ISYMOD+

(Projet de Recherche 2003)

Développement d'un système informatique de représentation et d'analyse des processus biologiques

http://www.lmgm.fr/Fichant/ISYMOD

CONTEXTE

Chez les bactéries, des processus moléculaires complexes sont impliqués dans la mise en place et le maintien de relations avec une grande diversité d'hôtes et, plus généralement, dans la colonisation d'une nouvelle niche écologique. Ces processus adaptatifs résultent de l'association de différentes protéines participant à un réseau complexe de relations, ces relations pouvant être d'ordre fonctionnel, physique, et évolutif. Si le séquençage de génomes complets et un premier niveau d'annotation donnent accès aux différents partenaires l'identification et la modélisation de leurs relations nécessitent le développement de nouvelles approches.

OBJECTIFS

Notre objectif est de construire un système informatique de représentation et d'analyse des processus biologiques impliqués dans l'adaptation des bactéries à leur environnement, par intégration des données issues du séquençage. Un premier travail réalisé en ce sens a mené au développement de la base de connaissances ISYMOD dédiée aux systèmes de transport ABC et aux systèmes de régulation à deux composants impliqués dans la transduction du signal.

Cette base, développée sous AROM, intègre connaissances du domaine et connaissances méthodologiques. Elle doit évoluer pour 1) accueillir les 3 types de relations évoqués et 2) faciliter leur analyse, via le développement d'un langage de description de propriétés algébriques de relations *n*-aires.

D'autre part, l'étape d'acquisition et de validation des données primaires étant le principal goulot d'étranglement, de nouveaux développements méthodologiques doivent être effectués afin d'accélérer l'ensemble du processus et d'automatiser les étapes d'apprentissage.

Parallèlement, le développement et l'évaluation de méthodes d'exploration des graphes modélisant les relations complexes entre objets est nécessaire.

Le système doit pouvoir intégrer automatiquement de nouveaux processus moléculaires sans nécessiter d'intervention humaine trop lourde, permettant ainsi de faire face à l'accroissement toujours plus rapide du flux de données.

PROJET

Amélioration et développement de méthodes d'acquisition et de classification des objets biologiques et de leurs relations.
Amélioration des méthodes d'analyse de la structure en sous-familles.
Extension du langage de représentation d'AROM.
Application à l'analyse des systèmes de transport chez les bactéries.

RESULTATS

- Nouvelle stratégie d'annotation hautement automatisée exploitant les ressources d'une grappe de PC. Généralisation à d'autres systèmes.
- Réalisation d'une plate-forme java pour la mise en œuvre d'algorithme de boosting d'alignements de séquences. Étude de la faisabilité avec comme classifieur faible BLASTP et extension à PsiBLAST.
- Etude de la relation « partie-tout » et de son comportement face à l'héritage et déduction d'une grammaire pour son expression dans AROM.
- Nouvelles méthodes de recherche de zones denses dans un graphe. Méthodes de simulation de graphes aléatoires et application aux partenaires des transporteurs ABC

CALENDRIER

2004-2005, C. Capponi temps partiel au LCB puis au LMGM (délégation CNRS).

2004 avril, réunion générale à Marseille, juin réunion de travail à Grenoble.

2004 régulièrement, discussions de travail à Marseille.

2005 juin, réunion à Toulouse, invitation de Pr.C. Froidevaux.

LABORATOIRES IMPLIQUES

LMGM/UMR5100/Toulouse
LIF/UMR6166/Marseille
Projet Helix/INRIA/Grenoble
IML/UMR6206/Marseille
LSR-IMAG/UMR5526/Saint Martin d'Hères

MOTS CLES

Base de connaissances, algorithmes d'apprentissage, boosting, algorithmes de classification, méthodes de partition de graphes, ABC transporteurs, génomique.

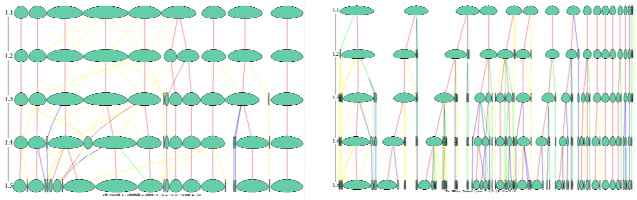
PUBLICATIONS PRINCIPALES

- J. Chabalier, C. Capponi, Y. Quentin, and G. Fichant. ISYMOD: a knowledge warehouse for the identification, assembly and analysis of bacterial integrated systems. Bioinformatics Advance Access published online on November 5, 2004.
- T. Colombo, A. Guénoche and Y. Quentin. Looking for high density areas in a graph. Application to orthologous genes. DAM submitted.
- C. Capponi, G. Fichant, Y. Quentin and F. Denis. Classification of Domains with Boosted Blast. ASMDA05, présenté a CAP et JOBIM 2005
- A. Guénoche. Comparing recent methods in graph partitionning, International Congress in Graph Theory, in press

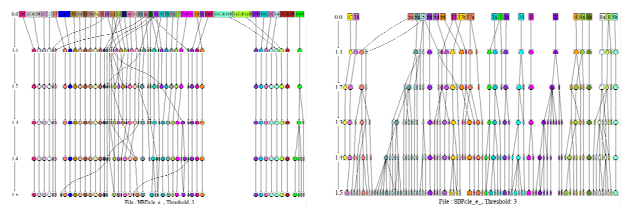
COMBIEN DE SOUS FAMILLES?

La présence de familles de gènes paralogues dans les génomes pose la question de l'existence ou pas d'une structuration en sous-familles. Généralement ce problème est abordé par l'utilisation de méthodes de reconstruction d'arbre évolutif. Néanmoins, ces méthodes se heurtent, d'une part au volume toujours croissant de données à analyser et d'autre part à la qualité des relations de similarité entre ces données. Une autre approche consiste à explorer le graphe de ces relations (les nœuds sont les protéines ou les gènes et les arêtes les relations d'homologie) à la recherche de zones particulièrement denses pouvant traduire la présence de sous-ensembles d'objets fortement liés.

Les figures suivantes illustrent ce type d'analyse avec les partitions obtenues, à des niveaux de plus en plus stringents, pour deux types de familles de protéines : les ATPases (à gauche) et les protéines affines (à droite) de transporteurs ABC.



L'analyse révèle une très forte structuration en sous-familles qui s'affine avec le durcissement du critère. Néanmoins, les deux familles présentent un comportement légèrement différent avec peu ou pas d'échanges entre les classes pour les protéines affines.



La superposition de ces partitions, avec les sous-familles précédemment décrites (code couleur), démontre la pertinence de l'approche. Nous pouvons cependant remarquer une plus faible résolution avec les ATPases, avec quelque échanges et de véritable éclatements avec certaines sous-familles de protéines affines.