

Exploration du contexte génétique d'un gène

Tristan Colombo, Belaïd Benhamou, Gwennaele Fichant, Yves Quentin

LCB-CNRS, 31 Chemin Joseph Aiguier 13009 Marseille

IML-CNRS, 163 Avenue de Luminy 13009 Marseille

LSIS, 39 rue Joliot-Curie, 13013 Marseille

{colombo, fichant, quentin}@ibsm.cnrs-mrs.fr, benhamou@lim.univ-mrs.fr

Résumé : Des études ont montré que dans de nombreux génomes, des gènes étaient conservés dans un même voisinage et qu'ils encodent la plupart du temps des protéines en interaction physique directe. Ces conservations de gènes ne sont pas ordonnées mais sont soumises à des contraintes d'ordre physique - éloignement par rapport au gène d'ancrage, et d'ordre évolutif - relations de similarité entre gènes. De plus, contrairement aux travaux existants [Snel, 2000], nous considérerons que le sens de transcription des gènes est quelconque. Pour résoudre ce problème, nous avons développé un algorithme basé sur les TCSPs (Temporal Constraint Satisfaction Problems - [Dechter, 1991]). Nous avons appliqué notre algorithme sur des gènes codant des protéines de systèmes de transport ABC - impliqués dans les échanges avec l'environnement extérieur. L'étude du voisinage de tels systèmes devrait permettre de cerner plus précisément leur activité (nature des molécules - ou substrat - transporté).

Mots Clés : Bioinformatique, Intelligence Artificielle.

1. INTRODUCTION

De nos jours, de nombreux génomes ont été entièrement séquencés - près d'une centaine permettant ainsi de disposer des séquences protéiques. Ces protéines sont les produits des gènes et possèdent une fonction biologique. On a pu constater qu'elles agissent bien souvent en interaction avec d'autres protéines, formant des *réseaux d'interactions*.

En nous appuyant sur la théorie de l'évolution, on prend pour hypothèse que les protéines codées par des gènes issus d'un gène ancestral commun possèdent des fonctions identiques ou similaires. Cette relation évolutive, appelée *orthologie* [Fitch, 1970], est donc une propriété fondamentale pour l'obtention des prédictions fonctionnelles.

Ce concept est relativement simple à énoncer et à manipuler en théorie, mais il est beaucoup plus difficile à mettre en oeuvre. En effet, nous n'avons pas directement accès aux trajectoires évolutives des gènes. Elles doivent être inférées à partir des données de séquences dont nous disposons. Pour cela, nous

estimons la distance évolutive entre les séquences appartenant à deux génomes différents G_1 et G_2 et retenons uniquement les paires de protéines présentant les plus petites distances lors de la comparaison de G_1 avec G_2 ; elles constituent des paires de *plus proches voisins réciproques*. Elles sont alors agglomérées dans un processus hiérarchique.

Le but de notre méthode est d'explorer le voisinage d'un gène donné - que nous appellerons *gène d'ancrage* - pour inférer ses liens fonctionnels avec d'autres gènes, les gènes *conservés*. Nous prenons pour hypothèse que plus les gènes conservés seront loin du gène d'ancrage, plus leur lien fonctionnel avec le gène d'ancrage sera ténu ; plus le nombre de gènes sans relation entre deux gènes conservés sera grand, plus la cohérence de l'ensemble des gènes conservés sera faible. Cette analyse devra donc respecter des contraintes liées à nos hypothèses : un gène ne sera considéré comme conservé que jusqu'à une certaine distance du gène d'ancrage, et dans ce voisinage on ne tolérera pas d'espacement entre gènes conservés - ou *NGR* pour Nombre de Gènes sans Relation - supérieur à un seuil δ .

Pour coder ces contraintes nous utiliserons un formalisme dérivé des CSPs (Constraints Satisfaction Problems) [Montanari, 1974], et nous appliquerons notre méthode à l'étude du voisinage de gènes codant pour une famille particulière de systèmes biologiques : les transporteurs ABC.

2. RECHERCHE DU VOISINAGE CONSERVE

La recherche du voisinage conservé se fera de part et d'autre du gène d'ancrage, définissant une *fenêtre*. Pour une fenêtre donnée de η nous considérerons les η gènes précédents le gène d'ancrage et les η gènes suivants. D'après cette définition, la distance physique séparant un gène du gène d'ancrage peut s'exprimer en nombre de gènes intercalés. De plus, d'après nos hypothèses, le lien fonctionnel décroît alors que la distance entre gènes conservés s'accroît. Le seuil δ indique le nombre maximum de gènes non conservés entre deux gènes conservés. La recherche s'effectue par paires de génomes. A ce point, le problème peut être énoncé

comme suit : Soit A_R un gène d'ancrage issu du génome de référence R et A_N le gène d'ancrage orthologue de A_R dans le génome N . Notre but est de trouver la liste des paires de gènes (G_{Ri}, G_{Ni}) en relation conservés dans la proximité de la paire de gènes d'ancrage (A_R, A_N) où le nombre de gènes sans relation entre deux gènes conservés est inférieur au seuil δ . On peut observer que la liste des gènes conservés au seuil $\delta-1$ sont inclus dans la liste des gènes conservés au seuil δ . Ainsi, en partant du seuil fixé δ , on pourra en déduire la conservation à des seuils inférieurs.

.1 Méthode par intersections

Nous avons commencé par implémenter un algorithme simple, résolvant le problème par intersections successives. Lorsque la taille de la fenêtre est fixée, on peut considérer les gènes comme des listes dans chaque génome. En phase initiale, les listes incluent les gènes i) dont la distance au gène d'ancrage est inférieure à η , la taille de la fenêtre et ii) appartenant à des paires de gènes en relation. Dans les phases suivantes, les gènes de part et d'autre des gènes d'ancrage sont retirés des listes jusqu'à ce que le nombre de gènes sans relation intercalés entre deux gènes en relation soit inférieur au seuil δ . Pendant cette *phase de révision*, effectuée par intersections successives des listes, les paires de gènes en relation sont retirées dès que l'un des deux membres se situe en dehors des limites de révision (ie. $NGR > \delta$). Cette mise à jour des paires de gènes considérées peut être vue comme une *étape de propagation*. Si l'un des partenaires d'une paire retirée se situe dans les limites, alors son retrait peut augmenter le nombre de gènes sans relation entre deux gènes et le processus de révision doit être réappliqué. Ainsi, les deux procédures de révision et de propagation sont appliquées jusqu'à convergence (ie. d'une étape à la suivante les listes de gènes conservés restent identiques). Nous obtenons ainsi le sous-ensemble des paires de gènes conservées autour d'un gène d'ancrage dans les deux génomes pour une fenêtre η et un seuil δ fixés. La conservation est calculée depuis le seuil maximum $\gamma=\delta$ et nous utilisons la liste calculée à l'étape $\gamma=\omega$ pour calculer les gènes conservés au seuil $\gamma=\omega-1$ et ce jusqu'à $\gamma=0$.

.2 Méthode basée sur les contraintes

La méthode précédente n'était pas assez efficace pour pouvoir traiter les jeux de données très importants issus de la biologie. C'est pourquoi nous avons développé une seconde méthode basée sur des CSPs particuliers : les STPs (Simple Temporal Problems) [Dechter, 1991]. Basiquement, un CSP est défini par un ensemble de variables prenant leurs valeurs dans un ensemble de domaines; ces variables sont liées par un ensemble de contraintes. Ici, il s'agit d'un modèle de réseau de contraintes temporelles de type quantitatif où les variables représentent des points du temps et où les contraintes sont exprimées en tant que distance entre les variables. Ce modèle peut être transposé pour représenter des données dans un espace à une dimension - telles que des gènes ordonnés sur un chromosome.

Dans un tel modèle, les variables G_{Ri} sont les différents gènes de la fenêtre centrée sur le gène d'ancrage A_R . Les contraintes sont i) C_d , le nombre de gènes apparaissant entre chaque gène et le gène d'ancrage, et ii) C_g , le plus grand nombre de gènes sans relation apparaissant entre chaque gène et le gène d'ancrage. La contrainte C_d est utilisée pour localiser des paires de gènes sur les fenêtres, et la contrainte C_g est utilisée pour décider si une paire de gènes est conservée ou supprimée de la liste. Ces contraintes représentent donc des distances et prennent leurs valeurs dans \mathbb{N} .

Les contraintes

Ces contraintes, traduction des hypothèses de départ, peuvent être exprimées sous forme d'équations linéaires.

- Contrainte de distance C_d :

$G_{Ni} - A_N = d_{Ni}$ avec $-\eta \leq i \leq \eta$ où d_{Ni} est la position relative ($-\eta \leq d_{Ni} \leq \eta$) du gène N_i dans le génome N . Cette distance est exprimée en nombre de gènes se trouvant entre N_i et le gène d'ancrage A_N (on peut noter que ces gènes ne sont pas forcément des gènes formant des paires de gènes en relation). Pour une paire de gènes, cette contrainte est composée de deux équations (une pour chaque gène de chaque génome). Les d_{Ni} ne sont calculés qu'une seule fois lors de la phase d'initialisation pour une fenêtre donnée de η gènes.

- Contrainte du nombre de gènes intercalés sans relation C_g :

$G_{Ni} - A_N = g_i$. Cette contrainte signifie qu'entre le gène G_{Ni} du génome N et le gène d'ancrage A_N , le nombre maximal de gènes sans relation intercalés est de g_i . Cette contrainte est susceptible d'être recalculée dans une étape de révision. C_g exprime la distance entre les gènes G_{Ni} et A_N en ne considérant que les gènes consécutifs sans relation.

Sur la Figure 1 sont représentés les deux génomes R et N . Les différents gènes sont représentés par des rectangles ; une paire de gènes en relation - ou orthologues - adopte un même motif.

Ainsi, les deux gènes d'ancrage A_R et A_N sont ils représentés par un même rectangle noir, le gène A_N étant orthologue au gène A_R . Les contraintes seraient alors les suivantes :

$$C_{d1} : G_{R1} - A_R = -4$$

$$C_{g1} : G_{R1} - A_R = 2$$

$$C_{d2} : G_{R2} - A_R = -1$$

$$C_{g2} : G_{R2} - A_R = 0$$

$$C_{d3} : G_{N1} - A_R = -1$$

$$C_{g3} : G_{N1} - A_R = 0$$

$$C_{d4} : G_{N2} - A_R = 2$$

$$C_{g4} : G_{N2} - A_R = 1$$

...

Déroulement de l'algorithme

Dans la phase d'initialisation, les données sont codées sous forme de STP en utilisant les contraintes décrites ci-dessus. Pour un seuil δ donné, on commencera par calculer la conservation à $\gamma=\delta$ jusqu'à $\gamma=0$. A chaque étape $\gamma=\gamma-1$, on révisé les données avec les nouvelles

valeurs. De même que dans l'algorithme naïf, les contraintes sont révisées en retirant les gènes dont les partenaires sont situés en dehors des limites fixées par le seuil. Lorsque cela est nécessaire, la révision est propagée aux gènes se situant entre le gène supprimé et la fin de la fenêtre considérée. L'ensemble des contraintes est révisé jusqu'à convergence : lorsque le seuil $\gamma=0$ est atteint ou alors lorsque pour un seuil γ considéré la liste de conservation est vide - donc plus aucune conservation possible pour un seuil inférieur à γ .

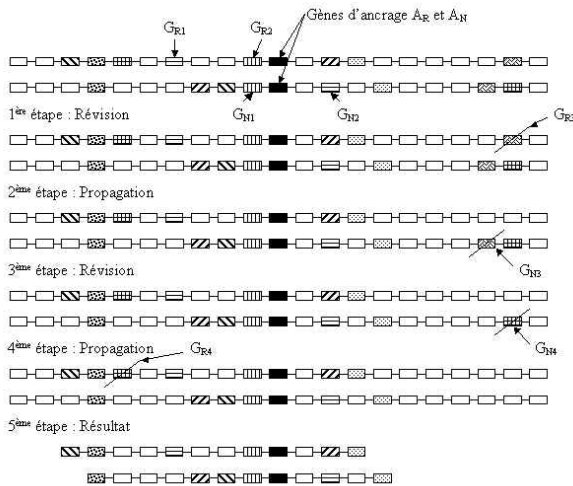


figure 1 : Recherche de voisinage conservé avec $\eta=10$ et $\delta=3$ - Etape 1 : Le nombre de gènes sans relation intercalés est trop grand (5 alors que le seuil est fixé à 3) : le gène G_{R3} est retiré de la liste des gènes conservés. - Etape 2 : Dans la première étape le gène G_{R3} brisait la contrainte C_g : son orthologue G_{N3} doit être supprimé de la liste. - Etape 3 : Le nombre de gènes sans relation intercalés est trop important après la suppression du gène G_{N3} lors de la 2ème étape : le gène G_{N4} est supprimé. - Etape 4 : Dans la 3ème étape le gène G_{N4} brisait la contrainte C_g : son orthologue G_{R4} doit être supprimé.

3 Complexité

Soit n le nombre de gènes étudiés.

- Dans la méthode par intersections, la phase d'initialisation est effectuée en n opérations et l'opération d'intersection entre deux listes de gènes est effectuée en $O(n^2)$ dans le pire des cas. Cet algorithme est donc en $O(n^2)$.

- Dans la méthode basée sur les contraintes, les opérations de révision et de propagation sont effectuées en $O(n)$ en considérant que tous les n gènes sont conservés à chaque itération. La complexité de cet algorithme est donc en $O(n)$ mais on peut souligner que en pratique n décroît très rapidement suivant γ .

Pour donner un ordre d'idée de la taille pratique de n , considérons que l'on veuille traiter 100 génomes (à l'heure actuelle il s'agit du nombre approximatif de génomes entièrement séquencés). Chacun de ces génomes contient en moyenne 3000 gènes. On obtient donc ici un n de 300 000. De plus, cette valeur

augmente très rapidement : en un an on est passé de 40 génomes entièrement séquencés à plus de 100.

4 Implémentation

Une interface web (Figure 3) a été développée en CGI Perl à partir de l'algorithme basé sur les contraintes. Cette interface utilise les données d'orthologie de la base COG [Tatusov, 2001] pour les relations intergénomiques. Les données issues de la base ABCdb [Quentin, 2002] sont utilisées pour la position des gènes sur le chromosome ainsi que leur fonction si cette dernière est connue. L'utilisateur doit désigner le nom du gène à étudier (ou le groupe d'orthologie COG auquel il appartient), ainsi que la taille de la fenêtre η et le seuil δ . Les résultats sont présentés de manière graphique avec la possibilité de restreindre la recherche sur certains génomes et d'obtenir une recherche automatique des gènes les plus représentatifs parmi les éléments conservés. L'utilisateur a également la possibilité de demander le tracé d'un dendrogramme des gènes conservés. Ce dendrogramme est tracé à l'aide des logiciels Neighbor - utilisant la "Neighbor Joining Method" [Saitou, 1987] de distances matricielles - et Drawgram du paquetage Phylip (<http://evolution.genetics.washington.edu/phylip.html>). Ainsi, l'utilisateur peut faire une première estimation sur la fonction globale et la proximité des gènes conservés.

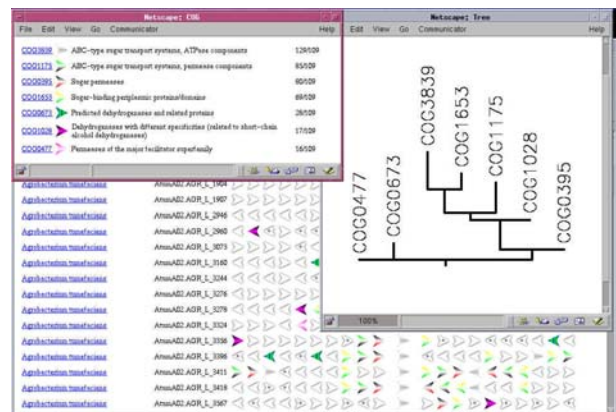


figure 3 : Exemple d'interface web.

Un autre programme a été développé dans le même but : le STRING web-server [Snel et al., 2000]. Ce serveur, à partir d'un gène donné, va rechercher les gènes conservés dans le voisinage mais au sein d'un même opéron (tous les gènes sont orientés dans le même sens sur le chromosome et sont très proches). Les résultats présentés par STRING sont en fait précalculés et stockés dans une base de données. Nous avons pu constater que dans le cas de certains gènes dont l'environnement était conservé mais sans contrainte d'orientation, les résultats fournis étaient trop incomplets. Ainsi pour le cas du gène *yclI* de *Bacillus subtilis*, connu pour être associé au gène *yclJ* au sein d'un système, est découvert par notre méthode alors qu'il ne l'est pas par STRING.

3. RESULTATS

Cette recherche des gènes conservés dans le voisinage d'un gène au cours de l'évolution a été appliquée à des systèmes intégrés (ensemble de protéines) particuliers : les transporteurs ABC qui sont des systèmes d'export ou d'import de molécules de diverses tailles - le *substrat*. L'analyse de la proximité de tels gènes peut révéler de nouveaux partenaires tels que des enzymes impliquées dans le métabolisme du substrat. Nous présentons ici les résultats obtenus en prenant comme gènes d'ancrage des gènes issus d'un groupe de transporteurs ABC impliqués dans le transport des oligosaccharides (famille *N5*). Comme le montre la Figure 2, l'analyse du voisinage de ces gènes nous permet de préciser la nature du substrat qu'ils transportent. Ainsi pour les gènes appartenant au groupe du gène d'ancrage *COG3839*, nous pouvons déduire de nos observations que les gènes conservés seraient impliqués dans des systèmes de transport du sucre.

4. CONCLUSION

Cet algorithme, rapide, permet d'effectuer des recherches de voisinages en s'affranchissant du génome de référence : tous les génomes possédant un gène orthologue au gène A_R seront considérés tour à tour comme génome de référence. De plus, l'interface développée, complète et simple d'utilisation, devrait fournir un puissant outil d'analyse à la communauté des biologistes.

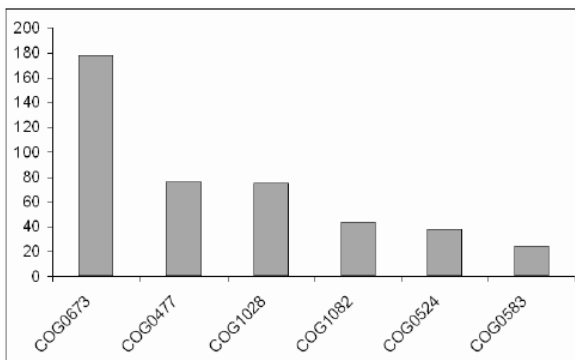


figure 2 : Recherche du voisinage des gènes du groupe COG3839 - Les fonctions des différents COG conservés sont : COG0673 Déhydrogénases - COG0477 Perméases - COG1028 Déhydrogénases - COG1082 Sugar phosphate isomerases / epimerases - COG0524 Sugar kinases, ribokinases - COG0583 Régulateur transcriptionnel

BIBLIOGRAPHIE

[Dechter, 1991] Dechter, R., Meiri, I., Pearl, J.: "Temporal Constraint Satisfaction Problems. Artificial Intelligence", Vol. 49, p. 61-95, (1991).
[Fitch, 1970] Fitch, W. M.: "Distinguishing homologous from analogous proteins", Syst. Zool., Vol. 19, p. 99-113, (1970).

[Montanari, 1974] Montanari, U.: "Networks of constraints : fundamental properties and application to picture processing", Information Sciences, Vol. 7, p. 95-132, (1974).

[Quentin, 2002] Quentin, Y., Chabalier, J., Fichant, G.: "Strategies for the identification, the assembly and the classification of integrated biological systems in completely sequenced genomes", Computers and Chemistry, (2002).

[Saitou, 1987] Saitou, Nei : "Neighbor Joining Method", Mol. Biol. Evol., Vol. 4, p. 406-425, (1987).

[Snel, 2000] Snel, B., Lehmann, G., Bork, P., Huynen, M. A.: "STRING : a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene", Nucleic Ac. Res., Vol. 28, p. 3442-3444, (2000).

[Tatusov, 2001] Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., Kiryutin, B., Galperin, M. Y.,

Fedorova N. D., Koonin, E. V.: "The COG database : new developments in phylogenetic classification of proteins from complete genomes", Nucleic Ac. Res., Vol. 29, p. 22-28, (2001).