation.

1 Introduction

MetaQuant (INRA UMR1319) is involved in the EC FP7 MetaHIT[1] (Metagenomics of Human Intestinal Tract) project, which involves the collaborative efforts of major genomics and bioinformatics institutes, such as BGI and EMBL. The project, started in 2008, has generated a large amount of data (13 TBytes) that has been organized in a large matrix of 3.3 millions rows by 800 columns. Since November 2010, a critical task of the MetaHIT project has been to develop a clustering process able to analyze and structure this data. The OpenGPU[2] collaborative research project, involving the French company AS+ and the MetaQuant platform, has initiated the design, coding and scale up for a multi-GPU computing cluster to accomplish this task.

Optimizations have been done to use in an efficient way the capability of the GPU processor. The optimized program has been benchmarked on the hybrid part (GPU/nVidia M2090) of the GENCI supercomputer named Curie installed at TGCC (Très Grand Centre de Calcul du CEA - http://www-hpc.cea.fr/en/complexe/tgcc-curie.htm). The benchmarking of the code on a real case onto the Curie supercomputer gives a result in less than an hour.

This code is named MetaProf and is currently released on request with a non disclosure license.

Acknowledgements

This work was supported by OPENGPU and Edari.

References

[1] www.metahit.eu

[2] www.opengpu.net/EN

Microbial *de novo* genome assembly:

Comparison of CLC Genomics & VELVET for assembly of contaminated but deeply covered Illumina 1.5 single reads.

Gregory FARRANT[1,2], Erwan CORRE[2], Mark HOEBEKE[2], Wilfrid CARRE[2], Christophe CARON[2], Frédéric PARTENSKY[1] and Laurence GARCZAREK[1]

[1] PROCARYOTES PHOTOSYNTHETIQUES MARINS, UMR7144 CNRS, Station Biologique, Place G. Teissier, 29680, Roscoff, France
[2] ABiMS, FR2424 CNRS-UPMC, Station Biologique, Place G. Teissier, 29680, Roscoff, France
gfarrant@sb-roscoff.fr

Abstract *The METASYN sequencing project aims at producing 25 novel genetically diverse marine* Synechococcus *cyanobacteria genomes using the Illumina 1.5 technology. For that purpose a pilot project on two strains produced two raw datasets of ~80 million reads for a final coverage around 2.300×. A first assembly using CLC Genomics led to ~50 contigs that were combined into a single scaffold with few N-gaps for both strains (genome size of ~2.5Mbp). These results were then compared with an assembly performed using the open source and widely used assembler VELVET. Our goal was to determine the optimal k-mer size and relevant data cleaning steps that would provide results comparable in quality to CLC Genomics.*

Keywords Microbial genome assembly, Next Generation Sequencing, short reads cleaning.

Assemblage de génomes microbiens *de novo* :

Comparaison de CLC Genomics et VELVET pour l'assemblage de courts fragments générés par Illumina 1.5 avec un fort taux de couverture mais incluant des contaminants

Résumé *Le projet METASYN vise à séquencer 25 nouveaux génomes de cyanobactéries marines couvrant une grande part de la diversité génétique existant au sein du genre* Synechococcus *à l'aide de la technologie Illumina 1.5. Un projet pilote, mené sur deux premières souches, a généré deux jeux de ~80 millions de fragments avec un taux moyen de couverture de séquençage de l'ordre de 2300×. Un premier assemblage, réalisé à l'aide de CLC Genomics a généré ~50 contigs par souche (taille du génome : ~2,5Mpb). Ces contigs ont été réunis en un scaffold unique, séparés par des N. Ces résultats ont été comparés à ceux obtenus avec un assembleur "open source" très utilisé : VELVET. Notre objectif était de déterminer la taille optimale des k-mer et les étapes de nettoyage adaptées permettant d'obtenir une qualité d'assemblage comparable à celle obtenue avec CLC Genomics.*

Mots-clés Assemblage de génomes microbiens, NGS, nettoyage de fragments.

1 Introduction

Marine photosynthetic organisms play a major role in primary production. Among these, two closely related cyanobacteria are particularly abundant and ubiquitous: *Synechococcus* and *Prochlorococcus*. Because of their ecological importance and limited genome size, these picocyanobacteria have been selected for many sequencing projects since the early 2000's. The METASYN project, in collaboration with the Genoscope, aims at sequencing 25 new *Synechococcus* genomes, selected to cover at best the large genetic diversity, pigmentation and geographical distribution known in this genus. A pilot study consisted in *de novo* sequencing of two strains: MINOS11 (Roscoff Culture Collection RCC 61) & BOUM118 (RCC 2421) using Illumina 1.5. For both strains, two datasets have been produced: ~80 million single 100 pb reads and ~90 million mate-pairs of 50 bp reads.

The challenge of *de novo* assembly of these reads was to provide a limited amount of contigs and then to exploit the paired reads to reorganize these contigs and assemble them into a single chromosome by inserting N-gaps in between. This step requires a majority of contigs with size larger than the distance between two paired reads (~10 kbp).

2 Methods

The CLC Genomics (CLC Bio, Aarhus, Denmark) assembly was performed after cleaning single reads using the CLC trimming tool with a threshold set at 0.005. Use of VELVET assembler [1] required a preliminary cleaning of raw data that were performed using the following approaches:

- AdaptiveTrim (adaptiveTrim.pl ©INRA 2010) to get rid of low-quality positions,
- JellyFish (University of Maryland, College Park) and the Quake k-mer cleaning and fixing pipeline [2] to eliminate sequencing errors by detecting k-mer coverage bias,
- Contaminants sequences were eliminated after identifying by BLAST of post-assembly contigs against 'nr' and selecting reads by negative mapping on contaminants genomes using Geneious [3]. Moreover, reads mapping on contigs showing a coverage rate distant from expected (as determined by mapping reads on a reference *Synechococcus* genome) were also eliminated.

Then, several assemblies using VELVET were performed using different k-mer sizes.

3 Results

Contigs obtained from the CLC Genomics assembler after trimming were as follows:

- BOUM118: 35 contigs, average length = 66,080 bp, average coverage = 2547x
- MINOS11: 81 contigs, average length = 28,041 bp, average coverage = 1,366x

The standard settings of VELVET on raw single data led to a very large number of contigs (~1,000) with a small average size (~4000 bp) and a low coverage (~800x). A slightly better assembly was obtained for k-mer size of about half the length of a read (i.e. 51 bp).

The main difficulty for a deBruijn graph-based assembler such as VELVET is to face an extremely large diversity of k-mers due to sequencing errors or to the presence of contaminant DNA in the sequenced pool. Elimination of low quality reads using AdaptiveTrim (minimal quality = 20, minimal length = 75 bp) led to a 4-fold drop of the mean coverage despite keeping about half of the original dataset (e.g. 56.4% for MINOS11) and no significant improvement of contigs length. Nevertheless, the summed size of all assembled fragments was higher than the actual estimated size of MINOS11 chromosome. Getting rid of sequencing errors using Quake (k-mers with coverage below 40x have been eliminated) led to somehow similar results but appeared to suppress the effect of k-mer size previously observed. However, the most significant improvement of the assembly was obtained by eliminating contaminants sequences (average size of ~3,500 and average coverage of ~560x).

In the near future, we will test other optimization steps including a more restrictive choice of minimal read length after quality-based trimming (95 to 100 pb), a read selection by negative mapping of the reads on contigs selected based on their coverage as well as elimination of obvious contaminants by BLAST of contigs against 'nr'.

Acknowledgements

This work was supported by the ANR génomique microbienne PELICAN (ANR-PCS-09-GENM-200).

References

[1] Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008 May;18(5):821-9. Epub 2008 Mar 18. PMID: 18349386

[2] Kelley DR, Schatz MC, Salzberg SL. Quake: quality-aware detection and correction of sequencing errors. *Genome Biol*. 2010;11(11):R116. Epub 2010 Nov 29.

[3] Drummond,A.J., Ashton,B., Burton,S., Cheung,M., Cooper,A., Heled,J., Moir,R., Stones-Havas,S., Sturrock,S., Thierer,T. et al. (2010) Geneious v5.1, http://www.geneious.com/

EVA: Exome Variation Analyzer

A tool for filtering strategies in medical genomics

Sophie Coutant [1], Chloé Cabot [2], Wassim Tair [3], Arnaud Lefebvre [2], Martine Léonard [2], Elise Prieur-Gaston [2], Dominique Campion [1], Thierry Lecroq [2], Hélène Dauchel [2]

[1] Rouen University, INSERM U1079 Molecular genetics of cancer and neuropsychiatric diseases, 76183 Rouen, France
{sophie.coutant, dominique.campion}@inserm.fr

[2] Rouen University, LITIS EA 4108 Computer science, information processing and systems laboratory, 76821 Mont-Saint-Aignan , France
{arnaud.lefebvre, martine.leonard, elise.prieur, thierry.lecroq, helene.dauchel}@univ-rouen.fr; chloe.cabot@etu.univ-rouen.fr

[3] Rouen University, CNRS UMR6085 Raphaël Salem Mathematics Laboratory, 76801 St Etienne du Rouvray, France
wassim.tair@etu.univ-rouen.fr

Keywords Clinical bioinformatics, exome, variation, filter, software, Alzheimer disease.

1 Introduction

Among next-generation sequencing technologies, targeted sequencing of all coding regions (called whole exome sequencing (WES)), has become in few years the strategy of choice to identify coding allelic variants for rare human monogenic disorder [1]. The exome-scale sequencing approach is a revolution in medical genetics history, impacting both fundamental research, and diagnostic methods leading to personalized medicine [2]. Numerous algorithms and software tools have been developed to ensure the variant discovery from WES data. Workflows include steps for quality control of the short reads, alignment of the short reads to a reference sequence, variation calling, and variation annotation [3]. However, the challenge remains in efficient filtering strategies to find, among ~20,000 candidates obtained per individual exome, the causal variant(s) and the corresponding gene for a rare disease [4]. For this purpose, additional analytical procedures which implicate various heuristic filtering strategies have emerged [5]. Usually, wide range common variations are firstly excluded. Then, the filters inspect the type of variations (focus on presumed deleterious allelic variants) and the functional effect of variations on gene products. Filtering strategies considere also the mode of inheritance of the disorder suggesting by pedigree. Finally, taking advantages of multiple individuals, intersection or differential exome strategies can drastically reduce the remaining variations to several genes. To deal with the expansion of WES in medical genomics individual laboratories, new convivial and versatile software tools have to implemente the unavoidable filtering steps of variations. Non-programmer biologists have to be autonomous to combine themselves different filtering criteria and conduct a personal strategy depending on their assumptions and study design.

With this aim, in parternship with and for medical geneticists, we developed EVA (Exome Variation Analyzer), a user-friendly web interfaced sofware dedicated to filtering strategies for medical projects investigated with WES [6]. EVA has been successfully used in a demonstrative case study allowing to identify a new candidate gene related to a rare form of Alzheimer disease [7].

2 EVA: overview and filtering strategies

For a given exome project corresponding to several individuals, EVA allows to manage data through different modules, as strictly condidential to the project owner. The first one, an online *Variation Integration module* takes in input standard raw Variant Calling Format files, and annotates the variations (Single Nucleotide Variation (SNV) and indel (microinsertion or microdeletion)) thanks to ANNOVAR and the Variation Effect Predictor Ensembl API. Annotated variations are then stored in a MySQL *database called ExomeDB*. On the web interface, the *Variation Statistics module* allows to display SNV and indel categories bar charts and pie charts, and amino-acid and nucleotide substitution matrix. It can be done for all the variations corresponding to a project or for a selection of individuals, chromosomes, regions or genes. The *filtering strategy module* proposes to combine multiple filters to drastically narrow down variations (see details below). The *Table Browser module* allows to explore whole exome data by project, individual, gene

or variation through sortable and interactive tables. A *Search module* offers a direct and quick access to a gene or a variation for a given project. It also provides caryotypic and graphical gene view.

The *Filtering Strategy module* integrates the current categories of filters based on common variations, molecular type of the variants, modes of inheritance, homozygous or heterozygous nature of the allelic variant and multiple individuals. First, EVA allows to compare the data to international catalogues of variations (dbSNP, HapMap Project, 1000 genomes Project, Complete Genomics and IntegraGen public data, Exome Sequencing Project). As a result, the set of all the variations is divided in known and unknown variations. Then, other filters offer to sift variations depending on their: (i) functional categories for SNV (synonymous, miss sense, stop loss and non sense) and indels (frameshift or non frameshift); (ii) genic region (UTR, CDS, intronic splice region); (iii) quality score. Finally, one of the strengths of EVA is the implementation of inheritance filters considering intersection or conversely differential exome strategies: (i) recurrence strategy for dominant or recessive independent familial cases; (ii) filters for homozygous, heterozygous or composite cases in intra-familial studies; (iii) and *de novo* strategy for sporadic cases or trio-family. EVA offers export files (cvs) and cross-links to external relevant databases and softwares for further functional effects inspection of the small subset of sorted candidate variations and genes.

3 Conclusions and perspectives

EVA is developed to be a user-friendly, versatile, and efficient-filtering assisting software for WES. It constitutes a platform for data storage and for drastic screening of clinical relevant genetics variations by non-programmer medical geneticists. Thereby, it provides a response to new needs at the expanding era of medical genomics investigated by this technology for both fundamental research and clinical diagnostics. The web address is public [http://bioinfo.litislab.fr/EVA/] but access is permitted through an authentication process using a project's specific login and a password given by the administrator. Version of the human genome, international variation catalogues, inheritance and other type of filters, organization of results tables and graphics are regularly updated. EVA will be soon available for free downloads. Future development include a more specific filtering strategy for somatic mutations and a new companion tool for functionally contextualize a small list of candidate variations.

Acknowledgements

This work has been partially supported by Grant PHRC GMAJ, Centre national de référence Malades Alzheimer jeunes.

References

[1] C. Garvey, A. Cosgrove, N. Attar, G. Bilsborough, T. Creavin (Eds) and J. Shendure (Guest Ed), Exome sequencing special issue In *Genome Biology* , 12 (9), 2011.

[2] M. Bamshad, S. Ng, A. Bigham, HK. Tabor, M. Emond, D. Nickerson, J. Shendure, Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* 2011, 12(11):745-55.

[3] J. Zhang, R. Chiodini, A. Badr and G. Zhang, The impact of next-generation sequencing on genomics, *J. Genet Genomics*, 38:95-109, 2011.

[4] N.O. Stitziel, A. Kiezun and S. Sunyaev, Computational and statistical approaches to analysing variants identified by exome sequencing, *Genome Biology* , 12 (9): 227-237, 2011.

[5] C. Ku, D. Cooper, C. Polychronakos, N. Naidoo, M. Wu and R. Soong, Exome sequencing: dual role as a discovery and diagnostic tool. *Ann Neurol* 2012, 71(1):5-14.

[6] S. Coutant, A. Lefebvre, M. Léonard, É. Prieur-Gaston, D. Campion, T. Lecroq and H. Dauchel, EVA: Exome Variation Analyzer, a convivial tool for filtering strategies, in R. Bellazzi and P. Romano (eds.), *Proceedings of the Eleven International Workshop on Network Tools and Application in Biology: Clinical Bioinformatics (NETTAB 2011)*, Pavia, Italy, pp 25-29. 2011.

[7] C. Pottier, D. Hannequin, S. Coutant, A. Rovelet-Lecrux, D. Wallon, S. Rousseau, S. Legallic, C. Paquet, S. Bombois, J. Pariente, C. Thomas-Anterion, A. Michon, B. Croisile, F. Etcharry-Bouyx, C. Berr, J.-F. Dartigues, P. Amouyel, H. Dauchel, C. Boutoleau-Bretonnière, C. Thauvin, T. Frebourg, J.-C. Lambert, D. Campion, collaborators PHRC GMAJ, High frequency of potentially causative SORL1 mutations in autosomal dominant early-onset Alzheimer disease. *Mol. Psychiatry* 2012, AOP, 3 April 2012: doi:10.1038/mp.2012.15.

A statistical approach to estimate DNA copy number from capture sequencing data.

Guillem RIGAILL[1,2], Roelof J.C. KLUIN[3], Zheng XUE[4], Rene BERNARDS[4], Ian J. MAJEWSKI[4] and Lodewyk F.A. WESSELS[1,2]

[1] Bioinformatics and Statistics, The Netherlands Cancer Institute, 1066 CX Amsterdam The Netherlands
g.rigaill@nki.nl
[2] Cancer Systems Biology Center (CSBC), The Netherlands Cancer Institute, 1066 CX Amsterdam The Netherlands
[3] Central Microarray Facility, The Netherlands Cancer Institute, 1066 CX Amsterdam The Netherlands
[4] Department of Molecular Carcinogenesis, The Netherlands Cancer Institute, 1066 CX Amsterdam The Netherlands

Abstract *Target enrichment, also referred to as DNA capture, provides an effective way to focus sequencing efforts on a genomic region of interest. Capture data is typically used to detect single nucleotide variants. It can also be used to detect copy number alterations (CNAs), which is particularly useful in the context of cancer, where such changes occur frequently. In copy number analysis of capture sequencing data it is common practice to determine logratios between test and control samples, but this approach results in a loss of information as it disregards the total coverage at a locus.*

We instead modeled the coverage of the test sample as a linear function of the control sample. This approach is able to deal with regions that are completely deleted, which are problematic for methods that use log ratios. To demonstrate the utility of our approach, we used capture data to determine copy number for a set of 600 genes in a panel of nine breast cancer cell lines. We found high concordance between our results and those generated using a SNP genotyping platform. When we compared our results to other methods, including ExomeCNV, we found that our approach produced better overall correlation with SNP data.

Keywords DNA copy number, capture sequencing, statistics

Interactome-Transcriptome Integration Uncovers More Stable and Better Performing Biomarkers in Breast Cancer

Maxime GARCIA[1,2,3,4], Pascal FINETTI[1,2,3,4],
François BERTUCCI[1,2,3,4], Daniel BIRNBAUM[1,2,3,4] and Ghislain BIDAUT[1,2,3,4]

[1] Inserm, U1068, CRCM, Marseille, F-13009, France
[2] CNRS, UMR7258, CRCM, Marseille, F-13009, France
[3] Institut Paoli-Calmettes, Marseille, F-13009, France
[4] Aix-Marseille Univ, Marseille, F-13284, France
maxime.garcia@inserm.fr, {finettip, bertuccif}@marseille.fnclcc.fr,
{daniel.birnbaum, ghislain.bidaut}@inserm.fr,

Keywords Breast cancer, large scale data integration, transcriptome, interactome, biomarkers.

1 Introduction

With the development of high-throughput gene-expression profiling technologies came the opportunity to define genomic signatures predicting clinical condition or patient outcome. However, such signatures show dependency on training set, lack of generalization and instability, due for some part to microarray data topology. Additional difficulties as we hypothesize are that subtle molecular perturbations in driver genes leading to cancer and metastasis (masked in classic microarray analysis) may provoke expression changes of greater amplitude in downstream genes (easily detected). Approaches combining gene-expression data and other sources of informations were devised to overcome these problems[1,2]. We are proposing an interactome-based algorithm: ITI (Garcia et al.[3,4]) 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 regions in the interactome whose expressions are discriminating for predicting 5 years relapse free survival in breast cancer. This method re-implements the algorithm proposed by Chuang et al. [1] with the added capability to extract a genomic signature from several gene-expression data sets simultaneously.

2 Methods

Two data types were 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 taking into account physical interactions and co-citation in articles (HPRD[5], Ramani[6], MINT[7], IntAct[8] and DIP[9]). All PPI were kept and 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 with affy and gcrma packages. Tumors without relapse information were removed, leading to a final compendium of 6 datasets containing 930 tumors [10,11,12,13,14,15]. Two studies were made by leaving out one dataset for cross-validation purpose with independent testing [10,14]. For each studies, two separate signatures were established on patients over-expressing Estrogen Receptor (ER+) and patients non-expressing Estrogen Receptor (ER-). A stratified 10-fold cross validation was performed to avoid potential bias in subnetworks selection. Interactome regions whose gene expression (Pearson correlation) were highly correlated with Distant Metastasis Free Survival [DMFS] status were then detected for each training dataset. Random distributions of score were drawn to assign p-values to the subnetworks and perform a statistical validation. Finally, the discriminative power of statistically significant subnetworks was tested against datasets held apart for independent testing.

3 Results and Discussion

We found two sets of 6 and 139 subnetworks signatures linked respectively to 5 years relapse free survival in breast cancer ER+ and ER-. These were confronted to Wang[15] and Van't Veer[16] signatures against independent datasets. Against Desmedt's dataset [10], in ER+/ER- the accuracy is respectively of 0.74/0.54, 0.41/0.44 and 0.60/0.38 for ITI, Van't Veer and Wang signatures. Against van de Vijver's dataset [14], in ER+/ER- the accuracy is respectively of 0.52/0.53, 0.62/0.53 and 0.63/0.56 for ITI, Van't Veer and Wang signatures. The performance is inferior in the second study and may reflect a bias toward Affymetrix induced by the training compendium. More interestingly, signatures found with ITI are characterized with a greater stability than the one measured between Wang and van de Vijver's signatures, with a respective overlap of 11.5% and 32.8% for ER+ and ER- ITI signatures. Wang and van de Vijver' signatures have only 3 genes in common (<5% overlap). Subnetworks composing our signatures were stored in a database, available from the ITI web site (http://bioinformatique.marseille.inserm.fr/iti). Intrinsic biology of the extracted subnetworks was examined using annotation information from the NCBI EntrezGene database and the Gene Ontology Consortium. We found that subnetworks formed complexes functionally 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. Several drivers genes were detected, including CDK1, NCK1 and PDGFB, some not previously linked to breast cancer relapse. This resource is the first of its kind to allow linking a human interactome to diseases or clinical situations. It can be mined for identification of potential drug targets to establish finer disease models.

Acknowledgements

Institut National du Cancer and Institut de la Santé et de la Recherche Médicale Grant 08/3D1616/Inserm-03-01/NG-NC (to G.B.); Ligue Nationale contre le Cancer (label D.B.). Support for the Beowulf cluster was obtained from Fondation pour la Recherche Médicale Young Team grant (To G.B.); Institut National de la Santé et de la Recherche Médicale - Région Provence-Alpes Côte d'Azur Doctoral Fellowship (to M.G.)

References

[1] H.Y. Chuang et al., Network-based classification of breast cancer metastasis. *Mol Syst Biol.*, 3:140, 2007.
[2] V. Lazar et al., A network-based, integrative study to identify core biological pathways that drive breast cancer clinical subtypes, *British Journal of Cancer*,106:1107-16,2012.
[3] 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.
[4] M Garcia et al., Interactome-Transcriptome integration for predicting distant metastasis in breast cancer. *Bioinformatics*, 8:672-678, 2012.
[5] T.S.K. Prasad et al., Human protein reference database-2009 update. *Nucleic Acids Res.*, 37:767-772, 2009.
[6] 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.
[7] A. Ceol et al., Mint, the molecular interaction database: 2009 update. *Nucleic Acids Res.*, 38:532-539, 2010.
[8] B. Aranda et al., The intact molecular interaction database in 2010. *Nucleic Acids Res.*, 38:525-531, 2010.
[9] L. Salwinski et al., The database of interacting proteins: 2004 update. *Nucleic Acids Res.*, 32:449-451, 2004.
[10] 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.
[11] S. Loi et al., Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen. *BMC Genomics*, 9:239, 2008.
[12] R. Sabatier et al., A gene expression signature identifies two prognostic subgroups of basal breast cancer. *Breast Cancer Res Treat*, 2010.
[13] 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.
[14] Van De Vijver et al., A gene-expression signature as a predictor of survival in breast cancer. *NEJM*, 25:1999-2009, 2002.
[15] Y. Wang et al., Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, 365(9460):671-679, 2005.
[16] Van 't Veer et al., Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530-6, 2002.

Influence of joint preprocessing of CGH data on downstream analysis

Émilie SOHIER[1], Bastien JOB[2], Natacha ENZ-WERLÉ[3], Nathalie GASPAR[1,4], Laurence BRUGIÈRES[1]and Cathy PHILIPPE[4]

[1]PEDIATRIE, INSTITUT G. ROUSSY, 114 rue Édouard Vaillant,94805, Villejuif
emilie.sohier@igr.fr

[2] PLATEFORME DE BIOINFORMATIQUE, INSTITUT G. ROUSSY, 114 rue Édouard Vaillant 94805 Villejuif

[3]PEDIATRIE ONCO-HEMATOLOGIE,CHRU HAUTPIERRE, avenue Molière, 67200, Strasbourg

[4] UMR8203 CNRS, INSTITUT G. ROUSSY, 114 rue Édouard Vaillant 94805 Villejuif

6th Release of HOGENOM, a Database of Homologous Genes in Complete Genomes

Simon Penel [1], Pascal Calvat[2], Vincent Daubin[1], Manolo Gouy[1], Vincent Miele[1], Guy Perrière[1], Rémi Planel[1], Laurent Duret[1]

[1] Laboratoire de Biométrie et Biologie Evolutive, CNRS UMR 5558, UCBL, Lyon I
43, Bd du 11 novembre 1918, 69622 Villeurbanne cedex -France
{simon.penel,manolo.gouy,guy.perriere,vincent.miele,remi.planel,
laurent.duret}@.univ-lyon1.fr

[2] Centre de Calcul de l'Institut National de Physique Nucléaire et de Physique des Particules - CNRS
27 Bd du 11 novembre 1918, 69622 Villeurbanne cedex -France
calvat@cc.in2p3.fr

Abstract: *Comparison of homologous genes in complete genomes is an essential step for studies related to molecular evolution. In this view we developed HOGENOM, a database of homologous gene families from completely sequenced organisms covering a wide range of phylogenetic distances. The 6th release (January 2012) of this database contains families of homologous genes from about 1500 fully sequenced bacteria, archaea and eukaryota.*

Keywords: *Genomics, Databases, Evolution, Comparative phylogenomics*

1 Introduction

HOGENOM (http://pbil.univ-lyon1.fr/databases/hogenom/) is composed of two databases, the first one containing the protein sequences in SWISSPROT format, and the second one containing the genome sequences in EMBL format. The databases are structured under ACNUC [1]. The first step of the construction is to harvest complete genome annotated data, from which a nucleic database is built. The CDS are then translated into protein and annotated proteome sets are generated including the information available for each CDS. A protein database is built, and its protein sequences are clustered into families as previously described [2]. Nucleic and protein databases are re-annotated with the family information. Finally alignments (CLUSTALOMEGA [3]) and phylogenetic trees (PHYML [4], FASTTREE [5]) are calculated.

2 Data Origin and Contents

HOGENOM provides families of homologous genes from various organisms covering a wide range of phylogenetics distances. The nucleotide database is thus built from several sources :

1) the NCBI Bacterial Genomes database (BG), 2) the EBI Genome Reviews database (GR) which provides an annotated view of completely deciphered genomes of bacteria, archaea and few eukaryota, for organisms not included into BG, 3) the Ensembl database for animal genomes, 4) the Fungi Ensembl, Plants Ensembl, Protists Ensembl and Metazoa Ensembl databases for Fungi, Plants, Protists and Metazoa respectively, 5) the NCBI Protozoan

Genomes database, 6) the EBI complete genome data for eukaryota which are not included into the latest databases and finally 7) NCBI, JGI, Sangers and TIGEr data for other peculiar genomes. All together, complete genomes from 1,233 bacteria, 97 archaea and 140 eukaryota are present in HOGENOM.

HOGENOM contains 7,278,659 proteins (and associated CDS) among which 85% are classified in 296,917 families, the other being orphans or unclassified.

3 Clustering of sequences

To build the families we perform a similarity search of all the proteins against each other with BLASTP2. Then, the results are processed with the SiLiX [6] software then-post processed with the HiFiX [7] software.

4 Usage and access

HOGENOM families can be retrieved and selected via the PBIL web server (http://pbil.univ-lyon1.fr), via the FamFetch (http://pbil.univ-lyon1.fr/software/famfetch.html) java tool, via an ACNUC client-server application or via the seqinR [8] package.

5 What's new in HOGENOM6

HOGENOM6 contains still more genomes including all Ensembl genomes, brand new clustering methods have been used avoiding sequence accumulation in clusters, and all isoforms from alternative splicing are now available.

References

[1] M. Gouy, C. Gautier, M. Attimonelli, C. Lanave, and G. di Paola, ACNUC--a portable retrieval system for nucleic acid sequence databases: logical and physical designs and usage. *Bioinformatics*, **vol. 1** pp. 167-172, 1985.

[2] S. Penel, A-M. Arigon, J-F. Dufayard, A-S. Sertier, V. Daubin, L. Duret, M. Gouy, and G. Perriere, Databases of homologous gene families for comparative genomics. *BMC Bioinformatics*, **vol. 10** pp. 1-13, 2009.

[3] F. Sievers, A. Wilm, DG. Dineen, TJ. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, JD. Thompson, DG. Higgins, Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7**, 2011.

[4] S. Guindon and O. Gascuel, PHYML: "A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood." *Systematic Biology*. **52**(5):696-704, 2003.

[5] MN. Price, PS. Dehal and AP. Arkin, FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE* **5**(3): e9490. doi:10.1371/journal.pone.0009490, 2010.

[6] V. Miele, S. Penel and L. Duret, Ultra-fast sequence clustering from similarity neytworks with SiLiX, *BMC Bioinformatics* **12**:116, 2011.

[7] V. Miele, S. Penel, V. Daubin, F. Picard D. Kahn and L. Duret. High-quality sequence clustering guided by network topology and multiple alignment likelihood, *Bioinformatics* (in press)

[8] D. Charif, J.Thioulouse, JR. Lobry and G. Perriere, Structural approaches to sequence evolution: *Molecules, networks, populations, Springer Verlag*, pp. 207-232, 2007

EvryRNA : bioinformatics platform for non-coding RNA

Fariza TAHI, Mederich BESNARD, Gabriel CHANDESRIS, Sébastien TEMPEL and Mikael TRELLET

IBISC - IBGBI, 23, Bd de France, 91034 EVRY, France
{fariza.tahi,sebastien.tempel}@ibisc.univ-evry.fr

Abstract *Non coding RNA structures play important roles in biological processes. Predicting their structure and finding them in genomes is therefore an important task, and computational methods are required. We developed a bioinformatics platform, called EvryRNA, that presents tools for the prediction, the search and the analysis of non-coding RNAs. This platform is available at the address: http://EvryRNA.ibisc.univ-evry.fr/.*

Keywords non-coding RNAs, micro RNAs, RNA structure prediction, RNA search, RNA analysis, webserver

Functional non-coding RNAs are often involved as regulators of gene expression at the post-transcriptional level. Because their detection by experimental techniques is difficult and expensive and requires a large amount of time, computational methods represent the first step in non-coding RNA identification and analysis. In our laboratory, we developed several methods and software for structure prediction, identification and analysis of non coding RNAs. To make them available for the community of biologists and bioinformaticians, we developed a website, called EvryRNA (Fig. 1) (http://EvryRNA.ibisc.univ-evry.fr/), allowing the use of these different software.

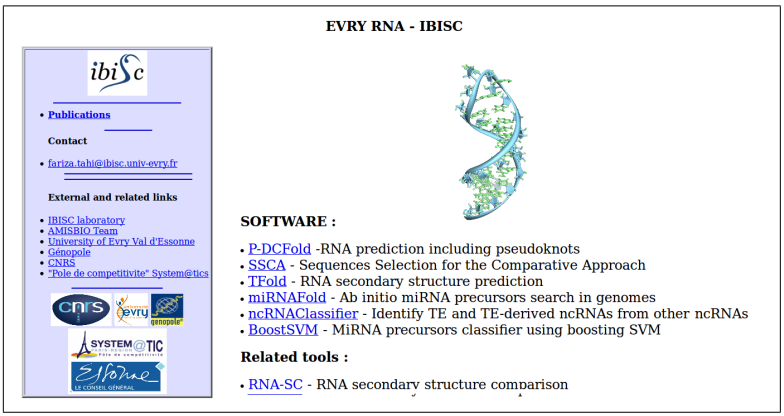


Figure 1. Homepage of the EvryRNA bioinformatics platform

On EvryRNA webserver, users can find:

- P-DCFold: software for predicting RNA secondary structures including pseudoknots [1].
- SSCA: software for selecting homologous sequences for secondary structure prediction by the comparative approach [2].
- TFold: software for predicting RNA secondary structures combining SSCA and P-DCFold [3].
- miRNAFold: software for searching for micro RNAs in genomes [4].
- BoostSVM: software for classifying pseudo and true miRNAs [6]
- ncRNAclassifier: software for classifying non coding RNA sequences based on their relationship with transposable element sequences [5].

P-DCFold predicts secondary structure of a given non coding RNA sequence, including pseudo-knots, using a set of homologous sequences [1]. Based on the comparative approach, it takes as input an alignment of RNA sequences in Fasta format. Because the prediction results depend on the homologous sequences used, we developed the algorithm SSCA (Sequence Selection for the Comparative Approach) for selecting from a set of sequences a subset of sequences that are the most appropriate to be used for the prediction of the secondary structure of a given RNA sequence [2]. SSCA was then efficiently integrated to P-DCFold in order to improve the prediction results of P-DCFold and especially to have an algorithm that is not sensitive to the homologous sequences used. The resulting algorithm is Tfold [3]. Tfold takes as input an alignment in Fasta where the user chooses the sequence for which the secondary structure prediction is required and puts optionally the known stems. Tfold returns then the stems of the predicted structure (giving several information about the stems, like their length, their position, their conservation, etc.) in different formats such as bracket format.

miRNAFold and BoostSVM are two ab initio microRNA prediction algorithms. miRNAFold is a fast algorithm for searching for microRNA precursors (pre-miRNA) in genomic sequences [4]. It is based on an original method where an approximation of miRNA hairpins are first searched, before reconstituting the pre-miRNA structure. The approximation step allows a substantial decrease in the number of possibilities and thus the time required for searching. BoostSVM is a machine learning algorithm allowing to classify pre-miRNA candidates as true or false pre-miRNAs [6]. It uses a boosting technique with weakened SVM component classifiers to deal with the imbalance in the training data. In case of miRNAFold, the user enters a genomic sequence in a Fasta format and the software returns the set of pre-miRNA candidates predicted in the sequence. Several parameters are set with default values but the user can change these values. In case of BoostSVM, the user enters one or several sequences corresponding to pre-miRNA candidates and the software classify each of these candidates as true or false pre-miRNA.

Finally, ncRNAclassifier [5] is an automatic method for classifying miRNA precursors into three categories, based on the percentage of TEs in their sequence and their dispersion in the genome: (i) precursors whose sequence is devoided of TE-derived sequences and not repeated nor dispersed to a significant extent in the genome: bona fide pre-miRNAs (or miRNA genes); (ii) precursors whose sequence corresponds to a small part of a known TE sequence and/or that are repeated and dispersed in the genome: TE derived miRNAs; and (iii) precursors whose sequence corresponds to a large part of a known TE sequence, either already annotated as such or identified by our method: mis-annotated miRNAs.

EvryRNA server provides also tools that we developed as needed for the project, e.g. RNA-SC for comparing a predicted secondary structure with a secondary structure of reference. This tool returns selectivity and sensitivity rates of the predicted structure when compared to the reference structure.

Acknowledgements

This work was funded by the Council of Essonne Region (Pôle System@tic, OpenGPU project). We would like to thank the underground students Thomas Letellier, Sabine Menigaud and Frédéric Merle, for their contribution in the improvement of EvryRNA webserver.

References

[1] F. Tahi, S. Engelen and M. Régnier. P-DCFold or How to Predict all Kinds of Pseudoknots in Rna Secondary Structures. *International Journal on Artificial Intelligence Tools*, 14(5):703-716, 2005.

[2] S. Engelen and F. Tahi, Predicting RNA secondary structure by the comparative approach: how to select the homologous sequences. *BMC Bioinformatics*, 8:464, 2007.

[3] S. Engelen and F. Tahi, Tfold: efficient in silico prediction of non-coding RNA secondary structures. *Nucleic Acids Res.*, 38:2453-66, 2010.

[4] S. Tempel and F. Tahi, A fast ab-initio method for predicting miRNA precursors in genomes. *Nucleic Acids Res.*, doi: 10.1093/nar/gks146, 2012.

[5] S. Tempel and F. Tahi. An automatic method for identifying TE-derived miRNAs. *JOBIM*, pages 245-252, 2011.

[6] Van Du Tran, S. Tempel, B. Zerath, F. Zehraoui and F. Tahi. BoostSVM: A miRNA classifier with high accuracy using boosting SVM. *JOBIM*, 2012.

SUMATRA : Fast and Exact Computation of Sequence Similarities

Céline MERCIER[1], Frédéric BOYER[1], Lucie ZINGER[1] and Eric COISSAC[1]
[1] LABORATOIRE, UMR5553 CNRS, LECA BP 53, 2233 Rue de la Piscine, 38041 Grenoble, Cedex 9, France
{celine.mercier, frederic.boyer, lucie.zinger, eric.coissac}@ujf-grenoble.fr

Keywords SIMD, multi-threading, lossless filter, sequence similarity.

SUMATRA : Calcul Rapide et Exact de Distances entre Séquences

Mots-clés Distance entre séquences, SIMD, multi-threading, filtrage sans perte.

Avec le développement important des techniques de séquençage à haut débit ces dix dernières années, le « *DNA barcoding* » est devenu applicable à de nombreuses études en taxonomie et en écologie [1]. Le principe du *DNA barcoding* repose sur l'utilisation d'un fragment d'ADN court, appelé marqueur ou code barre (barcode), pour assigner un organisme à un taxon. Par extension, le *metabarcoding* est une technique cherchant à identifier plusieurs taxons à partir d'un très grand nombre de séquences (plusieurs dizaines de milliers) issues de séquençage haut débit de l'ADN contenu dans des échantillons environnementaux [2].

Pour assigner ces séquences à des taxons, il est possible de calculer leurs distances avec des séquences d'une base de références et de les assigner au taxon de la séquence ayant la plus grande similarité. Une approche alternative, utilisée lorsqu'une base de séquences de référence n'existe pas, consiste à regrouper les séquences partageant une très forte similarité en groupes correspondant à des « unités taxonomiques » (MOTUs, Molecular Operational Taxonomy Units) [3]. Dans ce dernier cas, la stratégie communément utilisée consiste à ne considérer que les similarités dépassant un certain score puis à regrouper les séquences grâce à une méthode de classification.

Quelque soit la stratégie utilisée, il est nécessaire de calculer un très grand nombre de scores d'alignement entre ces séquences (banque *vs* référence ou banque *vs* banque). *Sumatra* a été développé dans l'optique de répondre à ce besoin en implémentant un algorithme à la fois exact et rapide pour calculer les scores LCS (Longest Common Subsequence) entre deux banques de séquences.

Sumatra permet à l'utilisateur de spécifier un score minimum d'alignement au dessous duquel les similarités de séquences ne doivent pas être reportées. Lorsque ce paramètre est spécifié, une procédure de filtrage *sans perte* est utilisée afin de n'aligner que les séquences ayant potentiellement une similarité supérieure à ce seuil. Ce filtre se base sur le nombre de mots communs chevauchants que les deux séquences doivent partager. Plus le seuil choisi est haut, plus l'utilisation du filtrage diminue le temps d'exécution du programme. En travaillant sur des jeux de séquences issus de metabarcoding, et donc d'une longueur moyenne comprise entre 50 et 100 paires de bases, nous avons observé que le filtrage le plus efficace est obtenu en choisissant une longueur de 4 nucléotides pour les mots conservés.

Pour illustrer les performances du filtrage implémenté dans *Sumatra*, la Figure 1A présente l'évolution du nombre d'alignements calculés en fonction du seuil d'identité minimum requis sur une banque de 146865 séquences du marqueur *trn*L-*gh* [4]. On observe qu'utiliser le filtrage réduit drastiquement le nombre de séquences alignées, et donc le temps d'exécution du programme, à partir d'un seuil d'identité minimum de 85%. On observe aussi que la proportion de couples de séquences possédant effectivement une identité supérieure au seuil choisi est 2 à 8 fois inférieure à l'estimation faite par le filtrage. Une amélioration de cette étape ou sa superposition avec une autre méthode de filtrage pourraient donc être intéressante.

Deux niveaux de parallélisation différents permettent à *Sumatra* de réduire les temps de calcul. Lorsqu'un couple de séquences passe le filtrage, le score LCS est calculé par programmation dynamique. Ces deux calculs sont optimisés grâce à l'utilisation d'instructions SIMD (Simple Instruction Multiple Data).

De plus, *Sumatra* peut s'exécuter sur plusieurs *threads* pour paralléliser les données. La Figure 1B présente l'influence de la parallélisation sur les performances du programme ; ces tests ont été effectués sur une machine équipée de 16 cores compatibles avec le jeu d'instructions SIMD SSE2.

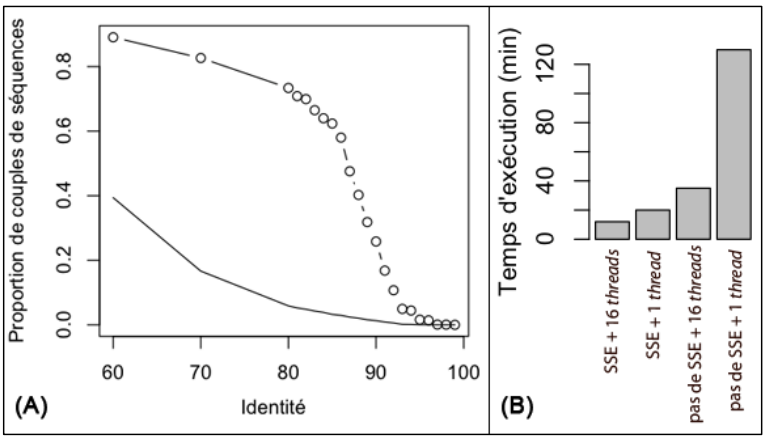


Figure 1. (A) La ligne avec les points représente la proportion de couples de séquences alignés en fonction du seuil d'identité choisi pour le filtrage. La ligne simple représente la proportion de couples de séquence dépassant effectivement ce seuil. **(B)** Temps requis en minutes pour l'exécution de *Sumatra* avec un filtrage à 95% d'identité, selon les parallélisations utilisées : instructions SSE ou non, et 1 ou 16 *threads*.

Sumatra rend en sortie un fichier texte tabulaire contenant les identifiants de séquences et les scores des similarités calculées, ce qui rend son intégration aisée au sein de la plupart des stratégies d'identification taxonomique ou de construction de MOTUs.

References

[1] P.D.N. Hebert, S. Ratnasingham and J.R. de Waard, Barcoding animal life : cytochrome C oxidase subunit 1 divergences among closely related species. *Proc. R. Soc. B*, 270: 596-599, 2003.

[2] A. Valentini, F. Pompanon and P. Taberlet, DNA barcoding for ecologists. *Trends Ecol. Evol.*, 24, 110-117, 2008.

[3] M. Blaxter, J. Mann, T. Chapman, F. Thomas, C. Whitton, R. Floyd and Eyualem-Abebe, Defining operational taxonomic units using DNA barcode data. *Philos. Trans. R Soc. Lond B Biol. Sci.*, 360, 1935–1943, 2005.

[4] P. Taberlet, E. Coissac, F. Pompanon, L. Gielly, C. Miquel, A. Valentini, T. Vermat, G. Corthier, C. Brochmann and E. Willerslev, Power and limitations of the chloroplast *trn*L (UAA) intron for plant DNA barcoding. *Nucleic Acids Res.*, 35(3): e14, 2007.

Towards improving docking-scoring functions in structure-based virtual screening

Lory MONTOUT[1], Michel PETITJEAN[1], Gautier MOROY[1], Leslie REGAD[1] and Anne-Claude CAMPROUX[1]

[1] Laboratoire MTi, UMR-S 973, Université Paris Diderot, 35, Rue Hélène Brion, 75205, Paris, Cedex 13, France
{lory.montout, michel.petitjean, gautier.moroy, leslie.regad, anne-claude.camproux}@univ-paris-diderot.fr

Keywords binding-sites, protein-ligand complexes, docking/scoring functions, multivariate data analysis.

Vers l'amélioration des fonctions de score dans le criblage virtuel basé sur la structure

Mots-clés Sites de liaison, complexes protéine-ligand, fonctions de docking/scoring, analyse de données multivariées.

1 Introduction

Structure-based virtual screening is now widely applied in early-stage of the drug discovery process and scoring functions are essential for analyzing the outputs of molecular docking. These scoring functions are often capable of identifying the correct binding pose of a ligand but have poor performance in estimating binding affinity and hence in ranking docked ligands. It is recognized that compounds selection based on calculated scores is not always sufficient [1]. Important challenges in the drug discovery process are 1) designing compound libraries more specific to one target before docking or 2) improving pose/compound selection after docking. This study focuses on the first point.

2 Dataset and methods

In an attempt to better understand the performance of scoring methods regarding the type of pocket-ligand pairs studied, we consider 483 proteins in complexes with drug-like ligands extract from the Astex set [2] and the PDBbind database [3] and perform a cross-docking analysis using Autodock vina [4]. The resulting predicted poses are re-scored using other scoring methods including Autodock4 score function [5] and X-score [6]. The performances of the scoring methods are then compared in terms of ligand or pocket similarity classification using a set of several properties developed in our laboratory [7]. These geometric and physico-chemical properties describe pockets, ligands and the protein-ligand pairs; they are relevant for the characterization of pocket-ligand complementarity. Besides classical descriptors such as volume, surface area, hydrophobicity, the amino acid composition or the polarity we also take into account information as rugosity or narrowness of the pocket [8-9]. We then use the DUD dataset [10] as a validation dataset to perform docking enrichment by using our descriptors as filters.

3 Results and discussion

Currently, new structure-based virtual screening approaches consisting for instance in using molecular interaction fingerprint to refine docking poses exist [11]. However we believe that explicit descriptors would be best suited than binary information for integration in docking functions. Our first results show that docking scores are strongly dependent of ligand properties; suggesting that taking into account topology and polarity information on pockets would improve docking/scoring methods. This study is a preliminary work to adapt docking/scoring methods in order to enhance docking results. Here we focus on designing

compound libraries more specific to targets extract from the DUD database resulting in enrichment analysis. Ultimately our goal is to find a way to integrate such descriptors directly into existing scoring functions as it as been proposed in the AutoShim method with pharmacophore points as descriptors incorporated in a scoring method [12].

References

[1] Cheng, T., Li, Q., Zhou, Z., Wang, Y., & Bryant, S. H., Structure-Based Virtual Screening for Drug Discovery: a Problem-Centric Review. *The AAPS journal*, 14(1), 2012.

[2] Hartshorn, M.J. et al., Diverse, high-quality test set for the validation of protein-ligand docking performance. *J. Med. Chem.*, 50(4):726-741, 2007.

[3] Wang, R. et al., The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem.*, 47(12):2977-2980, 2003.

[4] Trott, O. ,Olson, A. J., AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *J. Comp. Chem.*, 31:455-461, 2010.

[5] Huey, R., Morris, G. M., Olson, A. J. and Goodsell, D. S., A Semiempirical Free Energy Force Field with Charge-Based Desolvation. *J. Comp. Chem.*, 28:1145-1152, 2007.

[6] Wang, R.; Lai, L.; Wang, S., Further Development and Validation of Empirical Scoring Functions for Structure-Based Binding Affinity Prediction. *J. Comp.-Aided Mol. Des.*, 16:11-26, 2002.

[7] Pérot, S., Sperandio, O., Miteva, M., Camproux A.-C. and Villoutreix B., Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery. *Drug Discovery Today* 15(15-16):656-667, 2010.

[8] Pettit, F.K. and Bowie, J.U., Protein surface roughness and small molecular binding sites. *J. Mol. Biol.,* 285(4):1377–1382, 1999.

[9] Sugaya, N. and Ikeda, K., Assessing the druggability of protein-protein interactions by a supervised machine-learning method. *BMC Bioinformatics,* 10:263, 2009.

[10] Huang, N., Shoichet, B. K., & Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.*, 49(23), 6789-801, 2006.

[11] Marcou G., Rognan D., Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *J. Chem. Inf. Model.*, 47(1):195–207, 2007.

[12] Martin E.J., Sullivan D.C., AutoShim: empirically corrected scoring functions for quantitative docking with a crystal structure and IC50 training data. *J. Chem. Inf. Model.*, 48(4):861–872, 2008.

MetaMatch: Accurate and reliable algorithms for taxonomic assessment at species level.

Lenaïg KERMARREC[1,2], Philippe CHAUMEIL [3,4], Frédéric RIMET [1], Jean-Marc FRIGERIO [3,4], Agnès BOUCHEZ [1], and Alain FRANC [3,4]

[1] INRA, UMR CARRTEL, 74200 Thonon-les-Bains,
{Frederic.Rimet, Agnes.Bouchez}@thonon.inra.fr

[2] Asconit Consultants, 66350 Toulouges, France,

lenaig.kermarrec@asconit.com

[3] INRA, UMR BioGeCo, 33610 Cestas, France

[4] University of Bordeaux 1, UMR BioGeCo, 33400 Talence, France.

{Philippe.Chaumeil, Jean-Marc.Frigerio, Alain.Franc}@pierroton.inra.fr

Abstract *Accurate identification of the species in a community is a crucial step for biodiversity studies. Community ecology faces a new challenge as the next-generation sequencing approaches can yield data from hundreds of microbial community samples. This way, combined with accurate and reliable taxonomic assessment, yields hundreds of new data that will contribute to a better understanding of community assemblies formed under various environmental and historical conditions. Fast, reliable and accurate taxonomic assessment at species level, the building block of community ecology, from molecular data produced by NGS remains an open challenge. We present here a tool, called metaMatch, which aims at producing automatically a species inventory of a community in the context of metagenomics, from comparison of reads with an annotated reference library. The outcome of this tool is compared with morphological based taxonomy and with BLAST. Accuracy is better than with BLAST. A case study with diatoms in the context of biomonitoring (quality of freshwaters) is presented, including a benchmark on known communities. Results permit to envisage molecular based community inventories with 454 community pyrosequencing techniques and bioinformatics.*

Keywords molecular taxonomy, algorithm, NGS sequencing, HPC

The objective of this work is to produce an algorithm with a community inventory as an outcome, and an output of pyrosequencing and annotated reference library as inputs. Algorithms classifying sequences by comparison to a reference library are the most widely used tools for assessing community composition of environmental samples. We present here a tool, *metaMatch*, that can assess a sample with the best possible quality. It comprises two version, (i) perfect where perfect match between a read and a reference sequence is sought only and (ii) imperfect, where distances due to SNPs or short indels between a reference and a query are permitted. *MetaMatch*, both versions perfect and imperfect, permits algorithmic community taxonomic inventories with sets of reads from a mock community (perfect) or an environmental sample (imperfect) and a reference database as inputs. First, the quality of the tool has been assessed by matching reads from DNA of mock communities of diatoms, with perfect match (using a diatoms strain collection maintained at Thonon), with one answer `TRUE` or `FALSE`. Perfect match of a word within a longer one is a problem which has been addressed for decades, and very efficient and reliable algorithms are available.

We tested *metaMatch* by analysing three mock communities of diatoms, the exact composition of which was actually known, as the strains artificially assembled in a sample had been taken from a strain collection,

with known morphological and genetic characters. We observed that inventories are dependant of the algorithm used to assign species names to reads. Perfect *metaMatch* yields an inventory with a speed comparable to BLAST on a single CPU, and significantly better accuracy (fewer false positives). The way we suggest to use it is a first step with heuristic algorithms that can initially be run on a large reference database to determine the high taxonomic levels and as a second step *metaMatch* which can be used to obtain an expert taxonomic inventory within a particular candidate group.

However, perfect *metaMatch* may miss taxa when an environmental sample is analyzed, as within species genetic differences are very likely to exist between reference sequences and sequences from environmental samples. Imperfect *metaMatch* has been designed therefore, i.e. computing the edition or Levenstein distance between a query and a reference database on local alignment.

Several dynamic programming algorithms are available for matching a sequence against another: Needleman-Wusnch for global alignment and Smith-Waterman for local alignment being seminal approaches widely ameliorated afterwards namely for speed (both are indeed demanding in time). Here, we focus on local alignment, combining the two existing approaches for sequence alignment: search for anchors, i.e. perfectly conserved homologous regions, and dynamic programming in non homologous regions. Imperfect metaMatch is time demanding (roughly speaking, it is of order $O(rq)$ for one comparison if r is reference length and q query length. If we have to study Q query against R references, time needed is of order $QRqr$. However, this procedure can easily be distributed onto a large number of CPU, simply by splitting the query data set on as many sets as CPU, and send each set onto one CPU. The complete result with about 250 000 query reads tested against a 2 000 specimen reference data base has been obtained within less than 30 minutes on cluster “Avakas” (264 nodes of 2 hexacore units at 3.06 GHz, with 48 Go RAM each, or 3168 cores altogether) at Mésocentre de Calcul Intensif Aquitain (Bordeaux). This routine has been successfully sent on grid EGI as well, which accepts massively distributed jobs.

This procedure has been implemented on a data set part of a biomonitoring program. Diatoms have been selected as they are known to be excellent bioindicators, and diatom community inventories are routinely assessed in that objective. Our study tested the reliability of the 454 pyrosequencing to determine diatom inventories in environmental samples. Three markers were used: the SSU rDNA, the *rbc*L and the *cox*1. First, we studied reference libraries of the three markers to help defining thresholds between intraspecific and interspecific, and between intrageneric and intergeneric genetic distances. Thresholds were tested on mock communities (artificially built) and applied to four natural samples (one in duplicate) to assign taxa names to environmental sequences. For each sample, inventories obtained from pyrosequencing with *metaMatch* have been compared to the floristic list obtained by microscopy approach. Divergences were observed between morphological and molecular inventories. However some limits linked to the molecular method could be overcome extending our DNA reference libraries and others are not more important than the bias of the current method based on the morphology. In conclusion, 454 pyrosequencing could be used in biomonitoring programs after some optimizations.

The causal mediation analysis in genomic data

Going beyond simple correlations

Séverine AFFELDT[1], Param-Priya SINGH[1], Giulia MALAGUTI[1] and Hervé ISAMBERT[1]

UMR168 CNRS-UPMC, Institut Curie, 26, rue d'Ulm, 75248 Paris, France

{severine.affeldt, param-priya.singh, giulia.malaguti, herve.isambert}@curie.fr

Analysis of the structural conservation of β-strand irregularities: the β-bulges

Pierrick CRAVEUR[1], Agnel Praveen JOSEPH[2], Joseph REBEHMED[1] and Alexandre G. DE BREVERN[1]

[1] INSERM, UMR_S665, Dynamique des Structures et des Macromolécules Biologiques (DSIMB), Université Paris Diderot, Sorbonne Paris Cité, INTS, 6, rue Alexandre Cabanel, 75739, Paris, Cedex 15, France
pierrick.craveur@inserm.fr
{joseph.rebehmed, alexandre.debrevern}@univ-paris-diderot.fr

[2] NCBS, Tata institute of Fundamental Research, UAS, GKVK Campus, Bellary Road, Bangalore 560 065, India
agnelpj@ncbs.res.in

Keywords protein structures, protein folds, structural superimposition, structural alphabet.

1 Introduction

Protein structures are usually described as a succession of coil, repetitive secondary structures, *e.g.*, α-helix and β-strand, turns, less frequent repetitive structures, *e.g.*, Polyproline II helix, isolated E-strand, 3.10- and π-helix. Irregularities in these repetitive structures have been observed and directly associated to protein function [1]. β-sheet's irregularities are named β-bulge; they are defined as a region between two consecutive β-type hydrogen bonds, which include two to four residues on one strand opposite to a single residue (called residue X) on the other-strand [2]. They were classified in five different types [3]: classic, G1, wide, bent and special. This irregular motif, which formation remains poorly understood, produces two main changes in the structure of a β-sheet: (1) disrupts the classical alternation of side chain direction, (2) impacts the directionality of β-strands and accentuates the typical right-handed twist of β-sheets [3]. Several observations suggested their influence on deletions and insertions [4], and their implication in protein-protein interactions. They were, in particular, described as negative design to avoid β-strand aggregation into fibrillar structures in the case of several neurodegenerative pathologies [5].

2 Materials and Methods

Protein domains have been taken from a subset of the ASTRAL database version 1.75 (derived from SCOP [6]) with less than 95% identity [7]. PROMOTIF software version 3.01 [8] was used for the assignment of β-bulges and iPBA software v1.0 [9] to perform protein structural superimpositions.

12,436 protein structures were used to study the β-bulges structural conservation.

3 Results

β-bulges were already described but never on a large dataset (362 β-bulges for 170 proteins in 1993 [3]).

We observed near 21,477 β-bulges in our dataset. More than 90% are found on structure which are smaller than 400 residues and have less than 20 β-strands. We observed more than two β-bulges on average per proteins, and more than three concerning the all-β SCOP folds. Nonetheless, some interesting proteins, which have near 200 residues length, could have more than 15 β-bulges.

Thus we analyzed their amino-acid composition highlighting novel amino acid preferences in regards to the previous studies [3]. This amino-acid preferences depends on the β-bulge type, residues type (residue X or not) and β-sheet/coil localization. For example the position X of antiparrallel classic (AC type) β-bulges prefers residues V or W, while it prefers to be a C, D or N for antiparrallel G1 (AG type) β-bulges.

Using Protein Blocks [10], *i.e.*, local structure preferences, and β-sheet/coil localization, β-bulges were studied and the preferred positions at the N- and C-terminus of β-strand were highlighted.

The structural conservation of β-bulges was evaluated using the superimposition with GDT_TS [10] ≥ 15; beyond this value, the both superposed structures share the same global shape.

Through the around 790,000 structural superimpositions computed, we observed on average for each β-bulge, 33% of superimpositions composed by β-bulges which are at the exact same position (full superimpositions), 12% representing β-bulges which have slightly moved between equivalent fragments

(partial superimpositions) and 55% of no superimpositions. Thereby we observed on average, for each β-bulge, 10.2% of superimpositions with change of type (1.4% on full superimpositions and 8.8% on partial superimpositions).

As expected, the structural conservation of β-bulge is better for a higher GDT_TS score and for a higher sequence identity. Results are summarized in Table 1.

		Full superimposition	Partial superimposition	No superimposition	number of β-bul
ıuence	≤ 35 %	27 %	13 %	60 %	19,950
entity	> 35%	69 %	8 %	23 %	12,965
)T_TS	[15,40] (mean RMSD = 2.40 Å)	17 %	14 %	69 %	18,130
	[40,100] (mean RMSD = 1.42 Å)	55 %	9 %	36 %	18,610

Table 1. Average percentage superimposition for each β-Bulge based on the sequence identity and GDT_TS score

Our results show that conservation of β-bulges in structural homologues is not a dominant characteristic. Indeed there are a not negligible proportion of no superimpositions, even in case of very similar folds. However the structural conservation of β-bulges increases as protein structures are close in terms of sequence identity, or structural homology. Interestingly, taking structures from twilight zone (sharing less than 35% sequence identity) with similar fold (GDT_TS values in range [40,100]), we noticed close probabilities for each β-bulges to be superposed (47% and 10% respectively for full and partial superimpositions) than not superposed (43%).

Studying the β-bulges conservation by performing molecular dynamics simulations may explain the stable and instable characteristics of β-bulges.

Acknowledgements

This work was supported by University Paris Diderot, Sorbone Paris Cité, National Institute for Blood Transfusion (INTS), Institute for Health and Medical Research (INSERM) to PC, JR and AdB. PC acknowledges grant from Ministry of Research. JR acknowledges grants from ANR and APJ to HFSP program.

References

[1] Offmann B., Tyagi M., de Brevern A.G. Local Protein Structures. Current Bioinformatics, 2: 165-202, 2007.

[2] Richardson, J. S., Getzoff, E. D. & Richardson, D. C. The beta bulge: a common small unit of nonrepetitive protein structure. *Proc Natl Acad Sci U S A,* 75, 2574–2578, 1978.

[3] Chan, A. W., Hutchinson, E. G., Harris, D. & Thornton, J. M. Identification, classification, and analysis of beta-bulges in proteins. *Protein Sci.* 2, 1574–1590, 1993.

[4] D Shortle and J Sondek The emerging role of insertions and deletions in protein engineering. *Current Opinion in Biotechnology,* 6, no. 4: 387-393, 1995.

[5] Richardson, J. S. & Richardson, D. C. Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc. Natl. Acad. Sci. U.S.A.,* 99, 2754–2759, 2002.

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

[7] Chandonia JM, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE. The ASTRAL compendium in 2004. *Nucleic Acid Res., 32:*D189-D192, 2004.

[8] E. G. Hutchinson & J. M. Thornton PROMOTIF - A program to identify structural motifs in proteins. *Protein Science,* 5:212-220, 1996.

[9] J.-C. Gelly, Joseph A.P., Srinivasan N., de Brevern A.G. iPBA: A tool for protein structure comparison using sequence alignment strategies. *Nucleic Acid Res.,* 39:W18-23, 2011.

[10] Joseph, A. P. *et al.* A short survey on protein blocks. *Biophys Rev.,* 2, 137–147, 2010.

[11] Zemla, A. LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res.*,31,3370-3374,2003.

Gene networks classification based on topologicals properties and applications

Ian MISNER[1] and Cedric BICEP[2]

[1] Department of Biological Sciences, University of Rhode Island, 120 Flagg Road, RI 02881, Kingston
ianmisner@my.uri.edu

[2] Systématique Adaptation et Evolution, UMR CNRS 7138 UPMC, 7 Quai St-Bernard, 75005, Paris, France
cedric.bicep@snv.jussieu.fr

Frequent Subgraph Summarization by Substitution Matrices

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

[1] LIMOS, Blaise Pascal University - Clermont University, BP 10448, 6300 Clermont-Ferrand, France
[2] LIMOS - CNRS UMR 6158, Aubière 63173, France.
{dhifli, saidi, mephu}@isima.fr

Abstract *We propose a novel approach to select representative subgraph-motifs from a set of protein substructures by extending the DDSM method dedicated to sequences. Our method, termed F3SM is based on the evolutionary information of amino acids defined in the substitution matrices. The results issued from our experiments show that our approach decreases dramatically the number of motifs while enhancing their interestingness.*

Keywords F3SM, subgraph-motif selection, protein structures, classification, substitution matrix.

1 Introduction

Exploring protein structures may reveal relevant structural and functional information that protein sequences fail to expose. In this scope, proteins have been recently seen as graphs of amino acids and studied based on graph theory concepts. Indeed, algorithms of frequent subgraph discovery have been applied on protein structures to find motifs that could be interesting in any further analysis. However, when the support threshold is low, the number of frequent subgraphs is expected to be very large which may hinder rather than help. We claim that in the set discovered subgraph-motifs, there exist a subset of representative subgraph-motifs that can substitute several others and hence can summarize the whole set. This claim is based on the biological fact that some amino acids have similar properties and can thus be substituted by each other, without changing the structure or the function of the protein. This phenomenon is quantified by score matrices called *substitution matrices* [4]. In this paper, we propose the frequent subgraphs summarization using substitution matrices *F3SM*. For our purpose, we adapt an existing method termed *DDSM* dedicated to protein sequences [4]. We address the limitations of DDSM and we extend it to deal with protein structures in a graph perspective.

2 Contribution: F3SM

DDSM performs a substitution-matrix-based clustering among motifs by computing a *motif-similarity score*; then it keeps one motif per cluster. Yet, DDSM presents some limitations. Indeed, it is dedicated to protein sequences and does not consider structures. Besides, it neglects the negative scores in the substitution matrices and adopt a slow posteriori pruning. We extend this method and enhance it to address frequent subgraphs (substructures) summarization using substitution matrices (F3SM). Our method is claimed to be generic and can deal with structures and sequences. It takes also into account the negative scores and computes differently the motif-similarity score. In addition, it adopts different clustering and pruning approach that enables obtaining more independent and representative motifs in a shorter time. We briefly define our approach by the following formalization:

$$F3SM(\Omega, \mathcal{M}, \mathcal{T}) = \Omega^*$$

where Ω is the initial set of frequent subgraphs, $\mathcal{M}$ is a substitution matrix, $\mathcal{T}$ is a user-specified substitution threshold and Ω^* is the set of F3SM-motifs.

In order to experimentally evaluate our approach, we use already published protein structures datasets transformed into graphs using C_α method as in [3]. Table 2 summarizes the characteristics of each dataset. Generally, in a motif selection approach two aspects are emphasized, namely the number of selected motifs and their interestingness. In order to evaluate the number of selected motifs using our approach, we first use the state-of-the-art method gSpan [5] to extract frequent subgraphs with a frequency threshold of 30%. Then, we

Dataset	ID	Family name	G1	G2	# motifs
DS1	52592	G proteins	33	33	799 094
DS2	48942	C1 set domains	38	38	258 371
DS3	56437	C-type lectin domains	38	38	114 792
DS4	88854	Kinases, catalytic subunit	41	41	1 073 393

Table 1. Experimental data from [3]. **ID**: identifier in SCOP [1], **G1**: proteins sampled from a selected protein family, **G2**: proteins randomly sampled from PDB [2], **# motifs**: number of frequent subgraphs generated by gSpan.

use F3SM, using the substitution matrix $Blosum62$, on the set of frequent subgraphs to select representative ones. To evaluate the interestingness of F3SM-motifs, we perform a protein classification on each dataset using subgraph-motifs as features for the Naïve bayes (NB) classifier. Then, we compare the values of accuracy obtained using the initial set of frequent subgraphs and the F3SM-motifs.

The outcomes of our experiments are illustrated in Fig. 1 and Fig. 2. In Fig. 1, we notice that F3SM reduces dramatically the number of frequent subgraphs especially with lower substitution thresholds. Indeed, the number of F3SM motifs does not exceed 20% of the total number of frequent subgraphs generated by gSpan with all datasets for all substitution thresholds below 70%. This important reduction of dimensionality comes with a notable interestingness in terms of classification accuracy. This fact is illustrated in Fig. 2 which shows that F3SM motifs allow better classification performance compared to the original set of frequent subgraphs.

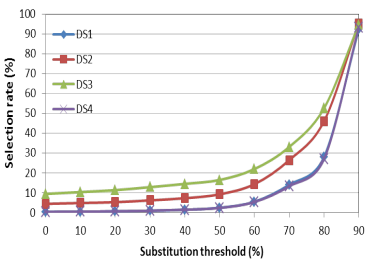


Figure 1. F3SM motif rate

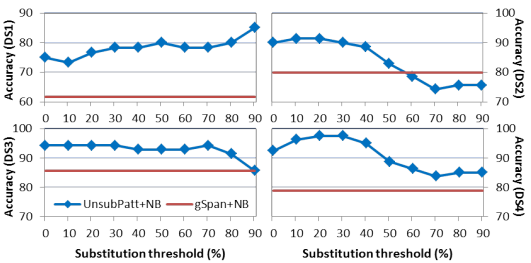


Figure 2. F3SM motif interestingness in classification

3 Conclusion

We proposed F3SM, a subgraph summarization method based on substitution matrices. Our method allows to reduce considerably the huge number of discovered frequent subgraphs to obtain an interesting and representative set of motifs enabling further exploration on proteins. The interestingness of F3SM motifs was evaluated in four protein classification cases and the experimental outcomes were promising. It is also worth mentioning that F3SM is not limited to classification tasks, but generic and can help with other subgraph-motifs-based analysis such as clustering, visual inspection, drug molecule prediction, etc.

References

[1] A. Andreeva, D. Howorth, J.-M. Chandonia, S. E. Brenner, T. J. P. Hubbard, C. Chothia, and A. G. Murzin. Data growth and its impact on the scop database: new developments. *Nucleic Acids Research*, 36:D419–D425, 2008.

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

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

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

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

Logical modelling of mesoderm specification

Abibatou MBODJ[1], Duncan BERENGUIER[1, 2] Guillaume JUNION[3],
Eileen FURLONG[3] and Denis THIEFFRY[4]

[1] TAGC, U1090 INSERM, 163 Avenue de Luminy, 13288, Marseille, Cedex 09, France
abibatou.mb@gmail.com

[2] IML, UMR6206 CNRS, 163 Avenue de Luminy, 13288, Marseille, Cedex 09, France
duncan.benrenguier@gmail.com

[3] EMBL, Meyerhofstraße 1, 69117 Heidelberg, Germany
{junion, furlong}@embl.de

[4] IBENS, UMR8197 CNRS and U1024 INSERM- ENS, 45 rue d'Ulm, 75005, Paris, France
thieffry@ens.fr

Keywords: drosophila development, mesoderm specification, regulatory network, logical modelling, GINsim.

During embryogenesis, complex signalling/regulatory networks control the formation of spatially and temporally refined patterns of gene expression. Focusing on early Drosophila development, we rely on a combination of published genetic data and high-throughput analyses (ChIP-chip, ChiP-seq, transcriptome [6-8]) to delineate a logical model encompassing the main transcription factors involved in mesoderm specification during stages 8 to 10 of development.

More specifically, we aim to account for the specification of mesoderm into four main presumptive territories (somatic muscles, visceral muscles, fat body and heart [1, 5]) in terms of alternative stable states, which are defined as specific combinations of transcription factors expressed in response to positional clues and associated to each presumptive territory.

Built with the software *GINsim* [2], our current logical model encompasses 55 components (most of them Boolean, excepting five multilevel nodes (four ternary and one quaternary) and 86 regulatory interactions, making it appealing and easy to grasp by biologists, but at the same time difficult to analyse, as the size of state transition graphs grow exponentially with the number of regulatory nodes. To ease the dynamical analysis of this comprehensive model, we apply a reduction model, which allow the selection of nodes to be made implicit and preserve essential dynamical properties such as stable states [3] (see also Chaouiya *et al.*, these proceedings).

Simulations of a reduced version of our model qualitatively recapitulates the wild-type behaviour, as well as about thirty reported mutant phenotypes (e.g. losses, extensions or replacements of specific tissues for single or combined gene loss-of-functions or ectopic expressions). Based on these encouraging results, we further designed a series of *in silico* multiple perturbation analyses using a dedicated function of *GINsim*, which eases the definition and storage of mutants. Starting from relevant initial states, *GINsim* is used to compute the reachable stable states for each genetic background. To ease the interpretation of simulations, we have defined a series of logical signatures corresponding to the set of expressed markers assigned to the formation of each presumptive tissue. These signatures are compared with each reachable stable state. Further automation has been achieved by defining a script calling *GINsim* iteratively for each initial state and for each genetic background considered. The results are integrated in the form of a web page containing a matrix displaying the resulting phenotypes in a graphical way, along with a list of expressed markers. This web page has proved to be very useful to discuss simulation results with biologists and select predictions for experimental validation at EMBL.

We are now extending our logical model to cover the developmental stages 10 to 12, focusing in particular on cardiac cell diversification [4]. We are currently able to qualitatively recapitulate wild-type cardioblast diversification, as well as the phenotypes of 17 reported mutants.

Acknowledgements

This work is supported by the EU EraSysBio + program (project ModHeart).

References

[1] Azpiazu N., Lawrence P.A., Vincent J.P., Frasch M. (1996). Segmentation and specification of the Drosophila mesoderm. *Genes Dev* 10: 3183–94.

[2] Chaouiya C., Naldi A., Thieffry D. (2012). Logical Modelling of Gene Regulatory Networks with GINsim. *Methods Mol Biol* 804: 463–79.

[3] Naldi A., Remy E., Thieffry D., Chaouiya C. (2011). Dynamically consistent reduction of logical regulatory graphs. *Theor Comput Sci* 412: 2207-18.

[4] Reim I., Frasch M. (2010). Genetic and genomic dissection of cardiogenesis in the Drosophila model. *Pediatric cardiology* 31:325-34.

[5] Riechmann V., Irion U., Wilson R., Grosskortenhaus R., Leptin M. (1997). Control of cell fates and segmentation in the Drosophila mesoderm. *Development* 124: 2915-22.

[6] Sandmann T., Jensen L. J, Jakobsen J. S., Karzynski M.M., Eichenlaub M.P., Bork P., Furlong E.E. (2006). A temporal map of transcription factor activity: mef2 directly regulates target genes at all stages of muscle development. *Dev. Cell* 10: 797- 807.

[7] Sandmann T., Girardot C., Brehme M., Tongprasit W., Stolc V., Furlong E.E. (2007). A core transcriptional network for early mesoderm development in Drosophila melanogaster. *Genes Dev* 21: 436-49.

[8] Wilczynski B. , Furlong E.E. (2010). Dynamic CRM occupancy reflects a temporal map of developmental progression. *Mol Syst Biol* 6: 383.

Posters, second appel

Les résumés présentés dans cette session ont étés reçus lors de la seconde vague d'appel à poster. Ils n'ont pas fait l'objet d'un processus de relecture par le comité.

QGP : Quantitative Genetics Platform

A high performance computing solution for quantitative genetics software

Misharl Monsoor[12], André Neau[3], Martin Souchal[4], Sylvie Nugier[4], François Laperruque[5], Eddie Iannuccelli[6], Pascale Le Roy[12], Edmond Ricard[5], David Robelin[6] and Olivier Filangi[12]

[1] INRA, PEGASE Physiologie, Environnement et Génétique pour l'Animal et les Systèmes d'Elevage, UMR1348, 35590 Saint-Gilles, France

[2] INRA, PEGASE Physiologie, Environnement et Génétique pour l'Animal et les Systèmes d'Elevage, UMR1348, 35000 Rennes, France

{misharl.monsoor, pascale.leroy, olivier.filangi}@rennes.inra.fr

[3] INRA, GABI Génétique Animale et Biologie Intégrative, UMR1313, 78532 Jouy-En-Josas,France

andre.neau@jouy.inra.fr

[4] INRA, CTIG Centre de Traitement de l'Information Génétique, US0310, 78532 Jouy-En-Josas, France

{martin.souchal, sylvie.nugier}@jouy.inra.fr

[5] INRA, SAGA Station d'Amélioration Génétique des Animaux, UR0631, 31326 Castanet Tolosan, France

{francois.laperruque, edmond.ricard}@toulouse.inra.fr

[6] INRA, LGC Laboratoire de Génétique Cellulaire, UR444, 31326 Castanet Tolosan, France

{eddie.iannuccelli, david.robelin}@jouy.inra.fr

1 Background

1.1 Genotype file format and conversion

Un des besoins croissants des généticiens quantitatifs est la manipulation des fichiers de génotypage. Ces fichiers contiennent les génotypes des individus, en général de quelques centaines à quelques milliers d'individus, en chaque marqueur le long d'un génome. Les premiers développements logiciels d'analyse génétique prévoyaient l'exploitation d'informations sur des marqueurs microsatellites, avec quelques centaines de marqueurs génétiques par individu. Aujourd'hui, les puces de génotypage à haute densité peuvent produire une quantité d'information bien plus élevée, allant par exemple aujourd'hui jusqu'à huit cent mille marqueurs génétiques (Single Nucleotide Polymorphisms) sur le génome des bovins. Les fichiers de données atteignent plusieurs Gigaoctet et sont difficilement manipulables par les tableurs et logiciels (R/SAS) communément utilisés par la communauté. De plus l'automatisation des processus de vérification et de cohérence des données, caractérisée par l'exécution de milliers de processus est devenue un préalable indispensable à l'analyse des caractères complexes.

1.2 Heterogeneous computing environment

Les méthodologies et outils se sont adaptés aux nouvelles architectures utilisés dans un contexte de calcul scientifique. Dans un souci de performance, l'utilisateur est amené à utiliser plusieurs modes d'exécution : soumission de job sur une grappe de serveurs, calcul déporté sur un processeur graphique (GPU), exécution en environnement multithreadé. Ces ressources sont la plupart du temps localisées dans plusieurs infrastructures, ce qui complexifie le workflow d'exécution. L'utilisateur doit, en outre, transférer ces fichiers de serveur en serveur.

2 Implementation

2.1 Galaxy : A Web unified interface

La plate-forme QGP (Quantitative Genetics Platform, https://qgp.jouy.inra.fr) a pour objectif de répondre à ces besoins. QGP implémente son environnement d'exécution sur le framework Galaxy [1,2,3]

qui fournit un ensemble d'outils d'analyse, de manipulation et de visualisation des données génomiques pour la communauté bio-informatique. Galaxy est capable de gérer des fichiers de séquence de grande taille comme le transfert, la suppression ou l'insertion de lignes ou de colonnes. Ces fichiers étant comparables aux fichiers de genotypage utilisés par les outils de génétique quantitative, les outils de manipulation de fichiers restent accessibles de la plate-forme QGP. Galaxy est conçu pour exécuter des chaînes de traitements prédéfinies (Workflow) et applicables aux nouveaux lots de données.

2.2 A variety of quantitative genetics software tools

Les logiciels hébergés par la plate-forme sont implémentés pour plusieurs types de serveurs de calcul et adressent plusieurs problèmes de génétique quantitative dans le domaine animal. L'analyse de la variabilité génétique des caractères complexes dans des populations expérimentales (QTLMap [4]), l'évaluation génétique (GS3, GABayes), l'analyse de parenté (PEDIG [5]), la détection d'incompatibilité de génotypes (Mendelsoft [6]). Un ensemble d'outils pour la conversion de format et la génération de graphiques sont également accessibles. Galaxy est capable de soumettre des exécutions de jobs sur des serveurs, distants et hétérogènes. Cette fonctionnalité permet de déployer des implémentations spécifiques, telles que les implémentations de QTLMap sur GPU [7], mais aussi des parallélisations naïves (indépendances des vérifications des génotypes aux marqueurs, reconstruction d'haplotypes par famille de demi ou plein frères,...) de façon simple et transparente à l'utilisateur.

3 Conclusion

L'utilisation de données sur des séquences complètes est d'ores et déjà évoquée pour l'analyse des caractères ou l'évaluation génétique des animaux d'élevage. Par ailleurs, il est probable que les phénotypes seront bientôt également difficiles à gérer, en quantité et en diversité, l'exemple du traitement des données de transcriptome ou de métabolome étant déjà une réalité pour beaucoup de généticiens. Le choix du framework Galaxy, imaginé à l'origine pour gérer des données de séquençage, permettra donc de suivre plus aisément cette évolution attendue du volume des données à traiter conjointement. De plus, il permet de profiter des evolutions de Galaxy (adaptation aux serveurs de calcul émergents, outils générique pour la manipulation de fichiers). Enfin, le choix de cet outil devrait également faciliter les connexions, devenues indispensables pour la biologie intégrative, entre le monde de la bioinformatique, où la culture est celle de l'intégration des connaissances et de la visualisation ou de l'extraction des résultats, et celui de la génétique quantitative, où la culture est celle de la modélisation et du calcul scientifique.

References

[1] Goecks, J, Nekrutenko, A, Taylor, J and The Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 2010 Aug 25;11(8):R86.

[2] Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J. "Galaxy: a web-based genome analysis tool for experimentalists". Current Protocols in Molecular Biology. 2010 Jan; Chapter 19:Unit 19.10.1-21.

[3] Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A. "Galaxy: a platform for interactive large-scale genome analysis." Genome Research. 2005 Oct; 15(10):1451-5.

[4] Filangi.O , Moreno C. , Gilbert H. , Legarra A. , Le Roy P. , Elsen JM. *QTLMap, a software for QTL detection in outbred populations*. 9th World Congress on Genetics Applied to Livestock Production, Leipzig. 2010.

[5] Boichard D., 2002. Pedig : a fortran package for pedigree analysis suited to large populations. 7th World Congress on Genetics Applied to Livestock Production, Montpellier, 19-23 août 2002, paper 28-13.

[6] Givry S.,Palhiere I.,Vitezica Z.G., Schiex T., Mendelian error detection in complex pedigree using weighted constraint satisfaction techniques, Workshop on Constraint Based Methods for Bioinformatics, Spain, Octobre 2005

[7] Chapuis G., Filangi O., Leroy P., Elsen JM., Lavenier D., *GPU Accelerated QTLMap*, The 15Th QTL-MAS Workshop, Rennes, 2011

ncPRO-seq: annotation and profiling analysis of ncRNAs from small RNA-seq

Chong-jian CHEN[12345 ⋆], Nicolas SERVANT[145⋆], Joern TOEDLING[123456], Alexis SARAZIN[7], Antonin MARCHAIS[8], Evelyne DUVERNOIS-BERTHET[7], Valérie COGNAT[9], Vincent COLOT[7], Olivier VOINNET[8], Edith HEARD[123 †], Constance CIAUDO[1238†] and Emmanuel BARILLOT[145†]

[1] Institut Curie, Paris, France
[2] CNRS UMR3215, Paris, France
[3] INSERM U934, Paris, France
[4] INSERM U900, Paris, France
[5] Mines ParisTech, Fontainebleau, France
[6] Institute of Molecular Biology gGmbH, Mainz, Germany
[7] Institut de Biologie de l'Ecole Normale Supérieure, CNRS UMR8197, INSERM U1024, Paris, France
[8] Swiss Federal Institute of Technology Zurich, Department of Biology, Chair of RNA biology, Zürich, Switzerland
[9] Institut de Biologie Moléculaire des Plantes, CNRS UPR2357, Université de Strasbourg, Strasbourg, France
bioinfo.ncproseq@curie.fr

Keywords ncPRO-seq, non-coding RNA, small RNA-seq, profiling analysis

Over recent years, deep sequencing technology has become a powerful approach for investigating small non-coding RNA (ncRNA) populations, i.e. small RNA-seq. It is now established that an increasing number of novel small ncRNA families distinct from microRNAs are generated over kingdoms from different coding/non-coding regions via various biogenesis pathways and might involve a great spectrum of biological processes. For example, two other major classes of endogenous small RNAs, Piwi-interacting RNAs (piRNAs) and endogenous small interfering RNAs (endo-siRNAs), have been identified and widely investigated in mammals [1]. Moreover, in other organisms like plants more classes of small ncRNA have been described indicating that a wide range of small ncRNAs exist [2].

However, most of the existing tools devoted to sRNA-seq analysis, are only based on miRNAs annotation and quantification, significantly neglecting other types or new types of small ncRNAs. Only two non-miRNA oriented approaches, SeqCluster [3] and DARIO [4], has been recently developped with the ambitions to process whole sRNA-seq data in an unbiased way. However, they just perform gene-based analysis, but not detailed family-based (profiling) analysis which is critically important to investigate known small ncRNA families and to identify novel small ncRNA families.

Here we present a comprehensive and flexible ncRNA analysis pipeline, ncPRO-seq (Non-Coding RNA PROfiling from sRNA-seq), which is able to interrogate and perform detailed profiling analysis on small RNAs derived from annotated non-coding regions in miRBase, Rfam and repeatMasker, as well as regions defined by users. We perform both gene-based and family-based detailed analyses of small RNAs. The ncPRO-seq pipeline also has a module to identify regions significantly enriched with short reads that can not be classified as known ncRNA families[5], thus enabling the discovery of yet unknown ncRNA families. The ncPRO-seq pipeline supports input read sequences in fastq, fasta and color space format, as well as alignment results in BAM format, meaning that small RNA raw data from the 3 current major platforms (Roche-454, Illumina-Solexa and Life technologies-SOLiD) could be analyzed with this pipeline. Finally, the ncPRO-seq pipeline can be used to analyze data based on genome from metazoan to plants. The current version proposes annotation files for fifteen different species.

The ncPRO-seq pipeline is a stand-alone pipeline, which can be easily installed in a local computer or cluster. We offer two ways to launch the pipeline, through either a command line or a user-friendly web interface.

⋆. These authors contributed equally to this work
†. These authors are co-last author

The ncPRO-seq pipeline allows users to specify different options at each analysis stage, from raw reads processing to ways to generate results, all of which can be done by either selecting parameters in the web page or manually editing a configuration file. The results are available through an HTML report. Users can directly view figures and tables in the result web page. Track files are generated for visualization in genome browsers. We deploy the ncPRO-seq pipeline in http://ncproseq.sourceforge.net, where users can find detailed information, such as basic descriptions, manuals, test dataset and example results.

References

[1] M. Ghildiyal, PD. Zamore. Small silencing RNAs: an expanding universe. *Nat Rev Genet.*, 10(2):94-108, 2009.

[2] P. Brodersen, O. Voinnet. The diversity of RNA silencing pathways in plants. *Trends Genet.*, 22(5):268-80, 2006

[3] L. Pantano, X. Estivill and E. Marti. A non-biased framework for the annoation and classification of the non-miRNA small RNA transcriptome.*Bioinformatics*, 27, 3202-3203, 2011.

[4] M. Fasold, D. Langenberger, H. Binder, PF. Stadler, S. Hoffmann. DARIO: a ncRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Res.*, 39:W112-7, 2011.

[5] J. Toedling, C. Ciaudo, O. Voinnet, E. Heard, and E. Barillot. Girafe–an R/Bioconductor package for functional exploration of aligned next-generation sequencing reads.*Bioinformatics*, 26, 2902–2903, 2010.

Accelerating QTL mapping with graphics cards

Guillaume CHAPUIS[1], Olivier FILANGI[2], Jean-Michel ELSEN[3], Dominique LAVENIER[1] and Pascale LE ROY[2]

[1] IRISA, UMR 6074, Campus de Beaulieu, 263 avenue du Général Leclerc - Bâtiment 12 35 042 Rennes, cedex, France
{Guillaume.Chapuis, Dominique.Lavenier}@irisa.fr
[2] INRA Agrocampus Ouest, UMR Pegase, 65 rue de Saint-Brieuc, INRA, 35042 Rennes, France
{Olivier.Filangi, Pascale.Leroy}@rennes.inra.fr
[3] INRA, UPR 0631, Chemin de Borde-Rouge - Auzeville BP 52627 31326 CASTANET-TOLOSAN, cedex, France
{Jean-Michel.Elsen}@toulouse.inra.fr

Keywords QTL mapping, GPGPU, GPU, linkage analysis, linkage disequilibrium, multiQTL analysis, QTLMap, Double-Precision, Cuda

1 Context

Most of the traits characterizing individuals are influenced by heredity. Geneticists are interested in detecting, localizing and identifying genes, the polymorphism of which explains a part of observed trait variability. Such genes are often called QTLs (for "Quantitative Trait Locus"), the term locus pointing to a physical position on the genome.

QTL detection procedures consist in a series of statistical hypotheses tests at successive putative locations on the genome. We focus here on regression approaches performed on sets of large families. These approaches were developed for exploiting both the linkage and the gametic association between loci observed on a per family basis and / or at the population level. QTLMap implements three types of QTL analyses: Linkage analysis (LA), Linkage Disequilibrium Analysis (LDA), and Linkage Disequilibrium Linkage Analysis (LDLA).

QTL mapping analyses are computationnally intensive. They often take weeks to run on modern computers and run times increase linearly with the density of available genetic markers.

Despite the computational burden that QTL mapping represents, few parallel tools exist. The first attempt was made by [3] with gridQTL. This tool is derived from QTLexpress ([2]), a popular web based tool for QTL analyses, and harnesses the power of computational grids to try and reduce run times. The previous version of QTLMap, developped by [1], takes advantage of modern CPUs by using all their cores simultaneously.

We have developped a new version of QTLMap, which takes advantage of Graphics Processing Units (GPUs). The empirical approach used in QTLMap makes it an ideal candidate for GPU computations. Statistical tests computed at each genome position and for each simulation are almost identical in terms of instructions and also completely independent. This type of data parallelism is a perfect fit for the single instruction multiple data (SIMD) architecture of GPUs. The new GPU implementation of QTLMap performs about 70 times faster than the previous multicore implementation, while maintaining the same level of precision.

2 Experiments and results

Tests were run on machines with two quadcore Intel® Xeon® E5420 (12M Cache, 2.50 GHz, 1333 MHz FSB) processors. Multicore cpu tests were run on the Genotoul platform (http://www.genotoul.fr). GPU tests were run on a machine equipped with an Nvidia® C2050 card. Each test consists of an LDLA analysis over a simulated dataset from the 2011 QTL-MAS workshop (https://colloque.inra.fr/qtlmas).

Fig. 1 shows the evolution of the speedup with respect to the number of studied genome positions between the previous multicore implementation and the GPU implementation in Double-Precision (DP). The openMP

version of QTLMap is not optimized for low numbers of genome positions ; therefore we observe a high speedup, when the number of genome positions is low. Fig. 1 shows a speedup of about 70 for the GPU version in Double-Precision over the openMP version.

Figure 1. Speedup of GPU-QTLMap / OpenMP-QTLMap (8 cores) over the number of genome positions

Reduced runtimes allow geneticists to consider more precise and time consuming analyses by increasing the number of simulations or the number of studied genome positions. All versions of QTLMap are available under CeCILL licences at http://www.inra.fr/qtlmap/.

Future work include the promotion and use of parallel computing in statistical genetics, focusing on two applications of the Single Nucleotide Polymorphism (SNP) chip technology:

- Dissection of the genetic architecture of characters through Genome Wise Association Studies (GWAS);
- Genomic Selection (GS).

SNP chip technology now makes possible the genotyping on millions of SNPs of tens or hundreds of thousands of individuals, thus increasing the demand for much faster computations. Faster computations are needed both for implementing more precise genetic models in research of trait genetic determinants, and for the industrial exploitation of genomic data, with production of statistical information at regular time intervals. Our aim is to produce, when needed, new algorithms better suited for parallel architectures (GPUs and/or clusters of computers).

Acknoledgments

This work is part of the AAP INRA/INRIA and is also supported by the region of Brittany, France.

References

[1] O. Filangi, C. Moreno, H. Gilbert, A. Legarra, P. Le Roy, and J. Elsen. Qtlmap, a software for qtl detection in outbred populations. In *Proceedings of the 9th World Congress on Genetics Applied to Livestock Production: 1-6 August; Leipzig*, number 787, 2010.

[2] G. Seaton, C. S. Haley, S. A. Knott, M. Kearsey, and P. M. Visscher. Qtl express: mapping quantitative trait loci in simple and complex pedigrees. *Bioinformatics applications note*, 18(2):339–340, 2002.

[3] G. Seaton, J. Hernandez, J. Grunchec, I. White, J. Allen, D. De Koning, W. Wei, D. Berry, C. Haley, and S. Knott. Gridqtl: a grid portal for qtl mapping of compute intensive datasets. In *Proceedings of the 8th World Congress on Genetics Applied to Livestock Production*, pages 13–18, 2006.

Elaboration d'un système d'aide à la décision à base d'ontologie dans le cadre des urgences odontologiques.

Faustine SERRAND-OBRY [a1], Valéry DONFACK GUEFACK [a], Régis DUVAUFERRIER [a], Jérémy LASBLEIZ [a], Valérie BERTAUD-GOUNOT [a]

[a] Unité Inserm U936, Faculté de Médecine, Université Rennes 1, France
faustineobry@hotmail.fr

PredAlgo, a new subcellular localization prediction tool dedicated to green algae

Marianne Tardif[1], Ariane Atteia[2], Michael Specht[3], Guillaume Cogne[4], Norbert Rolland[2], Sabine Brugière[1], Michael Hippler[3], Myriam Ferro[1], Christophe Bruley[1], Gilles Peltier[5], Olivier Vallon[6] and Laurent Cournac[5]

[1] Laboratoire Biologie à Grande Echelle, U1038, CEA/INSERM/UJF, 17 rue des Martyrs, 38054 Grenoble Cedex 9, France
{marianne.tardif, sabine.brugiere, myriam.ferro, christophe.bruley} @cea.fr;
[2] Laboratoire Physiologie Cellulaire Végétale, UMR5168, UMR1200, CEA/CNRS/INRA/UJF, 17 rue des Martyrs, 38054 Grenoble Cedex 9, France
ariane.atteia@ifr88.cnrs-mrs.fr; norbert.rolland@cea.fr
[3] Institute of Plant Biology and Biotechnology, University of Münster, Hindenburgplatz 55, 48143 Münster, Germany
micha.specht@gmail.com; mhippler@uni-muenster.de
[4] GEPEA UMR 6144CNRS/Univ.Nantes, 44602 Saint-Nazaire Cedex, France
guillaume.cogne@univ-nantes.fr
[5] Laboratoire de Bioénergétique et Biotechnologie des Bactéries et Microalgues, UMR 6191, CEA/CNRS/Univ Aix-Marseille, F-13108 Saint-Paul-lez-Durance, France
gilles.peltier@cea.fr; laurent.cournac@ird.fr
[6] Institut de Biologie Physico-Chimique, CNRS/Univ. Pierr et Marie Curie, UMR 7141, Paris, France
ovallon@ibpc.fr

Keywords Subcellular localization prediction, green alga, machine learning, neural networks

1 Introduction

Sequence analysis software allowing prediction of intracellular localization appears as an essential component of the genome annotation toolbox in eukaryotes [1]. The most popular tools simply use the characteristics of the N-terminal sequence as a proxy for protein localization, because most targeting signals are found at the N-terminus of the preproteins (Signal Peptide for proteins routed to the secretory pathway, Transit Peptides -mTPs, cTPs - for proteins delivered to the mitochondrial matrix or chloroplast stroma, respectively). The targeting peptide is cleaved concomitantly with import, generating a new N-terminus of the mature protein. Several programs are available that provide robust prediction for land plants. Yet, they are notoriously unreliable when used to predict the localization of algal proteins. We thus developed a new algorithm which specifically aims at predicting intracellular localization in green alga.

2 Training set and Neural Network for the development of PredAlgo

We used *Chlamydomonas reinhardtii* as the source of training sequences because it is the only green alga that can provide the necessary breadth of high quality experimental data. From the many peptides identified through the tandem Mass Spectrometry (MS/MS) technology generated by comprehensive surveys of whole mitochondrion and chloroplast organelles [2,3], we retrieved experimental protein N-termini of proteins and derived the targeting peptides for these proteins. A training set of 238 proteins was generated.

The first 150 aminoacids of the training set sequences were decomposed into 19-residue overlapping subsequences, each represented by an input matrix with 19 rows (one per position) and 20 columns (one per amino acid) where 1 encoded presence and 0 encoded absence. These subsequences were meant to train feedforward neural networks using SNNS software (http://www.ra.cs.uni-tuebingen.de/software/JavaNNS/) and produce a "predicted" triplet score (M_{sub}, C_{sub}, SP_{sub}) where each figure represents the probability for that subsequence to be part of a presequence targeting the protein towards either the mitochondrion, the chloroplast or the secretory pathway. In order to optimize network design, the original training set of subsequences was randomly divided into a preliminary training set (80% of the total) and a validation set (the remaining 20%) to use as test database. The final learning process, using the whole training database, resulted in the core of PredAlgo v.1.0. Results for the subsequences were used to compute a targeting

prediction triplet score for the proteins themselves (M_{prot}, C_{prot}, SP_{prot}). In order to give more weight to the start of the sequence, we calculated the average network outputs for subsequences starting at positions 1 to 10, 1 to 20, 1 to 30, 1 to 40 and 1 to 50, and defined the output for the protein themself as the sum of these values. When the three scores were below a certain cutoff, the protein was assigned to the "Other" category, otherwise to the compartment with the highest score.

3 Evaluation of PredAlgo

A benchmark set (577 Chlamydomonas proteins) completely distinct from that used in training, was used to evaluate the performances of the program and compare them to those of publicly available multi-sites prediction programs (TargetP http://www.cbs.dtu.dk/services/TargetP/ , Predotar http://urgi.versailles.inra.fr/predotar/predotar.html, Protein Prowler http://pprowler.imb.uq.edu.au/, WoLF PSORT http://wolfpsort.org/ and MultiLoc2 http://www-abi.informatik.uni-tuebingen.de/Services/MultiLoc2). The striking weakness of available predictors confirming their inadequacy to algal sequences, resided in the fact they largely mis-targeted chloroplast proteins towards the mitochondrion. In comparison, PredAlgo produced a highly improved discrimination between the chloroplast and mitochondria localization prediction, which is reflected by the achievement of 85% sensitivity for the chloroplast and of 72% precision for the mitochondria.

We also examined the quality of PredAlgo predictions on the proteomes of other green algal species. We chose the multicellular *Volvox carteri*, a close relative of Chlamydomonas in the group Chlorophyceae, two unicellular Trebouxiophyceae and three more distant relatives belonging to the group Prasinophyceae. Owing to the scarcity of intracellular localization data for these algae, we had to infer subcellular localization from orthology to *C. reinhardtii* proteins, i.e. to assume that orthologous protein pairs share the same intracellular localization. As expected, we found an excellent agreement between predictions in *Volvox* and in Chlamydomonas (92% correct prediction), to a less extent for Trebouxiophyceae, but not for Prasinophyceae.

4 Conclusion

PredAlgo appears to perform much better than the other publicly available programs to predict intracellular localization in the model alga *Chlamydomonas reinhardtii*, especially when it comes to distinguishing plastidial from mitochondrial targeting. With the advent of highly-efficient genomes annotation pipelines, it will become possible to better evaluate the predictors and if necessary, to build new ones for specific algal groups. In the field of algal research, improving the functional annotation of proteins constitutes a major challenge due to the strong interest in engineering the photosynthetic production of biofuels or other chemicals by microalgae. We hope to be able to improve PredAlgo in future releases.

Acknowledgements

This work was supported by the ANR "ALGOMICS", Bioenergies Program 2008-2012, and by CEA, INSERM, INRA, CNRS, French Universities and University of Münster. M.H. acknowledges support by the "Bundesministerium für Bildung und Forschung" Grant 0315265 C.

References

[1] K. Imai and K. Nakai, Prediction of subcellular locations of proteins: where to proceed? 2010. *Proteomics*, 10:3970-3983, 2010.

[2] A. Atteia, A. Adrait, S. Brugiere, M. Tardif, R. van Lis, O. Deusch, T. Dagan T, L. Kuhn, B. Gontero, W. Martin, J. Garin, J. Joyard and N. Rolland, A proteomic survey of Chlamydomonas reinhardtii mitochondria sheds new light on the metabolic plasticity of the organelle and on the nature of the alpha-proteobacterial mitochondrial ancestor. *Mol Biol Evol*, 26:1533-1548, 2009.

[3] M. Terashima, M. Specht, B. Naumann and M. Hippler.. Characterizing the anaerobic response of Chlamydomonas reinhardtii by quantitative proteomics, *Mol Cell Proteomics*, 9:1514-1532, 2010.

HTSstation

A web application for High Throughput Sequencing data analysis

Fabrice PA DAVID[1,2,*], **Solenne CARAT**[1,2,*], Julien DELAFONTAINE[1,2,*], Frederick ROSS[1,2], Gregory LEFEBVRE[1,2], Lucas SINCLAIR[1,2], Yohan JAROSZ[1,2], Marion LELEU[1,2] and Jacques ROUGEMONT[1,2]

[1] School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
{solenne.carat, jacques.rougemont}@epfl.ch
[2] Swiss Institut of Bioinformatics (SIB), Lausanne, Switzerland
* These authors contributed equally to this work

Abstract *HTSstation is a web tool with open-access libraries providing various applications of High Throughput Sequencing data analysis such as ChIP-seq, RNA-seq, 4C-seq and re-sequencing. It was built to be directly used by biologists through a user friendly interface with preselected options and filters, but also by bioinformaticians at the command line with optional configuration files to perform customised analysis. HTSstation provides graphical and tab-delimited results, and allows direct data visualisation in UCSC and GDV (own Genome Data Viewer)*

Keywords NGS, ChIP-seq, RNA-seq, SNP, 4C-seq

1 Introduction

High-throughput sequencing has produced a major shift in the amount and speed of data production in genomics. More importantly, the range of applications of this technology is constantly expanding, which increases the importance of bioinformatics and the necessity to implement efficient and reproducible analysis methods specifically adapted to each particular application (*e.g.* ChIP-seq, RNA-seq, 4C-seq, re-sequencing).

In the context of high pressure for quick scientific results, HTSstation was developed in order to gather in a simple portal all common steps of high-throughput sequencing analysis, allowing to allocate more time to scientific questions. This framework provides tools to construct specific pipelines, based on publicly available tools, with adapted and customised post-treatment. Result files can be directly visualised in UCSC or GDV genome browsers.

2 Methods

The HTSstation framework provides modular functionalities (Fig. 1) that can be combined into standardised analysis pipelines. These applications have been built upon two structural tools. First, GenRep, a genome data repository providing a common and consistent local mirror of stable genome assemblies, annotations and utility files (bowtie indexes, fasta and gtf file). Second, bein, a workflow managment system, aiming to track execution histories, results and errors.

- *Demultiplexing / Mapping*
  The Demultiplexing module separates multiplexed samples based on the presence of barcodes in reads. A final cleaning step filters specific sequences defined in a file describing primers used in the experiment. The Mapping module uses Bowtie [1] with default options (*-Sam 20 -best -strata -chunkmbs 512*). Specific user options can be given in a configuration file to customise the mapping. Then, the bam conversion can optionally include the removal of PCR duplicated reads. A final step converts the read alignment to genome-wide densities using a customised C program.
- *ChIP-seq*
  The ChIP-seq module uses MACS [2] to detect specific enrichment peaks, and a novel deconvolution model, which evaluates the shape of the peak, provides a more accurate estimation of binding site location and a lower number of false positives. Post-processing tools were designed to detect motif enrichement and to assign peaks to genes.

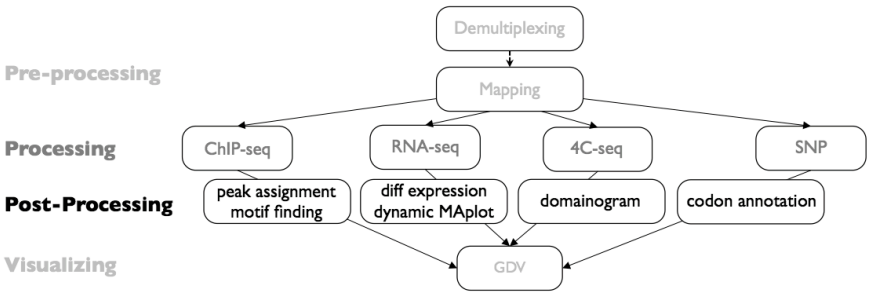


Figure 1. Global view of HTSstation modules

- *RNA-seq*
  The RNA-seq module, after mapping reads to the exonome, proposes an original optimisation procedure to properly infer transcripts expression from quantification at the exon level. It optimizes reads distribution among alternative transcripts, and then uses DESeq [3] to infer differential expression between samples.
- *4C-seq*
  Multiplex Circular Chromosome Conformation Capture (4C) coupled to sequencing is a recent technique to measure *in vivo* physical interactions between selected *loci* and the entire genome. For each selected locus (viewpoint), its corresponding genome-wide read density reflects the frequency of contact between every genomic position and the viewpoint. The densities have been smoothed, corrected for the trivial power-law component in the vicinity of the viewpoint (which contains a majority of reads) and clusters of high contact frequency were identified using a genome partioning algorithm.
- *Re-sequencing*
  The re-sequencing module detects relevant SNPs using samtools pileup, taking into account organism ploïdy. A post-processing step annotates selected SNPs according to their location relative to genes, with the corresponding amino-acid, in case of non-synonymous SNPs.

3 Results

HTSstation (http://htsstation.epfl.ch) provides many analysis tools covering major fields of high throughput sequencing data. Its web interface, specially designed for biologist users, its command line usage for bioinformaticians, and its open source code for developers made it useful for many usages.

Many modules of HTSstation were already used in several high-level studies, such as Noordermer *et al.* [4], Truman *et al.* [5] and Rey *et al.* [6].

References

[1] B. Langmead *et al.*, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10:R25, 2009

[2] L. Zhang*et al.*, Model-based analysis of ChIP-Seq (MACS). *Genome Biol*, **9**(9):R137, 2008.

[3] S. Anders and W. Huber, Differential expression analysis for sequence count data. *Genome Biology*, 11:R106, 2010

[4] D. Noordermeer *et al.*, The dynamic architecture of Hox gene clusters. *Science*, 334:222-225, 2011.

[5] RW. Truman *et al.*, Probable zoonotic leprosy in the southern United States. *The New England journal of medicine*, 364:1626-1633, 2011.

[6] G. Rey *et al.*, Genome-wide and phase-specific DNA-binding rhythms of BMAL1 control circadian output functions in mouse liver. *PLoS Biology*, 9:e1000595, 2011

Local Web Gui for Blast (LWBG)
A new Local interface for biological data treatment

Fida KHATER[1], Abdoulaziz MOUSSA[1]

[1] Master STIC pour la Santé, spécialité « Bioinformatique, Connaissance, Données »
Universités de Montpellier 1 & 2, Institut Telecom, TIC et Santé
CC 92000 Place Eugène BATAILLON 34095 Montpellier CEDEX 05
{fida.khater;abdelaziz.moussa}@etud.univ-montp2.fr

Abstract *Genomic and proteomic programs generate an increasing amount of sequences data. By now, one of the scientist's challenges is to confidentially analyze these data in lesser cost and duration. To date, there is no free software allowing using the features found in online tools such as NCBI.*
Local web Gui for Blast (LWBG) is a data analysis interface performing alignments and analysis of nucleic acid sequences via its Nucleotideblast tool. It offers to the users the possibility to create their own local databases via the makeblastdb tool and it constitutes a tool for optimizing results through its Local parser interface.

Keywords Blast, Local, Interface, Web, Genomic, proteomic

Développement d'une interface Graphique : Local Web Gui for Blast (LWBG), pour les traitements de données biologiques

Résumé *Les programmes de génomique, de protéomique, d'analyse de l'expression de gènes génèrent d'énormes volumes de données hétérogènes créées par les technologies à haut débit.*
L'analyse et le traitement confidentiel à faible cout et durée réduite de ces données brutes deviennent un challenge dans le monde scientifique. Aujourd'hui, il n'existe pas d'interface graphique libre permettant à l'utilisateur d'utiliser les fonctionnalités Bio-informatiques trouvées sur des outils en ligne tel que NCBI.
Local web Gui for Blast (LWBG) est un outil d'analyse de données permettant d'effectuer des alignements et des analyses de séquences nucléiques via Nucleotideblast. Il offre à l'utilisateur la possibilité de créer ses propres bases de données locales via l'outil makeblastdb et il constitue un outil d'optimisation de résultats via son interface Local parser.

Mots-clés Blast, Local, Intérface, web, Génomique, Protéomique

Contexte

Les programmes de génomique, de protéomique, d'analyse de l'expression de gènes manipulent d'énormes volumes de données hétérogènes créées par les technologies à haut débit. L'acquisition, le stockage, l'analyse et le traitement de ces données brutes, leur comparaison avec d'autres données de base extérieures généralistes ou spécifiques, voire leurs annotations a très vite nécessité la mise en place de collaborations interdisciplinaires, jusqu'alors peu existantes.

Aujourd'hui, des méthodes de recherche heuristiques utilisées en bioinformatique et dédiées au traitement des données biologiques sont accessibles en ligne [1,2]. Notons par exemple que l'outil nommé Blast permet de comparer une séquence nucléique ou protéique, dite requête, à une banque de séquences nucléiques ou protéiques. Cet outil reste le plus utilisé par la communauté scientifique.

Cependant, plusieurs inconvénients peuvent être énumérés tels que l'indexation de la séquence requête qui se traduit par un temps de calcul trop important lorsque l'on souhaite comparer des génomes complets ou des données haut débit issues des nouvelles technologies de séquençage et à ceci vient s'ajouter le fait que Blast ne permet que l'interrogation des données indexées par NCBI.

Aujourd'hui, une multitude de laboratoires de recherche se sont détournés de ces outils en optant pour une installation locale conséquente afin de garder la confidentialité des résultats en cours d'études et d'utiliser leurs propres banques de références pour les traitements de données, afin de minimiser le temps de calcul.

Cependant, les requêtes utilisées dans ces types d'outils ne sont réalisables qu'en lignes de commande d'abord compilées puis exécutées, ce qui s'avère être assez difficile pour des scientifiques n'ayant pas touché à la matière informatique. A ceci s'ajoute le fait qu'il y a peu d'outils présentant une documentation crédible (cas de commandes non illustrées) permettant aux utilisateurs de maitriser l'utilisation de l'outil.

De ces difficultés est née notre motivation de mettre en place un outil de traitement de données nommé Local Web Gui for Blast (LWBG) équipé d'une interface web simple d'accès qui permet même à des novices d'accéder en local à des outils dont blast.

En d'autres termes, cet outil permettra aux utilisateurs d'interroger des banques génomiques à partir de leurs machines locales ou de construire eux même leurs propres banques de référence (acides nucléiques, protéines) sur une machine locale. A ces deux fonctionnalités, nous avons doté l'interface d'un outil dédié au Parsing des résultats obtenus avec ces outils d'alignements.

Local web Gui for Blast (LWBG)

A ce jour, Local web Gui for Blast (LWBG) contient 3 outils fonctionnels. Le premier outil est appelé **NucleotideLW** et permet à l'utilisateur d'interroger les banques (Local NCBI Database) avec un query. Cet outil intègre des modules optionnels pour l'optimisation de l'algorithme blast ainsi que le résultat. Le deuxième outil dédié à la création de banque génomique est appelé **bioLogicaldbLW**. Cet outil permet la création d'une base de données (type nucléotide ou protéine) à partir d'un fichier fasta qui contient les séquences biologique. Et enfin, l'outil **Local Parser** est dédié à la recherche d'informations spécifiques (Parsing) dans un fichier contenant des alignements de séquences.

Discussion et Perspectives

Les avantages offerts par cet outil local à interface graphique sont multiples. Outre la confidentialité d'analyses que peut offrir cette interface, cet unique outil à interface graphique permet à l'utilisateur de créer des bases de données qui lui seront propres et d'effectuer des analyses à coût et durée beaucoup plus intéressants comparativement aux interfaces suggérées en ligne.

L'ensemble de notre code source ainsi que les outils (tel que Biopython [3]) nécessaires à l'installation de cet interface seront bientôt mis en ligne pour la communauté scientifique. Un manuel sera également mis à la disposition des utilisateurs afin de leur permettre d'accéder par étapes au téléchargement et au bon fonctionnement de cette interface.

Aujourd'hui, nous travaillons sur d'autres outils qui seront bientôt accessibles via cette interface tels que Blast pour protéines, Blast x etc. En ce qui concerne l'outil Local Paser, la prochaine mise à jour de cette interface portera sur le format des fichiers d'entrées pour un parsing.

Remerciement

Ce projet nous a permis d'être lauréats dans le cadre de l'appel d'offre « Les etudiants de l'UM2 sont entrepreneurs ». A ce titre, nous remercions l'Université de Montpellier2.

Nous tenons également à remercier le Labex NUMEV (http://www2.lirmm.fr/numev/) pour le financement de la participation à JOBIM 2012.

References bibliographiques

[1] Sayer Eric, Tanya Barrett, Dennis A Benson, Evan Bolton, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research.*, 40: D13-25, 2012.

[2] SF. Altschul, W. Gish, W. Miller, EW. Myers, DJ. Lipman, Basic local alignment search tool. *Journal of Molecular Biology*, 215 (3): 403-410, 1991.

[3] PJA. Cock, T. Antao, JT. Chang, BA. Chapman, CJ. Cox, A. Dalke, I. Friedberg, Biopython: freely available Python tools for computational molecular biology and bioinformatics ». *Bioinformatics*, 25: 1422-1423, 2009.

An *in silico* approach to model the assembly pathway of protein respiratory complexes in *Saccharomyces cerevisiae*

Annie GLATIGNY[1], Viet-Dung TRAN[1,] Philippe GAMBETTE[2] and Marie-Hélène MUCCHIELLI[1,3]

[1]CGM, UPR3404 CNRS, avenue de la terrasse, 91198, Gif sur Yvette, France
annie.glatigny@cgm.cnrs-gif.fr, viet-dung.tran@u-psud.fr, marie-helene.mucchielli@cgm.cnrs-gif.fr
[2]LIGM, UMR8049, Cité Descartes, Bât Copernic, 5 bd Descartes, 77454 Champs sur Marne
philippe.gambette@gmail.com
[3]UPMC, 5 place Jussieu, 75005, Paris, France

Abstract

The yeast Saccharomyces cerevisiae is a model for the analysis of the oxidative phosphorylation chain complexes. While the structure and the catalytic mechanisms of the complexes are well established, their biogenesis is far from understood. In a previous study we demonstrated how an in silico approach can be helpful to identify new factors involved in the biogenesis of the respiratory complexes [1]. Our method consists in using APID and BioGRID databases, then merging all the physical interactions concerning the core, the supernumerary subunits and the assembly factors of the complex III and constructing the resulting network via Cytoscape and its plugins [2]. To find sub-complexes the network was divided in highly interconnected sub-graphs with clusterONE, an algorithm that can produce overlapping classes and finally with MCODE that produces independent clusters. This approach allowed us to identify a protein interacting with complex III subunits Cor1p and Qcr2p. We then tested its function by genetic and biochemical approaches. The results indicated that the protein is indeed located within the mitochondria and is involved in the biogenesis of the respiratory complexes.

Using a similar approach, we are modelling the succession of events conducting to the final complex III. Our methodology to detect new assembly factors results in cliques which cannot be clustered anymore. Then it doesn't allow the detection of strongly interconnected sub-networks corresponding to assembly intermediaries. To resolve this problem, a first way consists in a hierarchical clustering of the subunits of the complex, starting from the weights of the edges between subunits. The weight of an edge between two nodes is computed as the number of nodes of the subnetwork resulting from the intersection of their direct neighbourhoods. This first method allows the identification of the order of assembly of the subunits of the complex. It does not permit the determination of assembly factors associated to the assembly intermediaries. To correct this drawback, we use another method consisting in clustering the weighted network to find all the subgraphs containing the assembly intermediaries and so the associated assembly factors.

The model built by using these two methods proposes new hypotheses on the order in which the sub-units and the assembly factors are involved in the biogenesis of the complex. These results are compared and discussed with previous models obtained by biochemical analysis [3].

Keywords : Protein complex assembly, Protein-Protein Interaction network, System Biology

References

[1] Glatigny, A., Mathieu, L., Herbert, CJ., Dujardin, G., Meunier, B., Mucchielli-Giorgi, MH. (2011) *BMC Syst Biol*, 5 (1) 173.
[2] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. *Genome Research* 2003 Nov; 13(11):2498-504
[3] Zara V, Conte L, Trumpower BL. *Biochim Biophys Acta.* 2009 Jan; 1793 (1):89-96.

The Universal Protein Resource (UniProt)

Reference Proteomes – New website

Benoît BELY[1], Rachel JONES[1], Sangya PUNDIR[1] and Maria JESUS MARTIN [1]

[1] The EMBL outstation - The European Bioinformatics Institue, Wellcome Trust Genome Campus, Hixton, Cambridge CB11 1SD, UK

{bbely, rjones, spundir, martin}@ebi.ac.uk

Keywords: Protein databases, uniprot.org, reference proteome, complete proteome.

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).

The UniProt Knowledgebase (UniProtKB) 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.

In the past year UniProt teams have worked on several specific projects, including the complete and reference proteome datasets and a new website design.

2 Collection of protein sets for complete genomes

With the significant increase in the number of complete genomes sequenced, it is essential to organize these data in a way that allows users to effectively navigate the corresponding increase in the number of complete proteome sequences available. The approach adopted by UniProt to meet this challenge is to designate a sub set of the UniProtKB complete proteomes as UniProtKB reference proteomes.

2.1 Complete proteomes

A complete proteome is defined as the entire set of proteins expressed by a specific organism. The majority of the UniProt complete proteome sets are based on the translation of a completely sequenced genome, and will normally include sequences that derive from extra-chromosomal elements such as plasmids or organellar genomes in organisms where these occur.

UniProt complete proteome sets may include both manually reviewed (UniProtKB/Swiss-Prot) and unreviewed (UniProtKB/TrEMBL) entries. Currently, the majority of UniProt complete proteomes () are based on translations of genome sequence submissions to the International Nucleotide Sequence Database Consortium (INSDC). A complementary pipeline for import of protein sequences has been developed in collaboration with Ensembl that provides proteome sequences for a number of key genomes of special interest for missing ORF annotation in INSDC.

2.2 Reference proteomes

UniProt has defined a set of reference proteomes that are ‘landmarks’ in proteome space. Reference proteomes have been selected to provide broad coverage of the tree of life, and constitute a representative cross-section of the taxonomic diversity to be found within UniProtKB. They include the proteomes of well-studied model organisms and other proteomes of interest for biomedical and biotechnological research. These are the proteomes that are preferentially selected for manual curation when resources permit.

Currently (release 2012_06), UniProt has defined 563 reference proteomes in close collaboration with Ensembl and NCBI Reference Sequence collection (RefSeq). The collaboration's goal is that all three resources provide the same consensus sets. The reference proteome set will be continuously reviewed as new proteomes of interest become available and as existing taxonomic classifications are revised. We would welcome and encourage input from our user community on our current list of reference proteomes and suggestions for new candidates.

2.3 Access and availability

The complete (http://www.uniprot.org/taxonomy/?query=complete%3ayes) and reference (http://www.uniprot.org/taxonomy/?query=reference%3ayes) proteomes have been available on the UniProt web and ftp sites since September 2011. In the future a new portal will also provide users with information and simple statistics for both complete proteomes and their individual components.

To make our reference proteome dataset more accessible we plan to include a new branch on our FTP site by the end of the year. The new repository should contain two fasta files per reference proteome; one with the canonical proteome set and an additional 'extended' set including all of the isoform products. We already provide a subset of reference proteomes in this format for the quest for ortholog project (QfO) ftp://ftp.ebi.ac.uk/pub/databases/reference_proteomes/

3 New website design.

3.1 UniProt Usability Review and Workshops

The UniProt web team have conducted two workshops (10-14 users) and 56 one-to-one usability tests, of the current website and new design, in many places around the E.U. and U.S.A. The users recruited were those who use UniProt (or have used UniProt databases in the past) and work in a wet lab or closely with wet lab personnel or Bioinformaticians. For this meeting we were not looking for programmers or database personnel that download or parse UniProt databases to include in their institutions internal computational resources and pipelines.

2.2 Consequences of the usability studies for the redesign of our website

The UniProt website (www.uniprot.org) already contains a large number of tools to do various tasks such as search from keyword/sequence, apply filter, customize view, download data/sequence/mapping, align sequences, give feedback, etc. The usability reviews revealed that lot of functionalities that users like to have are already included on the UniProt website, but that users can't find or see them. The main reasons for this were bad naming of tools, not noticeable or not intuitive. In view of these findings, we decided to redesign the UniProt website according to what users expect to have, with the hope that it will be more intuitive to use. New functionalities will be added to the website as a result of the usability studies, such as an isoform representation and feature browser.

The redesign of the UniProt website is also undergoing usability testing to validate the new design. A second round of tests and workshop will run on the new design and we encourage people to contact us if interested on these.

Acknowledgements

Informations in this document have been compiled from data and work done by: Borisas Bursteinas, Mark Bingley, Alan Da Silva, Jie Luo, Steven Rosanoff, Eleanor Stanley and Hermann Zellner. Thanks to Hélène Dauchel as well who kindly accepted our invitation to be part our E.U. workshop.

References

[1] The UniProt Consortium, Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res*., 40(D1): D71-D75, 2012.

Multiple Protein Structure Comparison using sequence alignment approaches

Sylvain LEONARD[1], Jean-Christophe GELLY[1], N. SRINIVASAN[2], Alexandre G. de BREVERN[1], Agnel Praveen JOSEPH[1]

[1] INSERM, UMR_S 665, DSIMB, Université Paris Diderot, Sorbonne Paris Cité 7, INTS, 6, rue Alexandre Cabanel, 75739 Paris Cedex 15, France.
{sylvain.leonard, jean-christophe.gelly, alexandre.debrevern, agnel.praveen}@univ-paris-diderot.fr

[2] MOLECULAR BIOPHYSICS UNIT, Indian Institute of Science, Bangalore 560012, India.
ns@mbu.iisc.ernet.in

Abstract. *Multiple structural alignments are essential for analysis of function, evolution and architecture of protein structures. In this study, we have developed a method for multiple structure comparison largely based on sequence alignment techniques. For this purpose, we propose a new web server called mulPBA (multiple Protein Block Alignment: www.dsimb.inserm.fr/dsimb_tools/mulpba/). This server implements a method based on a structural alphabet to describe backbone conformation of protein chain in terms of dihedral angles. A progressive alignment strategy similar to CLUSTALW was adopted for multiple PB sequence alignment (mulPBA). Highly similar stretches identified by the pairwise alignments are given higher weights during the alignment. The residue equivalences from PB based alignments are used to obtain a three dimensional fit of the structures followed by an iterative refinement of the structural superposition. The alignment quality is better than that of HOMSTRAD, MUSTANG and MULTIPROT in more than 85% of cases.*

Keywords. Protein Structure Comparison, Protein Blocks, Dynamic Programming.

1 Introduction

Multiple structure comparisons are essential to obtain a simultaneous comparison of a group of structures, which is also a critical step in many modeling and threading approaches. Most 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 (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. The pairwise alignment (*iPBA*) [3, 4] was carried out with the use of an anchor-based dynamic programming algorithm which first identifies all high scoring and structurally favorable local alignments (anchors) and then aligns the segments between them to obtain a global alignment. Here we extent the *iPBA* approach for multiple structure comparison and present a web-server for implementation of this method, namely *mulPBA*.

2 Methods

Atomic coordinate sets are first translated into sequence of Protein Blocks. The pairwise alignments are obtained using *iPBA* which performs an anchor based alignment by finding structurally conserved regions, identified as local alignment. This is followed by a progressive multiple sequence alignment strategy similar to CLUSTALW. A guide tree was used to guide the assembly of sequences based on the degree of similarity. The structurally conserved regions in the pairwise alignments are given higher weights during the progressive alignment. The alignment quality was calculated by different measures, one of them is N_{dist}, which is a weighted average of the number of columns associated with the distance cut-offs of 3.0Å, 4.0Å, 5.0Å and 6.0Å, calculated in a similar way as that of GDT score.

3 Results

The quality of alignments generated by *mulPBA* was compared with other popular methods available. An average gain of 84.7% in alignment quality was obtained with respect to the alignments in HOMSTRAD dataset. The alignments were also better than MULTIPROT in the same dataset, for about 87% of cases (Figure 1A). Assessments have also been carried out on a small dataset of 50 non-trivial cases randomly chosen from the twilight set in the SABMARK dataset. Alignments generated by methods like MUSTANG, MULTIPROT and 3DCOMB were used for comparison. 48 (96%) cases of alignments were of better quality than MUSTANG and 44 (88%) were better when compared to MULTIPROT. The difference was less striking with the recent 3DCOMB methodology with only 29 (58%) cases was of better alignment quality. In the SABmark dataset, about 6 of the alignments generated with mulPBA had large decline in the alignment quality (scores > 5) with respect to 3DCOMB. Most of the cases involved inherent flexibility of structures where the equivalences reflected in the PB alignments were not captured efficiently in the 3D structural fit. A dedicated *mulPBA* webserver is available at www.dsimb.inserm.fr/dsimb_tools/mulpba/ (see Figure 1B).

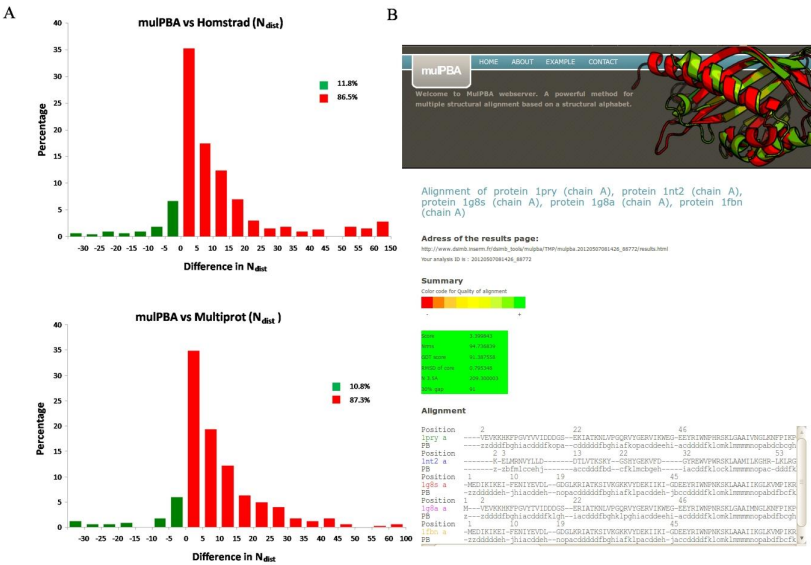


Figure 1. (A) Comparison of mulPBA with HOMSTRAD and MULTIPROT. Difference in N_{dist} scores in plotted. (B) mulPBA webserver output.

Acknowledgements

This work was supported by grants from CEFIPRA (3903E), the 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, G. Agarwal, S. Mahajan, J.-C. Gelly, L.S. Swapna, B. Offmann, F. Cadet, A. Bornot., M. Tyagi, H. Valadié, B. Schneider., C. Etchebest, N. Srinivasan and A.G. de Brevern, A short survey on protein blocks. *Biophys Rev*, 2:137-145, 2010.

[3] A.P. Joseph, N. Srinivasan and A.G. de Brevern, Improvement of protein structure comparison using a structural alphabet. *Biochimie*, 93:1434-1445, 2011.

[4] 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 Acids Res*, 39:W18-23, 2011.

SNPs and indels detection in resequenced genomes of Black truffle of Périgord

Thibaut PAYEN[1], Anaïs GIGANT[1], Emmanuelle MORIN[1],

Claude MURAT[1], Francis MARTIN[1]

[1] IAM, UMR 1136 INRA/Université de Lorraine, Champenoux, France

1 Introduction

Truffles (*Tuber melanosporum)* are hypogeous ectomycorrhizal fungi highly appreciated for their delicate organoleptic properties. Genetic diversity and population structure of this species is debated since many years. Taken advantage of the recent sequencing of the *T. melanosporum* genome [1], Murat and colleagues [2] developed microsatellite markers. These neutral markers highlighted an unexpected genetic diversity for this truffle but did not allow investigating adaptation of this species to symbiosis and environmental stresses. The aim of this study was to assess the genomic variability of *T. melanosporum* by resequencing the genome of six different isolates. Such an approach was innovative for a filamentous mycorrhizal and is the basis toward a future population genomic analysis.

2 Material and methods

In this project, six *T. melanosporum* isolates were harvested in different countries and ecological conditions (Spain, Italy and France). The genome of these six isolates was sequenced using the Illumina HiSeq in order to generate a coverage of 20x. The reference genome correspond to an ascocarp harvested in Provence (France) and named mel28. Its 125 Mb genome is composed of about 60% of repeated elements and 7500 protein-coding genes.

We have tested several mapping software (MAQ, Bowtie, BWA) and decided to use the combo BWA [3] -SAMTools [4] was chosen for sequence mapping of the six *T. melanosporum* genome reads and identification of SNPs/indels. For the mapping the parameters of BWA were the default parameters except for the number of mismatch allowed that was set up to 2. After the mapping, a list of SNPs and short indels was created using SAMTools. Different python scripts were written to extract information on the mapping (e.g. genome coverage), number of SNPs and indels and the localization of these polymorphisms in the different genomic regions.

3. Results and discussion

For each of the six *T. melanosporum* isolates 30 to 39 millions reads of 76 bp were generated by Illumina sequencing. Using BWA from 83% to 90 % of the reads mapped against the reference genome. The percentage of the reference genome non-

covered range from 2.1% to 2.6 %. The genomic region where reads did not map can correspond to polymorphic regions where the stringent criterion of 2 mismatches allowed by reads (less than 3%) did not allow the mapping of the reads or to genomic regions that are absent in the resequenced genomes, i.e. indels.

For each newly sequenced genome between 143,304 and 244,331 SNPs and between 2,094 and 2,965 indels were detected against the reference genome. Compiling all the data, a total of 526,792 SNPs and 6,220 indels were detected. The density of SNPs in the whole-genome is 4,213 SNPs per Mbp whereas there are 50 indels per Mbp. Transposable elements presented the highest density of polymorphism. Interestingly, indels were also frequent in UTR and introns confirming previous our results on microsatellites [2]. The density of SNPs and indels was lower in the coding regions indicating that the selective pressure is more important in these regions. However, 2,362 gene models were polymorphic in their coding regions.

Some important pathways have been searched for highly polymorphic genes and sulfur assimilation seems to have higher polymorphism level than other pathways. The sulfur assimilation is involved in the organoleptic qualities of truffles and this finding suggested that differences of taste among truffles can have a genetic origin.

4. Conclusion

To our knowledge this is the first study of large-scale polymorphism detection by genome resequencing for a mycorhizal fungus. The first challenge was to map 76 bp reads against the 125 Mbp genome of *T. melanosporum* that is rich in repeated elements (~60%). In the present study, a total of 526,792 SNPs and 6,220 indels were detected comparing the genome of seven *T. melanosporum* isolates. The density of these polymorphisms in the different genomic regions was variable suggesting a selection against polymorphism in coding regions. A more detailed investigation of the polymorphic genes allowed identifying several candidate genes that will be investigated to assess their involvement in truffle adaptation.

5. Acknowledgement

We would like to thank Henri Frochot, Mario Honrubia, Francesco Paolocci, Bernard Vonfli for providing us the *T. melanosporum* ascocarps. We are grateful to Stephane De Mita for its stimulating discussions and suggestions. This project was financed by Europeans Commission (Evoltree and Ecofinders projects) and Lorraine region.

6. References

[1] F. Martin et al. Périgord black truffle genome uncovers evolutionary origins and mechanisms of symbiosis. *Nature*, 464(7291):1033–8, Apr. 2010.
[2] C. Murat et al. Distribution and localization of microsatellites in the Perigord black truffle genome and identification of new molecular markers. *Fungal Genetics and Biology* 48.
[3] H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754--60, July 2009.
[4] H. Li et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078--9, Aug. 2009.

BioRepo : Biological Repository

A LIMS for HTS data

Yoann MOUSCAZ[1,2,*] , Adamandia KAPOPOULOU[1,2,*], Marion LELEU[1,2,*], Yohan JAROSZ[1,2], Jacques ROUGEMONT[1,2,†], Denis DUBOULE[1,†] and Didier TRONO[1,†]

[1] School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
{yoann.mouscaz, adamandia.kapopoulou, marion.leleu}@epfl.ch
[2] Swiss Institut of Bioinformatics (SIB), Lausanne, Switzerland
*, † These authors contributed equally to this work

Abstract *The rapid increase of genomics data generation, especially with the expansion of High Throughput Sequencing (HTS) technologies, brings biologists and bioinformaticians to find solutions to efficiently manage and store the resulting amount of data. BioRepo is a Biological data Repository which addresses those needs. Similarly to a Laboratory Information Management System (LIMS), its main goals are to store, to manage and to share data among collaborators, but also to allow their direct visualisation in genome browsers.*

Keywords biology, genomics, data repository, high throughput sequencing, database, storage, management, security, LIMS, TurboGears, python

1 Introduction

Biology and Bioinformatics are two fields in constant evolution. This evolution leads to the development of solutions to store data produced by High throughput sequencing (HTS).

A Laboratory Information Management System (LIMS) facilitates data management. However, a secure access needs to be implemented to keep data confidential until publication. In addition, HTS data often have to be visualised in genome browsers.

In this work we present BioRepo, a LIMS specifically designed for these HTS data but adaptable to other data types.

2 BioRepo functionalities

BioRepo offers a data-tracking system with a flexible architecture and a smart data exchange interface. These functionalities can be grouped in three categories : the storage, the management and the sharing of the data.

2.1 Storage

Users can store any file format and any kind of data in the database, but restrictions on the extensions and/or on the file size can be set up for a specific installation. In power terms, BioRepo can be installed on a relatively small server (on Linux, Mac or Windows).

2.2 Management

BioRepo management system was designed by bioinformaticians and biologists. The database has been designed to follow the life cycle of a typical HTS project, at the data level. Three main levels have been defined as illustrated on Figure 1: Project, Sample and Measurement. A measurement represents a final dataset and can be associated to one or several samples. A sample is attached to an unique project and contains one or several measurements. Finally, if a measurement combines several previous measurements, it is possible to keep track of their filiation.

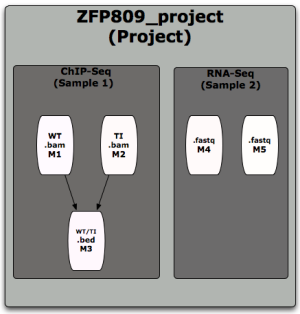


Figure 1. BioRepo management system

After login for the first time, BioRepo creates user-specific directories required on the server and associates them to the user account. To maximize storage space, data are organized in such a way that files are not duplicated as long as they are identical (The SHA1 cryptographic hash function was used to ensure this uniqueness). If two identical files are loaded, then a symbolic link is created.

2.3 Sharing and visualisation

BioRepo offers some file sharing and secured access functionalities. A file can be added to BioRepo either with an open access, meaning accessible to everybody, or with a restricted access. In that case, only some pre-defined users are allowed to perform certain operations such as export or visualisation of the file.

The current version is linked to Tequila, the EPFL login system, but it is possible to link BioRepo with any other security module.

In addition, BioRepo offers users the possibility to directly visualise one or several selected genomics data into a genome browser of their choice. Currently, links to two browsers are proposed: the UCSC genome browser and GDV, a Genomics Data Viewer developed at EPFL, but the list can be extended in future.

3 Search

An efficient search functionality has been developed to allow a full text search and/or a category-based search . It has been implemented with the JavaScript library, jQuery[1] which allows fast results visualisation sorted in a selectable datatable [2].

4 Technologies used

BioRepo is powered by TurboGears technology[3] which combines python language and SQLAchemy[4] for the database part. Moreover it is possible to use any type of database management system (e.g., SQLite, PostgreSQL, ...). In addition, some javascript code has been written using the jQuery library essentially for the search methods.

References

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

[2] http://www.datatables.net/

[3] http://www.turbogears.org/

[4] http://www.sqlalchemy.org/

Oncogenic mutations of KIT receptor differentially modulate tyrosine kinase activity and drug resistance

Isaure Chauvot de Beauchêne[1] , Elodie Laine[1] and Luba Tchertanov[1]

[1] BiMoDyM, LBPA, UMR8113 CNRS - École Normale Supérieure de Cachan, LabEx LERMIT, 61 avenue du Président Wilson, 94235, Cachan, France

{ichauvot,elodie.laine,luba.tchertanov}@lbpa.ens-cachan.fr

PSEUDOE: A computational method to detect Ψ-genes and explore PSEUDome dynamics in wine bacteria from the *OEnococcus* genus.

Laetitia BOURGEADE[1a,2], Tiphaine MARTIN[1b,2] and Elisabeth BON[1a,2]

[1] Laboratoire Bordelais de Recherche Informatique, 351 cours de la Libération, 33400 Talence, France
[1a]Univ. Bordeaux / [1b] CNRS : LaBRI, UMR5800.

[2] INRIA, Team MAGNOME (joint with LaBRI), 33400 Talence, France.

{Laetitia.bourgeade, tiphaine.martin, elisabeth.bon}@labri.fr

Abstract *By developing PSEUDOE, a suite of semi-automatic computational procedures, and by performing some genomic excavations, we analyzed the pseudogene (Ψ_{gene}) inventories in two distantly-related sequenced O. oeni strains with opposite oenological properties. Besides the discovery of the natural propensity of O. oeni genes to undergo pseudogenization, we shed the light on the first outlines of the pan-pseudome architecture and the rules that could govern gene disruption dynamics, providing additional insights into the shaping of gene/genome evolution from which future genome annotation will benefit.*

1 Introduction

Among lactic acid bacteria (LAB), the wine-making bacterium *Oenococcus oeni* is unique in terms of its remarkable tolerance to harsh wine conditions, and propensity for undergoing extensive genetic and phenotypic variations. Harsh conditions of life often lead to modifications in the utility of some bacterial genes. Pseudogenes (Ψ_{gene}) therefore constitute a useful "fossil" testimony of ecological niche pressure adaptation. To elucidate the role of Ψ_{gene} into the adaptation of *O. oeni* to wine, we developed a suite of computational procedures (PSEUDOE: *PSEUDOgenes finder for OEnococcus*) designed to identify Ψ_{gene} and we investigated for the first time the size, distribution and characteristics of the Ψ_{gene} populations, into two distantly-related sequenced *O. oeni* strains (~1.8 Mb, 38%-GC) both sharing opposite oenological properties. Besides computational contributions in Ψ_{gene} genomic excavations, the PSEUDOE results give us an unprecedented opportunity to crack the pseudome code and document the dynamics of gene exclusion during the process of reductive-evolution and *O. oeni* adaptation to wine.

2 Methodology

Genetic code sources: The complete genome sequences (gap-filled, published, primary annotated) of the *O. oeni* strains PSU1 (commercial starter, [1]) and ATCC BAA1163 (basic oenological traits, [2]) were used as matrices to crack the pseudome code.

Computational procedures: PSEUDOE combines *BLAST* genomic explorations with comparative analyses within a pair of related genomes to excavate Ψ_{gene} originating from point mutations and protein-coding gene (CDS) truncations. Written in *Python*, the workflow is designed into 4 modules (M1, M2, M3, M4): "*autoBLAST*" and reciprocal *Best BLAST Hit* search strategies in order to check primary genomic features (limits redefinition, functional re-annotation) and to recover undetected redundant or degenerated features. M1 identifies, through the neighbouring states, all possible "ambiguous blocks" ($ab1_{gene}$) i.e. artificial and temporary series of adjacent objects susceptible to be mergeable. M2 implements a reciprocal *Best BLAST Hit* search strategy on $ab1_{gene}$ in order to generate the final catalogue of all putative Ψ_{gene} of the target genome. In an original way, M3 was developed to evaluate the number of intra- and inter-genomic copies of a given Ψ_{gene} and, to investigate its redundancy within a given genome as well as its evolutionary conservation both at intra- or inter-species levels. Robustness of PSEUDOE method was evaluated against Ψ-Φ program [3].

Pseudome source code rules: CDS disrupting events (mutations, truncations, fusions) were crossed and used as criteria to define Ψ_{gene} populations. The prevalence of Ψ_{gene} was compared between *O. oeni* strains and

Vitamin K epoxide recognition by vitamin K epoxide reductase

Florent LANGENFELD[1], Isaure CHAUVOT DE BEAUCHENE[1], Etienne BENOIT[2] and Luba TCHERTANOV[1]

[1] Laboratoire de Biologie et Pharmacologie Appliquée, UMR8113 CNRS, 61 avenue du Président Wilson, 94235, Cachan, Cedex, France

{florent.langenfeld, isaure.chauvot-de-beauchêne, luba.tchertanov}@lbpa.ens-cachan.fr

[2] USC1233 INRA-Vetagro Sup, École Vétérinaire de Lyon, 69280 Marcy l'Étoile, France

e.benoit@vetagro-sup.fr

1 Introduction

The vitamin K cycle occurs in the endoplasmic reticulum membrane. Several glutamic acids of selected coagulation factors are γ-carboxylated by the γ-glutamyl carboxylase, which oxidizes vitamin K hydroquinone (VKH2) to vitamin K epoxide (VKO) to generate functional clotting factors [1]. Vitamin K epoxide reductase (VKOR) plays a key role in this cycle by catalyzing the regeneration of VKH2 from VKO.

VKOR reduces VKO into vitamin K quinone (VK1) and further VK1 into VKH2. These reduction steps are catalyzed through an electron transfer mediated by four cystein residues (C43, C51, C132 and C135) from a protein partner (TMX, TMX4, ERp18…) to the final electron acceptor, vitamin K [2]. The structural data of the human VKOR are not currently available and their absence impedes to confirm the putative reaction [3], which supposed that a covalent intermediate is formed between a cystein residue of the active site $C_{132}XXC_{135}$ motif and the VKO carbon 3 [Fig. 1 a]. In this study, we have carried out *in silico* study of the VKO/VKOR complex with two aims: first, to describe the VKO – VKOR recognition and second, to confirm that the covalent intermediate is likely to be formed for the reaction.

2 Methods

The homology model of the human Vitamin K epoxide reductase was generated from the crystal structure of its bacterial homolog reported by Li *et al.* [4] (PDB entry: 3KP9) and two others templates (PDB entries: 3IZS and 3D61). An induced-fit docking of the vitamin K epoxide onto VKOR was performed using Maestro (Schrödinger LLC). The obtained complex was embedded into a di-oleoyl-phosphatidyl-choline bilayer. Molecular dynamic (MD) simulations (26 ns) of constructed system were run using the 99sb [5] and ILDN [6] parameters sets of the amber force field implemented in GROMACS 4.5 [7].

3 Results

Our model presents four transmembrane helices (TM1 to TM4) connected by short coiled linkers, and one long loop (1/2-loop) between TM1 and TM2 that contains a short helix (1/2-helix) lying on the lipid surface [Fig. 1 b, c]. The loop with 1/2-helix positioned as a lid on the four-helix bundle is bound to TM2 helix by the H-bonds between N80 and W59 and/or G62 residue observed for 99 and 49% of the simulation time, respectively [Fig 1 c, d].

We evidenced that VKO is located in the VKOR cavity formed by the four transmembrane helices and the 1/2-helix shielding VKO from the endoplasmic reticulum space. The active site $C_{132}XXC_{135}$ is exposed towards VKO. The observed VKO binding mode is achieved by two H-bonds: Y139 interacts with one quinone oxygen atom of VKO and N80 interacts with the epoxide ring oxygen atom (70 and 80% of the simulation time, respectively) [Fig 1 d].

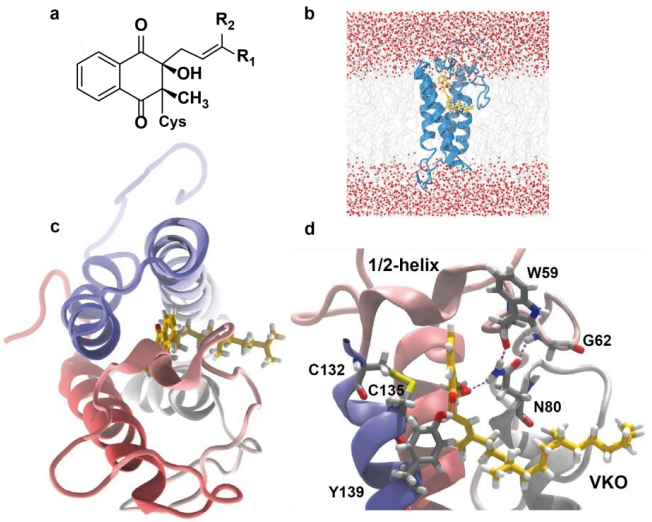


Figure 1. Molecular dynamic simulation of the VKO binding into VKOR. (a) The putative covalent intermediate formed during VKO reduction by VKOR; (b) General view of the VKO/VKOR complex on the lipid surface; (c) Structure of the VKO/VKOR complex; (d) Binding mode of VKO with VKOR. VKOR is shown as cartoon. VKO, catalytic residues (C132, C135) and residues involved in stabilizing H-bond interactions (W69, G62, N80 and Y139) are shown as sticks. The water molecules are drawn as ball-and-sticks and lipids are represented by thin lines. H-bonds are shown as the dashed lines.

These H-bond interactions influence on the VKO orientation respectively to the catalytic residues. Particularly, the VKO carbon atom in the 3rd position is positioned at 5 Å from the C135 sulphur atom. The epoxide ring oxygen atom is oriented outside from the active site $C_{132}XXC_{135}$, allowing the nucleophilic attack of VKO by the putative charged cystein residue according to the proposed reaction scheme.

Altogether, these results suggest that N80 plays a key role in the physiological function of VKOR. This residue is involved in the polyfunctional H-bond interactions which stabilize first, VKO in the binding site, secondly, the 1/2-helix as a lid respectively to the active site and third, the epoxide ring position outside the active site. *In vitro* studies confirm the essential role of N80 in the VKOR activity.

References

[1] B. Furie, B. A.Bouchard, B. C. Furie.Vitamin K-dependent biosynthesis of γ-carboxyglutamic acid. *Blood*, 93:1798-1808, 1999.

[2] D. Jin, J. K. Tie, D. W. Stafford. The conversion of vitamin K epoxide to vitamin K quinone and vitamin K quinone to vitamin K hydroquinone uses the same active site cysteins. *Biochemistry*, 46:7279-7283, 2007.

[3] C. H. Davis, D. Deerfield II; T. Wymore, D. W.Stafford, L. G. Pedersen, A quantum chemical study of the mechanism of action of Vitamin K epoxide reductase (VKOR) II. Transition states. *J Mol Graph Model.*, 26:401-408, 2006.

[4] W. Li, S. Schulman, R. J. Dutton, D. Boyd, J. Beckwith, T. A Rapoport. Structure of a bacterial homologue of vitamin K epoxide reductase. *Nature*, 463:507-513, 2010.

[5] V. Hornak,*R.* Abel, *A.* Okur, *B.* Strockbine, *A.* Roitberg, *C.* Simmerling. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins*, 65: 712–725, 2006.

[6] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, D. E. Shaw. Improved side-chain torsion potentials for the Amber99SB protein force field. *Proteins*, 78:1950-1958, 2010.

[7] B. Hess, C. Kutzner, D. van der Spoel, E. Lindahl. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comp*. 4:435–447, 2008.

Optimizing phenotypic plasticity in a fluctuant environment

Nicolas NOTTET[1], Patrick COQUILLARD[1], Alexandre MUZY[2] and Francine DIENER[3]

[1] SOPHIA AGROBIOTECH, UMR 7254 CNRS-INRA-Université de Nice, 400 route des Chappes, 06903, Sophia Antipolis, France. nicolas.nottet@gmail.com ; patrick.coquillard@unice.fr

[2] LIEU IDENTITE ESPACES & I3S, UMR 7271 BioInfo Team, Université de Nice, Bât. Algorithmes-Euclide, 2000 route des Lucioles, 06903, Sophia Antipolis, France. muzy@i3s.unice.fr

[3]LABORATOIRE MATHEMATIQUES J.-A. DIEUDONNE, UMR 6621 CNRS, Université de Nice, Parc Valrose, 06108, Nice Cédex 2, France ; francine.diener@unice.fr

The GEMO project: Genomic Evolution of the fungal pathogen *Magnaporthe oryzae*.

Ludovic MALLET[1], Cyprien GUÉRIN[1], Véronique MARTIN[1], Enrique ORTEGA-ABBOUD[2,3], Jonathan KREPLAK[4], Joelle AMSELEM[4,5], Annie GENDRAULT[1], Marc-Henri LEBRUN[5], Thomas KROJ[3], Didier THARREAU[2], Elisabeth FOURNIER[3] and Hélène CHIAPELLO[1]

[1] INRA, UR MIG, 78352 Jouy-en-Josas, France
helene.chiapello@jouy.inra.fr
[2] CIRAD, UMR BGPI, TA 54K, 34398 Montpellier
[3] INRA, UMR BGPI, TA 54K, 34398 Montpellier
[4] INRA, URGI, 78026 Versailles, France
[5] INRA, UMR BIOGER, 78850 Thiverval-Grignon, France

Keywords Comparative genomics, Evolutionary genomics, High Throughput Sequencing, Transposable elements, *Magnaporthe oryzae*.

Abstract

Magnaporthe oryzae is a successful pathogen of crop plants and a major threat for food production worldwide. This species gathers pathogens of different Poaceae, including rice and wheat and causing the main fungal disease of rice worldwide.The Evolutionary Genomics of *Magnaporthe oryzae* (GEMO) project, aims to characterize the genomic determinants and the evolutionary events involved in pathogenesis, host specificity and adaptation to the host. Eight strains of *M. oryzae* and one strain of the sister species *M. grisea* isolated from plants in different countries were selected [1] along with the public reference strain *M. oryzae* 70-15 [2] to represent different virulence phenotypes and genetic groups pathogenic of different species of Poaceae. Transposable elements (TE), horizontal transfers (HT), small secreted proteins (SSP) and gene expression are keys features for phenotype evolution and genome plasticity. Comparative and evolutionnary analyses at the whole-genome level will target, with the purpose to unveil the determinants of pathogenicity and host adaptation in this *Magnaporthe* species.

The nine genomes have been sequenced using HTS technologies (454 and Solexa/Illumina) and assembled by the Genoscope (Evry, France). The genes have been predicted by the EuGene software [3], trained, fitted and benchmarked with each one third of a known 300 genes/cDNA pairs. Sensitivity of the prediction is 81.8% for the genes and 89.4% for the exons, while specificity is 81.9% for the genes and 92.9% for exons. Results of annotation prediction are presented in Table 1 on the following page. Small secreted proteins may help a pathogen to take control of the host's metabolism and defense [4]. We used the SignalP 4.0 software [5] to predict such effectors and enhance the functional annotation. We found that 9 to 11% of the genes potentially contain a signal peptide and that 10 to 13% are SSP-like (see Table 1 on the next page). Taking into account strain genome size and gene content, SSP proportion appears to be quite similar in all the ten genomes. TE are currently being characterized and annotated with the REPET package [6]. Firsts analyses revealed that about 10-12% of the genomic DNA are mobile elements. Most frequent families are retro-transposons (class I TE) LTR but some class II transposons were also found. Predicted CDS were translated and the products were clustered with OrthoMCL [7]. Within the ten genomes, 20443 different groups were identified from which the core genome was found to represent 39%. The number of genome-specific genes is highly variable depending on the considered genome and ranges from 305 in TH16 to 1,550 in GY11. The contingency and distribution of accessory and genome-specific genes are presented in Table 1 on the following page. Unexpectedly, three strains of *M. oryzae* that were annotated with the same pipeline display a significant higher amount of accessory genes. These genes are indeed grouped together by orthoMCL and are common and specific of the three strains GY11, PH14 and TH12. As the boolean presence of this whole set of groups does not seems to be directly linked

to host specificity, further investigations will be led on the potential evolutionary origin and the implication on host adaptation of these specific genes.

A dedicated database is currently being developed using standard open source tools. Data and result summary will be provided in a genome browser and on a dedicated website.

Genomes	*70-15	†BR29	BR32	CD156	FR13	GY11	TH12	PH14	TH16	US71
Main host	*rice*	*rice*	*wheat*	*Eleusine*	*rice*	*rice*	*rice*	*rice*	*rice*	*Setaria*
Sequencing										
Size (Mb)	40.9	40.9	41.9	42.7	43.1	46.3	48.5	49.8	39.1	41.2
GC%	51.51	47.88	48.29	47.6	39.81	49.1	50.33	47.87	48.55	48.31
Nb contigs	216	9,644	6,044	26,535	79,619	13,188	9,908	11,772	4,114	7,398
Nb scaffolds	53	169	111	237	2,051	1,964	940	711	171	220
Depth coverage (x)	‡ 7	60	55	50	4	42	48	56	53	80
Scaff. N50 (kb)	6,607	955.4	1,760.5	1,066.4	101.6	187.3	590.5	597.1	938.5	813.9
% N	0.18	4.09	4.96	6.59	22.55	7.98	5.83	10.23	5.80	5.45
Annotation										
% Unannotated DNA	37.0	52.2	42.2	43.5	54.1	39.7	38.1	41.0	41.3	42.6
Predicted genes	12,827	12,616	14,781	14,415	15,035	20,621	19,811	20,067	13,725	14,013
Nb 5'UTR	9,084	1,637	5,070	5,068	3,989	4,808	5,042	5,049	5,090	5,126
Nb 3'UTR	6,666	1,581	6,674	6,722	5,440	6,376	6,678	6,678	6,731	6,773
Mono-exonic genes	2,507	3,755	4,184	4,268	3,094	6,122	7,297	6,695	3,800	2,376
CDS $\leq$ 300bp	1,274	1,537	2,541	2,473	3,095	3,580	2,827	3,029	2,296	4,172
Differential genomic										
Accessory genes	3,635	2,632	5,530	5,269	5,344	10,773	10,548	10,502	5,112	5,092
Specific genes	968	1,506	665	644	1,402	1,550	772	1,210	305	557
Small Secreted Proteins										
15 aa $\leq$ SP $\leq$ 30 aa	1,456	1,391	1,646	1,627	1,451	1,751	1,847	1,832	1,531	1,579
No SP, 1 N-Ter TMD	237	250	269	277	317	406	363	374	252	276
Total SSPs	1,693	1,641	1,915	1,904	1,768	2,157	2,210	2,206	1,783	1,855
% of genes	13.03	13.35	13.35	13.54	11.87	10.55	11.35	11.09	13.14	13.44

Table 1. Statistics and results for genome sequencing, gene prediction, comparative genomics and small secreted proteins. * The public reference strain 70-15 (MG8 assembly) was sequenced, assembled and annotated by the International Rice Blast Genome Consortium. † BR29 genome is from the *M. grisea* species, all the other genomes are from *M. oryzae* isolates. ‡ *M. oryzae* 70-15 has been sequenced with the Sanger technology. SP : Signal peptide. TMD : Transmembrane domain.

Acknowledgements

This work was funded by the French national agency for research (ANR 2009-GENM-029-01).

References

[1] Tharreau D, Fudal I, Andriantsimialona D, Santoso, Utami D, Fournier E, Lebrun M and Notteghem J. World population structure and migration of the rice blast fungus, *Magnaporthe oryzae*. In *Advances in genetics, genomics and control of rice blast disease / Wang Guo-Liang (ed.), Valent Barbara (ed.)*, pp. 209–215. Springer [Etats-Unis], New York 2009

[2] Dean RA, Talbot NJ, Ebbole DJ, Farman ML, Mitchell TK *et al.* The genome sequence of the rice blast fungus *Magnaporthe grisea*. *Nature*, 434(7036):pp. 980–986 2005

[3] Schiex T, Moisan A and Rouzé P. Eugène: An eukaryotic gene finder that combines several sources of evidence. In *JOBIM*, pp. 111–125 2000

[4] Klee EW and Ellis LB. Evaluating eukaryotic secreted protein prediction. *BMC Bioinformatics*, 6(1):p. 256 2005

[5] Petersen TN, Brunak Sr, von Heijne G and Nielsen H. Signalp 4.0: discriminating signal peptides from transmembrane regions. *Nature methods*, 8(10):pp. 785–6 2011

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

[7] Li L, Stoeckert CJJ and Roos DS. Orthomcl: identification of ortholog groups for eukaryotic genomes. *Genome Res*, 13(9):pp. 2178–89 2003

Improving the computation of guide trees for genome multiple alignments in Ensembl Compara API

Nicolas FIORINI[1], Paul FLICEK[2] and Javier HERRERO[2]

[1] UNIVERSITÉ MONTPELLIER 2 Sciences et Techniques, Place Eugène Bataillon, 34095 Montpellier cedex 5, FRANCE
nicolas.fiorini@etud.univ-montp2.fr

[2] EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UNITED-KINGDOM
{flicek, jherrero}@ebi.ac.uk

Keywords Phylogenies, genome multiple alignments, duplications.

1 Introduction

The Ensembl Compara database stores the results of genome-wide species comparisons calculated for each data release. There is a pipeline computing the guide trees and the genomic alignments called EPO (Enredo [1], Pecan [1,2], Ortheus [3]). This pipeline considers genome duplications, which is not the case of another pipeline of Compara: Pecan.

A part of the database provides conservation scores and constrained elements, which are determined from EPO or Pecan whole genome multiple alignments. GERP [4] is used to compute these scores, relying on an alignment and the guide tree for this alignment. Therefore, the conservation scores and the constrained elements identification might be inaccurate when the input data are unreliable. We noticed that sometimes, especially when there are duplications (only in EPO results then), the guide tree for the alignment is not reliable. In these cases, the alignment will be affected, as well as the GERP results.

We aim at improving the tree computing phase of EPO, in order to have more reliable results. The first step of EPO is to find the best guide tree for a specific set of sequences to be aligned. To do that, EPO starts by computing a draft tree using the Neighbour-Joining method. From this tree, an initial alignment is obtained from which a new tree is built by Maximum Likelihood (ML). The new tree is used to guide a new draft alignment. This process is repeated a few times until the tree is stable, unfortunately it is prone to local maximum (or local minimum logarithm). To understand how we could resolve these problems, we first have to better understand the tree computation on genomic alignments including cases where there are no duplications.

2 Control analysis

To evaluate the impact of duplications on the guide trees, it is important to know how common are the unexpected topologies/branch lengths whether or not there are duplications. There might be some discrepancies in a computed tree without duplication, due to long branch attraction (LBA), small set of data, etc. The first step of this analysis is to understand how phylogeny works on genome alignments, on the same set of species -12 Mammals- as we will use later for EPO. This control analysis is based on the Pecan pipeline, which does not consider duplications. This method uses the species tree as guide tree as it must be the best tree to align sequences. This is not possible for the EPO method, as the species tree does not contain any duplication.

By comparing the species tree to an inferred tree from the alignment (with a ML method, the same used as EPO), we are able to understand what happens when there are no duplications. We could identify three kinds of discrepancies. The first one concerns the Primates/Rodents/Laurasiatheria groups where we expect to see the Primates group close to the Rodents group. This happens rarely, in most cases the Rodents are close to the Laurasiatheria group, probably because of a LBA. In another common topology, the Opossum is close to the Platypus. However in these cases, we noticed that the Platypus sequence was really short. When the Platypus sequence is longer, we cannot observe this pattern. Therefore, this discrepancy may be due to the bad quality

of a set of data. Finally, we also saw some cases where the Horse and the Dog are not close to each other. It actually happens when there are long *indels* within the sequences. Moreover, the branch lengths in these cases are short, showing that this subtree is not reliable.

From this manual analysis, we created a tool to automatically analyze a large set of trees (around 1000 trees) and detect these discrepancies and other unseen patterns. At the moment, we are collecting the EPO data and filtering it. In fact, some alignments provided by the EPO pipeline cannot be used in this analysis. We focus on alignments that are long enough (so the alignment will be more informative) and contain duplication(s).

3 Future work and expected results

When we will have these data, it will be interesting to study two things. First, the differences between the statisctics of Pecan and those of EPO will be informative about how duplications can affect the identified phenomena. Second, the identification of specific patterns of EPO will help us to see how the topology and branch lengths can be affected when duplications are considered. Once these discrepancies are identified and understood, we will be able to predict and correct them in order to have a better alignment.

An effective way to overcome these identified discrepancies is to use the MCMCMC (*Metropolis-Coupled Markov Chain Monte Carlo*) algorithm, because the problem in most of our cases is the reaching of a local maximum. In an MCMCMC approach, two chains explore the range of possible trees. One chain is fine-grained, the other one is coarse-grained. The coarse-grained exploration would allow big randomly topology/branch length changes while the fine-grained would explore little changes. This way, the local maximum problem will be countered. So, with these corrected trees for EPO, we could realign the genome sequences and ultimately get better GERP analysis, so a better constrained elements identification.

Acknowledgements

We thank the EMBL-EBI for funding the internship of Nicolas Fiorini as well as the Labex NUMEV (http://www.lirmm.fr/numev/) for funding the JOBIM participation.

References

[1] Paten B., Herrero J., Beal K., Fitzgerald S. and Birney E., Enredo and Pecan: Genome-wide mammalian consistency based multiple alignment with paralogs. *Genome Research*, Nov;18(11):1814-28, 2008.

[2] Paten B., Herrero J., Beal K. and Birney E., Sequence progressive alignment, a framework for practical large-scale probabilistic consistency alignment. *Genome Research*, Feb 1;25(3):295-301, 2009.

[3] Paten B., Herrero J., Fitzgerald S., Beal K., Flicek P., Holmes I. and Birney E., Genome-wide nucleotide level mammalian ancestor reconstruction. *Genome Research*, Nov;18(11):1829-43, 2008.

[4] Cooper GM et al., Distribution and intensity of constraint in mammalian genomic sequence. *Genome Research*, 15:901-913, 2005.

Modeling cell signaling pathways with discrete dynamical systems

Application to Transforming Growth Factor β (TGFβ) signaling

Geoffroy ANDRIEUX[1,2], Michel LE BORGNE[2] and Nathalie THÉRET[1]

[1] INSERM U1085-IRSET, Université Rennes1, 2 av du Pr. Léon Bernard, 35043 Rennes, CS 34317, France
nathalie.theret@univ-rennes1.fr
[2] Centre de recherche Inria Rennes, 263, 35042 Rennes, Cedex, France
{geoffroy.andrieux, michel.leborgne}@irisa.fr

Keywords systems biology, guarded transitions, signaling network, model checking

1 Introduction

Signaling is a process by which a cell reacts to an external stimulus, like cytokines and growth factors produced by other cells. Signal usually starts with the activation of cell membrane receptors by ligand binding. It triggers cascade reactions such as protein phosphorylation, to ultimately regulates target genes. Cell signaling supports cell behavior and fate, including : proliferation, apoptosis and differentation.
A myriad of molecules are engaged during this process and the final answer of the cell often seems to be hardly predictible, according to intertwined pathways. Indeed, many signal pathways have been first described with a canonical pathway leading to a unique respons. But it has been shown that numerous receptors activates different pathways according to environment and protein status in the cell. Moreover these pathways are interconnected by positive and negative regulations. Despite of great advances in components identification of these pathways, understanding cell signaling as a whole network remains a difficult task. Therefore, a signaling network is assimilated to a complex system and cannot be understood by studying separated pathways. Systems biology attempts to understand living organism taking into account their complexity, using mathematical and computational methods. In most cases, quantitative data such as concentration or kinetic rates are not available for all components of a signaling network. Indeed quantitative models, such as differential equations, can not be used and the developement of discrete models is a challenging research [1]. Due to the complexity of living systems and current knowledge in biology, we still lack of concepts to model biological networks. Several discrete formalisms have been used such as Boolean networks [2], influence graphs, Petri nets [3] ... For example influence graphs represent each molecules as a node in a graph, and the arrow represent influence between two molecules. That viewpoint focuses on nodes, but in signaling pathways the major event is the signal transduction. The chosen formalism need to describe how the signal spreads over the cell.

2 Method

Signal transduction networks are composed by different level of reactions (complexation, activation, transcriptional regulation...) and different level of components (RNA, proteins, metabolites...). The question is to understand how the signal is transmitted ? What molecules are imply ? How different pathways act together? We do not want to approximate biochemical reactions, but to model biological facts. This abstraction defines a reaction under conditions, from input biomolecules to output biomolecules. The chosen formalism called guarded transitions, comes from the computational language UML (Unified Modeling Language) [4]. This high level language is more adapted to the representation of concurrency systems, as the different pathways of signal transduction processes. Biomolecules are represented by places taking two qualitative value : active or inactive. Places are connected by transitions with a logical formula as a guard. A transition means an association between molecules, such as activation, or translocation. The guard is composed by positive and negative regulators. This formalism allows discrete time simulations, a transition is fired if the input place is active and the guard is True. When a transition is fired, the input place is desactivated and the output place is activated.

We have implemented the guarded transition formalism into a user-friendly software called CADBIOM-Chart (http://cadbiom.genouest.org/). It also proposes a translation of Pathway Interaction Database (PID)[5], to directly create a model from curated data since PID clearly defines its concepts, separating modifications from regulations. Model analysis is another feature proposed by CADBIOM-Chart. Indeed dynamical systems are translated into propositional logic in order to use SAT solver to test the reachability or invariance of properties, such as the activation of target genes.

3 Application and Results

Complex signaling by the polypeptide transforming growth factor TGFβ is one of the most intriguing networks that govern complex multifunctional profiles and TGF- signaling pathways are the most documented. TGF-β was firstly described as a potent growth inhibitor for a wide variety of cells including epithelial cells. In addition to induce growth arrest, TGF-β affects apoptosis and differentiation thereby controlling tissue homeostasis [6]. At the opposite upregulation and activation of TGF-β has been linked to various diseases, including fibrosis and cancer through promotion of cell proliferation and invasion and of the epithelial-mesenchymal transition[7]. The TGFβ pleiotropic effect is related to different activated pathways : SMAD dependent (canonical) or independent (crosstalks)[8] but the accurate impact of these pathways is still dubious. Targeting the "bad effects" of TGF-β without affecting its physiological role is a common goal of therapeutic strategies and the aim of our project deals with the challenging modeling of TGF-β signaling pathways to identify new targets.

Using PID database content, we first created two models to study one of the most important effect of TGHFβ : cell cycle arrest. The first model represents the canonical pathway, the second one represents the crosstalks and is made up of the canonical one and p38, jnk, erk, PI3k pathways. In these models, the cytostatic effect of TGFβ is symbolized by the reachability of two model places : p15INK4b and p21CIP1. These are two TGFβ activated proteins that block the cell cycle. Using CADBIOM-Chart, we searched how initialize the model to reach or not the property, which returned many solutions. A solution is a list of places that have to be intialized to reach the property. We created a solution treatment to focus on the irreducible solutions called minimal activation conditions and compared them between both models. We find the same minimal activation conditions betwwen canonical and crosstalks models, meaning that TGFβ stop the cell cycle through its canonical pathway. Our analysis also underline an important role of p38 pathway in the inhibition of TGFβ cytostatic effect.

4 Conclusion

Guarded transition formalism is efficient to represent complex biological data like signaling pathway. Using PID, we benefit of a well curated source of knowledge. Model simulation and model questioning have been used to manage dynamical models with thousands nodes, which is a pretty good performance in the domain of dynamical discrete modeling. Moreover, the usable analysis with CADBIOM-Chart allows to learn new fairly good knowledge on intricate processes such as TGFβ pleiotropic effect.

References

[1] D. Machado, R. S. Costa, M. Rocha, E. C. Ferreira, B. Tidor and I. Rocha, Modeling formalisms in Systems Biology. *AMB Express*, 1:45, 2011
[2] S. Klamt, J. Saez-Rodriguez, J. A. Lindquist, L. Simeoni and E. D. Gilles, A methodology for the sturctural and functional analysis of signaling and regulatory networks *BMC Bioinformartics*, 7:56, 2006
[3] A. Sackmann, M. Heiner and I. Koch, Application of Petrinet based analysis techniques to signal transduction pathways. *BMC Bioinformatics*, 7:482, 2006
[4] J. Rumbaugh, I. Jacobson and G. Booch, The Unified Modeling Language. Reference Manual. *Pearson Higher Education*, ISBN 0321245628, 2004.
[5] C. F. Schaefer, K. Anthony, S. Krupa, J. Buchoff, M. Day, T. Hannay, K. H. Buetow, PID: the Pathway Interaction Database, *Nucleic Acids Res*, 37, 2009
[6] H. Ikushima, and K. Miyazono, Biology of Transforming Growth Factorβ. *Curr Pharm Biotechnol*, 2011.
[7] S. Dooley, P. T. Dijke, TGF-β in progression of liver disease, *Cell Tissue Res*, 2011
[8] Y. Mu, S. K. Gudey, M. Landstrom, Non-Smad signaling pathways, *Cell Tissue Res*, 2011

A knowledge-based system for diagnosis dedicated to inherited retinal dystrophies

Maxime HEBRARD[1,2,3], Gaël MANES[1,2,3], Béatrice BOCQUET[1,2,3,4], Isabelle MEUNIER[4], Isabelle MOUGENOT[3] and Christian P. HAMEL[1,2,3,4]

[1] INSERM U1051-INM, Hôpital Saint Eloi, 80 rue Augustin Fliche-BP74103, 34091 Montpellier, France
{maxime.hebrard, gael.manes, beatrice.bocquet, christian.hamel}@inserm.fr

[2] UNIVERSITE MONTPELLIER I, 5 bd Henri IV - CS19044, 34967 Montpellier Cedex 2, France

[3] UNIVERSITE MONTPELLIER II, Place Eugène Bataillon, 34095 Montpellier Cedex 5, France
isabelle.mougenot@univ-montp2.fr

[4] CENTRE DE REFERENCE MALADIES RARES MAOLYA, CHRU Montpellier, Service Ophtalmologie 1er étage bureau 1414, Hôpital Gui de Chauliac 80 rue Augustin Fliche, 34091 Montpellier cedex 5, France

Keywords: Clinical signs, diagnosis, thesaurus, ontology, knowledge-based system.

1 Introduction

Certaines spécialités médicales se trouvent confrontées à des groupes de pathologies pour lesquelles les symptômes sont très variés mais peuvent s'avérer assez proches d'une maladie à l'autre. Notre travail porte sur la modélisation fine des connaissances associées à un groupe de pathologies afin d'en proposer un système classificatoire. L'objectif est ensuite de fournir un environnement d'aide au diagnostic mettant en jeu des mécanismes de raisonnement sur le modèle de connaissance construit. Cette démarche théorique prend appui sur les dystrophies rétiniennes héréditaires mais peut être généralisée à d'autres domaines.

2 Terminologie et modèle

L'établissement d'un diagnostic médical se base sur l'observation de traits cliniques présents chez le patient. Chaque trait peut se révéler être un marqueur d'une maladie spécifique soit de manière individuelle, soit de manière corrélée avec d'autres traits. Un premier enjeu porte ainsi sur la définition d'un ensemble exhaustif de termes spécifiques venant décrire, de manière fine et standardisée, les traits cliniques impliqués dans la manifestation des dystrophies rétiniennes héréditaires. La définition d'un thésaurus dédié nous semble répondre à ces attentes. Les thésauri reflètent les choix délibérés des communautés, dans la désignation comme dans l'organisation des termes clés d'un domaine d'expertise. Dans ce sens, un thésaurus vient lever les ambigüités des langues naturelles et offre un contrôle et une clarification des échanges d'information au sein de la communauté. Cela permet alors aux scientifiques de mutualiser leurs connaissances et contribue à un renforcement du savoir collectif sur ces maladies.

Plusieurs thésauri et terminologies sont d'ores et déjà définis dans le domaine médicale, cependant ces schémas de termes/concepts, par trop généralistes, ne peuvent pas répondre entièrement aux besoins précis de notre spécialité. Un thésaurus spécifique a donc été mis en place. En complémentarité du vocabulaire dédié, un modèle ontologique, décrivant les concepts clés de la représentation d'un phénotype dans un contexte clinique, a été dégagé. Ce modèle formalise la connaissance associée à tout phénotype clinique, restant ainsi valable pour n'importe quelle pathologie.

3 Méthode

Afin de construire notre base de connaissance sur un socle solide, nous avons sélectionné un échantillon de patients dont les symptômes sont bien caractérisés et dont le diagnostic est certain. Le phénotype de ces patients a alors été décrit à l'aide du thésaurus dédié. Le modèle ontologique général ainsi que les données relatives aux patients sont exploités au sein d'un composant applicatif Java. La librairie de programmation Jena y facilite les activités de gestion et de manipulation de l'ontologie. La librairie de programmation JUNG sert de support à la représentation visuelle des connaissances. Ainsi, chaque patient est représenté sous la forme d'un graphe : l'ontologie est alors mise à contribution pour en extraire l'arborescence de données liée à un patient particulier. La portion de graphe ainsi obtenue

constitue un profil de patient. De manière analogue, l'application permet la construction et la visualisation d'un profil de maladie sous la forme d'une portion de graphe. L'ontologie est là aussi consultée afin de fournir les connaissances associées à une maladie donnée. Ainsi, le profil d'une maladie particulière est envisagé comme étant la superposition des profils des patients connus, atteints de cette maladie. Les nœuds patients forment alors un seul nœud maladie, les nœuds symptômes sont rassemblés en fonction de leur type. L'application pondère les arcs en fonction de la fréquence d'apparition de tel symptôme ou tel état causé par la maladie en question.

L'un des objectifs est donc de disposer de données provenant du plus large échantillon de patients possible, afin de disposer d'une collection de profils de maladies couvrant l'ensemble du spectre des pathologies considérées. Ces profils vont ensuite constituer la base de l'aide au diagnostic lors de la prise en charge de nouveaux patients. Pour ce faire, le profil du nouveau patient est dressé à l'aide du thésaurus puis confronté à l'ensemble des profils de maladies stockés dans le système. Une première méthode quantitative naïve a été retenue afin d'asseoir la comparaison deux à deux du graphe du patient et de celui de chacune des maladies du système. A cet effet, un modèle à base de score a été défini. Les arcs associés aux symptômes ou à leurs états communs aux graphes du patient et de la maladie cible sont relevés, puis les coefficients de pondération des arcs considérés sont additionnés. Le score ainsi calculé permet d'évaluer la correspondance entre le phénotype du patient et les symptômes causés par la maladie. La combinaison patient/maladie retournant le score le plus élevé indique quelle pathologie est la plus proche des signes cliniques du patient. La liste des maladies triées par score est alors proposée au médecin comme aide au diagnostic.

4 Résultats

Le thésaurus dédié aux pathologies rétiniennes héréditaires en cours de construction comprend aujourd'hui 25 classes de symptômes qui regroupent près de 70 traits phénotypiques observables et compte plus de 320 valeurs d'état associées. Un premier jeu de donnée a été constitué comprenant 51 patients dont le diagnostic est certain, réparties sur 14 pathologies distinctes. Il semble que la richesse de la description proposée soit suffisante pour discriminer les pathologies actuellement décrites par le système.

5 Perspectives

Pour permettre une utilisation conviviale et rapide de l'application, nous prévoyons la mise en place d'une interface utilisateur. Il s'agira pour l'utilisateur de décrire les symptômes d'un patient à l'aide d'un formulaire guidé, basé sur le thésaurus dédié. Le logiciel lui fournira directement les scores des maladies présumées que le médecin pourra confronter à ses propres connaissances. L'ophtalmologie étant une discipline s'appuyant fortement sur l'imagerie, nous souhaiterions proposer aux usagers des exemples picturaux en accord avec les symptômes et leurs états représentés au sein du système, ce qui constituerait une base de validation supplémentaire. Ce logiciel aura à terme des vertus pédagogiques, puisqu'il constituera un atlas illustré des différentes pathologies de la spécialité médicale.

Un effort doit être fourni pour s'assurer du bon niveau de représentativité de notre modèle de connaissance. Pour cela, un travail de fond est en cours, afin de disposer d'une bibliothèque de profils de patients la plus riche et complète possible. De plus, le seuil d'efficacité du système sera testé en lui présentant des cas plus difficiles à diagnostiquer.

Un objectif prometteur concerne l'ajout de données génotypiques et moléculaires au sein du profil de patient. Cela permettrait alors non seulement de discriminer les différentes pathologies sur la base du phénotype clinique, mais aussi de raffiner la collecte et le traitement de l'information jusqu'à obtenir une prédiction de diagnostic à l'échelle d'un gène spécifique ou même d'une mutation particulière.

Dans le futur, cette démarche pourra être reprise dans ce cadre d'aide au diagnostic pour d'autres familles de pathologies. Dans cette perspective, il suffit d'établir un thésaurus spécifique permettant la description des traits cliniques caractéristiques du nouvel ensemble de pathologies. Le cadre ontologique modélisant les maladies et leurs symptômes ainsi que la méthode d'analyse resteront quand à eux inchangés.

Relationships Between Structures and Sequences from a Super Secondary Structure Elements Approach.

Tristan BITARD FEILDEL[1] and Jean-François GIBRAT[1]

MIG, INRA, Bâtiment 233, Domaine de Vilvert, 78352, Jouy en Josas, Cedex , France
tristan.bitardfeildel@jouy.inra.fr
jean-francois.gibrat@jouy.inra.fr

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], Philippe Hupé[1,2,3,4]

1 Institut Curie, 26 rue d'Ulm, Paris, F-75248 France; 2 INSERM, U900, Paris, F-75248 France; 3 Mines ParisTech, Fontainebleau, F-77300 France; 4 CNRS UMR144, 26 rue d'Ulm, Paris, F-75248 France; 5 Institut Curie, Department of Surgery, 26 rue dUlm, Paris, F-75248 France.

Keywords Microarrays, Methylation, Spatial normalisation.

DNA methylation 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 methylated samples and 3 fully unmethylated DNA samples. We demonstrate that the spatial normalisation outperforms BeadStudio to predict the methylation state of a given locus.

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

The GAG database: a new resource to gather genomic annotation cross-references

Thomas OBADIA[1,3], Olivier SALLOU[2], Marion OUEDRAOGO[1], Gregory GUERNEC[1,4] and Frédéric LECERF[1]

[1] INRA/Agrocampus OUEST, UMR1348 PEGASE, F-35000, Rennes, France
lecerf@agrocampus-ouest.fr , marion.ouedraogo@rennes.inra.fr

[2] Genouest Platform, INRIA/Irisa – Campus de Beaulieu, F-35042, Rennes cedex, France
olivier.sallou@irisa.fr

[3] Present address : INSERM, UMR S 707, Paris, France
thomas.obadia@u707.jussieu.fr

[4] Present address : INSERM, UMR 1027, F-31000, Toulouse, France
gregory.guernec@inserm.fr

Keywords Annotation, sequence comparison, cross-references, database, prediction

1 Introduction

Hundreds of gigabytes of raw and processed genomics data are now available from several foundations, providing researchers with functional annotation related to coding and non-coding DNA and RNA. Although most of publicly available databases provide users with cross-references tables and link to external resources for each input, these tools only cover a limited amount of their content. Gathering as much annotation as possible for specified gene identifiers quickly becomes time-consuming.

The consensus coding sequence (CCDS)[1] is currently being hosted by the NCBI and regroups collaborating members from the EBI, the Wellcome Trust Sanger Institute, and the UCSC. This project aims at identifying a core set of human and mouse homolog protein-coding regions that are consistently annotated and of high quality. We propose a widened approach, which covers eight species chosen either for their status of reference species (human, mouse, rat), their agronomical interest (chicken, cow, horse, pig) or their medical interest (dog, used for oncogenic models). Furthermore, the GAG process focuses on the entire set of RNA transcripts rather than consistently annotated ones.

The whole process is hosted by the GenOuest platform to provide a public access to newly generated cross-references and to allow for regular updates (http://gag.genouest.org).

2 Realization

2.1 Data Preparation and Sequence Comparison

Retained datasets for the project include GenBank[2] and Ensembl[3] coding RNA datasets as the first layer for sequence comparison. To increase result robustness, we also included external resources such as UniProt/SwissProt[4] proteomics data and standardized HGNC[5] names.

We developed a set of PERL programs to handle parsing of available data, creation of the database tables, and researched of enriched cross-references. Sequence similarity is assessed by using the BLAST algorithm with its default scoring system: every single sequence from database “A” is aligned against the full database “B”, and vice-versa.

The main hypothesis is that matches should show a very high similarity score when compared to the multiple false positives. Subsequently, we developed a filtering process to extract what we believe are “true positives” from the BLAST outputs.

By comparing outputs with GenBank cross-references table, results are categorized: Known/Validated (GAG result confirmed by GenBank cross-references table); Known/Corrected (GAG result invalidated and

corrected according to GenBank table); Known/Forgotten (no cross-reference found by GAG, but official one existed in GenBank table); Predicted (cross-reference found by GAG, and no official data in GenBank table). Error-rate is assessed by the number of "Corrected" results. The various thresholds for this process were computed on a trial-error scheme: primary criterion is to minimize the number of "Known/Corrected" results, and secondary criterion is to maximize the "Predicted" in regard to the primary objective.

2.2 Increase in the Number of Available Cross-references

We achieve an average increase of 28.4% when comparing our filtered results with NCBI cross-references tables (see Table 1). Among newly discovered cross-references, the process highlights a lot a redundancy in the Ensembl dataset, where multiple gene identifiers link to a unique GenBank identifier. This redundancy seems to have accumulated with each new release of Ensembl database, and so this tool is useful to gather fragmented annotation data within databases. To provide user with a way of assessing filtered results, we also implemented a text comparison of the functional annotation between databases, and the process checks for common links within external resources (same homolog genes, same protein identifier, same gene name…)

Taxa	Generated cross-references (including new)	GenBank cross-references	Increase
Mus musculus (10090)	23075 (2334)	20396	+8.81%
Homo sapiens (9606)	20382 (2661)	18821	+5.54%
Gallus gallus (9031)	13493 (3310)	10159	+30.83%
Bos taurus (9913)	18434 (4884)	13229	+36.31%
Sus scrofa domesticus (9823)	16369 (5704)	9745	+57.69%
Equus caballus (9796)	16601 (3039)	13807	+21.98%
Canis familiaris (9615)	17222 (5495)	11633	+47.06%
Rattus norvegicus (10116)	20703 (3480)	17026	+18.95%

Table 1. Overall increase of cross-references compared to GenBank official files

3 Conclusion

Cross-references are of important use when it comes to analyze genomic sequences. Every major bioinformatics foundation now hosts its own database, and sometimes cross-referencing data is a time-consuming task. We implemented a new way to access enriched cross-references tables, and these data are available on the GenOuest bioinformatics platform. The complete process can be updated in a matter of hours if source data are updated. Furthermore, the database design allows for adding species very easily. As such, the GAG database provides a solid base for further developments involving functional genomics data, with queries made by a wide range of arguments (chromosomal localization, gene symbols or identifiers…)

References

[1] K.D. Pruitt, J. Harrow, R.A. Harte, C. Wallin, M. Diekhans, D.R. Maglott, et al., The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes, Genome Res. 19 (2009) 1316–1323.

[2] D.A. Benson, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, E.W. Sayers, GenBank, Nucleic Acids Res. 39 (2011) D32–37.

[3] X.M. Fernández-Suárez, M.K. Schuster, Using the ensembl genome server to browse genomic sequence data, Curr Protoc Bioinformatics. Chapter 1 (2010) Unit1.15.

[4] E. Jain, A. Bairoch, S. Duvaud, I. Phan, N. Redaschi, B.E. Suzek, et al., Infrastructure for the life sciences: design and implementation of the UniProt website, (07:26:05).

[5] R.L. Seal, S.M. Gordon, M.J. Lush, M.W. Wright, E.A. Bruford, genenames. org: the HGNC resources in 2011, Nucleic Acids Research. 39 (2011) D514.

COV2HTML: visualization and analysis tools of Bacterial NGS data for biologists

Marc Monot[1], Mickael Orgeur[2], Emilie Camiade[1], Clément Bréhier[1], and Bruno DUPUY[1]

[1] LABORATOIRE Pathogenèse des Bactéries Anaérobies, Institut Pasteur, 25 rue du docteur roux, 75015, Paris, France
mmonot@pasteur.fr;ecamiade@pasteur.fr;cleeeem@free.fr;bdupuy@pasteur.fr

[2] UNITE de Pathogénomique Mycobactérienne Intégrée, Institut Pasteur, 25 rue du docteur roux, 75015, Paris, France
morgeur@pasteur.fr

Keywords NGS, Mapping, Coverage, Visualization, Analysis.

1 Background

Next-generation sequencing (NGS) technologies are revolutionising genomics and their benefits are becoming widespread. Nowadays, they are widely used for understanding genomic variations and regulatory networks identification. Among others, the DNA-seq allows to incredibly increase the number of new genomes sequenced and as well to rapidly re-sequence existing genomes previously obtained by the traditional Sanger approach. Furthermore, regulation of gene expression is a fundamental process within every cell type, which allows organisms to control precisely the levels of the gene products in response to their environment. Deep transcriptome sequencing (RNA-seq) has recently emerged as a method enabling the study of RNA-based regulatory mechanisms in a genome-wide manner [1]. The RNA-seq that focus on transcriptionnal start site (TSS) have also demonstrated its effectiveness in accurate operon definition, discovery of non-coding RNAs, and correction of gene annotation [2].

All the NGS technologies pose several challenges in terms of results analysis but also in the visualization of the mapping coverage of the NGS data. There are many applications that have been developed for browsing, visualizing and interpreting these mapping files. For example Generic Genome Browser (GBrowse) [3] is designed to visualize genome assemblies, IGV [4] is a multiple genome browser and BamView [5] manage view and interpret short-read data. All these sofwares load large mapping file with a consequent amount of time. Futhermore managing these files are laborious and most of the analysis programs has to be installed, which is not completly obvious for the biologists. Currently, some scientist teams visualize and analyse large data sets from multiple NGS experiments by using available but often adapted bio-informatic tools [2]. However, we only reach a minimal part of the huge information obtained and analyzed in their publications. The remainder has to be interpreted, but it involves tools dedicated to medium skilled informatics users.

2 Results

We developed COV2HTML, an accessible visualization and analysis tools for biologists, which consists in two highly specialized sofwares that allow coverage visualization of the NGS data before to be studied (Figure 1). In order to ease both data loading and processing, COV2HTML uses an own coverage format instead of directly handling a huge alignment map or just a coverage file. To be done, after having been mapped against the reference genome, the NGS data are converted by the tool MAP2COV, which is provided with the visualization interface. MAP2COV extracts (i) the genome coverage from the vast mapping file, 1-10 Gigabytes (Go), and (ii) genetic element (genes and intergenic regions) informations from an annotation file to create a light result files of 1 Megabytes (Mo). This reduction greatly simplifies the data management of the NGS analysis and facilitates saving of data in classical computer. The second software is a visualization tool dedicated to study the mean coverage of genetic elements or their promotor region by the TSS experiments. We could analyse 2 conditions with a maximum of 4 replicates per conditions or visualize diverse experiments done on the same bacteria. The analysis is managed by filters that use simple criteria such as coverage level and fold change for 2 conditions comparison.

To demonstrate the utility of COV2HTML, the program was used on two NGS 2011's publications. The first one is a classical RNA-seq analysis of *Campylobacter jejuni* which consists of a comparison of 2

conditions with 2 replicates per conditions [1]. As only raw data were available, we align reads using bowtie [6] to get a mapping file in sam format [7]. Then we used MAP2COV with the *C. jejuni* genome sequence and the mapped file to obtain a COV2HTML readable file for each condition's replicates. After data integration, we performed the same author's analysis using COV2HTML's filters and get 26 out of 27 genes found by the authors whose expression are significantly altered in the mutant compared to the wild-type strains. The second publication contains stranded TSS and RNA-seq data on the Archaea *Sulfolobus solfataricus* with three conditions [2]. In this case authors give raw data files for TSS and WIG-like coverage files for RNA-seq. For a TSS case study we use bowtie [6] to create stranded combine condition mapping files. Then we launch MAP2COV to convert them into coverage files along with the *S. solfataricus* genome. Using our standard filter, it is worth noting that we found 80% of the published TSS however with two time more TSS than the authors. Finally we created a mixed visualization analysis of TSS and RNA-seq experiments, combining all RNA-seq WIG-like file into a single coverage file. All these data are available in the tutorial part of the COV2HTML main web page.

3 Conclusions

Because more and more NGS data sets produced are available, their analysis become more diverses and complexes and, as a consequence, need experienced and knowledgeable experts. In another hand the main bottleneck for the biologists to perform their own global and specific analysis is that they cannot use the tools dedicated without the help of the expert. Ideally, biologists should have a "plug and play" software for such studies. The COV2HTML interface offers a quick visualization of mapping coverage of the NGS data (DNA-seq, RNA-seq, ChIP-seq and TSS) performed in different prokaryotic organisms (bacteria, viruses...) or from 2 experimental conditions (such as mutant versus wild type strains or different growth states…) facilitating studies of global biological questions. The strength of the COV2HTML programs is to easily analyse and share data without software installation, login or a long training period. A web version is currently accesible at http://mmonot.eu/COV2HTML.

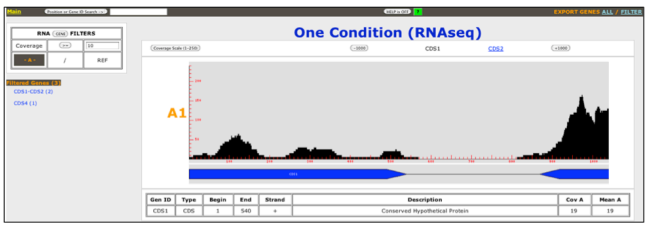


Figure 1. COV2HTML Visualization of one condition RNA-seq experiment.

References

[1] 1. Chaudhuri RR, Yu L, Kanji A, Perkins TT, Gardner PP, Choudhary J, Maskell DJ, Grant AJ: Quantitative RNA-seq analysis of the Campylobacter jejuni transcriptome. *Microbiology* 2011, 157(Pt 10):2922-2932.

[2] 2. Wurtzel O, Sapra R, Chen F, Zhu Y, Simmons BA, Sorek R: A single-base resolution map of an archaeal transcriptome. *Genome Res* 2010, 20(1):133-141.

[3] 3. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A *et al*: The generic genome browser: a building block for a model organism system database. *Genome Res* 2002, 12(10):1599-1610.

[4] 4. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP: Integrative genomics viewer. *Nat Biotechnol* 2011, 29(1):24-26.

[5] 5. Carver T, Harris SR, Otto TD, Berriman M, Parkhill J, McQuillan JA: BamView: visualizing and interpretation of next-generation sequencing read alignments. *Brief Bioinform* 2012.

[6] 6. Langmead B, Trapnell C, Pop M, Salzberg SL: Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009, 10(3):R25.

[7] 7. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009, 25(16):2078-2079.

Natural Selection in Pre-Eclampsia-Associated Genes

Lilian P.L. LAU and Hugues ROEST CROLLIUS
Institut de Biologie de l'Ecole Normale Supérieure (IBENS), Centre National de la Recherche Scientifique UMR8197, 46 rue d'Ulm, Paris, F-75005 France
{lilian.lau, hrc}@ens.fr

Keywords Pre-eclampsia, natural selection, positive selection, balancing selection

Sélection Naturelle dans les Gènes de Pré-éclampsie

Mots-clés Pré-éclampsie, sélection naturelle, sélection positive, sélection balancée

Pre-eclampsia is a medical condition affecting about 3-5% of all pregnancies worldwide, typically diagnosed by pregnancy-specific hypertension (blood pressure above 140/90mm Hg in previously normotensive women) and proteinurea (excess of 300mg protein per 24 hour urine collection) after 20 weeks of gestion. When left untreated, pre-eclampsia can proceed into severe complications, in particular haemolysis elevated liver enzymes and low platelets (HELLP) syndrome and eclampsia (i.e. seizures), leading to coma and maternal (and foetal) deaths.[1]

The exact pathophysiology of pre-eclampsia remains to be elucidated. It is widely accepted that pre-eclampsia develops over two stages: first involving poor/abnormal invasion of cytotrophoblast in the uterine wall, and secondly the manifestation of clinical symptoms as a result of the release of syncytiotrophoblast-derived factors from oxidatively stressed placenta. However, current poor understanding of the mechanism of pathogenesis means the only “treatment” effective to date is the delivery of the foetus and placenta.[2]

Pre-eclampsia is a major cause of maternal mortality, particularly in developing countries, implicated in some 50,000 cases annually (e.g. Ivory Coast has a 16% maternal mortality rate due to pre-eclampsia). In developed countries, it is through the advent of advanced medical care that the mortality rate is reduced to about 0.05% [3]. Together, these data suggest that should there be a genetic basis for predisposition to pre-eclampsia, the alleles should have long been eliminated in the course of human evolution before modern medical progresses catch up and enable adequate management of pre-eclampsia.

We hypothesised that continuous high prevalence of pre-eclampsia in populations may be due to the alleles of disease-associated genes undergoing balancing selection, where the alleles are either (a) maintained as heterozygotes that confer greater fitness over homozygotes; or (b) subject to diversifying frequency-dependent selection where as the alleles become rarer, the phenotype fitness improve.

The hallmarks of balancing selection are excess of polymorphic sites and excess of alleles at intermediate frequencies. A number of neutrality tests can be used to test balancing selection, and some frequently used methods include Tajima's *D*, Wright's F_{ST}, McDonald & Kreitman's (MK) test, and Hudson, Kreitman & Aguadé's (HKA) test.[4]

We are applying Wright's F_{ST} [5] on single nucleotide polymorphism (SNP) data from the 1,000 Genomes Project [6] to calculate the proportion of genetic diversity due to differences in allele frequency between test populations. We are also applying *K* statistic [7] (method inspired by HKA test) to identify loci with maintained variance. Loci identified from both tests as putatively under balancing selection would then be cross-referenced to a list of genes identified from previous differential expression studies as genes which expressions were altered in pre-eclampsia. Short-listed genes would be further validated using molecular methods on clinical samples consist of control and pre-eclamptic patients.

Acknowledgements

This work is supported by Fondation Pierre-Gilles de Gennes.

References

[1] J.M. Roberts and C.W.G. Redman, Pre-eclampsia: more than pregnancy-induced hypertension. *The Lancet*, 341:1447-1451, 1993.

[2] C.W.G. Redman, Preeclampsia: a multi-stress disorder. *La Revue de médicine interne*, 32S:S41-44, 2011.

[3] L. Duley, The global impact of pre-eclampsia and eclampsia. *Seminars in Perinatology*, 33(3):130-137, 2009.

[4] G.D. Weedall and D.J. Conway, Detecting signatures of balancing selection to identify targets of anti-parasite immunity. *Trends in Parasitology*, 26:363-369, 2010.

[5] K.E. Holsinger and B.S. Weir, Genetics in geographically structured populations: defining, estimating and interpreting FST. *Nature Reviews Genetics*, 10:639-649, 2009.

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

[7] D. Enard, F. Depaulis and H. Roest Crollius, Human and non-human primate genomes share hotspots of positive selection. *PLoS Genetics*, 6(2):e1000840, 2010.

Investigation of associations between human genetic variation and resistance to HIV-1 infection in highly exposed uninfected individuals with hemophilia A

Jérôme LANE [1], Lucy DORRELL[2], Kimberly PELAK[3], Kevin V SHIANNA[3], Mary CARRINGTON[4], James J GOEDERT[5], Barton F HAYNES[3], Andrew J MCMICHAEL[2], David B GOLDSTEIN[3] and Jacques FELLAY[1,3] for the NIAID Center for HIV/AIDS Vaccine Immunology (CHAVI)

[1] School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
{jerome.lane, jacques.fellay}@epfl.ch
[2] Weatherall Institute of Molecular Medicine, University of Oxford, UK
{lucy.dorrell, andrew.mcmichael}@imm.ox.ac.uk
[3] Duke Human Vaccine Institute, Duke University, Durham, NC, USA
kim.pelak@gmail.com, hayne002@mc.duke.edu, {k.shianna, d.goldstein}@duke.edu, jacques.fellay@epfl.ch
[4] Cancer and Inflammation Program, Laboratory of Experimental Immunology, SAIC Frederick, Inc., Frederick National Lab, Frederick, MD, USA and Ragon Institute of MGH, MIT and Harvard, Charlestown, MA, USA
carringm@mail.nih.gov
[5] Division of Cancer Epidemiology and Genetics, NCI, Bethesda, MD, USA
goedertj@mail.nih.gov

Keywords Genome wide association study, genetic variations, HIV-1, resistance.

1 Background

HIV-1 infects vital cells of the human immune system such as helper T cells, macrophages and dendritic cells and disease usually progress toward severe immunodeficiency in absence of treatment. However, a portion of exposed individuals will not be infected. Human genetic variations are involved in those differences. Many studies on genetic variations have found associations with host susceptibility to HIV-1 infection. However, the homozygous deletion of 32 base pairs in CCR5 gene (CCR5 Δ32) and a rarer mutation m303 are the only genetic variations certified to provide a near complete protection against HIV-1 infection [1, 2]. Hemophilia is a X-linked recessive bleeding disorder caused by the mutation of a clotting factor gene. To stop the bleeding, hemophiliacs were infused with concentrates of clotting factors prepared from large pools of donors and no viral inactivation of the concentrates was done until 1984. Therefore, hemophiliacs were highly exposed to HIV-1 infection before that date. CHAVI, the Center for HIV/AIDS Vaccine Immunology obtained samples from individuals with hemophilia A that were highly exposed to potentially contaminated factor VIII infusions. Thus, a genome wide association study (CHAVI014 study) was developed to find new genetic variations that could be implicated in the resistance to HIV-1 infection.

2 Methods

Individuals with moderate to severe hemophilia A and a documented history of factor VIII infusions before the introduction of routine viral inactivation procedures (i.e. 1979-1984) were recruited from 36 hemophilia treatment centers in 9 countries. CCR5 Δ32 and m303 genotypes were assessed by taqman and homozygous subjects were excluded from the GWAS. HIV-infected individuals included in other CHAVI host genetic studies and matched for ethnicity and gender were used as controls. The first part of the study consisted to look at single nucleotide polymorphisms (SNPs). DNA genotyping was performed on Illumina Human 1M or 1Mduo chips. A SNP imputation was performed to increase their number and therefore the power to detect genetic variations of interest using MACH [3] software. The second part of the study consisted in detecting copy number variants (CNVs) including large deletions and duplications from SNPs using PennCNV [4] software. After quality controls, an association analysis has been applied to selected SNPs, deletions, duplications and samples consisting in single-marker logistic regressions under additive, recessive and dominant genetic models.

3 Results

A total of 582 HIV-resistant cases were recruited: 36 (6.2%) were homozygous for CCR5 Δ32 or m303 and were excluded. Overall the 1,455 samples analyzed, 431 cases and 765 HIV-infected controls passed the quality controls as 1,081,435 SNPs that were tested for association with HIV-1 infection. The study had 80% power to detect associated variants with minor allele frequency (MAF) of 20% and genotype relative risk (GRR) of 2, or MAF of 5% and GRR of 2.94, under an additive genetic model. No SNP reached genome-wide significance of 9.4E-15 obtained after Bonferroni correction (Figure 1). Overall detected CNVs, 3,375 different regions corresponding to 1,193 deletions and 2,631 duplications were selected. CNV numbers were consistent with deletions and duplications detected in data from the 1000 Genomes Project. No significant association was found for CNVs (data not shown).

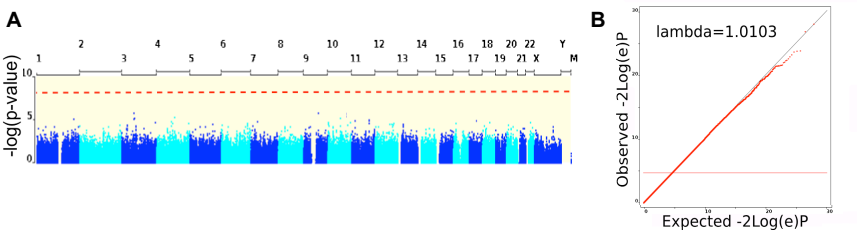


Figure 1. Manahattan plot of SNP association with the resistance against HIV-1 infection. [A] The Manhattan plot shows no significant association signal throughout the genome under the additive model. The dotted red line indicates tl significance threshold of 9.4E-15. [B] The QQ plot demonstrates that the observed distribution of p-values corresponds the expected distribution under the null hypothesis, indicating that potential confounders are well controlled.

4 Conclusions

Individuals with hemophilia who were highly exposed to potentially contaminated blood products, yet were not infected by HIV-1 in the early years of the pandemic form an ideal study group to investigate host resistance factors. All included cases had a documented history of treatment with FVIII concentrates with a high likelihood of HIV-1 contamination. Due to the severity of hemophilia, they received a relatively high number of FVIII infusions (median 51), each derived from pooled plasma from thousand of donors. Even if no association was found between common SNPs and HIV-1 susceptibility in this population, human genetic variants could still be involved in resistance phenotype. Genome wide association studies have been developed to detect common genetic variations whereas the causal genetic variant could be rare. Structural variants other than large duplications and deletions could also be responsible for the resistance phenotype. Epigenetic mechanisms such as DNA methylation and histone modification could play a role in the resistance by regulating gene expression. Sequencing studies are now required to comprehensively assess the role of human genetic diversity in HIV-1 acquisition.

Acknowledgements

We would like to thank all study participants and health care workers at the contributing hemoph treatment centers.

References

[1] J.J. Martinson, N.H. Chapman, D.C. Rees, Y.-T. Liu and J.B. Clegg, Global distribution of the CCR5 gene 32-base pair deletion. *Nat. Genet.*, 16: 100-103, 1997.

[2] C. Quillent, E. Oberlin, J. Braun, D. Rousset, G. Gonzalez-Canali, et al., HIV-1-resistance phenotype conferred by combination of two separate inherited mutations of CCR5 gene. *The Lancet*, 351: 14-18, 1998.

[3] Y. Li, C.J. Willer, J. Ding, P. Scheet and G.R. Abecasis, MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*, 34: 816-834, 2010.

[4] K. Wang, M. Li, D. Hadley, R. Liu, J. Glessner, et al., PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.*, 17: 1665-1674, 2007.

L'angiogenèse : d'un réseau d'interactions à un réseau biologique

Guillaume LAUNAY, Franck PEYSSELON, Romain SALZA and Sylvie Ricard-Blum
UMR 5086 CNRS - Université Lyon 1, Institut de Biologie et Chimie des Protéines, 69367 Lyon Cedex 07
guillaume.launay@ibcp.fr

Mots-clés Intégration de données biologiques hétérogènes, biologie des systèmes, réseau d'interactions

La formation de vaisseaux sanguins à partir de vaisseaux préexistants, appelée angiogenèse, est un processus indispensable à la croissance tumorale. Cela a suscité le développement de stratégies thérapeutiques anti-angiogéniques qui ciblent les cellules endothéliales vasculaires. De nombreuses molécules (des facteurs de croissance, des protéines extracellulaires, des récepteurs et des polysaccharides) régulent l'angiogenèse. Nous avons identifié par des puces à protéines et à sucres les interactions existant entre les régulateurs connus de l'angiogenèse [1], ainsi que de nouvelles interactions établies par ces régulateurs avec des molécules situées dans l'environnement vasculaire et à la surface des cellules endothéliales. Le réseau d'interactions construit à partir de ces données constitue un système complexe, multi-échelle (moléculaire et cellulaire) mais statique. Notre projet consiste donc à élaborer un réseau dynamique et à intégrer pour cela des données quantitatives et qualitatives, structurales et fonctionnelles, dans ce réseau qui comporte quelques centaines d'interactions. Sont intégrés les domaines protéiques et les régions intrinsèquement non structurées [2] des protéines présentes dans le réseau ainsi que les sites d'interactions déterminés expérimentalement ou prédits par modélisation moléculaire [3] pour discriminer les interactions mutuellement exclusives de celles qui se produisent simultanément. Les annotations de Gene Ontology sur la localisation, la fonction moléculaire et le processus biologique dans lesquels sont impliquées les protéines du réseau sont également ajoutées ainsi que des données d'expression (cellules endothéliales traitées ou non par des inhibiteurs de l'angiogenèse), de protéomique (sécrétomes de cellules endothéliales activées ou non) et des données du projet "Human Protein Atlas". Sont associés aux interactions les paramètres cinétiques (vitesse d'association et de dissociation) pour distinguer les interactions stables des interactions transitoires ainsi que l'affinité qui reflète la force des interactions et permet de les hiérarchiser. L'intégration des données et l'analyse du réseau annoté sont réalisées grâce à des outils bioinformatiques tels que Cytoscape et DAVID ("Database for Annotation, Visualization and Integrated Discovery") et grâce à des outils en cours de développement dans l'équipe. Les résultats obtenus permettront de déterminer l'ensemble des mécanismes moléculaires qui régulent l'angiogenèse de façon concertée et de modéliser le réseau pour prédire son comportement sous l'effet de stimuli physiopathologiques tels que la surexpression d'un facteur de croissance.

References

[1] C. Faye, E. Chautard, B.R. Olsen and S. Ricard-Blum. The First Draft of the Endostatin Interaction Network, *The Journal of Biol. Chem.*, 284:22041-22047, 2009.

[2] F. Peysselon, B. Xue, V. N. Uversky and S. Ricard-Blum. Intrinsic disorder of the extracellular matrix. *Mol. BioSyst.*, 7:3353-3365, 2011.

[3] G.Launay and T. Simonson, Recognizing protein-protein interfaces with empirical potentials and reduced amino acid alphabets, *BMC Struct. Biol*, 9:427-443, 2008.

Cyanorak v2

An information management system for annotating cyanobacterial orthogroups

Loraine Guéguen[1], Gregory Farrant[12], Gildas Le Corguillé[1], Erwan Corre[1], Wilfrid Carré[1], Laurence Garczarek[2], Frédéric Partensky[2], Christophe Caron[1] and Mark Hoebeke[1]

[1] ABiMS, FR2424 CNRS-UPMC, Station Biologique, Place Georges Teissier, 29680, Roscoff, France
{loraine.gueguen, gildas.le-corguille, erwan.corre, wilfrid.carre, christophe.caron, mark.hoebeke}@sb-roscoff.fr

[2] UMR7144 CNRS, Station Biologique, Place Georges Teissier, 29680, Roscoff, France
{gregory.farrant, laurence.garczarek, frederic.partensky}@sb-roscoff.fr

Keywords Database, comparative genomics, cyanobacteria, annotation, traceability, clustering.

Cyanorak v2

Système d'information pour l'annotation d'orthogroupes de cyanobactéries

Mots-clés Base de données, génomique comparative, cyanobactéries, annotation, traçabilité, clustering.

1 Introduction

Cyanorak v2 (http://abims.sb-roscoff.fr/cyanorak) est un outil destiné à l'annotation des cyanobactéries marines. Il reprend et complète les données présentes dans Cyanorak v1 [1], dont il élargit également les fonctionnalités. Cyanorak v2 contient les séquences et les annotations de 33 souches de cyanobactéries, ainsi qu'un ensemble de *clusters* d'orthologues regroupant les protéines de ces souches en fonction de leur similarité en séquence (à l'instar de COG [2]). Sa principale originalité réside dans sa capacité à assurer de manière automatique ou semi-automatique le transfert des informations associées aux *clusters* (contenu en gènes, et annotations) lors de l'ajout de nouvelles souches. Par ailleurs, afin d'améliorer la robustesse des processus de curation manuelle ou automatique, il garde un historique détaillé des modifications apportées aux contenus. Ce qui lui permet d'offrir un accès public à des "instantanés" reflétant un état d'avancement précis, tout en autorisant en continu les modifications "privées" effectuées par la communauté des curateurs.

2 Architecture du système d'information Cyanorak v2

Le système d'information Cyanorak v2 a été conçu autour d'une architecture modulaire, à trois niveaux : un niveau d'accès aux données reposant d'une part sur une base de données relationnelle munie de *mappings* permettant son exploitation dans des langages à objet, et d'autre part sur un entrepôt non-structuré pour la conservation des séquences brutes, ou d'autres données générées à la volée. Un ensemble de modules capables d'échanger avec ces sources de données, et effectuant les différents étapes d'importation de données et de calcul, constitue le second niveau. Ces outils en ligne de commande se chargent par exemple de l'extraction des séquences protéiques, de leur *clusterisation* et de l'importation des *clusters* dans la base. Ils implémentent également la logique de décision permettant la mise en correspondance de jeux de *clusters* issus de deux exécutions de *clustering* successives sur des jeux de données différents. Enfin, le troisième niveau prend la forme d'une application Web offrant à la fois un accès public à des versions expressément estampillées comme telles par les curateurs, ainsi qu'un accès restreint permettant à ces derniers de modifier les données de manière continue.

3 Clustering

Le passage de Cyanorak v1 à Cyanorak v2, s'est traduit en premier lieu par le doublement du nombre de

génomes pris en compte. Un nouvel ensemble de *clusters* d'orthologues a été reconstruit avec OrthoMCL [3] à partir de la totalité de leurs séquences protéiques. Par ailleurs, un travail de curation manuelle considérable avait été réalisé pour l'annotation des *clusters* de Cyanorak v1, et la nécessité de transposer au mieux ces informations d'une version de Cyanorak à la suivante a conduit à la définition d'une classification des liens entre *clusters* produits par deux *clustering* successifs, basée sur le recouvrement relatif de leurs contenus en gènes. Ces différentes classes d'association ont servi de base à la construction d'un jeu de règles pour mettre en correspondance les *clusters* entre deux versions de *clustering* (N et N+1). Dans le cas trivial où un *cluster* issu du *clustering* N+1 a exactement le même contenu en protéines qu'un *cluster* issu du *clustering* N (*clusters* typés "EXACT-MATCH"), seul le *cluster* N est conservé, avec ses annotations. Pour un *cluster* N+1 qui contient, en plus de l'ensemble des protéines d'un seul *cluster* N, des protéines appartenant aux nouveaux génomes (*clusters* typés "SUPERSET"), les annotations sont transférées de N vers N+1 avec un niveau de confiance élevé. En revanche, dans le cas plus complexe où le contenu en protéines d'un *cluster* N+1 se répartit dans plusieurs *clusters* N, qui eux-mêmes partagent leur contenu en protéines avec d'autres *clusters* N+1, le transfert automatique d'annotations n'est pas possible. Les protéines sont alors toutes regroupées dans un même *cluster* qui devra être passé en revue par un curateur. Celui-ci, aidé d'informations supplémentaires telles que l'arbre phylogénétique des protéines, choisira au cas par cas la validation ou l'éclatement du *cluster*, ainsi que l'attribution des annotations des *clusters* N.

4 Premiers résultats

Le *clustering* des 33 génomes de Cyanorak v2 a généré un ensemble de 14 004 *clusters*. On y dénombre 6 154 *clusters* (44 % du total) ne contenant que des protéines issues des 19 nouveaux organismes, et donc non impliqués dans des associations avec des *clusters* de Cyanorak v1. Dans les 7 850 *clusters* restants, 2 566 sont de type EXACT-MATCH, et 3 363 sont de type SUPERSET : ces *clusters* (42 % du total) se voient donc automatiquement affecter l'annotation du *cluster* v1 avec lequel ils sont en relation. Le reste des *clusters* v2 associés à au moins un *cluster* v1, soit 1 921 *clusters* (14 % du total), nécessite un examen plus approfondi.

5 Conclusions & perspectives

Cyanorak v2 constitue à ce jour un outil précieux pour l'annotation des cyanobactéries marines. En effet, la mise au point et l'implémentation d'un jeu de règles capable de mettre en correspondance des *clusters* construits à partir de deux jeux de données différents, représentent un gain important dans le processus de curation en évitant nombre de revisites manuelles de *clusters*. Cette aide à l'annotation sera encore améliorée par la mise en place prochaine de nouvelles fonctionnalités dans l'application Web, comme la représentation des gènes dans leur contexte génomique, ou la visualisation des arbres retraçant les proximités en séquences des protéines de chaque *cluster*.

Remerciements

Le développement de Cyanorak v2 se fait dans le cadre de l'ANR génomique microbienne PELICAN : ANR-PCS-09-GENM-200.

Références

[1] A. Dufresne, M. Ostrowski, D. J. Scanlan, L. Garczarek, S. Mazard, B. P. Palenik, I. T. Paulsen, N. T. de Marsac, P. Wincker, C. Dossat, S. Ferriera, J. Johnson, A. F. Post, W. R. Hess, F. Partensky. Unraveling the genomic mosaic of a ubiquitous genus of cyanobacteria. *Genome Biology, 9:R90, 2008.*

[2] R.L. Tatusov, N.D. Fedorova, J.D. Jackson, A.R. Jacobs, B. Kiryutin, E.V. Koonin, D.M. Krylov, R. Mazumder, S.L. Mekhedov, A.N. Nikolskaya, B.S. Rao, S. Smirnov, A.V. Sverdlov, S. Vasudevan, Y.I. Wolf, J.J. Yin, D.A. Natale. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics 4: 41, 2003*.

[3] C. Habib, E. Le Corguillé, E. Corre, L. Brillet, M. Hoebeke, W. Carré, L. Garczarek, F. Partensky, C. Caron. PELICAN: Orthologous Groups and Gene Lateral Transfers for Comparative Genomic Analysis of Marine Cyanobacteria. *Actes JOBIM 2011*.

A local search approach for determining the insertion potential of the amino acids

Sami Laroum[1], Béatrice Duval[1], Dominique Tessier[2], and Jin-Kao Hao[1]

[1] LERIA, 2 Boulevard Lavoisier, 49045 Angers, France
{laroum,bd,hao}@info.univ-angers.fr
[2] UR 1268 Biopolymères Interactions Assemblages,
INRA, 44300 Nantes, France
{tessier}@nantes.inra.fr

Abstract *Recent and important discovery to understand the machinery responsible of the insertion of membrane proteins was the results of Hessa experiments [3]. The authors developed a model system for measuring the ability of insertion of engineered hydrophobic amino acids segment in the membrane. The main results of these experiments are summarized in a new biological hydrophobicity scale where each amino acid is represented by a curve. The scale shows that the contribution of each amino acid in the process of insertion of membrane proteins depends on its position along the curve. In this paper, we introduce a predictive method that models the biological behaviors of proteins that cross the translocon machinery. The method uses a local search approach to optimize an insertion curve for each amino acid. The experiments performed on a dedicated dataset and on reference protein datasets show a very good discrimination between signal peptides and transmembrane segments.*

Keywords membrane protein, peptide signal, transmembrane segment, insertion curves, local search.

1 Introduction

Subcellular localization of proteins is important for the understanding of protein function. Proteins transported across the endoplasmic reticulum (ER) membrane include soluble proteins and membrane proteins. Recent studies have led to a better understanding of the transport mechanisms of these proteins [6,9]. A targeting signal localized in the N-terminal sequence and called signal peptide (SP) guides the nascent protein to the ER membrane. Next, the nascent protein gets into the sec61 translocon, a protein complex located in the ER membrane. The translocon discriminates between the proteins which cross the ER membrane and are released in the ER lumen, and the proteins which get inserted in the ER membrane. When membrane proteins lack discrete signal peptides, the first transmembrane sequence directs the nascent protein to the membrane like a signal peptide. In this case, the first transmembrane segment is called a signal anchor (SA). The recognition inside the translocon channel is based on identifying the ”right key”. If the segment of amino acids contains the code, the translocon opens sideways and the protein fits in the membrane. Otherwise, the protein is fully translocated across the ER membrane and released into the ER lumen.

There exist several methods using both experimental and statistical data that are designed to predict the insertion of membrane proteins. Some of them are based on experimental works that try to elucidate precisely how membrane proteins get inserted or secreted through the ER membrane. Scampi [1] is such a prediction method using published experimental results of the energetics of insertion of a single transmembrane (TM) segment into the ER membrane [3]. MINS [8] and MINS2 [7] use computational methods for predicting the membrane insertion free energies of protein sequences. Recently, several experiments were designed to identify the specific ’sequence-coding’ for membrane insertion. Hessa *et al.* [2] carried out a series of in vitro experiments which assess the contribution of each amino acid in different positions along the membrane. The experiments revealed that the amino acid position plays a determining role during targeting by the translocon. So, Hessa et al. suggest a ‘biological hydrophobicity scale’ derived from their experiments.

The purpose of the present study was to elaborate a new "in silico" scale, where the insertion profile of each amino acid is represented by a curve that gives the amino acid contribution according to its position inside the translocon channel. To achieve this goal, we propose to study the insertion phenomenon on two sets of protein segments which cross the translocon and share the same chemical hydrophobic profile : SP and TM segments. We need to build appropriate datasets to learn the curves for each amino acids. The curves which we learned could benefit from the data stored in the protein databases and consequently could be much more precise than the scales derived by in vitro experiments. Also, we need to define a classifier that can discriminate between signal peptide sequences and TM segment sequences. The classifier is determined by 20 curves that represent the insertion profiles of the 20 amino acids. Finally, we have to use an efficient method to optimize the curves. In this work, we decide to use a local search (LS) approach [5] which relies on a neighborhood relation to explore effectively the search space of candidate solutions. Starting from an initial solution that can be chosen randomly or according to some relevant knowledge, LS iteratively moves from the current position to a neighboring solution with the objective to improve the solution quality that is measured by an evaluation function. The neighbors of a solution s are solutions slightly different from s, that are obtained by particular modifications applied to s. The process ends up when no improving neighbor can be found for the current solution which is therefore a local optimum of the problem.

The experiments conducted on a different datasets show that our approach improves the initial solution to reach a good accuracy. The performance results are comparable to several methods dedicated to the discrimination between signal peptides and the transmembrane segments in the literature.

Acknowledgements

This research was partially supported by the region Pays de la Loire (France) for the regional project BIL.

References

[1] A. Bernsel, H. Viklund, J. Falk, E. Lindahl, G. von Heijne, and A. Elofsson. Prediction of membrane-protein topology from first principles. *Proceedings of the National Academy of Sciences of the Unites States of America*, 105(20):7177–7181, 2008.

[2] T. Hessa, H. Kim, K. Bihlmaier, C. Lundin, J. Boekel, H. Andersson, I. Nilsson, S.H. White, and G. von Heijne. Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature*, 433(7024):377–381, 2005.

[3] T. Hessa, N. M. Meindl-Beinker, A. Bernsel, H. Kim, Y. Sato, M. Lerch-Bader, I. Nilsson, S.H. White, and G. von Heijne. Molecular code for transmembrane-helix recognition by the Sec61 translocon. *Nature*, 450(7172):1026–U2, 2007.

[4] T. Hessa, N.M. Meindl-Beinker, A. Bernsel, H. Kim, Y. Sato, M. Lerch-Bader, I. Nilsson, S.H. White, and G. von Heijne. Molecular code for transmembrane-helix recognition by the Sec61 translocon. *Nature*, 450(7172):1026–U2, 2007.

[5] H. H. Hoos and T. Stutzle. *Stochastic Local Search—Foundations and Applications*. Morgan Kaufmann Publishers, San Francisco, CA, 2004.

[6] E.C. Mandon, S.F. Trueman, and R. Gilmore. Translocation of proteins through the sec61 and secyeg channels. *Curr Opin Cell Biol*, 21:501–507, 2009.

[7] Y. Park and V. Helms. MINS2: Revisiting the molecular code for transmembrane-helix recognition by the Sec61 translocon. *Bioinformatics*, 24(16):1819–1820, 2008.

[8] Y. Park and V. Helms. Prediction of the translocon-mediated membrane insertion free energies of protein sequences. *Bioinformatics*, 24(10):1271–1277, MAY 15 2008.

[9] T. A. Rapoport. Protein transport across the endoplasmic reticulum membrane. *Febs Journal*, 275(18):4471–4478, 2008.

Differential expression analysis and SNP detection from RNA-Seq data in *Asobara tabida*, a parasitoid of *Drosophila*

Marie-Christine CARPENTIER[1], Janice KIELBASSA[1,2], Vincent LACROIX[1], Marie-France SAGOT[1,2] 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] INRIA Grenoble Rhône-Alpes, France

Keywords RNA-seq. De novo assembly. SNP. Insect symbiosis

Wolbachia, a maternally transmitted intracellular bacterium, is widely distributed in arthropods and generally facultative for its hosts. In contrast, in the hymenoptera *Asobara tabida*, a parasitoid of *Drosophila* larvae, *Wolbachia* has become obligatory for oogenesis. Two phenotypes are observed in the absence of *Wolbachia*: either females can not produce any eggs, or females produce some eggs, but they are sterile. This difference is due to the genetic variability of the host.

To understand the functional basis of this polymorphism, the ovarian transcriptome of infected and uninfected individuals from two *Asobara* lineages with the two different phenotypes have been sequenced (RNA-Seq, Illumina single-end 76pb). As no reference genome is available for *Asobara*, we assembled the transcriptome de novo and quantified the expression level of genes (velvet[1] and oases[2]). Then, the differentially expressed genes between the infected and uninfected individuals for the two phenotypes were identified (R package DESeq[3]).

Furthermore, we searched for single nucleotide polymorphism (SNP) between the two lineages. To this end, we ran KisSplice[4], a local de-bruijn graph assembler, on the same datasets, in order to analyze the relationship between the sequence variation and the ovarian phenotype. We observed that SNPs are more often found in differentially expressed genes than in the genes that were not. In future work, we will be looking for mutation patterns (i.e. synonymous and non synonymous mutations) and then characterize polymorphism in other lineages in order to perform association analyses between the specific alleles and the oogenesis phenotype.

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] M.H. Schulz et al., Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, 2012.

[3] S. Anders and W. Huber, Differential expression analysis for sequence count data. *Genome Biology*, 11:R106, 2010.

[4] Sacomoto et al., KisSplice: de-novo calling alternative splicing events from RNA-seq data.*BMC Bioinformatics*, 2012.

Providing Bioinformatics Services on Cloud

Christophe BLANCHET[1], Clément GAUTHEY[2] and Charles LOOMIS[2]
[1] Infrastructure Distribuée pour la Biologie, IBCP-CNRS FR3302, 7 passage du Vercors, 69007, Lyon, France
{christophe.blanchet, clement.gauthey}@ibcp.fr
[2] LAL, UMR8607 CNRS, Bât. 200 BP 34, lieu, 91898, Orsay, Cedex, France
loomis@lal.in2p3.fr

Abstract *Improvements of experimental technologies forces biologists to face a deluge of data that require relevant tools and sufficient resources to be analyzed. The cloud helps bioinformatics experts to define virtual appliances with pre-installed tools and workflows, and helps scientists to deploy them, on demand, on research infrastructures.*

Keywords Bioinformatics services, Predefined appliances, Cloud computing.

1 Introduction

Biologists and bioinformaticians simultaneously use many of the bioinformatics tools from the arsenal of thousands available from the international community. Most of the time, they also need to combine multiple software packages to study their data with public or their own analysis pipelines [1]. For intensive usage, they have access to computing clusters through command line interfaces. The typical usage, however, is through web portals and services for the ease of use and for the capacity to compose these tools into pipelines [2]. For several years, the focus has been put on providing such composable services and defining standards. The European project EMBRACE has led to online public programmatic services [3] like those of the French platform 'Infrastructure Distributed for Biology' (IDB, http://gbio-pbil.ibcp.fr/ws). But these bioinformatics applications can process gigabytes of data stored in flat-file databases like UNIPROT, GenBank, PDBseq or genome-specific files. And they require access to reference databases in a POSIX manner (like NFS) to the cloud storage containing the biological data.

2 Bioinformatics Services on Cloud

The platform in collaboration with the StratusLab EU-FP7 project provides cloud services to the bioinformatics community with the StratusLab framework [4]. They have developed several appliances, predefined virtual machines, with installed and configured bioinformatics tools (See Fig. 1). Two examples of them are the 'biocompute' appliance (with BLAST, ClustalW2, Clustal-Omega, FastA, etc.) and the 'biodata' appliance (with databases Swiss-Prot, PROSITE, etc.) providing easy access to common tools and data.

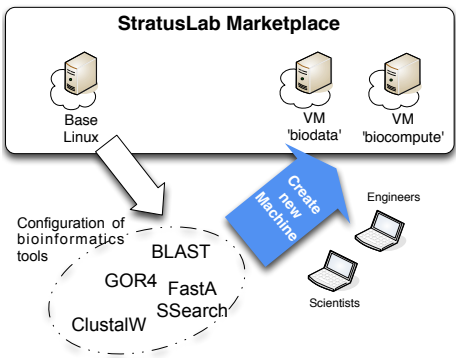


Figure 1. Creation of bioinformatics appliances with standard bioinformatics tools and workflows.

As identifying and running the relevant bioinformatics appliances could be difficult, we have developed a custom web interface to the cloud (See Fig. 2). This portal is coupled with the StratusLab Marketplace where

the bioinformatics are referenced and tagged with RDF metadata. These metadata can be used to help to select the right bioinformatics appliance according to the desired tools (BLAST, ClustalW, etc.) or the kind of analysis to perform (sequence or structural analysis, assembling/mapping, etc.). The bioinformatics portal aims to make the cloud easy to use by non-computing and non-cloud-specialist scientists, like biologists and bioinformaticians. The available features allow the creation and termination of the virtual machines, the management of the persistent disks, and provide assistance for bioinformatics appliance selection and instance contextualization.

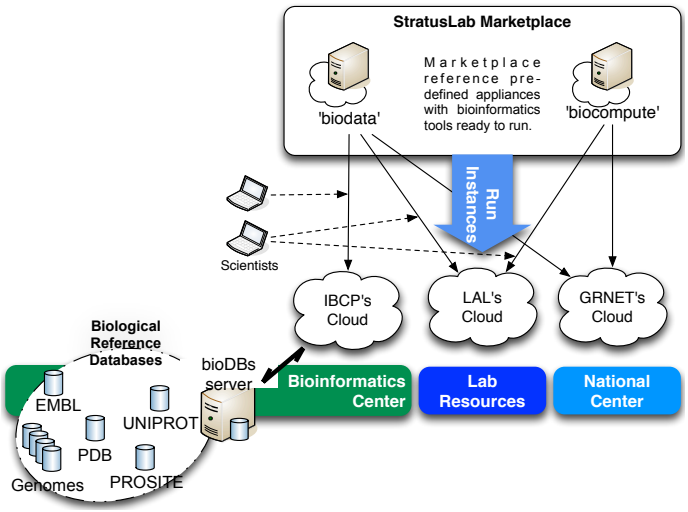


Figure 2. Deployment on demand of bioinformatics appliances on research infrastructures.

3 Conclusions

The adoption of clouds for bioinformatics applications will be strongly correlated to the capability of cloud infrastructures to provide ease-of-use of common bioinformatics tools and access to reference biological databases. In that way, clouds for bioinformatics have to be connected with public bioinformatics infrastructures and StratusLab is collaborating with the French Bioinformatics Network RENABI (www.renabi.fr) to help fulfill the requirements from the Bioinformatics community.

Acknowledgements

IDB and LAL acknowledge co-funding by the European Community's Seventh Framework Programme (INFSO-RI-261552), and IDB also by the French National Research Agency's Arpege Programme (ANR-10-SEGI-001).

References

[1] M. Kalas, P. Puntervoll, A. Joseph, E. Bartaseviciute, A. Töpfer, P. Venkataraman, S. Pettifer, J. C. Bryne, J. Ison, C. Blanchet, K. Rapacki, and I. Jonassen, BioXSD: the common data-exchange format for everyday bioinformatics web services. *Bioinformatics*, vol. 26, no. 18, pp. i540–6, Sep. 2010.

[2] G. Benson, Editorial. Nucleic Acids Research annual Web Server Issue in 2011. *Nucleic Acids Research*, vol. 39, no. Web Server, pp. W1–W2, Jun. 2011.

[3] S. Pettifer, J. Ison, M. Kalas, D. Thorne, P. McDermott, I. Jonassen, A. Liaquat, J. M. Fernández, J. M. Rodriguez, INB-Partners, D. G. Pisano, C. Blanchet, M. Uludag, P. Rice, E. Bartaseviciute, K. Rapacki, M. Hekkelman, O. Sand, H. Stockinger, A. B. Clegg, E. Bongcam-Rudloff, J. Salzemann, V. Breton, T. K. Attwood, G. Cameron, and G. Vriend, The EMBRACE web service collection. *Nucleic Acids Research*, vol. 38, no. Web Server issue, pp. W683–8, Jul. 2010.

[4] C. Loomis, M. Airaj, M. E. Bégin, E. Floros, S. Kenny, and D. O'Callaghan, StratusLab Cloud Distribution. *In European Research Activities in Cloud Computing, Dana Petcu and Jose Luis Vasquez Poletti (Ed.)* 2012 271.

GPCR_AlignDB: a database of aligned sequences of G-protein-coupled receptors

Jean-Michel BECU, Chloé RAIMBAULT, Julien PELE, Matthieu MOREAU and Marie CHABBERT

Laboratoire BNMI, UMR CNRS 6214 – INSERM U1083, Faculté de Médecine, 3 rue Haute de reculée, 49045 ANGERS, France

jean-michel.becu@etu.univ-rouen.fr; marie.chabbert@univ-angers.fr

Abstract *With more than 85 000 reported sequences, G-protein-coupled receptors (GPCRs) constitute one of the largest receptor families, involved in numerous physiological functions and diseases. The analysis of the evolutionary information hidden within the reported sequences will help understand the sequence-function relationships of these receptors. However, this analysis requires a considerable effort in the classification and alignment of these sequences. We developed a manually curated database of aligned sequences of GPCRs, GPCR_AlignDB, with a phylogeny-based classification. This database facilitates the evolutionary analysis of the different sub-families through a wide range of organisms from cnidarians to mammalians and should help understand the diversification of the GPCR family.*

Keywords GPCR, database, classification, multiple sequence alignment

1 Introduction

With more than 85 000 reported sequences, the G-protein-coupled receptors (GPCRs) form one of the largest transmembrane receptor families, involved in numerous physiological functions and diseases. The analysis of the evolutionary information hidden within the reported sequences will help understand the sequence-function relationships of these receptors. However, this analysis requires a considerable effort in the classification and alignment of the sequences. The GPCR_DB database reports more than 30 000 sequences but classifies them according to the nature of the ligands [1]. The GRAFS classification system [2], which is now widely adopted, is based on phylogeny and classifies vertebrate receptors in 5 classes. The main class A is further classified into a dozen of sub-families. This classification facilitates the analysis of evolutionary information [3,4]. We thus developed a database, GPCR_AlignDB, to handle GPCR sequences and their alignment with phylogeny-based classification.

2 General Organization

The aim of the database is the easy selection of *aligned* GPCR sequences based on either their phylogenetic classification or specific sequence patterns. The database has been organized to facilitate the manual curation of secondary data and the updating of multiple sequence alignments.

The database contains primary information, extracted from Uniprot, and secondary information processed from sequence analysis. Secondary data include quality information, clustering information, classification, alignments and specific sequence patterns. Key steps in the secondary data processing are curated manually in order to insure the quality of the sequences and of the alignments.

Data are extracted from UniProt and subsequently processed by species and GPCR families. Only species whose complete proteome is available are considered. These species correspond to the keyword “complete proteome” in the Uniprot entries. Presently, the database is limited to the repertoire of class A GPCRs. This family includes 90% of human GPCRs and corresponds to the PS50262 profile in UniProt.

The analysis of the GPCR repertoire requires the clustering of redundant sequences. Clustering is included within the database. The selection of a representative sequence from each cluster is carried out manually with the help of a visual tool included in the database. All the subsequent secondary data processing, including sequence alignments, is carried out on the representative sequence of each cluster.

Alignments of full length sequences are done by species. These alignments are used to determine the sequence quality, the sub-family classification and specific sequence patterns. These data and the full length alignments are uploaded into the database. When sub-families or sequence patterns are searched for in the database, and the hits originate from different species, the alignment of the hit transmembrane regions is generated and can be downloaded for further analysis.

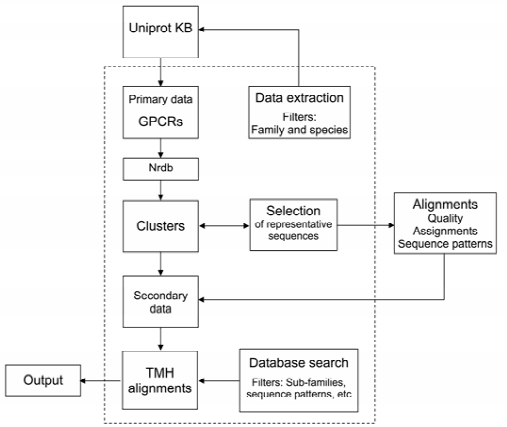


Figure 1. General organization of GPCR_AlignDB

3 Conclusions

We have developed a GPCR database that facilitates the handling of aligned sequences. Presently, GPCR_AlignDB contains more than 4000 sequences of non olfactory class A GPCRs from eight species (*H. sapiens*, *M. musculus*, *D. rerio*, *M. melanogaster*, *C. intestinalis*, *B. floridae*, *C. elegans* and *N. vectensis*). These sequences correspond to a non redundant set of 2300 high quality sequences that are assigned and aligned. The analysis of this sequence set will give insights into the evolutionary history of specific sub-families or sequence patterns and will help understand the sequence determinants that have led to the divergence and specialization of the GPCR family.

Acknowledgements

This work was supported by institutional grants from CNRS, INSERM and University of Angers and by a grant from the Agence Nationale de la Recherche (ANR-11-BSV2-026). We thank NEC Computers SA (Angers) for the kind availability of a multi-processor server. JP was supported by a fellowship from Conseil Général de Maine-et-Loire. JMB was supported by studentships from CHU of Angers and CNRS.

References

[1] F. Horn, E. Bettler, L. Oliveira, F. Campagne, F.E. Cohen and G. Vriend, GPCRDB information system for G protein-coupled receptors. *Nucleic Acids Res, 31*:294-7, 2003.

[2] R. Fredriksson, M.C. Lagerstrom, L.G. Lundin and H.B. Schioth, 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.

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

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

*my*ProMS, a Software for Mass Spectrometry-based Proteomics Data Collaborative Processing and Analysis

Florent YVON[1], Guillaume ARRAS[2], Falaye CAMARA[1], Damarys LOEW[2], Emmanuel BARILLOT[1] and Patrick POULLET[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
{florent.yvon, guillaume.arras, patrick.poullet}@curie.fr
[2] Mass Spectrometry and Proteomics platform, Institut Curie, 26 rue d'Ulm, 75248 Paris, Cedex 05, France

Keywords Proteomics, mass spectrometry, software tools.

Mass Spectrometry (MS) applied to Proteomics produces large amount of data. Protein identification data generated by MS-search engines such as Mascot need to be curated, analyzed and interpreted in order to extract meaningful biological information. These processes require the skills of both bioinformaticians, MS specialists and biologists.

*my*ProMS has been developed to facilitate these different steps of data processing. It is a Perl-CGI web software that relies on a MySQL database. It uses a selective access level management to data processing depending on the user skills. *my*ProMS supports Mascot and Proteome Discoverer output files that can be easily imported in the software database. Imported data can be accessed by MS specialists for automatic or manual curation of the protein identifications given by the search engine. Post-translational modification site attributions can also be corrected, including phosphorylations with the included PhosphoRS software [1]. Validated data are shared with biologists, which can analyze their results with several *my*ProMS integrated tools. These tools include sample comparison, computation and visualization of quantitative data with interactive graphical displays, data interpretation using Gene Ontology with GO::TermFinder enrichment analysis tool [2] or individual analysis of each protein with multiple external links.

*my*ProMS is used by multiple laboratories and is continuously improved based on user feedbacks. The software can be evaluated and downloaded freely for academic users at http://myproms-demo.curie.fr.

References

[1] T. Taus, T. Köcher, P. Pichler, C. Paschke, A. Schmidt, C. Henrich and K. Mechtler, Universal and Confident Phosphorylation Site Localization Using phosphoRS, *Journal of proteome research*, 10:5354-5362, 2011

[2] E.I. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J.M. Cherry and G. Sherlock, GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, 20:3710-3715, 2004

PARYS, a Web Server for Managing Reverse-Phase Protein Array Platform Data

Patrick Poullet[1], Stéphane Liva[1], Sylvie Troncale[1], Leanne de Koning[2], Fanny Coffin[1], Belilei He[1], Philippe Hupé[1], and Emmanuel Barillot[1]

[1]U900 INSERM - Mines ParisTech - Institut Curie, Bioinformatics and Computational Systems Biology of Cancer, Institut Curie, 26 rue d'Ulm, 75248 Paris cedex 05, France

[2] Translational Research Department - Institut Curie, Hôpital Saint-Louis, 75475 Paris cedex 10, France

{patrick.poullet, stephane.liva}@curie.fr

Keywords Protein array, protein quantification, phosphorylation, antibody.

1 Introduction

The Reverse-Phase Protein Array (RPPA) is an emerging high-throughput technology relying on highly specific antibodies to quantify proteins in nanogram-amounts of spotted cell or tissue lysates. This technology has thus 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. Typical research projects utilizing this platform focus on finding new therapeutic targets by examining the expression and activation (eg. phosphorylation state) levels of multiple proteins in hundreds of tumour biopsies.

To fully exploit the potential of RPPA technology, we have developed PARYS (Protein Array Server), a comprehensive bioinformatics environment for the platform.

2 Methods

PARYS is developed in Perl-CGI, HTML and JavaScript. It relies on an Apache web server coupled with a MySQL database. Data analysis is performed using R statistical language.

3 Results

PARYS is organised into two main sections:

- 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 spot quantification.

- A Project section to coherently organize, mine and analyze the quantitative data generated.

PARYS implements an extended data annotation scheme coupled with differential and exploratory analysis tools to provide the platform collaborators with means to extract relevant biological information from their data.

PARYS is now fully integrated to our RPPA platform activity. To date, over 1400 antibodies targeting critical biological pathways are referenced with detailed information on their suitability for RPPA experiments. Use of the PARYS server has significantly reduced the delay between raw data generation and their biological interpretation.

EMA - A R package for Easy Microarray data Analysis

Nicolas SERVANT[123]⋆, Eleonore GRAVIER[1234]⋆, Pierre GESTRAUD[123], Cecile LAURENT[123678], Caroline PACCARD[123], Anne BITON[1235], Jonas MANDEL[123], Bernard ASSELAIN[123], Emmanuel BARILLOT[123] and Philippe HUPÉ[1235]

[1] Institut Curie, Paris, France
[2] INSERM U900, Paris, France
[3] Mines ParisTech, Fontainebleau, France
[4] Institut Curie, Departement de Transfert, Paris, France
[5] CNRS, UMR144, Paris, France
[6] CNRS, UMR3347, Orsay, France
[7] INSERM, U1021, Orsay, France
[8] Université Paris-Sud 11, Orsay, France
ema-support@curie.fr

Keywords R package, Microarray, Transcriptome 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**) 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**, or **mostclust** 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) 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.
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.

References

[1] Bertoni, A. and Valentini, G. Model order selection for bio-molecular data clustering. *BMC Bioinformatics*, 8(2):S7, 2007.

[2] Servant, N. and Eleonore, G. and Gestraud, P. and Laurent, C. and Paccard, C. and Biton, A. and Brito, I. and Mandel, J. and Asselain, B. and Barillot, E. and Hupé, P. EMA - A R package for Easy Microarray data Analysis. *BMC Research Notes*, 3:277, 2010.

[3] Le, S. and Josse, J. and Husson, F. FactoMineR: an R package for multivariate analysis. *Journal of statistical software*, 25:1-18, 2008.

[4] Falcon, S. and Gentleman, R. Using GOstats to test gene lists for GO term association. *Bioinformatics*, 23:257-258, 2007.

[5] Tusher, V. G. and Tibshirani, R. and Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*, 98:5116-21: 2001.

⋆. These authors contributed equally to this work

A pipeLine Dedicated to Oligonucleotides design (ALDOv1.2)

A workflow for molecular detection 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 *Design of primers and probes is an important step in the conception of molecular detection reagents. Bioinformatics is very helpful both for the design of these oligonucleotides and to validate candidates in silico before in vitro testing, which is an expensive and time-consuming task. ALDOv1.2 is a bioinformatics pipeline dedicated to the conception of primers and probes for qPCR applications . Moreover, a part of the pipeline may be used for monitoring assay performance to cope with the emergence of new variants of targeted alleles. To our knowledge, this tool is the first one able to automatically adapt primers and probe sequences in response to new evolving target alleles.*

Keywords primers and probes, assay design, *in silico* validation, qPCR simulation.

1 Introduction

Advances in genome sequencing technology led to the availability of many pathogen genome sequences, making sequence-based detection assays (combination of two primers and a probe) an attractive option in comparison to culture- or immuno-based assays. Molecular *in vitro* detection consist to analyze biological samples in the aim to detect and measure the presence of targeted species. Primers and probes are used in molecular detection reagents, with different technologies such as qPCR, to detect organisms by targeting species-specific genomic sequences. On the one hand, some bioinformatics tools and programs were developed for designing candidate primers and probe. On the other hand, other programs validate these oligonucleotides *in silico* with sequence-based simulations of PCR.

Our objective was to develop a pipeline integrating both the steps of designing primers and probes and their subsequent validation *in silico*. In this context, we propose a thermodynamics-based simulation model. The bioinformatics workflow delivers assay candidates ranked according to their performance score. qPCR primers and probes designed via ALDOv1.2 have several types of applications. Among this, we can cite molecular diagnostics of infectious diseases, molecular detection of species in environmental samples, or detection of cancer-related fusion transcripts in human samples.

2 Principle

According to the complexity of the project, two main tools comprise the design part of ALDOv1.2 : ALDO [1] for the design of primers and probes for simplex reaction conditions (one target) and ALDOplex for the design for multiplex conditions (more than one target). ALDOv1.2 relies both on internally-developed programs as well as public programs such as Primer3 [2], EMBOSS utilities [3], NCBI eUtils programs [4] and Fastagrep [5].

2.1 ALDO for simplex assay design

ALDO is used for relatively well conserved genetic sequences. The design step is performed by SLv8 program, searching for locally conserved regions into the multiple sequence alignment (MSA) of the targeted sequences. SLv8 program encapsulates Primer3. Candidates primers and probes are selected by SLv8 in accordance with more than 20 qPCR-specific design rules. Each set of primers and probes (qPCR assay) becomes a candidate assay whose performance is computed *in silico*.

2.2 Candidate Assay Validation (eNv4 model)

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, eNv4 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 of non-targeted species).

2.3 ALDOplex for multiplex assay design

For more complex targets, where MSA construction is difficult or even impossible due to highly dissimilar target sequences, ALDOplex is used. For each individual target sequence, an individual design step is performed and candidate assays are designed. Subsequently, performance of each assay candidate is computed globally on all targeted sequences. Finally, combinations of best multiplexes assays candidates are kept, if they allow to detect the maximum number of the targeted sequences with a minimum set of primers and probes.

2.4 Assay Performance Surveillance

A recent survey [6] showed that two thirds of published primers are not able to properly detect every genetic variant of a gene. Therefore, a surveillance tool based on the eNv4 model has been developed. Regularly, a qPCR simulation reaction against an updated database is launched to check the performance of specific assay candidates. The global performance of an assay is evaluated both in term of inclusivity (ability to detect targeted species) and exclusivity (ability not to cross-detect non-targeted species). The program is able to warn the user in case of degradation of the assay performance, due to the emergence of new alleles, and automatically improve the primers and probes sequences in order to restore the optimal performance.

3 Results and conclusion

In the context of a project dedicated to the detection of highly drug-resistant bacteria producing Klebsiella pneumoniae carbapenemases (KPC), a design was performed using ALDO. The design was based on 10 different alleles of the KPC gene (KPC2 to KPC11). Three best candidates assays proposed by the tool were tested *in vitro* and yielded good results. Later, new alleles KPC12 and KPC13 were published and were added to our sequence collection. One of our assays was impacted, as a mismatch occurred in the 3' end of its reverse primer. This mismatch degraded the performance of the test, as the score calculated by eNv4 decreased regarding the total number of sequences to be detected. This assay candidate was discarded, and only the other non-impacted assays were are kept.

ALDOplex is currently used to design assays for other on-going multiplex-needing projects.

The design modules (ALDO and ALDOplex) have been validated but the validation of the auto-improvement of assays by the surveillance module is still on-going.

Acknowledgements

This work was supported both by bioMérieux and the University of Rouen. We thank Dr. Fritz Schwarzmann for reviewing this paper.

References

[1] I. Rabearivelo, F. Paillier, *A pipeLine Dedicated to Oligonucleotides design (ALDO).* JOBIM 2011 poster abstract #127.

[2] 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.

[3] P. Rice, I. Longden, and A. Bleasby, *EMBOSS: The European Molecular Biology Open Software Suite.* Trends in Genetics 16, (6) pp276-277, 2000.

[4] E. Sayers, D. Wheeler, *Building Customized Data Pipelines Using the Entrez Programming Utilities (eUtils).* NCBI eUtils http://eutils.ncbi.nlm.nih.gov/

[5] L. Kaplinski, *Fastagrep*. http://bioinfo.ebc.ee/download

[6] J. Gardès PhD thesis. *Etude et conception in silico d'amorces PCR pour l'identification des principaux pathogènes bactériens.*

LTR75 and LTR75B retrotransposon in mammals: over-representation on X chromosome and co-regulation of nervous system genes

Sébastien TEMPEL[1], Kenji KOJIMA[2] and Jerzy JURKA[2]

[1] IBISC - IBGBI, 3ème étage, 23, Bd de France 91034 EVRY, France
{sebastien.tempel}@ibisc.univ-evry.fr
[2] Genetic Information Research Institute, 1925 Landings Dr., Mountain View, CA 94043, USA
{kojima, jurka}@girinst.org

Abstract *LTR75 and LTR75B are families of long terminal repeats (LTR) introduced in ancient mammals. These families are over-represented on X chromosomes, especially nearby the nervous-related genes. LTR75 and LTR75B sequences contain consensus transcription factor binding sites that are known to participate in regulation of the 'Nerve Genes'.*

Keywords transposable elements, nervous system, co-regulation

1 Introduction

Some transposable element (TEs) families, such as L1, are over-representated in mammalian X chromosome and are linked to the X-chromosome inactivation [1]. Old TE families such as SINE, MIR or LTR can supply Transcription Factor Binding Sites [2] in mammalian genomes. They can work as enhancer and regulate distant genes [3]. TFBS from multiple TEs can create a regulatory network [4] that targets a specific tissue [5] such as nervous system [6]. In this article, we analyze LTR75 and LTR75B families that are over-represented on the X chromosome in the dog, human, mouse and rat genomes. We found that the orthologous genes belonging to the MeSH term 'Nerve' were over-represented. Moreover, the LTR75 sequences close to these orthologous genes contain numerous TFBS related to the nervous system regulation.

2 Materials and Methods

The annotations and genomic sequences of the dog, human, mouse and rat genomes came from the NCBI website. We used Censor [7] to detect the occurrences of LTR75 and LTR75B (LTR75/75B) in the four mammalian genomes. We analyzed the closest genes of each LTR75 and LTR75B (LTR75/75B) elements, defined as "genic neighborhood". A gene function list of the neighboring genes was downloaded from the Gene Ontology website. For each genic neighborhood, we extracted all MeSH terms in the 'Biological Process' and the 'Cellular Component' categories. We used the χ^2 statistical test to determine whether or not the LTR75/75B occurrences are over-represented in any kind of MeSH terms for the total set of neighboring genes.

3 Results

We identified 238 occurrences of LTR75 and 274 occurrences of LTR75B families in the human genome. In the dog genome there were 198 and 186 occurrences of LTR75 and LTR75B elements, respectively. In the mouse genome the corresponding occurrences were 91 and 48, and in the rat genome there were 80 LTR75 and 47 LTR75B elements. In the four genomes, we obtained an over-representation of LTR75/75B occurrences in mammalian X chromosomes. For example, 39 LTR75B occurrences in human X chromosome correspond to more than 15% of the total LTR75B occurrences.

We analyzed the genic neighborhood to all LTR75/75B occurrences. We found 58 orthologous genes shared among all four genomes in the neighborhood of LTR75/75B occurrences. We also found 79 genes shared among three of the four species (Fig. 1).

We extracted the Mesh Terms of these orthologous genes. We obtained 35 'Nerve Genes' for a total of 137 orthologous genes. The calculated χ^2 score of these 35 orthologous 'Nerve Genes' assuming independence

ACADL	DRP2	MERTK	THOC2	FLRT2	TEX13B	GRM4	ANAPC10	LUZP2
ACBD7	EFCAB6	MOSPD2	TRDN	IGFBP7	TNKS	HTR1A	ANK3	LYPLAL1
AFF2	EHHADH	PAPD4	TRIM25	KIAA1486	TRHDE	IPO11	ASB18	MED14
ANAPC1	ELOVL4	PCYT1B	USP9X	LIPC	ZBTB34	KCMF1	ATXN10	MORC4
AP1AR	FAM3C	PDZRN4	VAMP4	LRRC55	ACO2	LIMD1	C2ORF67	NCAM2
ASB9	GPR112	PPFIA2	VIT	LRRN3	E230019M04RIK	MAN2A2	CDC5L	SOX3
ATP1B4	HOMER1	PRRG1	XIAP	MAGED2	PRKRIP1	MATN1	CDH18	TIFA
CCDC59	HPRT1	PTPLAD1	ZC3H6	MAL2	PTP4A1	NOVA1	CHST4	USP11
CNKSR2	IBSP	PTPRZ1	ZC3H8	MID2	SH2B2	NUDT11	DOCK4	ZFAT
CNTN1	IQCF6	SCML4	ZDBF2	PLAC1	TCEAL3	PCM1	GRID1	
COL6A3	LAMP2	SLAMF7	ABI2	QRFPR	ADAM23	TNIK	GRM7	
COPS8	LDOC1	SLC15A1	APIP	RPRM	CDH9	XRN1	HHIP	
CSMD3	LGSN	SLC6A20	ARL6IP6	SCGB3A2	ERC1	ZCCHC12	IGFBP3	
CYP27A1	LY9	SLC30A10	C1GALT1C1	SEL1L	FES	ZNF24	KHDRBS3	
DGKE	LZTFL1	SUPT3H	COLEC10	SPINK1	FOXJ3	ZNRF3	KIF6	
DPP8	MEPE	TAF7L	EHF	STRN	GK5	AGTPBP1	KLHL13	

Gene	Orthologous in four genomes	Gene	Orthologous in dog, human and rat genomes
Gene	Orthologous in human, mouse and rat genomes	Gene	Orthologous in dog, human and mouse genomes
Gene	Orthologous in dog, mouse and rat genomes	Gene	Nerve related genes

Figure 1. Orthologous genes in the whole dog, human, mouse and rat genomes. The red names are 'Nerve Genes'.

between 'Nerve Genes' and LTR75/75B occurrences was 23.04. These results showed that the LTR75/75B occurrences have a strong relationship with the 'Nerve Genes'.

We selected the LTR75/75B occurrences which are near the orthologous genes and we used the PROMO website [8] to detect the transcription factor binding sites (TFBS) present in these occurrences. The 30 of 35 orthologous LTR75/75B occurrences near 'Nerve Genes' have at least one 'Nerve TFBS'. These results show the LTR75/75B sequences could regulate with their 'Nerves TFBS' the neighboring 'Nerve Genes'.

4 Conclusion

LTR75 and LTR75B were old transposable elements that were over-represented on the X chromosome in the dog, human, mouse and rat genomes. The analysis of genes neighboring the occurrences showed the preferential accumulation of these elements close to the 'Nerve Genes'. The LTR75/75B sequences contain TFBS that can regulate the orthologous 'Nerve Genes' and/or regulate a sub-network of 'Nerve Genes'.

References

[1] Lyon, MF, The Lyon and the LINE hypothesis. *Semin Cell Dev Biol*, 14:313-8, 2003.

[2] Conley AB and Jordan IK, Identification of transcription factor binding sites derived from transposable element sequences using ChIP-seq. *Methods Mol Biol.*, 674:225-40, 2010.

[3] Sasaki, T and Nishihara, H and Hirakawa, M and Fujimura, K and Tanaka, M and Kokubo, N and Kimura-Yoshida, C and Matsuo, I and Sumiyama, K and Saitou, N and Shimogori, T and Okada, N, Possible involvement of SINEs in mammalian-specific brain formation. *P.N.A.S.*, 105:4220-5, 2008.

[4] Wang, J and Bowen, NJ and Marino-Ramirez, L and Jordan, IK, A c-Myc regulatory sub- network from human transposable element sequences. *Mol Biosyst*, 5:1831-9, 2009.

[5] Jjingo D and Huda A and Gundapuneni M and Marino-Ramirez L and Jordan IK, Effect of the transposable element environment of human genes on gene length and expression. *Genome Biol Evol.*, 3:259-71, 2011.

[6] Vedrine, SM and Vourc'h P, and Tabagh, R and Mignon, L and Hofflin, S and Cherpi-Antar, C and Mbarek, O and Paubel, A and Moraine, C and Raynaud, M and Andres, CR, A functional tetranucleotide (AAAT) poly- morphism in an Alu element in the NF1 gene is associated with mental retardation. *Neurosci Lett.*, 491:118-21, 2011.

[7] Kohany, O and Gentles, AJ and Hankus, L and Jurka, J, Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics*, 7:474, 2006.

[8] Farre,D and Roset,R and Huerta,M and Adsuara,J.E and Rosello,L and Alba,M.M and Messeguer,X, Identification of patterns in biological sequences at the ALGGEN server: PROMO and MALGEN. *Nucleic Acids Res.*, 31:3651-3653, 2003.

Reconstruction of ancestral isochores in Boreoeutheria chromosomes

Céline BON and Hugues ROEST CROLLIUS
Ecole Normale Supérieure, Institut de Biologie de l'ENS, IBENS, CNRS UMR8197, 46 rue d'Ulm, 75005, Paris, Cedex 05, France
{celine.bon, hrc}@ens.fr

Keywords Genome evolution, Vertebrate genomics, ancestral sequences reconstruction, GC-content

Reconstitution des séquences codantes ancestrales de Boreoeutheria

Mots-clés Evolution des génomes, Génomique des vertébrés, Reconstitution de séquences ancestrales, Contenu en GC.

Most amniote genomes differ from other vertebrates because they display large variations in base composition over large scales, called isochores[1]. Isochores shape mammalian chromosomes as a mosaic of regions of homogeneous base composition and are usually associated with genomic properties, such as a higher gene density and less introns in GC-rich isochores. It has been suggested that isochores are due to GC-biased gene conversion (gBGC), that lead to an increase of AT → GC mutations in high-recombination hotspots [2]. Because one crossover at least must appear per chromosome arm and per meiosis, the average rate of recombination is correlated to the size of the chromosome arm. Indeed, the average GC-content in short chromosomes is typically higher than in those of longer chromosomes.
It has been proposed that GC-rich isochores originated in the ancestral Amniota genome and are now disappearing from mammalian genomes. For example, genomic rearrangements that occurred in the mammalian lineage may have had an impact on the isochore structure, by leading to a homogenization of GC-content in lineages with frequent rearrangements. Until now, all studies addressing the problem of ancestral isochore evolution have relied on estimated ancestral GC3 values from small numbers of genes [3] or reasoned on the average value over the entire genome [4]. Reconstructing the ancestral isochore landscape of ancestral mammalian chromosomes would directly reveal the pattern of isochore evolutions, in terms of homogeneisation, structuration and overall GC-content variation.

Here, we took advantage of reconstructed ancestral gene orders in mammals [5] to represent individual GC3 values predicted in ancestral genes along their respective chromosomes.

First, we benchmarked two state-of-the-art ancestral sequence reconstruction methods using simulated data. Simulations, performed using *evolver* from the PAML package [6] consisted in the generation of 10 dataset of 6,000 bp sequences and 12 species, under 3 rates of divergence between sequences. We next tested the performances of the *codeml* program from PAML and the *prequel* program from the PHAST package [7] in reconstructing ancestral sequences. Interestingly, both approaches perform similarly well (0.3% to 0.03% error rates at the Boreoeutheria node for both). However, since *prequel* also resolves the occurrence of indels during evolution, while *codeml* does not take deletion into account, *prequel* was chosen for our analysis.

Gene families were downloaded from *Ensembl* [8]. Sequences for each family were aligned using M_coffee [9], and tree reconstruction was performed with Treebest [10]. Alignments and trees were then used for

ancestral sequence reconstruction by *prequel*. Here we present the analysis conducted on the 17,784 ancestral Boreoeutheria gene-coding sequences present in the families.

We find that the Boreoeutheria average GC-content is 54.0% and that the GC3 is 62.2%, which is higher than human GC and GC3 content (respectively, 46.8 % and 49.1%). We calculated the average GC3 for human chromosomes and the 20 longest (chromosome size) Boreoeutheria ancestral blocks, and observed the same trend. For each gene, we plotted the GC3 of the human gene against the value for the Boreoeutherian ortholog, and show that the GC3 decrease in human lineage is mainly due to a decrease in the GC-richest Boreoeutheria genes.
We also plotted the gene GC-content along the ancestral Boreoeutheria blocks, and evidenced variations of base composition on several blocks, that may be interpreted as ancestral isochores.

Our results favor a heterogeneous GC-rich Boreoeutheria ancestral genome, with isochore-like structuring. To analyze the origin and evolution of isochores in mammalian genomes, we now plan to compare Boreoeutheria genome with other ancestral node in order to recount the history of isochores along the Mammalian lineage.

References

[1] G. Bernardi. Isochores and the evolutionary genomics of vertebrates. Gene 241(1):3-17, 2000

[2] L. Duret, N. Galtier. Biased Gene Conversion and the Evolution of Mammalian Genomic Landscape. *Annu. Rev. Genomics Hum. Genet.* 10:285-311, 2009

[3] E.M.S. Belle, L. Duret, N. Galtier, A. Eyre-Walker. The Decline of Isochores in Mammals: An Assessment of the GC Content Variation Along the Mammalian Phylogeny. *J Mol Evol*. 58(6):653-60, 2004

[4] J. Romiguier, V. Ranwez, E.J.P. Douzery, N. Galtier. Contrasting GC-content dynamics across 33 mammalian genomes : relationship with life-history traits and chromosome sizes. Genome Res. 20:1001-09, 2010

[5] M. Muffato, A. Louis, C.E. Poisnel, H. Roest-Crollius. Genomicus: a database and browser to study gene synteny in modern and ancestral genomes. *Bioinformatics*. 26(8):1119-21, 2010

[6] Z. Yang. PAML4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24(8): 1586-91

[7] M.J. Hubisz, K.S. Pollard, A. Siepel. PHAST and RPHAST: phylogenetic analysis with space/time model. *Brief Bioinform.* 12(1):41-57, 2011

[8] P. Flicek, M.R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, S. Coates, S. Fairley, S. Fitzgerald & al. Ensembl 2012. *Nuc. Ac. Res.* 40(Database issue):D84-90, 2012

[9] C. Notredame. Computing multiple sequence/structure alignments with the T-coffee package. *Curr Protoc Bioinformatics* Chapter 3:Unit3.8:1-25, 2010

[10] A.J. Vilella, J. Severin, A. Ureta-Vidal, L. Heng, R. Durbin, E. Birney. EnsemblCompara GeneTrees: Complete, duplication- aware phylogenetic trees in vertebrates. *Genome Res.* 19(2):327-35, 2009

The SPROUTS Submission Workflow

Ruben ACUÑA[1], Zoé LACROIX[1], AND JACQUES CHOMILIER[2,3]

[1]SCIENTIFIC DATA MANAGEMENT LABORATORY
Arizona State University, P.O. Box 875706, Arizona, 85282-5706, Tempe, USA
{ruben.acuna,zoe.lacroix}@asu.edu
[2]INSTITUT DE MINERALOGIE ET DE PHYSIQUE DES MILIEUX CONDENSES and CNRS
Université Pierre et Marie Curie
75252 Paris cedex 05, France
[3]RESSOURCE PARISIENNE EN BIOINFORMATIQUE STRUCTURALE
15 rue Hélène Brion 75 013 Paris, France

Abstract: *The Structural Prediction for pRotein fOlding UTility (SPROUTS) Workflow Submission Server executes a workflow that correlates static and dynamic properties of protein structure. They include the prediction of special positions (Most Interacting Residues), fragments (Tightened End Fragments), and stability under the effect of a point mutation. The results are stored in a database and can be accessed as text, or visualized with 2D and 3D modes. The Workflow Submission Server extends the SPROUTS database by the ability to submit a protein through its PDB identifier and thus gaining access to a variety of structural analyses at one place and with strong integration. It is a unique resource to visualize characteristics of protein folding and analyze the effect of point mutations on protein structure and to provide experimentalists with the information needed to avoid mutations at positions suspected to be critical for the folding nucleus. SPROUTS workflow is available at http://bioinformatics.engineering.asu.edu/springs2012/Sprouts/query.php.*

Keywords: SPROUTS, workflow, database, server, protein structure, mutation, analysis.

The production of point mutation in a sequence is now routinely performed in molecular biology laboratories thanks to the development of protein engineering techniques. In the field of fundamental research, it is widely used in order to verify whether a given amino acid belongs to the folding nucleus supported by the ΦF value determination initially proposed by Fersht [1,2]. Indeed mutations may have unexpected yet significant impact. For example, over expression of eukaryotic sequences in E. coli may produce inclusion bodies instead of soluble globules. One way to avoid this problem is to create random mutations, hoping that the solubility will be increased. However one has to check whether the proposed mutations have possibly dramatic effects such as greater instability which may lead in some cases to an unfolded protein or to inclusion bodies. The automated prediction of stability may support critical applications of both fundamental and experimental research.

The workflow is an implementation of the complete process that populates the SPROUTS database. The process shown in Figure 1 invokes five tools accessed by the workflow as black boxes: DFIRE version 2.0 [3], I-MUTANT 2.06 [4], I-MUTANT-DSSP 2.06 [4], MUpro version 1.1 [5], and MIR version 1.0 [6]. This website is open to all users and there is no login requirement. The input to the server is a valid PDB ID and an email address for completion messages. The first step consists in checking whether the user input is already in the SPROUTS database. In this case the workflow is not executed and the user receives a message with a link to the results. The workflow automatically calls methods that correlate static and dynamic properties of the protein structure. The basic process begins with a user provided PDB. We then locate the appropriate data for our tools, determine which tools have the necessary input, process data according with respect to the internal limitations of each tool, and then run the tools. After the tools have terminated, we validate their output, upload it to the databases, and send a message to the user. Data analysis with TEF [7] is done when a user queries the database as opposed to the other five tools.

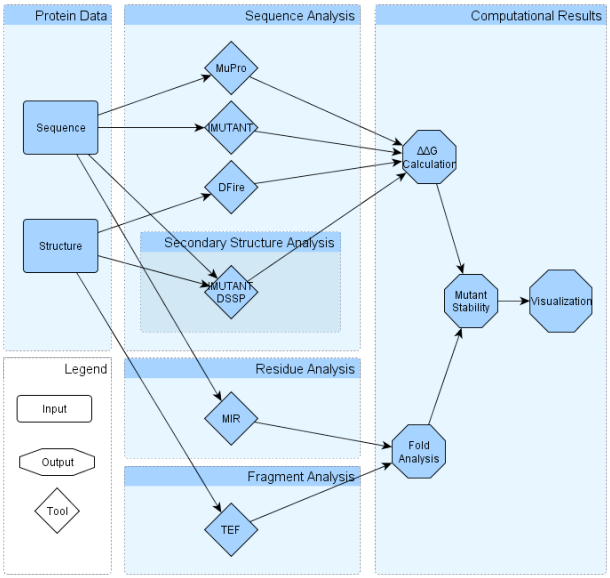


Figure 1. Workflow process overview. Time progression from left to right.

Acknowledgements

We thank our former collaborators Mathieu Lonquety and Christophe Legendre who respectively developed the initial SPROUTS database [8] and worked on the first draft of the workflow. Many thanks for Nikolaos Papandreou for his comments and valuable suggestions and the authors of the different software devoted to stability changes and especially to Jean-Marc Kwasigroch for his help on using the PoPMuSiC software. We also want to acknowledge Pierre Tufféry for his help on using the RPBS resources to compute the MIR calculations. This research was partially supported by the National Science Foundation[1] (grants IIS 0431174, IIS 0551444, IIS 0612273, IIS 0738906, IIS 0832551, and CNS 0849980) and by an invitation of the Université Pierre et Marie Curie in May 2012.

References

[1] A. R. Fersht, Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol.*, 7(1): 3-9, 1997.
[2] A. Fersht, S. Sato, Φ-value analysis and the nature of protein folding transition states. *PNAS*, 101, 7976-7981, 2004.
[3] H. Zhou, Y. Zhou, "Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction," Protein Sci., vol. 11, 2002, pp. 2714–2726, doi: 10.1110/ps.0217002.
[4] E. Capriotti, P. Fariselli, R. Casadio. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.*, 33:W306–W310, 2005.
[5] J. Cheng, A. Randall, P. Baldi. Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins*, 62:1125–1132, 2006.
[6] N. Papandreou, I. N. Berezovsky, A. Lopes, E. Eliopoulos, J. Chomilier, Universal positions in globular proteins: observation to simulation. *Eur. J. Biochem.*, 271:4762–4768, 2004.
[7] M. Lamarine, J.-P. Mornon, N. Berezovsky, J. Chomilier, "Distribution of tightened end fragments of globular proteins statistically match that of topohydrophobic positions: towards an efficient punctuation of protein folding?" Cell. Mol. Life Sci., vol. 58(3), Mar. 2001, pp. 492–498.
[8] M. Lonquety, Z. Lacroix, N. Papandreou, J. Chomilier, SPROUTS: a database for the evaluation of protein stability upon point mutation. *Nucleic Acids Res.*, 37(Database issue):D374-D379, 2009.

1. Any opinion, finding, and conclusion or recommendation expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Sequence polymorphism detection.

Experience sharing, properties and limits of the current methods.

Leila Bastianelli[1], Frédéric Bigey[1], Jean-Luc Legras[1], Virginie Galeote[1], and Sylvie Dequin[1]

[1] UMR1083 Sciences Pour l'Oenologie, INRA, F-34060 Montpellier, France
frederic.bigey@supagro.inra.fr

Keywords *Saccharomyces cerevisiae*, sequence polymorphisms, variant detection

Recherche de polymorphisme de séquence.

Partage d'expérience, propriétés et limites des méthodes actuelles.

Mots-clés *Saccharomyces cerevisiae*, polymorphisme de séquence, détection de variants

1 Contexte et objectifs

La détection de polymorphismes à l'échelle d'un génome entier est un processus pour lequel plusieurs méthodes bioinformatiques ont été développées et publiées, et sont maintenant reconnues. Cependant, quand on s'intéresse à des organismes non modèles ou pour lesquels il n'existe pas de données de référence, il devient difficile d'évaluer la pertinence des données générées, afin par exemple de respecter les seuils souhaités de sensibilité tout en limitant les faux positifs.

Dans le cadre de l'étude de deux populations de *S. cerevisiae* (17 souches au total), on s'intéresse aux polymorphismes nucléotidiques simples (SNP) retrouvés au sein de chacune de ces deux populations, avec pour but d'obtenir des résultats précis et exhaustifs. Deux pipelines de génotypage existants (GATK [1] et SAMtools/BCFtools [2]) ont été appliqués, en suivant précisément la procédure décrite dans la documentation disponible [3][4]. Les génotypages obtenus devraient permettre d'une part d'établir les différences entre les 2 méthodes testées, mais aussi, couplés à des sources extérieures, d'obtenir un génotypage pouvant servir de référence ainsi qu'un protocole répétable.

2 Méthodologie proposée

2.1 Comparaison des deux protocoles testés

On utilise comme départ les données issues du séquençage haut débit de chaque souche (Illumina 1.5), mappées sur le génome de référence de la souche de *S.cerevisiae* S288c. Les deux méthodes de génotypage sont appliquées à chaque population (traitée en « pool »), en suivant point par point les procédures décrites dans la littérature mise à disposition par leurs auteurs. Lorsque des paramètres variables sont à préciser, on a utilisé les valeurs par défaut sauf préconisation contraire dans la documentation. On obtient des résultats très différents puisque le nombre de SNPs détectés varie du simple au double entre SAMtools et GATK : pour une population on passe de 53670 positions variables détectées avec SAMtools (sur un génome d'environ 12 Mbases) à 99472 avec la procédure GATK, pour l'autre de 52638 à 113582. Après comparaison, on constate que les sets sont en fait chevauchants, avec le plus petit presque entièrement inclus dans le plus grand (toutes les positions détectées par SAMtools sont inclues dans celles obtenues par GATK sauf 528 pour la première population, 503 pour la seconde). Il est donc probable que les génotypages obtenus présentent des sensibilités différentes (soit une des deux méthodes est très peu sensible, soit l'autre beaucoup trop), et que l'intersection des deux permettrait déjà de constituer un premier jeu de données assez robuste.

En l'absence de critère objectif d'évaluation de ces données, on choisit de s'intéresser au pourcentage de SNP hétérozygotes présentes dans les deux jeux obtenus. En effet, la majorité des souches étant diploïdes homozygotes, on a pourtant, pour les deux méthodes de génotypage, un pourcentage non négligeable de sites où le génotype détecté est hétérozygote et représente donc des faux-positifs. Il ne s'agit pas d'une contamination des échantillons : ces positions sont détectées pour l'ensemble des souches. De plus, on a également une souche dont on sait qu'elle a été contaminée (détermination par d'autres méthodes) et les positions hétérozygotes sont conséquemment plus nombreuses dans ce cas.

Aucune de ces 2 méthodes, utilisée seule, ne permet donc d'obtenir des résultats satisfaisants. On va donc explorer quelques pistes qui permettront de filtrer ou complémenter ces résultats.

2.2 Filtres sur les LOD-scores et qualité des génotypages

Les deux génotypeurs ont des méthodes d'attribution de score aux variants détectés (selon différents paramètres dont la couverture, la qualité des reads à cette position...). Une première approche peut consister à filtrer les sets selon des seuils (empiriques) sur ces scores. Dans les 2 cas, un score de qualité par génotype est disponible, et filtrer les données selon ce score nous a permis d'éliminer certains des faux-positifs… mais pas tous. Le framework GATK fournit en plus un LOD-score sur chacune des positions. De la même façon, il permet d'améliorer un peu les données, sans que ce soit satisfaisant.

2.3 Données extérieures

On peut également obtenir d'autres jeux de données par assemblage *de novo* des séquences puis alignement face à la séquence de référence. Cette méthode donnera un jeu de SNPs complémentaire dont la comparaison (intersection, union, …) avec les autres pourra éventuellement permettre d'en confirmer une partie voire de les compléter.

2.4 Exclusion de certaines régions

Certaines zones du génomes sont plus variables (transposons, séquences répétées, télomères...) et peuvent poser un problème lors du mapping, et donc générer des erreurs de génotypage. Exclure les positions polymorphes détectées dans ces zones (lorsqu'elles sont connues) pourrait également permettre d'augmenter la robustesse des résultats.

Acknowledgements

Nous remercions l'INRA (UMR1083 Sciences pour l'Oenologie) pour le financement du stage de Leïla Bastianelli ainsi que le labex NUMEV (http://www.lirmm.fr/numev/) pour le financement de la participation à JOBIM.

References

[1] DePristo, M., Banks, E., Poplin, R., Garimella, K., Maguire, J., Hartl., C., Philippakis, A., del Angel, G., Rivas, M.A, Hanna, M., McKenna, A., Fennell, T. Kernytsky, A., Sivachenko, A, Cibulskis, K., Gabriel, S., Altshuler, D. and Daly, M. *A framework for variation discovery and genotyping using next-generation DNA sequencing data.* Nature Genetics. 2011 Apr; 43(5):491-498.

[2] Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) *The Sequence alignment/map (SAM) format and SAMtools.* Bioinformatics, 25, 2078-9.

[3] *Best Practice Variant Detection with the GATK.* http://www.broadinstitute.org/gsa/wiki/index.php/Best_Practice_Variant_Detection_with_the_GATK_v3

[4] *Calling SNPs/INDELs with SAMtools/BCFtools* http://samtools.sourceforge.net/mpileup.shtml

Identification de voies biologiques dérégulées dans les sarcomes à génétique complexe.

Céline BRULARD[1] and Frédéric CHIBON[1]

Institut Bergonié, INSERM U916, génétique et biologique des sarcomes, France
{c.brulard, f.chibon}@bordeaux.unicancer.fr

Abstract Genetically complex sarcomas are rare and heterogenous tumors which do not display recurring chromosomic or genetic alterations. We investigated the hypothesis that there are common biological pathways implicated in sarcomas oncogenesis but that different genes are targeted. Bioinformatic methods were established to select genes with genomic alterations which have an effect on gene expression, then to identify biological pathways and associated genes. We present approximately sixty pathways common to all tumors and identify genes depending on histotype, localization and prognosis. The importance of the chromosomic localization of these genes was [illegible] to the group.

Keywords Sarcomas, genomic, transcriptomic, pathways

1 Introduction

Les sarcomes des tissus mous de l'adulte (STMA) sont des tumeurs rares, 1 à 2% des cancers, qui sont très hétérogènes aussi bien en terme de localisation, d'histologie, de profil moléculaire et de pronostic. Les sarcomes dits à génétique complexe représentent 50% de ces tumeurs et sont caractérisés par un caryotype très complexe, mais sans altération chromosomique/génétique récurrente ou spécifique identifiée. A partir des données cliniques, génomiques et transcriptomiques de 108 sarcomes à génétique complexe, nous avons étudié l'hypothèse selon laquelle ces tumeurs ne partagent pas de gènes altérés communs mais altèrent des voies biologiques communes soit à toute les sarcomes, soit à des sous-groupes de sarcomes dépendant de l'histotype, de la localisation ou du pronostic.

2 Méthodes

Les données génomiques proviennent de la puce Genome-Wide Human SNP Array 6.0 d'Affymetrix normalisées et contrôlées avec la Genotyping Console 2.0, et les données transcriptomiques de la puce GeneChip Human Genome U133 Plus 2.0 Array d'Affymetrix normalisées avec la méthode GCRMA.

Figure 1. Schéma du projet.

On considère un gène altéré comme étant un gène qui a un nombre de copies anormal (perte ou gain) et qui a également une expression dérégulée (sous-expression ou sur-expression). Le point de départ étant l'altération génomique, l'expression est contrôlée en considérant qu'un gène pour lequel il y a deux copies a

Setting Galaxy on the ATGC ReNaBi platform at Montpellier

Integration of the crac mapping and annotation tool

Marine ROHMER[1], Thérèse COMMES[2], Vincent LEFORT[3] and Alban MANCHERON[3]

[1] Master STIC pour la Santé, spécialité « Bioinformatique, Connaissance, Données »
Universités de Montpellier 1 & 2, Institut Telecom, TIC et Santé
CC 92000 Place Eugène BATAILLON 34095 Montpellier CEDEX 05
marine.rohmer@etud.univ-montp2.fr

[2] Centre de Recherche de Biochimie Macromoléculaire, UMR5237
CC091, Université Montpellier 2, Place Eugéne Bataillon, 34095 Montpellier CEDEX 5
therese.commes@crbm.cnrs.fr

[3] Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, UMR5506
CC06001, 95, rue de la Galéra, 34095 Montpellier CEDEX 5
{vincent.lefort, alban.mancheron}@lirmm.fr

Abstract NGS *(Next Generation Sequencing) technologies produce increasingly widespread data, so that scientific community needs the development and the share of bioinformatics analysis tools.* Galaxy *is a worldwide spread workflow interface, which allows making bioinformatics pipeline analysis. The* ATGC *bioinformatics platform, which belongs to the french bioinformatics platforms network ReNaBi (national network of bioinformatics platforms), provides its own new* Galaxy *service to access a* NGS *tools range that includes* crac *(a new mapping/annotation tool).*

Keywords Galaxy, crac, mapping, NGS, ReNaBi.

Mise en place de Galaxy sur la plateforme ReNaBi ATGC de Montpellier

Intégration de l'outil de mapping et d'annotation crac

Résumé *Les techniques* NGS *(Next Generation Sequencing) génèrent un volume de données tel qu'il nécessite le développement et la mise à disposition d'outils de bioinformatique pour la communauté scientifique.* Galaxy *est une interface de workflow à visée internationale, qui permet de réaliser des pipeline d'analyse bioinformatique. La plateforme de services bioinformatiques* ATGC*, qui est fédérée par le reseau francais ReNaBi (Réseau National des plates-formes Bioinformatiques), offre son nouveau service* Galaxy *pour accéder à une panoplie d'outils* NGS *qui inclue* crac *(un nouvel outil de mapping/annotation).*

Mots-clés Galaxy, crac, mapping, NGS, ReNaBi.

1 Contexte

À l'intérieur de la communauté bioinformatique, deux systèmes de *workflow* s'imposent en standard, à en juger par leurs larges et grandissantes communautés : Taverna [1] et Galaxy [2] [1–3]. Ces deux systèmes sont efficaces, *open source*, et offrent la possibilité d'être installés localement ou sur un réseau. Galaxy est utilisé par les bioinformaticiens et les biologistes pour son grand panel d'outils proposés et sa simplicité d'utilisation. De fait, l'utilisateur peut enchaîner à sa guise plusieurs de ces outils, disposer de l'historique des actions effectuées, revenir à une étape en particulier ou bien encore enregistrer un *pipeline* particulier (*workflow*). Parmi les outils proposés se retrouvent principalement des outils de manipulation de données NGS. Galaxy permet aussi une utilisation du *cloud computing* ou d'un *cluster* de calcul.

Crac est un outil de manipulation de données NGS principalement dédié à l'analyse du transcriptome [4]. Cet outil [3] est développé conjointement par l'équipe MAB du LIRMM et par l'équipe BONSAI du LIFL. Il

1. http://www.taverna.org.uk/
2. http://galaxyproject.org/, http://getgalaxy.org/
3. Article en cours de soumission

permet non seulement de réaliser une étape de *mapping* sur un génome de référence, mais aussi de détecter les différentes sources d'erreurs/variations (*e.g.*, erreur de séquençage, jonction d'épissage, *Single Nucleotide Polymorphism*, *Single Nucleotide Variant*, chimère, ...). À ce jour, `crac` n'est utilisable que localement par ses développeurs, mais sera disponible après publication au téléchargement. À l'instar de ses concurrents, cet outil est dépourvu d'interface graphique, et requiert d'importantes ressources de calcul et de mémoire .

La **plateforme de services bioinformatiques ATGC** (http://www.atgc-montpellier.fr/) est membre du réseau français ReNaBi et fédère les équipes régionales de recherche en bioinformatique et coordonne les plateformes du sud de la France. ATGC valorise et distribue des outils pour la biologie moléculaire à grande échelle et propose déjà plusieurs dizaines d'outils, que l'utilisateur peut utiliser à distance ou télécharger.

2 Déploiement de `Galaxy` sur la plateforme ATGC

Notre objectif est double. D'une part, il s'agit de proposer un service supplémentaire sur la plateforme ATGC en offrant la possibilité à la communauté d'utiliser `Galaxy` et d'autre part de rendre `crac` accessible à la communauté, et notamment auprès des utilisateurs non adeptes de la ligne de commande.

Préalablement à l'installation de `Galaxy` sur ATGC, plusieurs étapes sont nécessaires. Une première installation de ce service a été installée en local afin d'établir son paramétrage. Ensuite, la même configuration est appliquée sur une machine de développement qui est un clone de la machine de production, ceci afin d'assurer la compatibilité des configurations. C'est seulement à l'issue de cette étape que `Galaxy` est disponible sur le serveur de production, à savoir ATGC.

L'intégration de `crac` dans le panel d'outils de `Galaxy` requiert une configuration spécifique. En effet, l'objectif *in fine* est de diffuser `crac` par le *Tool Shed* (sorte de dépôt similaire à l'Apple Store ou à l'Android Market). Celui-ci permet de déployer des outils ou des *workflows* sur n'importe quelle plateforme `Galaxy`. Pour diffuser `crac` ainsi, il est nécessaire de fournir les fichiers de configuration, *wrappers*, dépendances,..., nécessaires au fonctionnement de `crac`. Il faut également réaliser des tests de robustesse comprenant notamment la vérification des données d'entrées/sortie qui se doivent d'être compatibles avec les autres outils présents sur l'interface. Cela passe par l'utilisation de jeux de données réels (privés ou publiques), mais intègre également l'utilisation de fichiers aberrants.

Enfin, pour rendre `Galaxy` et `crac` accessibles sur ATGC, il convient de gérer différents droits d'accès (anonyme ou authentifié). De surcroît, l'exécution des tâches par `Galaxy` (effectuée localement par défaut) est configurée pour utiliser les ressources d'un *cluster* de calcul pour les membres authentifiés ou d'un *cloud* pour l'ensemble des utilisateurs. Ce choix permet de délocaliser l'exécution des tâches nécessitant d'importantes ressources de calcul, permettant ainsi de ne pas saturer l'accès à la plateforme.

Remerciements

Nous remercions le Labex NUMEV (http://www2.lirmm.fr/numev/) pour le financement du stage de Marine ROHMER, ainsi que de sa participation à JOBIM 2012.

Références

[1] Goecks, J, Nekrutenko, A, Taylor, J and The Galaxy Team. : *"Galaxy : a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences."* Genome Biol. 2010 Aug 25 ;11(8) :R86.

[2] Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J. : *"Galaxy : a web-based genome analysis tool for experimentalists".* Current Protocols in Molecular Biology. 2010 Jan ; Chapter 19 :Unit 19.10.1-21.

[3] Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A. : *"Galaxy : a platform for interactive large-scale genome analysis."* Genome Research. 2005 Oct ; 15(10) :1451-5.

[4] Nicolas Philippe : *Développement de méthodes et d'algorithmes pour la caractérisation et l'annotation des transcriptomes avec les séquenceurs haut débit.* Thèse de doctorat, Université de Montpellier 2, France, 2011.

Functional description and inference of carbohydrate-active enzymes

Elodie DRULA[1], Vincent LOMBARD[1], Corinne RANCUREL[1], Marie-Line GARRON[1], Pedro M COUTINHO[1] and Bernard HENRISSAT[1]

[1] Architecture et Fonction des Macromolécules Biologiques, UMR7257 CNRS, Aix Marseille Université, 163 Avenue de Luminy, 13288, Marseille, Cedex 9, France

{elodie.drula, vincent.lombard, corinne.rancurel, marie-line.garron, Pedro.coutinho, bernard.henrissat}@afmb.univ-mrs.fr

Abstract *Carbohydrates play a central role in life. The extraordinary diversity of glycans is mastered by a whole array of enzymes. These typically correspond to 1-5% of the genes of a living organism. The CAZy database describes the families of structurally-related catalytic of enzymes that cleave and build glycosidic bonds. Within CAZy, enzyme specificity is attributed to proteins based on experimental evidence. Our expert biocuration of sequence and particularly functional information lead to the creation of a hierarchy of functional descriptors. An automated functional annotation pipeline uses these descriptors and manages different degrees of ambiguity in order to mimic manual approaches.*

Keywords Glycobiology; Enzyme Specificity; Enzyme Classification; Functional Prediction.

Description et inférence fonctionnelle des enzymes actifs sur les glucides

Résumé *Les glucides jouent un rôle central dans la vie sur Terre. L'extraordinaire diversité des glycanes est maîtrisée par une panoplie d'activités enzymatiques dédiées. Ces enzymes représentent 1 à 5% des gènes d'un organisme. La base de données CAZy classifie en familles les enzymes catalysant la synthèse et la dégradation des liaisons glycosidiques. La spécificité de substrat des enzymes est assignée à une protéine seulement en présence d'évidence expérimentale. Notre biocuration experte de données de séquences et en particulier fonctionnelles a abouti à la création d'une hiérarchie des descripteurs fonctionnels. Ces descripteurs sont utilisés par un pipeline d'annotation fonctionnelle automatique qui mime l'approche utilisée manuellement et décrit différents degrés d'ambiguïté.*

Mots-clés *Glycobiologie, Spécificité enzymatique, Classification des enzymes, Prédiction fonctionnelle.*

Introduction

Les glucides sont omniprésents comme composants structuraux, de réserve et d'échange d'énergie. L'extraordinaire diversité des glycanes est maîtrisée par une panoplie d'activités enzymatiques pour leur biosynthèse, dégradation et modification. La classification des CAZymes (pour *Carbohydrate Active enZymes*) est essentielle pour rationaliser leurs caractéristiques biochimiques et leurs applications. Avant 1980, la classification des enzymes (EC pour *Enzyme Classification*) qui prévalait pour ces enzymes était celle du Comité de Nomenclature de l'IUBMB, reposant sur leur spécificité de substrat. Ces variations s'appuyaient ainsi sur des données empiriques à défaut d'aspects moléculaires. En 1990, une classification des Glycoside Hydrolases (GHs) fondée sur leurs séquences a été proposée **[1,2]**. Cette classification avait l'avantage de ne pas se restreindre aux activités présentes au sein d'une famille et a ainsi rapidement été adoptée car elle permettait d'associer des observations expérimentales avec des aspects moléculaires particuliers. Les principes de la classification des GHs ont été étendus aux glycosyltransférases (GTs), aux polysaccharide lyases (PLs) **[3,4]**, aux estérases de glucides (CEs) et aux modules auxiliaires d'adhésion aux glucides (CBMs) **[3,5]**. Depuis 1998, toutes ces classifications basées sur la séquence sont hébergées dans la base de données CAZy (http://www.cazy.org) **[3,5]**.

Pour les CAZymes, déterminer la fonction revient à déterminer la spécificité de substrat (et parfois de produit). Chaque famille de CAZyme est créée à partir de protéines caractérisées expérimentalement et est peuplée par des séquences significativement similaires provenant de base de données publiques. La présence simultanée d'enzymes de spécificités différentes dans une même famille a induit des efforts d'identification de sous-groupes structuralement et fonctionnellement homogènes à l'intérieur de la famille. Les informations de structure et de spécificité sont régulièrement extraites de la littérature et de la PDB (*Protein Data Bank* http://www.rcsb.org/pdb/) disponibles. La classification (i) met l'accent sur les particularités structurales des enzymes, (ii) aide à mettre en lumière les relations évolutives et (iii) fournit un cadre pour décrypter les propriétés mécaniques (méthode de repliement, etc.). Le but est de capturer l'ensemble des connaissances dans le domaine des enzymes actives sur les glucides. CAZy est disponible depuis plus de 13 ans et a été

utilisé pour augmenter la qualité de la prédiction de fonctions de nombreux projets.

Hiérarchisation des Connaissances Fonctionnelles

L'identification de tous les niveaux d'ambiguïté est déterminante pour organiser l'information fonctionnelle des CAZymes. Dans CAZy, les activités (spécificités de substrat) sont reliées aux protéines sous la forme de numéros EC. Les numéros EC présentant des limites **[6]**, une hiérarchie alternative des descripteurs fonctionnels des CAZymes a été mise en place. Cette structure permet la résolution des ambiguïtés mais offre aussi un cadre pour la prédiction d'activités enzymatiques. Etant donné que de nombreuses CAZymes portent des CBMs, notre système constitue un support idéal pour afficher la spécificité d'adhésion découlant de la proximité d'activités enzymatiques. Une nomenclature appropriée aux modules CBMs a été établie à partir des descriptifs conçus pour l'adhésion des lectines **[7]**. Ainsi, plus de 600 évidences d'adhésion ont été collectées et suivent cette classification (cf Table I). Au final, l'ontologie que nous avons ainsi instaurée nous permet de décrire des activités enzymatiques et/ou d'adhésion même pour les cas où la classification EC officielle est inadaptée.

Table I – État des différentes classes de CAZy (en date de Mai 2012). Le nombre de caractérisations expérimentales d'enzymes (EC), et de CBMs (BC pour *binding classification*) est donné.

Classe	GH	*GT*	*PL*	*CE*	*CBM*
Modules	*184937*	*125530*	*6001*	*21903*	*42032*
EC	*6190*	*1820*	*262*	*288*	*24*
BC	*26*			*6*	*573*

Prédiction de Fonction

La similitude de séquence d'une protéine avec celles d'enzymes caractérisés constitue le mode traditionnel d'inférence de fonction. Ainsi la prédiction fonctionnelle que nous réalisons est basée sur un jeu de données de référence constituée exclusivement de séquences de modules fonctionnels dont l'activité a été établie expérimentalement. Notre pipeline d'annotation fonctionnelle automatique est développé à partir d'un protocole mimant l'approche utilisée manuellement. Chaque module de la protéine étudiée est comparé séparément contre nos données de référence. Un descripteur est attribué en fonction de la proximité et de l'abondance des séquences caractérisées similaires. Ces descripteurs sont combinés de façon à proposer un descripteur unique. Dans ce schéma d'annotation, la nomenclature utilisée pour décrire les différents degrés d'ambiguïté comprend (par similarité décroissante) des activités "candidates", des activités "apparentées" ou des activités "lointainement apparentées". La prédiction peut être limitée à un descripteur plus général lorsque la distance avec un membre caractérisé expérimentalement est trop grande ou ambigüe.

References

[1] Henrissat B. (1991) A classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem J* 280:309-316

[2] Henrissat B, Bairoch A (1993) New families in the classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem J* 293:781-788

[3] Coutinho PM, Henrissat B (1999) Carbohydrate-active enzymes: an integrated database approach (1999) In *Recent Advances in Carbohydrate Bioengineering*. H.J. Gilbert, G. Davies, B. Henrissat and B. Svensson eds., The Royal Society of Chemistry, Cambridge, pp. 3-12.

[4] Lombard V, Bernard T, Rancurel C, Brumer H, Coutinho PM, Henrissat B (2010) A hierarchical classification of polysaccharide lyases for glycogenomics. *Biochem. J.* 432:437-444

[5] Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res* 37:D233-238

[6] Degtyarenko K (2004) Controlled vocabularies and ontologies in enzymology, *in Experimental Standard Conditions of Enzyme Characterizations.* Hicks MG. and Kettner C, eds., Beilstein Institute, Frankfurt, pp. 143-173

[7] Gallagher JT (1984) Carbohydrate-binding properties of lectins: A possible approach to lectin nomenclature and classification. *Bioscience Rep*. 4:621-632

BioSpring: an interactive and multi-resolution software for flexible docking and for mechanical exploration of large biomolecular assemblies

Nicolas FEREY[1], Olivier DELALANDE[2] and Marc BAADEN[3]

[1]LIMSI, UPR3251 CNRS, Bât. 502 bis, 508 et 512, Université Paris XI, 91403, Orsay, Cedex 91403, France
nicolas.ferey@limsi.fr

[2] Equipe RMN-ILP, Université de Rennes 1, UMR6026 CNRS , CS 34317, Rennes, Cedex 35043, France
olivier.delalande@univ-rennes1.fr

[3] IBPC LBT, Université-Paris 7, UPR9080 CNRS, 13, rue Pierre et Marie Curie, 75005, Paris, France
marc.baaden@ibpc.fr

Keywords: Interactive Molecular Simulation, Spring Network Model, Multi-Resolution Model

1 Introduction

Recent advances in experimental techniques allow us to solve larger and larger biomolecular 3D structures. However, even if structure is known to be strongly linked to biological function, static states often lack in providing dynamical information that are crucial for the understanding of the subtle mechanisms occurring at the molecular level. Thus, molecular simulations are nowadays used to complete experimental biostructural studies, especially to better understand the dynamic behaviour and the fundamental mechanisms involved in a biomolecular complex. In spite of the increasing computational resources, classical modeling methods are still not well adapted to quickly characterize biomechanics of biomolecular assemblies, mainly because of the limited timescale accessible to all-atom simulations. For these reasons, it's necessary to develop new simulation approaches especially designed to study very large biological structures formation and stability. We present ***BioSpring***, an unconventional and innovative software, implementing a method based on a **spring network model** adding **non-bonded interactions**, guided by an **electrostatic map**, and using **advanced human interaction techniques** dedicated to **interactive molecular modeling**. The interactive simulation tool BioSpring allows a user to quickly study the biomechanical properties, by interactively highlighting rigidity, flexibility, and allosteric effects, in order to provide new hypotheses about a biomolecular system. Finally, our approach is also designed to help a user in the complex task of modeling large biomolecular complexes before using more time-consuming and costly classical simulation tools.

2 Method

2.1 Spring Network Simulation Model

$$\vec{F}_{spring}(p) = \sum_{p' \in Springs(p)} k_{stiffness} \vec{u}_{pp'} (d_{pp'} - e_{pp'})$$

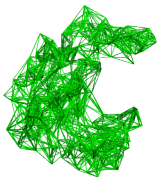
2.2 User-defined non bonded interaction

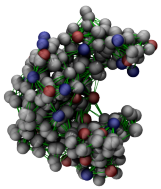
$$s_{pp'} = (r_p + r_{p'}) - d_{pp'}$$

$$\vec{F}_{linearsteric}(p) = \begin{cases} \vec{0} & \text{if} \quad s_{pp'} \leq 0 \\ \sum_{p' \neq p} -k_{steric} \vec{u}_{pp'} s_{pp'} & \text{else} \end{cases}$$

$$\vec{F}_{coulomb}(p) = \sum_{p' \neq p} -\vec{u}_{pp'} \frac{q_p q_{p'}}{4\pi\epsilon_0 d_{pp'}^2}$$

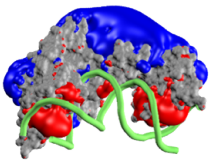
$$\vec{F}_{lennardjonessteric}(p) = \sum_{p' \neq p} \vec{u}_{pp'} 4\epsilon_{pp'} \left[\left(\frac{\sigma_{pp'}}{9 d_{pp'}} \right)^9 + \left(\frac{\sigma_{pp'}}{7 d_{pp'}} \right)^7 \right]$$





2.3 Interaction of the flexible structure and a potential map

$$\vec{F}_{map}(p \in V_{i,j,k}) \simeq \begin{bmatrix} \frac{E_{i+1,j,k} - E_{i-1,j,k}}{2\Delta_x} \\ \frac{E_{i,j+1,k} - E_{i,j-1,k}}{2\Delta_y} \\ \frac{E_{i,j,k+1} - E_{i,j,k-1}}{2\Delta_z} \end{bmatrix}$$



2.4 human/molecule interaction model

$$\vec{F}_{control}(p \in Selection) = -k_{control}[\vec{P}(tool) - \frac{1}{|Selection|} \sum_{p' \in Selection} \vec{P}(p')]$$

$$\vec{F}_{feedback}(tool) = -k_{feedback}[\vec{P}(tool) - \frac{1}{|Selection|} \sum_{p' \in Selection} \vec{P}(p')]$$

3 Applications

3.1 Modeling DNA sequence recognition within RecA nucleoprotein filaments [2]

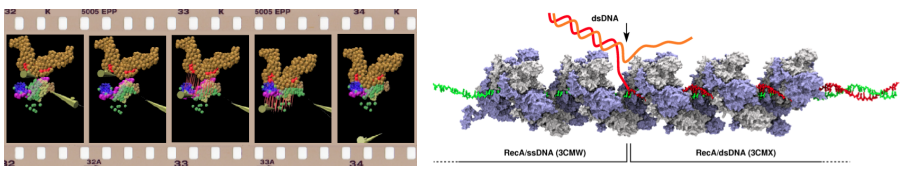


Figure 1. Bimanual haptic interactive modelling (left), and description of RecA-induced DNA strand exchange (right)

3.2 Adenylate Kinase (AK) interactive closure

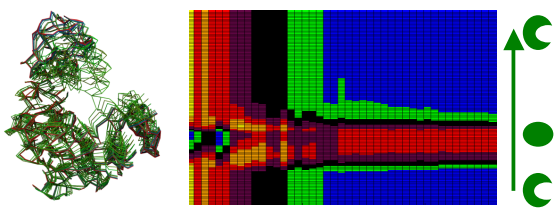


Figure 2. Bimanual haptic interactive study of AK closure (left) and RMSD results (right)

4 Conclusion

BioSpring was especially designed for molecular sketching and modeling. The main goals of ***BioSpring*** is to provide a tool to get interactively a quick overview of biomechanical properties, to help user in the complex task of modeling large biomolecular complexes, in order to perform an interactive flexible docking, and also to Providing/Exploring/Testing new hypothesis prior to more complex investigations. This software can be plugged with ***VMD*** using the ***MDDriver*** [1] library, with or without force feedback using a haptic device. BioSpring should be used to quickly explore biomechanical properties of biomolecules, for interactive molecular modeling and docking, prior to further investigation using classical simulation tools.

References

[1] Delalande, Férey, Grasseau, Baaden. Complex Molecular Assemblies at Hand via Interactive Molecular Simulation. *J. Comput. Chem., 30(15), 2375-87,* 2009

[2] Saladin, Amourda, Poulain, Férey, Baaden, Zacharias, Delalande, & Prévost, Modeling the early stage of DNA sequence recognition within RecA nucleoprotein filaments*, Nucleic Acids Res.*, 38, 6313-23, 2010

Representation of cis-regulatory region in a model organism database

Florent DUMOND[1], Patrick LEMAIRE[2] and Cyril MARTIN[2]

[1] Université Montpellier 2, Place Eugène Bataillon, 34095 Montpellier, Cedex 5, France
flodumond@gmail.com

[2] Équipe Lemaire, CNRS-UMR 5237, 1919 Route de Mende, 34293 Montpellier, Cedex 5, France
{patrick.lemaire, cyril.martin}@crbm.cnrs.fr

Keywords Cis-regulatory Region, ascidian, online model organism database.

1 Introduction

ANISEED [1] is a database designed to offer a representation of ascidian embryonic development at the level of the genome (cis-regulatory sequences, spatial gene expression, protein annotation), of the cell (cell shapes, fate, lineage) or of the whole embryo (anatomy, morphogenesis).

Our objective is to understand how the linear sequence of the genome directs the formation of a complex multicellular animal, using embryos as a model system of a close relative of vertebrates, the ascidian *Ciona intestinalis*.

Helped from ongoing technological revolutions, we decided to implement this database with data from experiments of ChIP-on-chip and ChIP-seq, nucleosomes, and conservation data. By the comparison of all these data, we hope to identify cis-regulatory regions.

2 Data analysis

Data analysis is necessary to determine where a cis-regulatory can be found, and must be done before saving anything on database. Additionally, we can not save all the data, we have to make a peak extraction and compare these peaks between them.

The ChIP-on-chip is a microarray-based method to measure genome wide binding sites for proteins of interest. With Kubo A. and Satou Y. data from the *ghost database* [2], we first made a data normalization using *CoCAS* [3] peaks caller software, specialized to Agilent microarray analysis. The method used to normalize is the *Linear Correction and Weighted Loess method* [4], with a substract background correction method.

ChIP-sequencing is also used to analyze protein interactions with DNA, but combines chromatin immunoprecipitation with massively parallel DNA sequencing. These data were obtained by our team, and analysed with MACS [5], which uses a dynamic Poisson distribution to effectively capture local biases in the genome sequence.

Ciona intestinalis was compared to a closer species, *Ciona savignyi*, to determine highly conserved regions. By comparing theses regions with untranslated regions, we can assume the potential existence of a cis-regulatory region. The analysis was done fixing two thresholds: the first determines where the peak begins, the second fixes the minimal value for a peak.

Nucleosomes data was used to determine where a transcription factor can not bind on the genome.

3 Data Integration

The current developed version of ANISEED is using *Jelix*[1] framework with *Chado*[2] as database system. A data integration pipeline has been made within the team to incorporate data into ANISEED. Because of storage constraints, we choose to save only the peak's information : the contig, the start and end position, and a scalar value.

Chado database system allows to organise data and link them together : for example, ChIP-on-Chip data were obtained from an experience, on an Agilent array, and then analyzed with a software. We can also link this type of data to a transcription factor, like Brachyury.

In the end, we hope to compare all of these data to make relations between cis-regulatory regions and target genes.

Acknowledgements

The authors thank Lemaire's team, and NUMEV[3] for financial support.

References

[1] Tassy, O., Dauga, D., Daian, F., Sobral, D., Robin, F., Khoueiry, P., Salgado, D., Fox, V., Caillol, D., Schiappa, R., Laporte, B., Rios, A., Luxardi, G., Kusakabe, T., Joly, J. S., Darras, S., Christ, *The ANISEED database: digital representation, formalization, and elucidation of a chordate developmental program*. Genome Res, 20:1459-68, 2010.

[2] Atsushi Kubo, Nobuhiro Suzuki, Xuyang Yuan, Kenta Nakai, Nori Satoh, Kaoru S. Imai, Yutaka Satou , *Genomic cis-regulatory networks in the early Ciona intestinalis embryo* , Development , 137:1613-1623 , 2010.

[3] Benoukraf T, Cauchy P, Fenouil R, Jeanniard A, Koch F, Jaeger S, Thieffry D, Imbert J, Andrau JC, Spicuglia S, Ferrier P., *CoCAS: a ChIP-on-chip analysis suite*, Bioinformatics, 25(7):954-5, 2009.

[4] Peng S, Alekseyenko AA, Larschan E, Kuroda MI, Park PJ., *Normalization and experimental design for ChIP-chip data*, BMC Bioinformatics, 8:219, 2007

[5] Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS., *Model-based analysis of ChIP-Seq (MACS)*, Genome Biol., 9(9):R137, 2008

1. *Jelix,* open-source PHP framework*:* http://jelix.org/
2. *Chado,* a relational database schema*:* http://gmod.org/wiki/Chado_-_Getting_Started
3. *NUMEV:* http://www.lirmm.fr/numev/

Graph Algorithms and Software Framework for Interactive RNA Structure Modelling

ANR project "AMIS ARN"

Fabrice Jossinet [⋆1], Alexis Lamiable[⋆2,3], Philippe Rinaudo[⋆3,2,5], Liza Al-Shikhley[1], Franck Quessette[2], Sandrine Vial[2], Dominique Barth[2], Eric Westhof[1] and Alain Denise[3,4,5]

[1] Architecture et Réactivité de l'ARN, Université de Strasbourg, IBMC du CNRS
[2] PRISM, Univ Versailles Saint-Quentin, CNRS, Versailles, F78000
[3] LRI, Univ Paris-Sud, CNRS, Orsay, F91405
[4] IGM, Univ Paris-Sud, CNRS, Orsay, F91405
[5] AMIB group, INRIA, Saclay, F91405

Abstract *We outline the AMIS-ARN project and present its current results. The project aims to make significant progress in automatic or semi-automatic RNA three-dimensional structure modelling. It gathers three groups of bioinformaticians and computer scientists.*

Keywords RNA, 3D structure, modelling, prediction, graph algorithms

Algorithmes de Graphes et Plateforme logicielle pour la Modélisation Interactive des Structures d'ARN

Projet ANR "AMIS ARN"

Résumé *Nos présentons les grandes lignes et les résultats actuels du projet AMIS-ARN, dont le but est de faire des avancées substantielles dans la modélisation automatique ou semi-automatique des structures d'ARN. Ce projet rassemble trois équipes de bioinformaticiens et informaticiens.*

Mots-clés ARN, structure 3D, modélisation, prédiction, algorithmique des graphes

Le problème de la prédiction de la structure tridimensionnelle d'une molécule d'ARN est appelé la *modélisation* de structure d'ARN. Les rares programmes de modélisation automatique ne sont efficaces que pour de petites molécules, de l'ordre de quelques dizaines de nucléotides. Au-delà, la prédiction de la structure tridimensionnelle requiert les compétences d'un expert du domaine. C'est notamment la très grande plasticité des molécules d'ARN et l'accumulation de domaines additionnels au cours de l'évolution qui rend cette tâche particulièrement difficile. Dans ce contexte, notre but est de faire des avancées substantielles dans la modélisation automatique ou semi-automatique des structures d'ARN. Le projet se découpe en trois volets principaux : (1) Développement d'approches informatiques pour résoudre des étapes clé pour la modélisation ; (2) Intégration de ces méthodes dans la plateforme logicielle développée à l'IBMC ; (3) Application de ces méthodes sur des molécules d'intérêt.

Schématiquement, la modélisation de grandes structures d'ARN s'appuie sur deux approches complémentair
La donnée est une (ou plusieurs) séquence(s) dont on veut prédire la structure. Les deux approches sont :

- L'approche par homologie, utilisée lorsque la structure d'une séquence homologue à la séquence d'intérêt a déjà été résolue expérimentalement. Le principal problème est alors de savoir calculer un alignement de la séquence sur la structure, qui permettra d'en déduire aussi bien que possible la structure asssociée à la séquence. La difficulté réside bien sûr dans la nécessité d'une grande précision de l'alignement.
- L'approche *ab initio*, utilisée lorsqu'aucune structure homologue n'est connue pour la séquence d'intérêt. Elle est utilisée aussi, complémentairement à l'approche par homologie, afin de prédire la structure des parties de la séquence qui ne s'alignent pas sur la structure connue. En effet, les deux séquences, tout en étant homologues, peuvent avoir subi des mutations de très grande ampleur au cours de l'évolution.

⋆. The first three authors should be considered co-first-authors

La Fig. 1 présente le *workflow* de modélisation mis en place. Le logiciel S2S est spécialisé dans la pro-

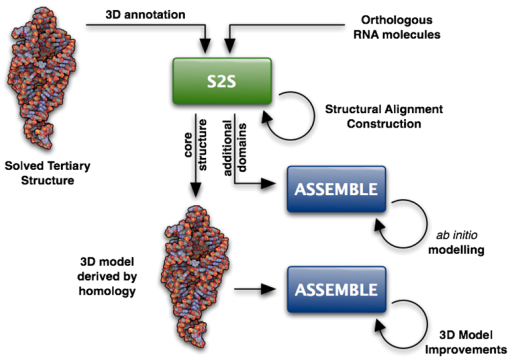


Figure 1. Schéma général du *workflow* de modélisation.

duction d'alignements structuraux. Le logiciel Assemble effectue la part de modélisation et visualisation tridimensionnelle des molécules. Ces deux logiciels, interactifs, automatisent un certain nombre d'étapes clé, mais font encore un appel intensif à l'expertise de leur utilisateur en modélisation. Notre but est d'aller plus loin dans l'automatisation, en développant des algorithmes efficaces et précis pour certaines tâches difficiles qui demandent actuellement une intervention de l'utilisateur expert.

Voici un bref résumé des avancées effectuées jusqu'ici durant le projet :

- **Approche par homologie.** Pour résoudre le problème d'alignement séquence-structure en considérant les interactions non canoniques et des classes générales de pseudo-noeuds, nous utilisons la *décomposition arborescente* de graphes. Cette technique aboutit à un schéma de programmation dynamique intuitif, découpant le problème en deux parties successives : la décomposition arborescente du modèle puis l'alignement générique entre une décomposition arborescente modèle et la séquence cible [3].
- **Approche ab initio.** Une approche de construction de structure 3D consiste à prendre la structure secondaire d'une molécule, à construire des hélices "type" et à les placer relativement les unes par rapport aux autres. La forme des hélices est bien connue, mais la forme des jonctions entre hélices est plus variable. Nous avons développé des méthodes automatiques de classification des jonctions en familles topologiques à partir de leur structure secondaire [2].
- **Plateforme logicielle.** Elle est composée des logiciels Assemble et S2S. Ces deux logiciels s'appuient sur une couche logicielle qui gère les communications avec les services web (bases de données et logiciels externes) et la modélisation des concepts biologiques. Depuis le début du projet, la plateforme a été améliorée dans plusieurs directions. Notamment, en lien avec les deux points développés plus haut, une attention particulière a été portée à la manipulation des alignements structuraux et l'identification des jonctions et des motifs dans les structures [1].

Remerciements. Ces travaux ont été financés par le projet AMIS ARN ANR-09-BLAN-0160 ainsi que par le projet Digiteo PASAPAS et le projet PASAPRES du PRES UniverSud Paris.

Références

[1] F. Jossinet, T. E. Ludwig, and E. Westhof. Assemble : an Interactive Graphical Tool to Analyze and Build RNA Architectures at the 2D and 3D Levels. *Bioinformatics*, 2010.

[2] A. Lamiable, D. Barth, A. Denise, F. Quessette, S. Vial, and E. Westhof. Automated Prediction of Three-Way Junction Topological Families in RNA Secondary Structures. *Computational Biology and Chemistry* 37, 2012.

[3] Ph. Rinaudo, Y. Ponty, D. Barth, and A. Denise. Tree decomposition and parameterized algorithms for RNA structure-sequence alignment including tertiary interactions and pseudoknots. Poster submitted to JOBIM 2012.

Tree Decomposition for RNA Structure-sequence Alignment Including Tertiary Interactions and Pseudoknots (Abstract)

Philippe RINAUDO[1,2,4], Yann PONTY[3,4], Dominique BARTH[2] and Alain DENISE[1,3,4]

[1] LRI, Univ Paris-Sud, CNRS, Orsay, F91405
[2] PRISM, Univ Versailles Saint-Quentin, CNRS, Versailles, F78000
[3] IGM, Univ Paris-Sud, CNRS, Orsay, F91405
[4] AMIB group, INRIA, Saclay, F91405
[5] Ecole Polytechnique, LIX, Palaiseau, F91128

Abstract *We present a general setting for structure-sequence comparison in a large class of RNA structures that unifies and generalizes a number of recent works on specific families on structures. Our approach is based on* tree decomposition *of structures and gives rises to a general parameterized algorithm, where the exponential part of the complexity depends on the family of structures. For the previously studied classes, the complexity of our algorithm is the same as previous ad hoc algorithms, and it generalizes to far more general classes of structures.*

Keywords tree decomposition, structure-sequence alignment, pseudoknots, tertiary interaction

Décomposition Arborescente pour l'Alignement Structure-séquence avec interactions tertiaires et pseudo-nœuds (Résumé)

Résumé *Nous présentons un cadre général pour l'alignement structure-séquence pour une large classe de structure d'ARN qui unifie et généralise un certain nombre de travaux récents sur des familles spécifiques. Notre approche est basée sur la* décomposition arborescente *des structures et conduit à un algorithme général à complexité paramétrée, où la partie exponentielle dépend de la classe de la structure. Notre algorithme a la même complexité que les algorithmes ad hoc précédemment développés pour chacune des classes. De plus, il s'applique à des classes de structures bien plus générales.*

Mots-clés décomposition arborescente, structure-séquence alignement, pseudo-nœuds, interaction tertiaire

Rechercher un ARN dans un grande banque de données ou prédire précisément sa structure à partir de celle d'un homologue représente les enjeux majeurs de l'alignement structure-séquence d'ARN, c'est à dire l'alignement entre un ARN de structure connue et un autre ARN de structure inconnue. Non pris en compte dans les algorithmes standard, les pseudo-nœuds et les interactions non-canoniques contiennent néanmoins beaucoup d'informations essentielles et doivent être impérativement considérées si l'on veut que les fonctions associées à ces types d'éléments structuraux soient représentées dans les alignements. Malheureusement, le problème d'alignement qui est polynomial si on le restreint aux structures secondaires (sans pseudo-nœud) a été montré NP-Difficile dès lors que l'on autorise des croisements dans la structure [1]. Dans ce contexte, de nouvelles approches ont été développées afin de prendre en compte certains types de structures comprenant des pseudo-nœuds [2,3] et interactions non-canoniques [4], chacun de ces algorithmes visant un type bien spécifique de famille structurelle.

Nous avons développé une puissante alternative, consistant à "formater" la structure par une approche de décomposition arborescente de graphe avant d'appliquer un algorithme de programmation dynamique pour résoudre le problème de l'alignement. Ce pré-formatage reporte la difficulté du problème sur la décomposition arborescente alors que l'alignement qui s'en suit est résolu par un unique algorithme à complexité paramétrée par la largeur de la décomposition arborescente. Or, nous avons montré que les classes de structures d'ARN

précédemment étudiées dans le cadre de l'algorithmique de l'alignement peuvent être décomposées avec un unique algorithme, aboutissant à une décomposition arborescente de faible largeur. De fait, cette approche par décomposition arborescente unifie et de plus généralise les précédents algorithmes, dont chacun était spécifique à une classe de structure particulière. Notamment, nous prenons en compte les structures où chaque nucléotide peut être en interaction avec n'importe quel nombre de partenaires, considérant ainsi tous types d'interactions non-canoniques. Pour chaque famille, notre algorithme à la même complexité que l'algorithme spécifique existant. Ainsi nous avons défini trois nouvelles classes de structures : les structures standard, les structures non-standard simples, et les hélices triples standard étendues. Chacune de ces classes peut ensuite être utilisée comme brique de base pour former un structure complète, comprenant pseudo-nœuds et interactions tertiaires et dont l'alignement final résulte de la décomposition indépendante de chacune de ses briques. Cette approche nous permet donc de résoudre le problème d'alignement structure-séquence avec un unique algorithme pour une très large classe de structure, représentant plus de 80% des structures secondaires avec pseudo-nœuds, en autorisant de plus d'autres structures comme les triple hélices avec interactions tertiaires.

Structures	Time complexity	multiple interactions	Reference
Secondary Structures (Pseudoknot-free)	$O(nm^3)$	no	[1]
Standard Pseudoknots	$O(nm^k)$	no	[2]
Standard Structures	$O(nm^k)$	yes	×
Simple Non-Standard Pseudoknots	$O(nm^{k+1})$	no	[3]
Simple Non-Standard Structures	$O(nm^{k+1})$	yes	×
Standard Triple Helices	$O(nm^3)$	yes	[4]
Extended Standard Triple Helices	$O(nm^3)$	yes	×
Embedded Standard Pseudoknots	$O(nm^{k+1})$	no	[2]
2-Level Recursive Simple Non-Standard Pseudoknots	$O(nm^{k+2})$	no	[3]
Recursive Classical Structures	$O(nm^{k+2})$	yes	×

Table 1. Panorama des algorithmes existants pour l'alignement structure-séquence. Les classes de structures que nous avons défini sont marquées d'une croix (case référence). Les classes de structures sont regroupées de telle sorte qu'à l'intérieur de chaque groupe (entre lignes grises) la dernière classe (marquée d'une croix) inclut celle(s) situées au-dessus. Notation : k est le degré du pseudo-nœud/(simple) structure standard.

Remerciements. Ce travail a été financé par le projet AMIS ARN ANR-09-BLAN-0160.

Références

[1] T. Jiang and G. Lin and B. Ma and K. Zhang, A General Edit Distance between RNA Structures, *Journal of Computational Biology*, 9 :371-388, 2002

[2] B. Han and B. Dost and V. Bafna and S. Zhang, Structural Alignment of Pseudoknotted RNA, *Journal of Computational Biology*, 15 :489-504, 2008.

[3] T.K.F. Wong and T.W. Lam and W.K Sung and B.W.Y Cheung and S.M. Yiu, Structural Alignment of RNA with Complex Pseudoknot Structure, *Journal of Computational Biology*, 18 :97-108, 2011.

[4] T.K.F Wong and S M Yiu, Structural alignment of RNA with triple helix structure, *Journal of Computational Biology*, 19 :365-378, 2012.

Comparative analyses of 454 pyrosequencing fungal ITS sequences processing methods

Juliette Lengellé[1], Marc Buée[1], Claude Murat[1], Emmanuelle Morin[1] and Francis Martin[1]

[1] INRA de Nancy, UMR1136 INRA/UHP, Interactions Arbres/Microorganismes, 54280 Champenoux, France
{jlengelle, buee, claude.murat, emmanuelle.morin, fmartin}@nancy.inra.fr

Keywords Metagenomics, sequencing errors, OTU, fungal ITS, NGS, pipeline, trimming/denoising programs, clustering

Analyses comparatives de méthodes de traitement de séquences ITS fongiques issues du pyroséquençage 454.

Mots-clés Metagenomique, ITS fungiques, nouvelles technologies de séquençage, erreurs de séquençage, OTU, 'pipeline', logiciels de 'trimming/denoising', 'clustering'

Pour tenter de comprendre le fonctionnement des écosystèmes, il est nécessaire de répertorier les organismes qu'ils hébergent afin de caractériser la dynamique spatio-temporelle des communautés, leurs rôles et leurs interactions potentiels. Grâce au séquençage à haut débit, tel que le pyroséquençage 454, il est possible de déterminer la composition taxonomique des communautés microbiennes de façon exhaustive. Cependant, les erreurs de séquençage inhérentes à ces nouvelles technologies, le pyronoise étant la plus importante, favorisent la création d'OTUs (Operational Taxonomic Units) artéfactuelles [1]. Ces artéfacts induisent également des biais dans l'identification des OTUs (redondance importante des OTUs même après l'étape de 'clustering'), entrainant une surévaluation de la diversité taxonomique réelle [2]. En particulier, l'abondance des séquences uniques, ou singletons, qui peuvent représenter plus de 50% des OTUs, ne correspond pas aux connaissances que nous avons de la « rare biosphère » en écologie. Il est donc indispensable d'identifier et d'éliminer les erreurs de séquençage avant toute assignation taxonomique. De nombreux programmes, tentant de remédier à ces problèmes, sont disponibles pour l'étude de la biodiversité bactérienne utilisant le gène codant l'ADN ribosomique 16S. En revanche, peu d'outils sont actuellement disponibles pour l'étude de la biodiversité fongique utilisant l'espaceur intergénique ribosomique, l'ITS. Le but de cette étude était de combiner différentes procédures d'analyses bioinformatiques destinées à nettoyer les séquences 454 afin de réduire la surestimation et/ou la redondance artéfactuelle des OTUs sans altérer l'estimation de la diversité fongique supposée du milieu étudié.

Plusieurs de ces outils de 'trimming' et/ou de 'denoising', essentiellement ceux de Mothur [3] et d'AmpliconNoise [4], ont été associés en six pipelines distincts pour traiter ces données. Le pipeline utilisant le script perl trimSeqs.pl a été utilisé comme témoin car Dickie (2010) considère qu'il crée un trop grand nombre d'OTUs artéfactuels [2]. Leur efficacité à réduire la redondance des OTU sans diminuer de manière artéfactuelle la diversité des espèces trouvées a ensuite été testée sur trois jeux de données d'environ 100,000 séquences chacun. Ces trois lots d'ADN amplifiés provenaient de trois sites différents : une forêt tropicale, une forêt tempérée et une truffière qui a été inoculée spécifiquement avec *Tuber Melanosporum*. Ce dernier jeu de données a servi de référence car il ne peut présenter au maximum que deux écotypes de *Tuber Melanosporum*, connu comme l'espèce la plus abondante de la truffière. Après extraction spécifique de la région ITS fongique à l'aide du programme Fungal ITS extractor [5], deux autres étapes ont également été ajoutées à l'ensemble des six pipelines testés afin de diminuer encore la redondance. La première étape consiste à effectuer 6 clusterings successifs à 97% d'homologie avec le logiciel Uclust à partir des régions ITS fongiques afin d'obtenir une réduction maximale du nombre d'OTUs. La seconde étape consiste, dans un premier temps, à assigner taxonomiquement chaque OTU, grâce à un Balstn (1^{e}-05) de sa séquence

consensus contre la base nt du NCBI. Puis, dans un second temps, à regrouper les OTUs ayant un 'gi number' (numéro indexation sous NCBI) identiques. Plusieurs analyses statistiques ont été effectuées sur les différents pipelines : le nombre de séquences sélectionnées après nettoyage, le nombre d'OTUs générés, le nombre de singletons trouvés, nombre moyen de séquences supportant chaque OTU, taille moyenne des séquences consensus des OTUs. La diversité a également été étudiée à l'échelle de l'ordre et de l'espèce avec MEGAN.

Les résultats des trois jeux de données sont similaires. Nous présenterons principalement les données obtenues à partir des échantillons de la truffière (site de référence), car la « calibration » sur *Tuber melanosporum* est la plus robuste. Les résultats obtenus pour les six pipelines sont très différents pour ce site. Par exemple, le nombre moyen de séquences supportant chaque OTU est très faible pour les pipelines utilisant plusieurs programmes de 'denoising', qu'il s'agisse de ceux de Mothur ou d'Ampliconoise (moins de dix). Le nombre de séquences sélectionnées après nettoyage est également très variable suivant le programme utilisé (entre 7 et 43% des séquences). Le nombre de singleton généré est plus important pour les pipelines faisant plusieurs étapes de 'denoising'. La diversité va du simple au double suivant la méthode de nettoyage utilisée et les dix espèces les plus abondantes ne sont pas toujours les mêmes. Dans ce dernier cas, *Tuber Melanosporum* n'est pas systématiquement l'OTU le plus abondant suivant le pipeline utilisé.

Après comparaison des différents résultats, il semble que le pipeline utilisant trim.flows de Mothur pour l'étape de 'triming/denoising' soit un bon compromis car il donne les résultats les plus satisfaisants suivant nos critères d'évaluations.

Acknowledgements

This work was supported by European ECOFINDERS program, the Lorraine Region and the FEDER

References

[1] V. Gomez-Alvarez, TK. Teal and TM. Schmidt, Systematic artifacts in metagenomes from complex microbial communities. *The ISME Journal*, 3:1314-1317, 2009.

[2] IA. Dickie, Insidious effects of sequencing errors on perceived diversity in molecular surveys. *New Phytologist*, 188:916-918, 2010.

[3] P. Schloss, SL. Westcott, T. Ryabin, JR. Hall, M. Hartmann, EB. Hollister, RA. Lesniewski, BB. Oakley, DH. Parks, CJ. Robinson, JW. Sahl, B. Stres, GG. Thallinger, DJ. Van Horn and CF. Weber, Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and environmental microbiology*, 75:7537-7541, 2009.

[4] C. Quince, A. Lanzen, RJ. Davenport and PJ. Turnbaugh, Removing Noise From Pyrosequenced Amplicons. *BMC Bioinformatics*, 12:20-38, 2011.

[5] RH. Nilsson, G. Bok, M. Ryberg, E. Kristiansson and N. Hallenberg, A software pipeline for processing and identification of fungal ITS sequences. *Source Code for Biology and Medicine*, 4:1, 2009.

Adapting comparative genomics to MapReduce

Natalia GOLENETSKAYA[1,2,3], David James SHERMAN[1,2,3]

[1] MAGNOME, INRIA Bordeaux Sud-Ouest, 351 cous de la Libération, Talence
[2] LaBRI UMR 5800 CNRS (*ditto*)
[3] PRES Université Bordeaux (*ditto*)
{natalia.golenetskaya,david.sherman}@inria.fr

Keywords Comparative genomics, data mining, distributed computation, MapReduce

1 Introduction

Comparative genomics seeks to understand the history and function of genomes by large-scale analysis of the *relations* between genomic elements, as contrasted with analyzing of those elements individually[1,3,7].

Beyond the cost of storing increasingly voluminous genome data[12], comparative genomics produces a significant strain on bioinformatics: for every linear increase in basic genome data, there is a geometric increase in the relations, represented as n-tuples, between genome elements. From a practical standpoint, scaling large-scale analyses to extreme data volumes is increasing addressed by NoSQL and MapReduce technologies, both associated with cloud computing [6,2]. However, adopting these technologies imposes new constraints on algorithms, and there is as yet an insufficient real body of best practices for their use in bioinformatics.

In this work we attempt to define some ground rules for successful conversion of global bioinformatic analyses to highly scalable MapReduce deployments. We define a criterion that can be used to guide redefinition of bioinformatics algorithms so that they can be efficiently deployed on NoSQL and MapReduce infrastructures, with a reasonable guarantee of scaling out to new data volumes.

We implemented a Cassandra schema of column families, derived from use-case analysis of the Génolevures data resource in [8], that indexes serialized objects representing genomic elements, and to encode the n-ary relations between these elements. We illustrate our approach with two successful conversions of global comparative genomics analyses.

2 Systematic identification of gene fusion and fission events.

One consequence of genome remodeling in evolution is the modification of genes, either by fusion with other genes, or by fission into several parts. By mapping these events to phylogeny we can paint a global picture of remodeling across many species simultaneously. In [4] we developed a method that is appropriate for highly redundant eukaryote genomes. Briefly, paralogous groups of genes are encoded by HMM, that are systematically aligned using hhsearch [9] and filtered to obtain a graph of relations between genome segments. This graph is then mined for subgraph motifs, each containing at least two collinear segments aligned to non-collinear segments in another genome. Each of these motifs is an "Event", corresponding to a fusion, a fission, or some combination thereof. Events are mapped onto the species phylogeny using a parsimony argument.

The bottleneck of this algorithm is the graph analysis. The input data for this part is a list of relations between composite "C-groups" (HMMs with 2 or more aligned segments) and elementary "E-groups" (HMMs with 1 aligned segment). To express the graph analysis using the MapReduce paradigm we use the knowledge that the graph is bipartite, so to find all events we can decompose the graph into sub-graphs with of the longest path 2. For each E-group that can participate in an event we calculate (in parallel, using grouping by C-groups, and E-groups) all nodes that are reachable using path of length 1 or 2.

3 Iterative sampling for consensus clustering for multi-genome protein families

Protein families in related genomes provide key information for phylogenetic analysis, functional annotation, exploring of functional diversity, and inferring metabolic networks. Many complementary methods have been developed for computing families. In [5] we developed a tailored algorithm for consensus clustering to reconcile complementary partitions of proteins into families, based on a compact encoding of the confusion matrix and an election procedure to find the best consensus from competing partitions. The 4500 families in [10], for example, were computed by reconciling 12 competing partitions of 45000 proteins (thus 10^9 pairwise sequence comparisons) .

The conflict regions of the consensus procedure are independent and can be naturally processed in separate map functions. Beside this natural decomposition, there is another scaling issue. With the linear growth of proteins we have quadratic growth in the size of the distance matrices. These rapidly reach to the point when the existing tools (i.e. MCL clustering[11]) cannot work with these large distance matrices.

We cluster large matrices using an iterative sampling strategy: each consensus clustering operation is performed only on a part of data. Then we randomly redistribute proteins into chunks with the condition that proteins which was found in the same family will go to the same sample chunk. In the next iteration we recluster proteins again in samples etc., until the total size of families does not change. In each iteration the family partition is better than in the previous are, because the chance that similar proteins appeared in the same sample increases. The process converges in a small number of iterations.

References

[1] J. Aitchison and T. Galitski. Inventories to insights. *The Journal of cell biology*, 161(3):465–469, May 2003.

[2] D. Avresky, M. Diaz, A. Bode, C. Bruno, and E. Dekel, editors. *Cloud Computing: First International Conference, CloudComp 2009*, LNICST 34, Munich, October 2009. Springer-Verlag.

[3] R. Barriot, J. Poix, A. Groppi, A. Barré, N. Goffard, D. Sherman, I. Dutour, and A. de Daruvar. New strategy for the representation and the integration of biomolecular knowledge at a cellular scale. *Nucleic Acids Research*, 32(12):3581–3589, July 2004.

[4] P. Durrens, M. Nikolski, and D. Sherman. Fusion and fission of genes define a metric between fungal genomes. *PLoS Comput Biol*, 4(10):e1000200, 10 2008.

[5] M. Nikolski and D. J. Sherman. Family relationships: should consensus reign?- consensus clustering for protein families. *Bioinformatics*, 23:e71–e76, 2007.

[6] S. Sakr, A. Liu, D. Batista, and M. Alomari. A survey of large scale data management approaches in cloud environments. *Communications Surveys & Tutorials, IEEE*, 13(3):311–336, 2011.

[7] D. Searls. Data integration: challenges for drug discovery. *Nature Reviews Drug Discovery*, 4(1):45–58, Jan. 2005.

[8] 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 Research*, 37(suppl 1):D550–D554, 2009.

[9] J. Söding. Protein homology detection by HMM-HMM comparison. *Bioinformatics*, 21(7):951–960, Apr. 2005.

[10] J.-L. Souciet et al. Comparative genomics of protoploid Saccharomycetaceae. *Genome Research*, 19:1696–1709, 2009.

[11] S. M. van Dongen. *Graph Clustering by Flow Simulation*. PhD thesis, University of Utrecht, The Netherlands, 2000.

[12] K. Wetterstrand. DNA sequencing costs: Data from the NHGRI large-scale genome sequencing program available at: www.genome.gov/sequencingcosts., 2010.

Enzyme survey and how to find new ones

Maria Sorokina[12], Adam Alexander Thil Smith[1], Mark Stam[1], Karine Bastard[1], Claudine Médigue[1] and David Vallenet[1]

[1] CEA, IG, Genoscope, 2 rue Gaston Crémieux CP5702, F-91057 Evry, France,
CNRS, UMR 8030, 2 rue Gaston Crémieux, CP5706, F-91057 Evry, France,
Université d'Evry, F-91057 Evry, France.
[2] Université Paris-Sud 11, F-91405 Orsay, France.
msorokina@genoscope.cns.fr

Keywords orphan enzymes, protein families, genomic context, metabolic context

1 Introduction

Of all biochemically characterized metabolic reactions, a significant fraction has yet to be associated with a nucleic or protein sequence, *i.e.* these reactions are are sequence-orphan enzymatic activities or orphan enzymes for short. Enzyme activities are classified by the IUBMB using Enzyme Commission (EC) numbers which are specific numerical identifiers. In 2004, Peter Karp [1] called for the first time for an enzyme genomics initiative, raising the orphan enzyme problem. Since then, several number of surveys have already been published, for example: Lespinet *et al*. in 2006 [2], Chen *et al*. [3] and Pouliot *et al*. [4] in 2007 and Hanson *et al*. [5] in 2010. Here, we present an update of these observations and some new analyses such as the study of the relationship between enzymatic activities and protein families. In order to reduce the gap between enzymatic activities and available sequence data, we propose a new method to find candidate genes for orphan activities, named CanOE (Candidate genes for Orphan Enzymes).

2 Enzyme survey

The first generation of protein sequencing technology allowed association of some protein sequences with activities. This task became easier with the development of molecular biology technologies (PCR, expression cloning and DNA sequencing), and a huge number of activities were discovered and associated to genes. Surprisingly, genomics and Next Generation Sequencing (NGS) do not preserve this tendency at present time.

There are currently about 4852 validated EC numbers. By this time, the proportion of orphan activities is about 23% (see Fig 1). Since 2004, this percentage decreased from around 40% to 23%. The call for an "enzyme genomics initiative", which goal was to associate a least one sequence to an EC number, has thus been heard. However, this decreasing is more related to database curation efforts than new experiments.

We define here two kinds of orphan activities: global and local ones. A global orphan enzymatic activity is an activity that has no known associated genes in any organism. A local one is an activity that has no known associated genes in a given clade, but has been observed in it. In contrast to the proportion of global orphan EC numbers (23%), the proportion of local orphans is around 36% at the superkingdom level (Fig 1). This higher level of local orphans raises questions about the lack of knowledge in a specific kingdom which could be explained by non-detectable homologs or the presence of uncharacterised proteins that evolved convergently. The attribution of an EC number for a newly discovered activity is a quite laborious task which requires a reviewing process and solid experimental evidences. Thus, the entire universe of known biochemical reactions is not (yet) covered by this classification. This present survey is conducted using the EC classification and could be completed by integrating other sources of reactions.

Another interesting fact to highlight is that the percentage of global orphan is nearly identical to the percentage of protein families with no known function. Paradoxically, new activities are not found into new protein families (Fig 2), but rather into already known protein families.

3 How to Find New Activities

The discovery of the various metabolic functions catalyzed by enzymes encoded by the genes from the exponentially increasing number of sequenced genomes is one of the main focuses of bioinformatics tools today. We have developed CanOE (Candidate genes for Orphan Enzymes) [6], a four-step bioinformatics strategy that proposes ranked candidate genes for orphan enzymes. The first step locates "genomic metabolons", i.e. groups of co-localized genes coding proteins catalyzing reactions linked by shared metabolites, in one genome at a time. In the second step, they are used to generate candidate associations between un-annotated genes and gene-less reactions. The third step integrates these gene-reaction associations over several genomes using gene families, and summarizes the strength of family-reaction associations by several scores. In the final step, these scores are used to rank members of gene families which are proposed for metabolic reactions. These associations are of particular interest when the metabolic reaction is a sequence-orphan enzymatic activity. Our strategy found over 60,000 genomic metabolons in more than 1,000 prokaryote organisms, generating candidate genes for many metabolic reactions, of which more than 70 distinct orphan reactions. The results are available at this URL: http://www.genoscope.cns.fr/agc/microscope/metabolism/canoe.php.

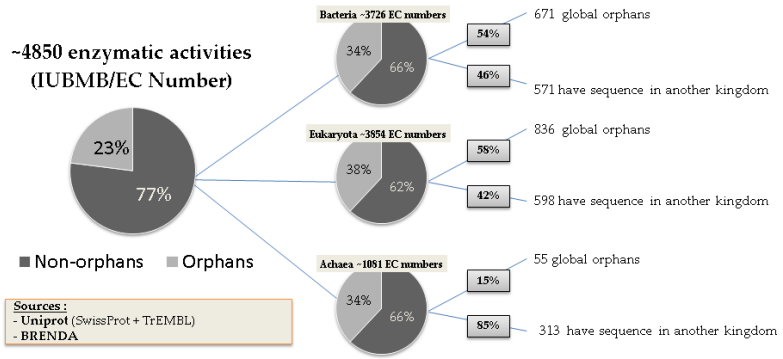


Figure 1. Distribution of local orphan activities across superkingdoms

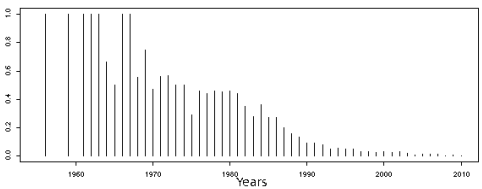


Figure 2. Proportion of sequenced enzymes with at least one new Pfam domain by year of discovery

References

[1] P. D. Karp. Call for an enzyme genomics initiative. *Genome Biol.*, 5:401, 2004
[2] O. Lespinet and B. Labedan. Puzzling over orphan enzymes. *Cell. Mol. Life Sci.,* 63:517-523, Mar 2006.
[3] L. Chen and D. Vitkup. Distribution of orphan metabolic activities. *Trends Biotechnol.,* 25:343-348, Aug 2007.
[4] Y. Pouliot and P. D. Karp. A survey of orphan enzyme activities. *BMC Bioinformatics,* 8:244, 2007.
[5] A. D. Hanson *et al.,* 'Unknown' proteins and 'orphan' enzymes : the missing half of the engineering parts list – and how to find them. *Biochem. J.,* 425(1):1-11, Jan 2010.
[6] A.A.T. Smith *et al.*, The CanOE strategy: integrating genomic and metabolic contexts across multiple prokaryote genomes to find candidate genes for orphan enzymes. *PloS Comp. Biol.*, (in press), 2012.

Development of analysis tools for transcriptomic data studying *Microcebus murinus* brain in relation with age or neurodegenerative pathology

Pierre-Antoine JEAN[1], Anne LAURENT[2], Jean Michel VERDIER[3] and Gina DEVAU[3]

[1] Master STIC pour la Santé, spécialité « Bioinformatique, Connaissance, Données »
Universités de Montpellier 1 & 2, Institut Telecom, TIC et Santé
CC 92000 Place Eugène BATAILLON 34095 Montpellier CEDEX 05
pierre-antoine.jean@etud.univ-montp2.fr

[2] Labotatoire d'Informatique, de Robotique et de Microélectronique de Montpellier,
UMR5506 Université de Montpellier 2, 34095 Montpellier CEDEX 5
anne.laurent@univ-montp2.fr

[3] Institut National de la Santé et de la Recherche Médicale, UMR S710
CC 105 Université Montpellier 2, place Eugène BATAILLON, 34095 Montpellier CEDEX 5
{gina.devau , jean-michel.verdier}@univ-montp2.fr

Abstract *Procedures for transcriptomic data analysis still are not that easy. We provide tools to facilitate sorting, extraction, interpretation and visualization of results. Samples used for this work are from Microcebus murinus brain.*

Keywords Alzheimer disease, *Microcebus murinus*, transcriptome, data analysis.

Développement d'outils d'analyse de données transcriptomiques étudiant le vieillissement cérébral physiologique et pathologique de Microcebus murinus

Résumé *Les procédures d'analyse de données transcriptomiques restent encore difficiles, nous proposons des outils pour faciliter le tri, la discrimination, l'aide à l'interprétation et la visualisation des résultats. Ce travail a été mené à partir de données transcriptomiques issues d'échantillons de cerveau de Microcebus murinus.*

Mots-clés Maladie d'Alzheimer, Microcebus murinus, transcriptome, analyse de données

1 Introduction

Le vieillissement est le premier risque de maladies neurodégénératives comme la maladie d'Alzheimer (MA). Cependant les mécanismes moléculaires impliqués soit dans le vieillissement cérébral physiologique soit dans la MA sont encore très mal connus. Actuellement, les approches transcriptomiques permettent de mieux appréhender ces mécanismes biologiques complexes. Par ailleurs, les modèles rongeurs, classiques en laboratoire ne présentent pas de manière naturelle la MA, donc des modèles animaux plus proches de l'Homme et porteurs des signes histologiques de la MA sont nécessaires. C'est pour cela que le modèle primate lémurien, *Microcebus murinus,* est un modèle pertinent pour l'étude du vieillissement cérébral physiologique et de la MA [1]. Une étude a donc été menée avec des puces Affymetrix HG U133 plus2, à partir du cortex temporal, d'animaux jeunes adultes, âgés sains et âgés porteurs des signes histologiques de MA.

2 Objectifs

Le but de notre travail est d'améliorer les procédures d'analyse et d'interprétation de données transcriptomiques massives.

2.1 Les tris de données

La première étape a été d'automatiser les tris de données issues des biopuces. Pour en augmenter l'efficacité et la fiabilité tout en diminuant le temps d'exécution par rapport à une approche manuelle utilisant un tableur. Pour cela nous avons conçu un *pipeline* permettant le tri des gènes en fonction de leur présence dans les échantillons biologiques. Nous avons envisagé trois algorithmes de tri différents et un outil pour réunifier des fichiers sans occurrences. Il est codé en Python, ce langage nous offre la facilité de manipuler les fichiers textes et une bibliothèque standard riche.

2.2 Exploitation des motifs graduels

La deuxième partie se focalise sur le traitement et l'exploitation des motifs graduels extraits des données transcriptomiques filtrées. Les corrélations positives et négatives entre transcrits nous permettent de représenter les modifications dynamiques qui peuvent réguler les gènes d'intérêt.

2.3 Visualisation des résultats

Afin de permettre l'interprétation biologique de ces résultats, la troisième étape de ce travail a été la conception d'un *pathway* sur le vieillissement et de compléter le *pathway* existant sur la maladie d'Alzheimer avec les résultats issus de nos analyses. Pour réaliser cette tâche, nous avons récupéré les ontologies connues sur nos gènes d'intérêts à partir des banques de données. Les procédures pour obtenir ces informations peuvent être longues. C'est pourquoi nous avons conçu, un petit *pipeline* artisanal, pour automatiser les requêtes qui sont constituées de nombreux gènes. De plus nous concevons un petit logiciel pouvant contenir et croiser ces informations afin de voir les liens et les interactions entre ces différents acteurs moléculaires. L'étude sur les *pathway* est toujours en cours. Tout au long de ce projet, nous travaillons pour développer une interface des logiciels facile d'utilisation pour les biologistes. Chaque *pipeline* est testé avec un fichier source contenant 54 675 *probesets*, et les résultats obtenus sont comparés manuellement avec un tableur. Les outils seront utilisables pour toutes les données transcriptomiques, si le format de fichier est normalisé.

Remerciements

Nous remercions le Labex NUMEV (http://www2.lirmm.fr/numev/) pour le financement de la participation à JoBIM.

References

[1] R. Abdel Rassoul, S. Alves, V. Panstesco, J. De Vos, B. Michel, M. Perret, N. Mestre-Francés, JM. Verdier, G. Devau, Distinct Transcriptome Expression of the Temporal Cortex of The Primate Microcebus murinus during Brain Aging versus Alzheimer's Disease-like Pathology. PloS ONE 5(9) : e12770. Doi:10.1371/journal.pone.0012770, 2010

Thalia

A database dedicated to association genetics in plants

Yannick De Oliveira[1], Guy-Ross Assoumou-Ella[1], Julien Cornouiller[1], Johann Joets[1] and Alain Charcosset[1]

[1] UMR Génétique Végétale, INRA – Université Paris-Sud - CNRS, Ferme du Moulon, 91190, Gif-sur-Yvette, France

ydeoliveira@moulon.inra.fr

Abstract *Association genetics studies aim at revealing significant association between phenotypic traits and genotypic data, taking into account the possible confounding effect of admixture and relatedness.*

In order to facilitate this diagnosis, managing germplasm and related information within a database is essential. For this purpose Thalia project aims to offer a database to store , maintain and extract data in view of association genetics analysis.

Keywords Association genetics, genetic resources, genotyping, phenotyping.

1 Introduction

Diversity and association genetics studies lead to manipulate a large number of individual, lines, clones and/or populations. Moreover, emergence of high-throughput technologies for both genotyping and phenotyping generates a large amount of data. These need to be stored and managed in order to perform requests and organize datasets to conduct association genetics studies.

The Thalia database has been developed with this aim. It manages genetic resources, phenotyping and genotyping data, and also population structure information. Thalia enables data extraction in formats used by genetic association software.

2 Data Structure

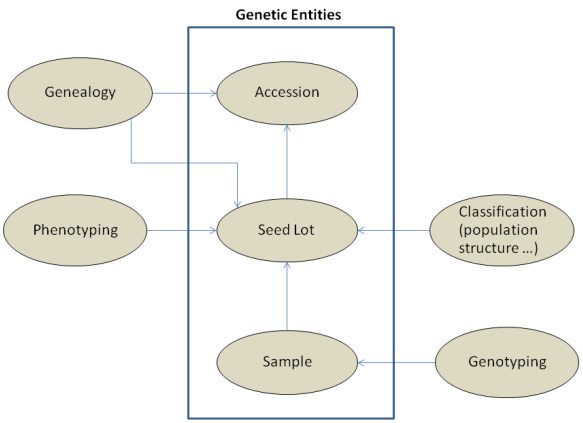


Figure 1. Database structure

Thalia schema enables dynamic description of accession (an introduction within the collection) and seed-lot types. Thus, users can describe accession types like inbred lines, population, etc. Images can be linked to an accession.

This dynamic description is also available for markers. Thus, users can manages SNP, SSR, RFLP or any other kind of marker. A (DNA) sample is characterized for a locus in a given experience (conducted by a person in an institute, both documented in the database). One or more alleles are observed with given frequencies. Alleles are described following a referential used for the experiment. A correspondence option allows to bind heterogeneous data (observed with different referential).

Phenotyping data are stored as expertised data relative to a seed lot observed in a given environment (average values or other statistical estimation). Raw data are stored in compressed files.

Classifications are expertised results concerning the assignation of a seed lot to classes of a population structure analysis [1]. All data in Thalia are managed in projects. A user can access to the data concerning projects in which he is involved. Some users can have an administrator status, which give them rights to insert data, and link data and users to projects.

3 Requesting and Analyzing Data

Each user can request and extract data concerning projects he is involved in. Genotyping and phenotyping data can be requested separately, but it is also possible to cross those data to extract information for association genetics studies like in Tab.1. Classification results can also be requested.

Seed Lot	Vgt1 MITE A	Vgt1 MITE B	Days to pollen	Northern Flint	Andes	Caribbean	Corn belt dent	Europe	Mexican	Italian
PPS48	0,00	1,00	57,9	0,9140	0,0104	0,0088	0,0092	0,0162	0,0106	0,0308
PI217460	0,21	0,79	72,3	0,9156	0,0056	0,0088	0,0412	0,0084	0,0096	0,0110
PI218143	0,52	0,48	85,8	0,0852	0,0178	0,0164	0,6516	0,0178	0,2016	0,0088
PI213719	0,80	0,20	81,8	0,0810	0,0136	0,0268	0,8024	0,0150	0,0204	0,0410
PPS774	1,00	0,00	77,2	0,0090	0,0664	0,2494	0,0174	0,0410	0,0898	0,5274

Tab. 1. Cross view and classification extraction for association genetics studies.

Thalia enables a data analysis module. At the moment this part of the application includes a genetic distance algorithm.

A Google map viewer has been integrated to Thalia. For accessions associated with geographical coordinates, this viewer makes it possible to display classification, allele frequencies, trait or accession repartition on the map.

4 Next development

The main functionality that will be available in Thalia in next month is integration of NGS data management, a web-service for some external tool request, trait ontology integration to manage phenotyping data, data extraction tools to interact with SniPlay [2] and GnPAsso databases.

References

[1] L. Camus-kulandaivelu, J-B Veyrieras, D. Madur, L. Combes, M. Fourmann, S. Barraud, P. Dubreuil, B. Gouesnard, D. Manicacci, A. Charcosset, Maize Adaptation to Temperate Climate: Relationship Between Population Structure and Polymorphism in the Dwarf8 Gene. *Genetics* 172, 2449-2463, 2006.

[2] A. Dereeper, S. Nicolas, L. Le Cunff, R. Bacilieri, A. Doligez, J-P Peros, M Ruiz, P This, SNiPlay: a web-based tool for detection, management and analysis of SNPs. Application to grapevine diversity projects. *BMC Bioinformatics* 12:134, 2011.

Systems analysis of yeast fermentation and respiration: transcriptomic, proteomic and interactomic changes.

Emmanuelle BECKER[1], Aurélie LARDENOIS[1], Régis LAVIGNE[2], Yuchen LIU[1], Bertrand EVRARD[1], Charles PINEAU[2] and Michael PRIMIG[1]

[1] IRSET, U1085 INSERM, University of Rennes 1, 35042 Rennes, France
{emmanuelle.becker, aurelie.lardenois, yuchen.liu, michael.primig}
@univ-rennes1.fr
[2] IRSET Proteomic Platform, U1085 INSERM, University of Rennes 1, 35042 Rennes, France
{regis.lavigne, charles.pineau}@univ-rennes1.fr

Abstract S. cerevisiae *is a model organism for biological processes such as mitotic growth and meiotic development. We used the transcriptome, proteome, and interactome of diploid yeast cells in two different metabolic states: fermentation of glucose and respiration of acetate, and integrated these different types of data together. We thus present a multi-scale comparison of yeast undergoing fermentation and respiration.*

Keywords Yeast, Fermentation, Respiration, Transcriptome, Proteome, Protein-protein interaction network.

1 Introduction

S. cerevisiae is a model organism for biological processes such as mitotic growth and meiotic development. We used the transcriptome, proteome, and interactome of diploid yeast cells in two different metabolic states: fermentation of glucose and respiration of acetate, and integrated these different types of data together. We thus present a multi-scale comparison of yeast undergoing fermentation and respiration.

2 Systems analysis of yeast fermentation and respiration

We used transcriptomic data previously published by our group [1], proteomic data generated specifically for this study by the IRSET Proteomic Platform (see below), and protein-protein interaction data from the gold-standard yeast binary interactome [2].

2.1 Determination of the global Yeast Proteome for fermentation and respiration

Duplicate total protein extracts from logarithmically growing diploid cells cultured in rich medium with glucose (YPD, fermentation) or acetate (YPA, respiration) were run on an SDS-polyacrylamide gel, each lane was cut into 30 bands which were digested with Trypsin. Samples were then processed by LC-MS/MS analysis for protein identification. The resulting information was integrated with the output of our own whole-genome expression profiling experiments using tiling microarrays as well as DNA sequencing data and protein network data available via certified public repositories (see [3] for method description).

Within each of these four independent experiments (fermentation in duplicate: YPD1 and YPD2, and respiration in duplicate: YPA1 and YPA2), we identified more than 4000 proteins. The results are presented in Table 1. In both conditions, the overlap between replicates is important: 3528 out of 4967 proteins found with both YPD1 and YPD2 for fermentation (71%), and 3744 out of 5035 proteins found with both YPA1 and YPA2 for respiration (74%). Taken together, 5513 proteins out of the 6713 theoretical ones are recognized in at least one experiment (82%).

	YPD1	YPD2	YPA1	YPA2
Proteins identified	4416	4079	4379	4400
Proteins identified in both replicates	3528		3744	
Proteins identified at least once	4967		5035	

Table 1. Number of proteins identified in the 4 experiments (YPD1, YPD2, YPA1, YPA2). The theoretical yeast genome, based on Saccharomyces Genome Database, contains 6613 proteins.

2.2 Pathway analysis, Interactome modules analysis

We examined the impact of shifting from fermentation to respiration for different biological processes. To model biological processes, we used two orthogobal approaches: (i) Gene Ontology Biological Processes, and (ii) functional modules identified from a gold-standard yeast interaction network with Overlapping Cluster Generator [4].

The idea was to identify clearly the processes affected by this switch and those whose actors are stable. The poster will present several examples and global tendancies. Figure 1 shows an example of a functional module globally up-regulated in respiration, from both transcriptomic and proteomic data.

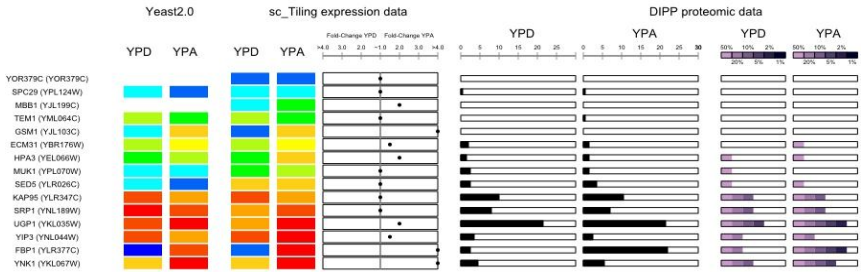


Figure 1. Graphic View. Transcriptomic results are presented with a color scale from blue (lowest transcription decile) to red (highest transcription decile). Transcriptomic results include data from Yeast 2.0 arrays and sc_Tiling experiments. The fold-change was computed from the sc_Tiling data. The proteomic data are presented with one scale indicating the gel saturation, and another scale counting the peptides identified, from the 0.50 percentile to the 0.99 percentile (steps : 0.50 - 0.80 - 0.90 – 0.95 – 0.98 – 0.99).

Acknowledgements

This work is supported by CREATE grant from the Région Bretagne awarded to MP.

References

[1] Lardenois A., Liu Y., Walther T., Chalmel F., Evrard B., Granovskaia M., Chu A., Davis R.W., Steinmetz L.M. and Primig M. Execution of the meiotic noncoding RNA expression program and the onset of gametogenesis in yeast require the conserved exosome subunit Rrp6. *Proc Natl Acad Sci USA.*, 108(3):1058-1063, 2011

[2] Yu H., Braun P., Yildirim M.A., Lemmens I., Venkatesan K., Sahalie J., Hirozane-Kishikawa T., Gebreab F., Li N., Simonis N., Hao T., Rual J.F., Dricot A., Vazquez A., Murray R.R., Simon C., Tardivo L., Tam S., Svrzikapa N., Fan C., de Smet A.S., Motyl A., Hudson M.E., Park J., Xin X., Cusick M.E., Moore T., Boone C., Snyder M., Roth F.P., Barabási A.L., Tavernier J., Hill D.E. and Vidal M. High-quality binary interaction map of the yeast interactome network. *Science*, 322(5898):104-110, 2008

[3] Lavigne R., Becker E., Liu Y., Evrard B., Lardenois A., Primig M., and Pineau C. Direct iterative protein profiling (DIPP) – an innovative method for large scale protein detection applied to budding yeast mitosis. *Mol. Cell. Proteomics*, 11(2):M111, 2012.

[4] Becker E., Robisson B., Chapple C.E., Guénoche A., and Brun C. Multifunctional proteins revealed by overlapping clustering in protein interaction network. *Bioinformatics*, 28(1):84-90, 2012.

Analysis of fluorescently labeled combed DNA molecule images

Leila MUREŞAN[1], Jean-François MARIET[2], Xavier DARZACQ[2] and Hugues ROEST CROLLIUS[1]

Ecole Normale Supérieure, Institut de Biologie de l'ENS, IBENS, CNRS, UMR 8197, Paris, F-75005 France

leila.muresan, mariet, darzacq, hrc@ens.fr

Keywords Combed DNA molecules, microscopy image analysis, segmentation, profile construction

1 Introduction

This work focuses on detection and profile analysis of elongated structures, such as DNA molecules, in microscopy images. The fluorescence marked molecules are combed on a coverslip and imaged via TIRF microscopy. They appear as thin, curvilinear structures, ideally bright ridges on a uniform, dark background. However, depending on the labeling technique, the concentration of the fluorescent dyes might considerably vary, leading to dark gaps in the image of one molecule. Moreover, shot noise and readout noise further hamper the quality of the microscopy image (also excess noise if EMCCD camera is used).

The challenge consist of correctly identifying the entire molecule, even if it is weakly labeled and to uniformly sample the fluorescence intensity along the molecule for use in subsequent analysis. An example of fluorescent image of combed molecules can be seen in Fig. 1.

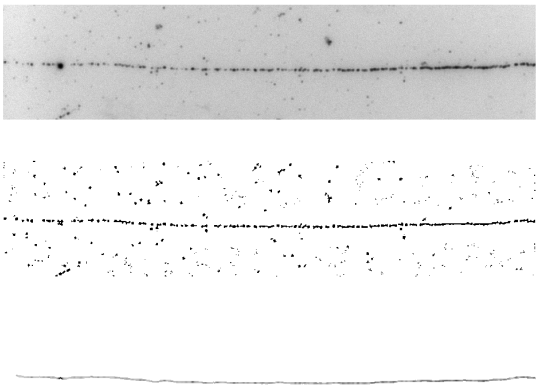


Figure 1. Up: Inverted intensity DNA molecule image (corresponding to approximately 150 Kb). Middle: Ridge detection result. Down:Molecule detected after tensor voting completion. Image intensities are inverted for better visualization.

2 Detection of DNA molecule support

The first step in the analysis consists of detecting the bright pixels corresponding to the image of the DNA molecules. The approach for ridge detection in the literature ([6,4,1]) is the computation of the largest absolute eigenvalue of the Hessian matrix of the image. An alternative approach based on beamlets was used by Berlemont *et al.* in [3].

We shall use the computed eigenvalues of the Hessian matrix after a variance stabilization transform and a regularizing Gaussian blur were applied. In case of very sparse labeling, the detection of a molecule results in short fragments or blobs (see Fig. 1, middle) that have to be subsequently combined. In order to correctly reconstruct the entire support of the DNA molecule a completion step is applied, which relies on tensor voting, a method that performs completion of contours by making use of the coherence of orientations of contour elements in the neighborhood of these elements (see [2,5] for description and implementation, respectively).

3 Profile construction

Once the support of the DNA molecule is detected, the profile information can be built based on the same fluorescent intensity which served for the detection of the molecules or on an additional imaging channel. We perform resampling of the curve representing the molecule based on cord-length and arc-length parametrization, and for the fluorescence intensity estimation we use the Taylor expansion at the re-sampled locations. In Fig. 2 are shown the computed profiles of the molecule in Fig. 1. Although the differences are moderate, they can be significant when the application at hand requires high localization accuracy. Further study of the properties of the profile construction step is planned in the future.

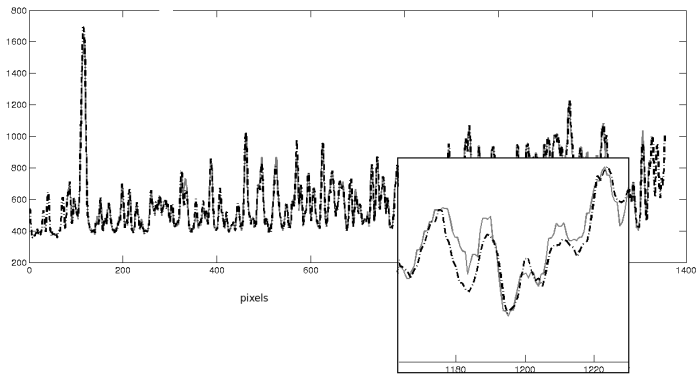


Figure 2. Computed profiles of the combed DNA molecule from Fig. 1 (thin line: chord-length, dotted line: arc-length parametrization) and inlay of zoomed detail

Acknowledgements

The authors thank David Tschumperlé for fruitful discussions and the Hyrien lab (IBENS) for the combed DNA molecule image samples used to illustrate this work.

References

[1] M. Jacob and M.Unser, Design steerable filters for feature detection using Canny-like Criteria, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004.

[2] G. Guy and G.G. and Medioni, Inferring Global Perceptual Contours from Local Features, *IJCV*, 20:113-133, 1996

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

[4] C.Steger, An Unbiased Detector of Curvilinear Structures, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20, 2:113-125, 1998

[5] T.Linton, Tensor Voting Framework, *http://www.mathworks.com/matlabcentral*, 2008

[6] T.Lindeberg, Feature detection with automatic scale selection, *IJCV*, 30, 2:77-116,1998

Feature selection for microRNA prediction

Benjamin ZERATH[1], Sébastien TEMPEL[1] and Fariza TAHI[1]

IBISC - IBGBI, 3ème étage, 23, Bd de France 91034 EVRY, France
{benjamin.zerath, sebastien.tempel, fariza.tahi}@ibisc.univ-evry.fr

Keywords microRNAs, microRNA features, feature selection, microRNA ab initio prediction

1 Introduction

Micro-RNAs (miRNA) are a family of non coding and tiny sized (˜22 nucleotides) RNAs involving in post-transcriptional regulation by targeting mRNAs for cleavage or translational repression. Mature miRNAs are made from ˜90 nucleotides-sized precursors (pre-miRNAs), characterized by a hairpin-type structure. More than 18000 miRNAs have been discovered over more than 140 species, of which about 2000 in Homo sapiens. However, recent studies estimate that a huge number of miRNAs have not been discovered yet.

Today, numerous miRNA prediction algorithms are proposed, based on varied techniques. They can be based on homology or on comparative approach methods that need detailed genome information. Ab initio methods, that are able to work on non-annotated genomes, are also proposed. They use a set of features that describe miRNAs based on their structure or their nucleotidic composition. These features play an important role in the efficiency of the prediction: all features are not always beneficial (some give weak information, others are redundant, . . .). Thus, it is necessary to select the most adapted and the most representative feature set.

2 Methods

We first considered 22 features that we created and used in the miRNAFold ab initio pre-miRNA prediction algorithm we developed in the laboratory [10]. These features describe palindromes, non-exact palindromes and hairpins and concern for instance the maximal number of consecutive bulges and the maximal number of consecutive same nucleotides [10].

We extracted 165 features from several pre-miRNA prediction algorithms. Among them, we used the 32 structural features from Triplet-SVM [9] which considers the structure of 3 adjacent nucleotides depending on the middle one, like “U(.(” or “C.((”. We used 63 features from microPred [2], 14 features from miRabella [6], and some others from miRank [8] and miPred [4]. These features correspond to the primary sequence of pre-miRNA such as ratio of dinucleotides [2,7], and the secondary structure of pre-miRNA such as the number of appariements [6] and the average size of internal loops [2].

Thus, we tested a bench of 187 features giving information on the structure and the sequence of the precursor miRNA. To classify these 187 features, we used the algorithms Best First, Linear Forward Selection, Greedy Stepwise, Scatter Search and Genetic Search, five feature selection algorithms proposed by the statistics workbench Weka [3].

3 Materials

We used several data set tests, each one representing human genome, mouse genome, or a combination of species from miRbase. These sets have been constructed on an analogous method: we gathered the known pre-miRNAs from mirBase [11] which are the true positives (about 1300 for human, 500 for mouse) to whom we added other ARNs known for not being pre-miRNAs (true negatives) but which have the same size. Also, none of the true negatives have more than 80% similarity with another one, to avoid over-fitting. The different data sets have same true positive miRNAs but different true negative miRNAs, which are based on several

databases. In the first set, about 4900 true negatives for human and 2900 for mouse were used, and 35000 for the combination of different species (from mammals, insects, trees).In the second one, 4700 true negatives were used for human, 1400 for mouse. The third data set gathers only the true negatives that have a hairpin-type secondary structure: 1500 for human, 1700 for mouse.

4 Results

All of these data sets result in the selection of similar numbers of features, from 10 to 20 features. For each data set, theses features describe the structures and the sequences. About 3 of our features are selected (except one human set), among 11 original features selected among other features validated in literature.

Among the literature's features selected, MFE and delta-G are always selected, known for being very important features in pre-miRNA classification. Also, 6 features from Triplet-SVM are also selected, as many other features. Moreover, four features are selected in all of the data sets: ratio of the number of matching nucleotides from the hairpin, ratio of the number of nucleotides in the assymetric loops or bulges in the hairpin, MFE ajusted and delta-G ajusted. This could permit, with further researches like tests on machine learning algorithms, to validate our new features. Finally, the combinations of species in the data sets permitted us to observe different selections depending on the species. In the data set composed by different species, very few features are on common with mouse and human, when a set of human + mouse results in a combination of about all of the features from human and mouse sets separated. This indicates that a consensus could exist between species in a same group (mammals, insects) but that these groups do not share the same characteristics.

	human	mouse	human + mouse	combination
set 1	14(4)	16(1)	N/A	15(4)
set 2	10(1)	13(3)	15(2)	N/A
set 3	19(0)	18(2)	N/A	10(1)

Figure 1. Features selected for each set. Numbers between parenthesis correspond to the number of our features that are selected.

Acknowledgements

This work was supported by the Council of Essonne Region (Pôle System@tic, OpenGPU project)

References

[1] D. Bartel, MicroRNAs: genomics, biogenesis, mechanism and function. *Cell*, 116:281-197, 2004.

[2] R. Batuwita and V. Palade, microPred: effective classification of pre-miRNAs for human miRNA gene prediction. *Bioinformatics*, 25:989-995, 2009.

[3] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, IH. Witten, The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11:1, 2009.

[4] P. Jiang, H. Wu, W. Wang, W. Ma, X. Sun and Z. Lu, MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Res.*, 35:W339-344, 2007.

[5] LP. Lim, ME. Glasner, S. Yekta, CB. Burge, DP. Bartel, Vertebrate MicroRNA Genes. *Science*, 299:1540, 2003.

[6] 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.

[7] Y. Wang, X. Chen, W. Jiang, L. Li, W. Li, L. Yang, M. Liao, B. Lian, Y. Lv, S. Wang, S. Wang, X. Li, Predicting human microRNA precursors based on an optimized feature subset generated by GA-SVM. *Genomics*, 98:73-8, 2011.

[8] Y. Xu, X. Zhou, W. Zhang, MicroRNA prediction with a novel ranking algorithm based on random walls. *Bioinformatics*, 24:i50-i58, 2008.

[9] C. Xue, F. Li, T. He, G.P. Liu, Y. Li, X. Zhang, Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics*, 6:310, 2005.

[10] S. Tempel, F. Tahi, A fast ab-initio method for predicting miRNA precursors in genomes. *Nucleic Acids Res.*, 2012.

[11] S. Griffiths-Jones, H.K. Saini, S. van Dongen and A.J. Enright, miRBase: tools for microRNA genomics. *Nucleic Acids Res.*, 36:D154-D158, 2008.

Network based approach for robust transcriptomic data analysis

Rémy NICOLLE[1,2], Mohamed ELATI[1] and François RADVANYI[2]

[1] iSSB, Institute of Systems and Synthetic Biology, CNRS UPS3509, University of Evry-Val-d'Essonne EA4527, Genopole Campus 1, Genavenir 6, 5 rue Henri Desbruères, 91030 Evry, France
remy.nicolle, mohamed.elati@issb.genopole.fr
[2] Institut Curie, CNRS UMR 144, 26 rue d'Ulm, 75248 Paris, cedex 05, France
francois.radvanyi@curie.fr

Abstract Our project aims at analysing transcriptomic data through the influence of master regulators. This methodology can be used to greatly reduce the number of features in gene expression data sets and making data analysis much more reproducible across data sets.

Keywords gene regulation, gene expression data, network inference, feature selection, dimension reduction

Digital representation of embryonic development

ANISEED system and its application

Cyril MARTIN[1], Delphine DAUGA[1] and Patrick LEMAIRE[1]

[1] Centre de Recherche de Biochimie Macromoléculaire (CRBM), UMR5237 CNRS, 1919 route de Mende, 34293, Montpellier, Cedex 5, France

cyril.martin@crbm.cnrs.fr

Abstract *ANISEED (Ascidian Network for In situ Expression and Embryological Data) is a generic system designed to offer a representation of embryonic development at the level of the genome (cis-regulatory sequences, gene expression, protein annotation), of the cell (morphology, fate, induction, lineage) or of the whole embryo (anatomy, morphogenesis).*
An additional module called 3D Virtual Embryo allows to manipulate 3D embryo models and permits to quantitatively describe the shapes and arrangement of cells. Combined with ANISEED, this module can be used to relate this quantitative description of cell shape and behaviour with molecular or embryological information. http://www.aniseed.cnrs.fr

Keywords database, data integration, gene expression, annotations, curation, ascidians.

1 Introduction

To understand how cell behaviours are individually executed and subsequently integrated to generate a final morphological structure, molecular, cellular and embryological data must be confronted. To bridge the gap between these different types of information, we developed a generic open source software named "3D Virtual Embryo" integrated as a module of an advanced model organism database, ANISEED [1]. This software allows to manipulate 3D embryo models and permits to quantitatively describe the shapes and arrangement of cells. Moreover, the system offers a large number of tools to query the database, to curate and annotate data and to manage the database. The integration of three-dimensional (3D) embryonic models with the arsenal of molecular and expression tools present in ANISEED discloses novel possibilities such as the understanding of the regulatory networks, the prediction *in silico* of the consequences of rearrangements and ablations of individual cells on the global embryonic landscape [2].

2 Results

2.1 The ascidian embryo: a simplified model of chordate development and evolution

Ascidians, which are marine invertebrate chordates, are model organisms of choice because their embryos develop with a small number of cells and an invariant lineage, allowing their study a cellular level of resolution. Moreover, ascidians display a larval body plan similar to that of vertebrates suggesting that the study of these simple embryos will shed light on the more complex vertebrates.

2.2 General architecture of the system

The ANISEED system is composed of a database (using Chado [3], a relational database schema), interfaced to both a classical web interface, and the 3D Virtual Embryo module [4]. In addition to these mining tools, a manager tool facilitates the creation of parallel databases for additional model organisms, manages users, and centralises updating and bulk upload scripts. Finally, the curator website permits validation or amendment of submitted data. The database itself is organised into five main parts, each corresponding to a given view of development: anatomy, fate specification processes, molecules, functional gene annotation, and gene expression data.

2.3 Resources in ANISEED

The system represents developmental processes at the anatomical, embryological and molecular levels. Features include a rich description of the anatomy of developing embryos including geometrical data from 3D embryo models, a functional annotation pipeline for molecular data, and powerful flexible mining interfaces. The annotation pipeline forms the basis for the Gene Ontology classifications of the gene model present in ANISEED. InterProScan, InParanoid and PsiBlast were performed to allow this classification. ANISEED hosts over 4,200 anatomical entities, 30,000 *in situ* hybridisation profiles, 12,000 functionally annotated genes, 1.3 million of ESTs and 410,000 cDNA. The majority of patterns and images come from the "Ghost" database, created by the Satoh laboratory (Kyoto University; http://ghost.zool.kyoto-u.ac.jp/indexr1.html) [5], while the data for Halocynthia embryos come from the MAGEST consortium (http://www.genome.jp/magest/).

2.4 ANISEED tools

The web interface of ANISEED contains a useful set of powerful tools. The set of search tools allows users to search gene models or clone by biological name, ID, Gene Ontology terms or InterPro terms. The visitors can also search for *in situ* gene expression patterns in *Ciona intestinalis* and *Halocynthia roretzi* embryos at different stages of development, as well as in Ciona juveniles. Another interesting option called "STEPBlast" allows users showing similar or antisimilar expression profiles.

Users can use the traditional BLAST. A GBrowse-based genome browser is cross linked to the database and allows aligning various information on the genome. The visitors can also explore all the literature present in the database and retrieve data from a given article.

Finally, the 3D virtual Embryo module uses interactive three dimensional digital embryo reconstructions to display information from the database as well as to enter into the database quantitative descriptions of the geometry of individual embryonic territories and their interactions.

3 Conclusion

I presented a brief description of the strategy that presided over the design of a generic digital environment for the representation of embryonic development. The system offers several important novelties.

First it supports a rich anatomical representation of the embryo anatomy that includes both neighbourhood relationships and a set of mathematical descriptors of the shape of each tissue or cell. Second, it allows to represent embryological concepts such as inductions and cell autonomous process. Finally, it allows a large panel of tools to query, to curate and to manage data in ANISEED.

References

[1] Tassy, O., Dauga, D., Daian, F., Sobral, D., Robin, F., Khoueiry, P., Salgado, D., Fox, V., Caillol, D., Schiappa, R., Laporte, B., Rios, A., Luxardi, G., Kusakabe, T., Joly, J. S., Darras, S., Christ, The ANISEED database: digital representation, formalization, and elucidation of a chordate developmental program. *Genome Res*, 20:1459-68, 2010.

[2] A. Di Gregorio, AK Hadjantonakis, The multidimensionality of cell behaviors underlying morphogenesis: a case study in ascidians. *Bioessays.* 2006 Sep;28(9):874-9.

[3] Christopher J. Mungall, David B. Emmert and The FlyBase Consortium, A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics* (2007) 23 (13): i337-i346.

[4] O. Tassy, F. Daian, C. Hudson, V. Bertrand, P. Lemaire, A quantitative approach to the study of cell shapes and interactions during early chordate embryogenesis. *Curr Biol.* 2006 Feb 21;16(4):345-58.

[5] KS. Imai, K. Hino, K. Yagi, N. Satoh, Y. Satou . Gene expression profiles of transcription factors and signaling molecules in the ascidian embryo: towards a comprehensive understanding of gene networks. *Development*. 2004 Aug;131(16):4047-58. Epub 2004 Jul 21.

Prediction of Amyloidogenic Motifs in Human Proteins

Mathieu ÉMILY[1,2], Claire SCHIRMER[3], Antoine JAN[3], Anthony TALVAS[1,3], Cyrille GARNIER[3] and Christian DELAMARCHE[3]

[1] IRMAR, UMR6625 CNRS, campus Beaulieu, Bat 13, 35042, Rennes Cedex, France
[2] Université Rennes 2, Place du recteur Henri Le Moal, 35043, Rennes, France
mathieu.Emily@univ-rennes2.fr, antony.talvas@irisa.fr

[3] IGDR, Université de Rennes 1, UMR6290 CNRS, campus Beaulieu, Bat 13, 35042, Rennes Cedex, France
claire.schirmer@univ-rennes1.fr, antoine.jan@etudiant.univ-rennes1.fr,
cyrille.garnier@univ-rennes1.fr, christian.delamarche@univ-rennes1.fr

Keywords Protein motif discovery, logistic regression, conformational diseases, Alzheimer's disease, Amyloid Code.

Biological Background

Conformational diseases are a major group of human diseases. Misfolded proteins or proteolytic fragments self-associate into highly ordered fibers (amyloid fibrils) forming toxic deposits in a variety of organs. The accumulation of amyloid aggregates has been identified in some thirty different pathologies: Alzheimer's, Huntington's, amyotrophic lateral sclerosis, type II diabetes, prion diseases, medullary thyroid cancer, systemic amyloidosis of kidney and heart, etc [1]. Although native proteins vary in sequence, function and 3D structure, amyloid fibrils share a common property at the molecular level which consists in stacks of parallel or anti-parallel beta-sheets named protofilaments. Then, the protofilaments twist around a central axis, grow to form mature fibrils of cross-beta structure, with diameters of 10-15 nm and up to several microns long. Many experimental studies show that the formation of beta-sheets involves protein patterns, as short as six amino acids, often called "hot spots". These short segments are not necessarily in that structural form in native proteins, but may result from conformational changes during the misfolding processes. Hot spots have clearly no known consensus sequence, except for some physico-chemical properties of the constituent amino acids. Hence discovering of the "Amyloid Code" that governs the composition of hot spots is crucial in understanding the underlying cause of many human diseases.

Prediction of Hot Spots

The last decade has led to the implementation of multiple methods combining various biological properties to predict hot spots in proteins. The large number of published algorithms reflects that biological mechanisms involved in amyloid fibrils formation are complex. Moreover, performance evaluation of tools is questionable, in part because experimental knowledge is biased by the over-representation of information coming from a limited number of model peptides. Given this situation, it is likely that some methods are complementary and that a proper combination of several programs might improve the global predicting performances. Consequently, we recently developed a meta-predictor based on a logistic regression model. Our program, called MetAmyl for METa-predictor for AMYLoid proteins, aims to combine eleven predictors by searching the best linear combination of their individual amyloid scores [2]. As a result four methods were statistically selected (Salsa [3], Pafig [4], FoldAmyloid [5] and Waltz [6]).

We computed the score of the 64,000,000 possible hexapeptides for the four selected methods, and the values were stored in tables of a database. Thus, building the profile of an input sequence consists in querying the database instead of calculated the score of each hexapeptide for each method.

Amyloidogenic Motifs in Human Proteins

The identification of amyloid motifs in human proteins is very important for the development of therapeutic strategies. For instance, rational design of molecules directed against hot spots could inhibit amyloid formation [7].

We scanned the human proteome with two objectives:

- Identification of hot spots in a set of 69 human proteins known to form amyloid [8], then compositional analysis of these segments.
- Identification of hot spots in the complete proteome to predict new amyloid proteins, to evaluate the distribution of hexapeptides among the 64,000,000 possibilities, etc.

Poster Content

Here we succinctly describe evaluation of MetAmyl on a data set including known amyloid and non-amyloid peptides. Compared to other methods MetAmyl is proved to be more sensitive and more specific. Moreover, MetAmyl, available at http://metamyl.genouest.org/, is suitable for large-scale analysis.

Some statistical results on the human proteome are shown. For instance, it was recently reported that the signal sequences, when present, tend to be detected as amyloidogenic regions [9]. An exhaustive analysis of human proteins containing a signal sequence confirms this observation with MetAmyl. This surprising result could be due to the amino-acid composition of the signal sequence. However, another hypothesis involved in the molecular mechanism of maturation by chaperones or signal peptidases.

Please do not hesitate to contact us for questions and/or collaborations.

Acknowledgements

Claire Schirmer is supported by a grant from the Région Bretagne.

References

[1] F. Chiti F and C.M. Dobson, Protein misfolding, functional amyloid, and human disease. *Annu Rev Biochem.*, 75:333-366, 2006.

[2] M. Emily, A. Talvas and C. Delamarche, MetAmyl: a new METa-predictor for AMYLoid proteins. *Submitted.*

[3] 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.

[4] J. Tian, N. Wu, J. Guo and Y. Fan, Prediction of amyloid fibril-forming segments based on a support vector machine. *BMC Bioinformatics*, 10:S45, 2009.

[5] 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, 2009.

[6] 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.

[7] T. Härd and C. Lendel, Inhibition of Amyloid Formation. *J Mol Biol. In press*, 2012.

[8] S. Pawlicki, A. Le Béchec and C. Delamarche, AMYPdb: a database dedicated to amyloid precursor proteins. *BMC Bioinformatics.* 10:273-284, 2008.

[9] K. Frousios, V. Iconomidou, C.M. Karletidi and S. Hamodrakas, Amyloidogenic determinants are usually not buried. *BMC Structural Biology*, 9: 44-53, 2009.

De novo comparison of metagenomic data.

A new tool: Compareads

Nicolas MAILLET[1], Claire LEMAITRE[1], Rayan CHIKHI[2], Dominique LAVENIER[1] and Pierre PETERLONGO[1]

[1] INRIA Rennes - Bretagne Atlantique/IRISA, EPI GenScale (Symbiose), Campus universitaire de Beaulieu, 35042, RENNES Cedex, France
{nicolas.maillet, claire.lemaitre, dominique.lavenier, pierre.peterlongo}@inria.fr

[2] ENS Cachan/IRISA, EPI GenScale (Symbiose), UMR 6074 CNRS, Campus universitaire de Beaulieu, 35042, RENNES Cedex, France
rayan.chikhi@irisa.fr

Keywords Comparative metagenomics, Next-generation sequencing, Bloom filter

Comparaison *de novo* de données métagénomique

Un nouvel outil : Compareads

Mots-clés Métagénomique comparative, Séquençage de nouvelle génération, filtre de Bloom

La métagénomique étudie l'ensemble des génomes, ainsi que leurs interactions, au sein d'un même biotope. Une très grande proportion des micro-organismes n'est pas cultivable en milieu contrôlé (plus de 99,9% dans l'eau de mer[1]). C'est pourquoi la métagénomique va consister à séquencer tous les génomes présents dans un échantillon sans passer par une phase de culture ou de différenciation des espèces en amont. À l'issu d'un séquençage, on récupère des centaines de millions de fragments (des *lectures*, ou *reads*) qui représentent chacun une fraction de l'ADN des divers organismes présents dans cet échantillon. Cette discipline récente, supportée par l'évolution rapide des technologies de séquençage à haut débit, amène naturellement de nouvelles problématiques comme, par exemple, la mesure de similarité entre deux métagénomes. Une manière d'évaluer cette similarité peut être de compter le nombre de *reads* communs entre deux métagénomes. Deux échantillons contenant des espèces différentes partageront peu de *reads* en commun. À l'inverse, deux métagénomes de compositions identiques auront un grand nombre de *reads* similaires. Cette mesure approximative ne reflète probablement pas une réalité effective, elle permet simplement d'établir une mesure de similarité relative entre plusieurs métagénomes. La mise en œuvre de cette mesure de similarité métagénomique n'est cependant pas immédiate. Une comparaison classique de tous les *reads* d'un échantillon A contre ceux d'un échantillon B représente un volume de calcul très conséquent. À titre d'exemple, deux *runs* de type Illumina de 2×10^8 *reads* impliqueraient 4×10^{16} comparaisons individuelles de *reads*. Avec une hypothèse (optimiste) de 100ns pour comparer deux *reads*, le calcul prendrait 4×10^9 secondes, soit 15 ans sur un processeur actuel (8 cœurs).

Notre méthode pour comparer deux métagénomes repose d'abord sur une manière simple et rapide d'exprimer une similarité entre deux *reads*. Deux *reads* sont considérés comme similaires s'ils partagent au moins P mots non chevauchant de K caractères. Le principe de la comparaison de deux métagénomes A et B s'effectue alors de la manière suivante :

– 1. Tous les mots de K caractères du métagénome A sont indexés. La structure de données utilisée stocke en mémoire tous les mots différents de K caractères présents dans le jeu de données A. L'interrogation de cette structure de données avec un mot quelconque de K caractères indique rapidement si ce mot y est présent ou non.

– 2. Pour chaque *read* du métagénome B, on test la présence de tous ses mots non chevauchant de K caractères dans la structure de données (*i. e.* le métagénome A). Si au moins P mots sont trouvés, le *read*

est retenu et stocké dans un ensemble $B*$. À la fin de cette étape, l'ensemble $B*$ représente donc tous les *reads* de B qui ont au moins une occurrence dans A.

– 3. On répète l'étape 1., en utilisant cette fois-ci le métagénome B.

– 4. On répète l'étape 2., en utilisant cette fois-ci le métagénome A. Les *reads* retenus sont mis dans l'ensemble $A*$.

Pour éviter des comparaisons inutiles, optimiser le processus et réduire encore les temps de calcul, l'étape 3. indexe uniquement l'ensemble $B*$. En effet, si un *read* de B n'est pas présent dans A, il est inutile d'effectuer des tests de comparaison avec ce *read* à l'étape 4..
Le résultat de la comparaison est un couple de nombres (X, Y) où X est le cardinal de $A*$ et Y le cardinal de $B*$. X représente donc le nombre de *reads* de A ayant au moins une occurrence dans B, et Y, le nombre de *reads* de B ayant au moins une occurrence dans A. Dans la structure d'indexation mise en place, l'information qui relie mots et *reads* est perdue. Ainsi, aux étapes 2. et 4., on teste simplement l'appartenance d'un mot à un métagénome, ce qui génère des faux-positifs. Si pour un *read* donné de A, au moins P mots qui le composent appartiennent au métagénome B, rien n'indique que ces mots appartiennent au **même** *read* de B. En pratique, pour des mots suffisamment longs, le nombre de faux-positifs reste faible.
Dans cette approche, la structure d'indexation est centrale. On doit représenter en un minimum d'espace mémoire un très grand nombre de mots de K caractères (plusieurs milliards). De plus, cette structure est extrêmement sollicitée et doit donc être rapide. Notre implémentation repose sur un index probabiliste basé sur le concept de filtre de Bloom[2]. L'inconvénient de cette structure probabiliste est qu'elle génère également des faux-positifs.

Le logiciel Compareads est une première implémentation (mono-cœur) réalisée en C pour valider cette approche. La comparaison de deux métagénomes comprenant 10^8 *reads* chacun dure environ 5 heures. De nombreuses optimisations, dont de la parallélisation sur multi-cœurs, sont encore possibles pour réduire significativement ce temps.
Une validation fonctionnelle portant sur 15 métagénomes, représentant 3 conditions différentes d'un même milieu, a été conduite afin de vérifier que cette mesure de similarité permet rapidement de positionner des métagénomes les uns par rapport aux autres. Le calcul des similarités deux à deux des 15 métagénomes (105 comparaisons) a ainsi permis de retrouver aisément la hiérarchisation relative aux 3 conditions initiales.
Une seconde validation a consisté à filtrer 4 jeux de données métagénomique fortement contaminés par des séquences d'ADN humain. Chaque échantillon a été comparé au génome humain afin d'en éliminer les séquences similaires. Les résultats des 4 intersections sont venus confirmer de précédentes études réalisées à l'aide de BLAST[3] et de MG-RAST[4] en terme de pourcentage de contamination.

La méthode mise en pratique dans le logiciel Compareads a plusieurs avantages. Premièrement, cette méthode permet de donner un score de similarité entre plusieurs métagénomes. Deuxièmement, cette méthode peut servir de filtre numérique sur des données métagénomique, afin par exemple de retirer une espèce d'un métagénome. Enfin, la consommation mémoire n'est pas dépendante de la quantité de données à traiter, quantité qui peut être très fluctuante suivant les technologies de séquençage utilisées.

Références

[1] R. I. Amann, W. Ludwig and K. H. Schleifer, Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol. Rev.*, 59 :143-169, 1995.

[2] B. H. Bloom, Space/time trade-offs in hash coding with allowable errors *Communications of the ACM*, 13 :422-426, 1970.

[3] S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman, Basic local alignment search tool *J. Mol. Biol.*, 215 :403-410, 1990.

[4] F. Meyer, D. Paarmann, M. D'Souza, R. Olson, E. M. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening and R. A. Edwards, The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes *BMC Bioinformatics*, 9 :386, 2008.

A new insight in DNA topoisomerases classification

Large Scale Classification Using an Alignment-free Technology

Damien Correia[1], Florence Vogliolo[1], Hélène Debat[2], Marc Nadal[2], Vladimir Daric[2] , Claire Kuchly[2] and Claudine Devauchelle[1]

[1] LABORATOIRE STATISTIQUE ET GENOME, UMR8071 CNRS, IBGBI, 23 bvd de France, 91000, Evry, France
claudine.devauchelle@genopole.cnrs.fr

[2] INSTITUT DE GENETIQUE ET MICROBIOLOGIE, UMR8621 CNRS, Bât 400 et 409, 15 rue Georges Clémenceau, 91405 Orsay Cedex, France
{helene.debat, marc.nadal, vladimir.daric, claire.kuchly}@igmors.u-psud.fr

Keywords DNA topoisomerases IA classification, comparative genomics, microbiology.

1 Introduction

Les ADN topoisomérases de type IA sont des enzymes uniquitaires présentes dans l'ensemble du règne vivant. Elles ont un rôle fonctionnel très important puisqu'elles participent au maintien de l'ADN dans une topologie correcte lors de tous les processus cellulaires qui déplacent des machineries le long de l'ADN (transcription, traduction, réplication, recombinaison...). S'il existe plusieurs types d'ADN topoisomérases, seules les topoisomérases de type IA sont ubiquitaires présentant un intérêt majeur en terme d' évolution.

Nous nous sommes intéressés plus particulièrement aux topoisomérases IA des bactéries et des archées. Nous avons donc développé un filtre de détection automatique de cette famille d'enzymes dans les génomes bactériens et archéens. Nous avons ensuite classé ces enzymes - en y rajoutant les topoisomérases eucaryotes - afin de déterminer les sous-familles existantes. Nous avons pour cela utilisés des méthodes de comparaisons sans alignement qui sont compatibles avec des analyses à large échelle.

Nous présenterons les résultats obtenus par notre filtre automatique de détection sur 1937 séquences génomes bactériens et 123 génomes d'archées, nous comparerons ces résultats avec ceux dont les protéomes sont annotés. Nous présenterons également la base de données répertoriant les ADN topoisomérases de type IA détectées et expertisées ainsi que leur classification. Cette base de données TopoDB sera prochainement mise à disposition de la communauté des microbiologistes (et des annotateurs) sur la plateforme e-bio.

2 Automatic detection of Type IA Topoisomerases

Les études évolutives portant sur la famille des ADN topoisomérases de type IA nécessitent un examen exhaustif de toutes les copies de ce gène dans les génomes complets. Nous avons donc développé un flitre automatique qui permet de détecter ces enzymes et nous avons comparés nos résultats aux annotations actuellement disponibles pour les protéomes bactériens et archéens.

Le filtre bioinformatique que nous avons développé consiste à analyser systématiquement les génomes de bactéries et d'archées disponibles à l'aide du logiciel blastn en utilisant un jeu de topoisomérases de références. Les résultats de blastn sont ensuite parsés de manière à retrouver les zones du génome d'intérêt qui ont matché avec les séquences de références, en autorisant des insertions/délétions et pour une zone de recouvrement maximale. Nous avons testé plusieurs valeurs des paramètres de notre filtre : taux de recouvrement, distance minimum entre 2 HSP pouvant être fusionnés, E-value maximale pour qu'un HSP soit considéré. Nous présenterons les taux de FP et de FN obtenus par notre filtre en les comparant aux annotations des protéomes.

3 Automatic classification of type IA topoisomerase

La classification d'un grand nombre de séquences n'est pas sans poser de problèmes. Non seulement pour des raisons de temps de calcul, de mémoire et de représentation des résultats, mais également du fait de la dégradation des alignements multiples sur lesquels reposent la majorité des méthodes de comparaison de séquences lorsque le nombre de séquences inclus dans ces alignements est grand. Nous avons fait le choix délibérés d'utiliser des méthodes sans alignement qui reposent sur des similarités locales pour organiser la connaissance sur la famille des DNA topoisomérases IA. En effet certaines sous-familles (les gyrases inverses) possèdent une grande extension N-terminale alors que d'autrtes séquences ont des intéines (introns protéiques) qui "trompent" certains programmes d'alignement multiples. La méthode que nous utilisons au laboratoire a été développé initialement par Gilles Didier [1] et a été utilisé avec succès pour sous-typer les génomes rétroviraux [2,4] ainsi que les ADN topoisomérases [3].

La classification des topoisomérases de type IA de 93 organismes, pour un total de 410 séquences par la méthode MS4 [4] montre l'organisation de cette super-famille d'enzymes en 5 sous-classes : les Topo III eucaryotes, lesTopo III archées, les Topo I bactériennes, les Topo III bactériennes et les gyrases inverses. Ces résultats ont été précédement décrits pas Forterre *et al.* [5] sur des jeux de données choisis et comprenant peu de séquences. Nous présentons ici une analyse exhaustive et nous discuterons des séquences difficilement classables.

4 Conclusion

En conclusion le filtre bioinformatique que nous avons développé permet de détecter semi-automatiquement les topoisomérases IA des génomes complets de bactéries et d'archées. L'organisation en sous-familles ainsi que l'expertise des séquences détectées sera mise à disposition des microbiologistes et des annotateurs dans une base de données *TopoDB* developpée en utilisant le framawork web Django et respectant l'architecture REST[6]. *TopoDB* sera prochainement disponible sur la plateforme e-bio de l'IGM à Orsay. La mise à jour de cette base sera facilitée par le filtre bioinformatique qui présente l'avantage de mettre l'accent sur les faux négatifs qui peuvent être ensuite expertisés par les microbiologistes responsables du Wiki associés à la base de données.

Acknowledgements

This work was supported by the Pres UniverSud.

References

[1] Didier, G. and Pupin, M. and Laprevotte, I and Hénaut, A., *Local decoding sequence and alignment-free comparison*, Journal of Computational Biology. Vol 13, pp 1465-1476, 2006

[2] Didier, G. and Debomy, L. and Pupin, M. and Zhang, M. and Grossmann, A. and Devauchelle, C. and Laprevotte, I., *Comparing sequences without alignments: application to HIV/SIV subtyping***.** BMC Bioinformatics Vol. 8 pp. 1., 2007.

[3] Corel, E. and El Feghali, R. and Gérardin, F. and Hoebeke, M. and Nadal, M. and Grossmann, A. and Devauchelle, C. *Local Similarities and Clustering of Biological Sequences : New Insights from N-local Decoding***.** The First International Symposium on Optimization and Systems Biology. Vol. Lecture Notes in Operations Research No. 7 pp. 189-195, 2007

[4] Corel, E. and Pitschi, F. and Laprevotte, I. and Grasseau, G. and Didier, G. and Devauchelle, C , *MS4 - Multi-Scale Selector of Sequence Signatures: An alignment-free method for classification of biological sequences*. BMC Bioinformatics Vol. 11 pp. 406, 2011.

[5] Forterre P., and Gribaldo S., and Gadelle D., and Serre M.C., Or*igin and evolution of DNA topoisomerases* , Biochimie 89 (2007) 427-446.

[6] REpresentational State Transfer - http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm

Listes et index

Liste des conférences invitées

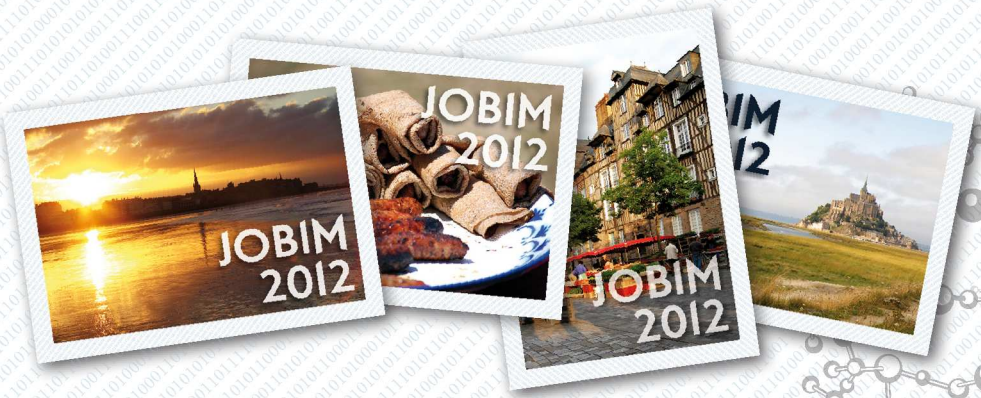
Liste des articles originaux

Liste des résumés étendus

Liste des posters

Liste des présentations industrielles

Liste des contributeurs



Jobim est le rendez-vous annuel de la communauté bio-informatique francophone, ouvert à toutes les personnes travaillant aux frontières de la biologie, de l'informatique, des mathématiques et de la physique.

La treizième édition de ces rencontres, Jobim 2012, est organisée à Rennes du 3 au 6 juillet 2012 par le centre Inria Rennes - Bretagne Atlantique et sous l'égide de la Société Française de Bio-Informatique (SFBI).

Cet ouvrage rassemble les articles des 33 communications orales et 112 affiches exposées, ainsi que les résumés des 7 conférences invitées et des 5 communications industrielles.

Avec le soutien de

GdR BiM · ReNaBi · Université de Rennes 1 · INRA · aviesan alliance nationale pour les sciences de la vie et de la santé · Région Bretagne · Métropole Rennes · Biogenouest · fondation Rennes action Michel Métivier · ministère éducation nationale enseignement supérieur recherche · ministère délégué à l'enseignement supérieur et à la recherche · Rennes · FASTERIS · panasas · NVIDIA · INGENUITY SYSTEMS · KoriLog BIOINFORMATICS solutions · Centre de Recherche Commun Inria Microsoft Research · Genostar Bioinformatics solutions

Éditeur : Inria Rennes - Bretagne Atlantique
ISBN-13 978-2-7261-1301-1