Un système pour l'extraction de corrélations linéaires dans des données de génomique médicale

Arriel Benis*, Mélanie Courtine*

*LIM&Bio- Laboratoire d'Informatique Médicale et de Bioinformatique - E.A. 3969
Université Paris Nord, 74 rue Marcel Cachin, 93017 Bobigny Cedex, France
{benis,courtine}@limbio-paris13.org
http://www.limbio-pari13.org

1 Contexte et Problématique

L'aide à la découverte de biomarqueurs permettant le diagnostic et la prédiction dans le cadre de maladies complexes telles que l'Obésité ou le Cancer, repésente un enjeu important en terme de Santé Publique. Les outils telles que les puces à ADNc (Schena et al. (1995)) issues des recherches en Génomique Fonctionnelle permettent de fournir des données afin de contribuer à cet objectif.

Notre objectif est de découvrir des relations globales ou partiellement linéaires. Peu de travaux s'intéressent de manière spécifique à la découverte automatique de corrélations linéaires (Chiang et al. (2005)).

Nous proposons une méthode, nommée DISCOCLINI, afin de réaliser de manière automatique, avec ou sans *a priori*, l'exploration d'un grand nombre de relations entre des données numériques d'expression génique (quelques dizaines de milliers par individu) et biocliniques (quelques dizaines par individu), chaque relation est calculée pour au maximum quelques dizaines d'individus. Ainsi, ce système permet à l'expert en un temps réduit d'explorer un grand nombre de relations.

2 DISCOCLINI : Un flux d'aide à la découverte

DISCOCLINI consiste en un flux constitué de cinq grandes étapes :(1) définition des sources de données biocliniques et d'expression génique issues de puces à ADNc ; (2) extraction depuis les sources des données relatives aux individus des informations à inclure dans l'étude corrélationnelle ; (3) calculs sur les ensembles (3a) univariés définis précédemment et (3b) bivariés correspondant à la mise en relation d'un attribut issu de l'ensemble des données biocliniques et d'un attribut issu de l'ensemble des données d'expression génique ; (4) exploration visuelle des résultats des calculs sur les ensembles bivariés et sélection des relations potentiellement « intéressantes » ; (5) validation biologique de ces résultats par l'expert du domaine.

Ainsi, l'approche proposée avec textscDsicoClini est objective tant au niveau de l'analyse que de l'exploration des données, où aucun *a priori* en terme de connaissances dans le domaine

d'application, n'est requit. Dans DISCOCLINI, les relations sont des valeurs de corrélations non-paramétriques entre les deux types de données. Notre méthode permet donc à l'utilisateur de disposer de résultats d'analyse sous une forme synthétique et facilement exploitable : (1) un diagramme de Hasse regroupant les relations intéressantes au regard des seuils de valeurs statistiques définies automatiques ou pour l'utilisateur et (2) un tableau associant pour chaque relation des données statistiques et une représentation graphique « compacte ». Ce mode de restitution des résultats permet à l'expert de visualiser simultanément un ensemble de relations potentiellement intéressantes.

Différentes expérimentations ont permis de valider DiscoClini et de produire des résultats qui ont contribué à des avancées biomédicales dans le domaine de l'Obésité (Viguerie et al. (2004) ;Clément et al. (2004) ;Taleb et al. (2005)).

Les données de génomique Fonctionnelle que nous utilisons sont des données bruitées et lacunaires. Nous avons donc développé une approche pour détecter automatiquement les valeurs singulières dans les ensembles de données univariées et multivariées composés de peu d'individus. Cela permet d'améliorer la qualité des résultats communiqués à l'expert. Ces informations lui permettent d'accroitre ou de relativiser la confiance qu'il peut avoir dans les résultats et sur certaines potentielles découvertes.

Références

Chiang, R., C. Cecil, et E. Lim (2005). Linear correlation discovery in databases : a data mining approach. *Data and Knowledge Engineering 53*(3), 311–337.

Clément, K., N. Viguerie, C. Poitou, C. Carette, V. Pelloux, C. Curat, A. Sicard, S. Rome, A. Benis, J. Zucker, H. Vidal, M. Laville, G. Barsh, A. Basdevant, V. Stich, R. Cancello, et D. Langin (2004). Weight loss regulates inflammation-related genes in white adipose tissue of obese subjects. *FASEB J 18*(14), 1657–1669.

Schena, M., D. Shalon, R. Davis, et P. Brown (1995). Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science 270*(5235), 467–70. 0036-8075 Journal Article.

Taleb, S., D. Lacasa, J. Bastard, C. Poitou, R. Cancello, V. Pelloux, N. Viguerie, A. Benis, J. Zucker, J. Bouillot, C. Coussieu, A. Basdevant, D. Langin, et K. Clément (2005). Cathepsin s, a novel biomarker of adiposity : relevance to atherogenesis. *FASEB J 19*(11), 1540–2.

Viguerie, N., K. Clément, P. Barbe, M. Courtine, A. Benis, D. Larrouy, B. Hanczar, V. Pelloux, C. Poitou, Y. Khalfallah, G. S. Barsh, C. Thalamas, J. D. Zucker, et D. Langin (2004). In vivo epinephrine-mediated regulation of gene expression in human skeletal muscle. *J Clin Endocrinol Metab 89*(5), 2000–14. 0021-972x Journal Article Validation Studies.