BlastSets

(Projet de Recherche 2003)

Représentation et intégration des données biologiques

http://cbi.labri.fr/BlastSets/ BlastSets

CONTEXTE

Les fonctions cellulaires résultent de mécanismes moléculaires étudiés individuellement par combinaison d'approches expérimentales de type séquençage et post-séquençage. La compréhension des interactions entre ces mécanismes à l'échelle de la cellule nécessite des méthodes et des outils efficaces pour l'intégration de grandes quantités de données hautement hétérogènes. Ces données incluent les annotations fonctionnelles et structurales sur les séquences, les profils d'expression des gènes, les interactions moléculaires entre biomolécules, la littérature, etc. L'intégration de ces données permettra de progresser vers une vision globale du fonctionnement de la cellule.

OBJECTIFS

Le projet BlastSets porte sur la mise au point d'une nouvelle méthode permettant de rapprocher les informations hétérogènes disponibles à l'échelle des génomes (ou protéomes) entiers. Il s'agit d'une intégration généraliste qui vise à faire apparaître et à exploiter des correspondances nouvelles entre les données. Le principe fondamental de cette méthode réside en un parti pris dans la représentation des connaissances qui consiste à ramener systématiquement les connaissances au niveau des ensembles de séquences (gènes ou protéines). Il s'agit donc de se focaliser sur les relations entre entités biologiques plutôt que sur ces entités prises individuellement. La représentation des connaissances en ensembles de séquences partageant certaines propriétés biologiques permet, grâce à un modèle probabiliste, de comparer les ensembles produits. Ces comparaisons correspondent au recoupement et à la confrontation des données existantes.

La confrontation des données devrait permettre de :
- donner une indication sur le rôle d'un gène ou d'une protéine
- accélérer l'analyse de données provenant d'expériences haut-débit
- identifier des phénomènes biologiques à l'échelle des génomes

PROJET

Mise en valeur et exploitation des données publiques disponibles
Intégration des données sous forme de voisinages afin de pouvoir les confronter dans le but de :
- donner une indication sur le rôle d'un gène ou d'une protéine
- accélérer l'analyse de données provenant d'expériences haut-débit
- identifier des phénomènes biologiques à l'échelle des génomes

RÉSULTATS

• Nouvelle méthode pour la représentation et l'intégration des données de biologie moléculaire à l'échelle de la cellule [1,2]
• Mise en place de services Web pour utiliser le système BlastSets [3]
• Utilisation du système BlastSets pour la recherche de corrélations entre les interactions physiques des protéines et le niveau d'expression des gènes [4]
• Stratégie pour l'analyse et l'intégration des données d'expression [5]
• Génération efficace des ensembles pertinents à partir d'un voisinage pour leur comparaison à un ensemble donné [6]

CALENDRIER

2004	Janvier (Bordeaux) Démarrage
2004	Février (Bordeaux) Réunion générale
2004	Novembre (Bordeaux) Réunion générale
2005	Juin (Bordeaux) Réunion générale

MOTS CLÉS

Génomique fonctionnelle, intégration de données hétérogènes, analyse de données d'expression, génomique comparative.

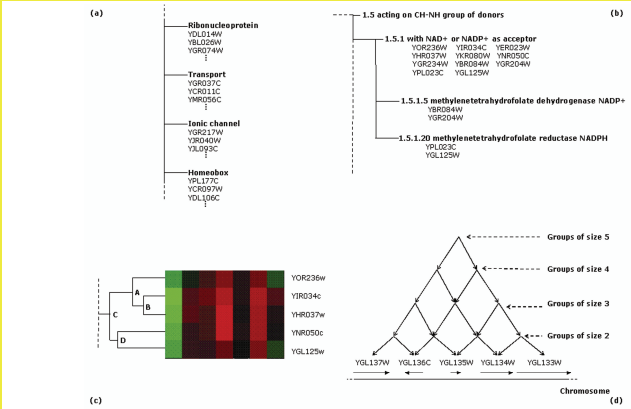
LABORATOIRES IMPLIQUÉS

Centre de Bioinformatique de Bordeaux (de Daruvar) / Bordeaux
Laboratoire Bordelais de Recherche en Informatique (Dutour) / UMR 5800 / Bordeaux
Laboratoire Statistique Mathématique et ses Applications (Poix) / EA 2961 / Bordeaux
Unité de Génétique des Génomes Bactériens (Danchin) de l'Institut Pasteur / Paris
UMR Génomique Développement Pouvoir Pathogène (Blanchard)/ IFR 103 / Bordeaux

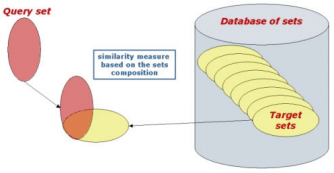
PUBLICATIONS PRINCIPALES

• [1] Barriot, R., Poix, J., Groppi, A., Barré, A., Goffard, A., Sherman, D., Dutour, I., de Daruvar, A., New strategy for the representation and the integration of biomolecular knowledge at a cellular scale, *Nucleic Acids Res.*, 32(12):3581-3589
• [2] Barriot, R., Poix, J., Gaugain, C., Dutour, I., de Daruvar, A., UK Integration of functional genomics data using sets of biological entities, *DBiBD 2005*, Edinburgh (UK).
• [3] Barriot, R., Lamiable, A., Web services and client packages as a framework for data mining on the neighborhood of organized sets of biological sequences, *ECCB-ISMB 2004,* Glasgow (UK).
• [4] Gaugain, C., Goffard, N., Groppi, A., Barriot, R., de Daruvar, A., Correlating expression profiles and physical interactions, *en cours de sélection*
• [5] Groppi, A., Stratégie pour l'analyse et l'intégration des données d'expression, *Journées RNG (réseau national des Génopoles) Transcriptome et bioinformatique 03/2005,* INRA site d'Auzeville (Toulouse, France).
• [6] Barriot, R., Dutour, I., Sherman, D., Efficient generation of pertinent target sets for the BlastSets system, *JOBIM 2005*, Lyon (France).

VOISINAGES ET RECHERCHE D'ENSEMBLES SIMILAIRES



Exemples d'intégration de données hétérogènes disponibles sur la levure (Saccharomyces cerevisiae) en voisinages (ensemble d'ensembles de séquences). En (a) chacun des mots-clés des annotations UniProt définit un ensemble de séquences. En (b) les codes d'enzymes de la nomenclature des enzymes sont utilisés pour créer les ensembles. En (c) les ensembles correspondent aux nœuds internes de l'arbre issu du clustering hiérarchique des profils d'expression des gènes. En (d) les ensembles correspondent aux gènes adjacents sur les chromosomes (paires, triplets, etc.). ■



Les ensembles de séquences appartenant aux différents voisinages sont stockés dans une base de données. La confrontation des données correspond à l'identification d'ensembles similaires : on distingue un ensemble requête, pour lequel on cherche à identifier des ensembles similaires dans la base de données d'ensembles cibles. L'ensemble requête est comparé à ceux de la base et une mesure de similarité entre deux ensembles permet d'identifier les ensembles cibles qui lui sont similaires.