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