DiscoClini : Une méthodologie pour l'extraction de relations linéaires dans des données de génomique médicale

Arriel Benis[1,2⋆], Jean-Daniel Zucker[1,2]

[1] LIM&Bio- Laboratoire d'Informatique Médicale et de Bioinformatique - E.A.3969
Université Paris Nord, 74 rue Marcel Cachin, 93017 Bobigny Cedex, France
{benis,zucker}@limbio-paris13.org et

[2] Nutriomique - INSERM U.755 et E.A. 3502 - Université Pierre et Marie Curie
Service de Nutrition, Hôtel-Dieu, 1 place du Parvis Notre-Dame, 75004 Paris, France

DiscoClini est une méthode qui permet de mettre en évidence de manière automatique des relations entre 2 ensembles de données numériques. Notre domaine d'application est la génomique fonctionnelle. Nos sources de données sont ainsi des données d'expression génique issues de puces à ADN et des données biocliniques. Le volume de données à explorer est conséquent car nous disposons de dizaines de milliers de mesures réalisées simultanément sur les puces à ADN pour quelques dizaines d'individus et des dizaines de paramètres cliniques pour chaque individu.

Notre objectif est de faciliter la découverte de relations globales ou partiellement linéaires. Peu de travaux s'intéressent de manière spécifique à la découverte automatique de corrélations linéaires.

Nous proposons un environnement ayant pour but de réduire les *a priori* sur les calculs effectués et les temps d'exploration des données par l'expert.

Le flux de DiscoClini est le suivant : (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 que l'on souhaite inclure dans une 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 ; (5) validation biologique des résultats par l'expert du domaine.

Il permet aux biologistes (dans notre contexte applicatif) de découvrir sans *a priori* des relations linéaires entre les deux types de données. Notre méthode peut être assimilée à une suite d'approximations et de reformulations, qui permettent à l'utilisateur de disposer de résultats d'analyse sous une forme synthétique et facilement exploitable. En (3b) les corrélations entre les sous-ensembles sont calculées. Les meilleurs résultats

⋆Nous tenons à remercier l'Institut Benjamin Delessert pour son soutien via le « Prix de Projet de Recherche 2004 »

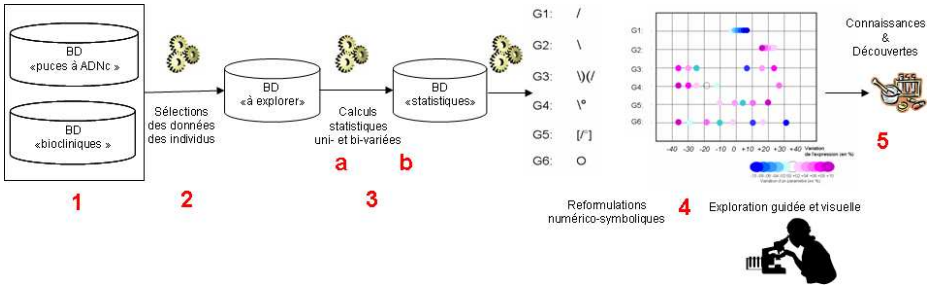

FIG. 1 – Flux de fouille de données de DiscoClini.

sont proposés à l'expert sous la forme d'un tableau associant pour chaque relation des données statistiques et une représentation graphique « compacts ». Ce mode de restitution des résultats permet à l'expert de visualiser simultanément un ensemble de relations potentiellement intéressantes.

Nous proposons avec DiscoClini une approche objective, au niveau de l'analyse et de l'exploration des données, où aucun *a priori*, en terme de connaissances dans le domaine, n'est requit.

Différentes expérimentations ont permis de valider DiscoClini et de produire des résultats qui ont contribué à des avancées biomédicales (Viguerie *et al.* (2004) ;Clément *et al.* (2004) ;Taleb *et al.* (2005)).

Les données que nous utilisons sont des données bruitées et lacunaires. Nous avons donc développé une méthode, actuellement en cours d'évaluation, permettant de détecter automatiquement les valeurs suspectes afin d'améliorer la qualité des résultats fournis à l'expert. L'intégration de ces nouvelles informations dans l'étape de visualisation est un des problèmes qui se pose aujourd'hui à nous.

Références

CLÉMENT K., VIGUERIE N., POITOU C., CARETTE C., PELLOUX V., CURAT C., SICARD A., ROME S., BENIS A., ZUCKER J., VIDAL H., LAVILLE M., BARSH G., BASDEVANT A., STICH V., CANCELLO R. & LANGIN D. (2004). Weight loss regulates inflammation-related genes in white adipose tissue of obese subjects. *FASEB J*, **18**(14), 1657–1669.

TALEB S., LACASA D., BASTARD J., POITOU C., CANCELLO R., PELLOUX V., VIGUERIE N., BENIS A., ZUCKER J., BOUILLOT J., COUSSIEU C., BASDEVANT A., LANGIN D. & CLÉMENT K. (2005). Cathepsin s, a novel biomarker of adiposity : relevance to atherogenesis. *FASEB J*, **19**(11), 1540–2.

VIGUERIE N., CLÉMENT K., BARBE P., COURTINE M., BENIS A., LARROUY D., HANCZAR B., PELLOUX V., POITOU C., KHALFALLAH Y., BARSH G. S., THALAMAS C., ZUCKER J. D. & LANGIN D. (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.

Benis A., Zucker Jean-Daniel

DiscoClini : une méthodologie pour l'extraction de relations linéaires dans des données de génomique médicale

In : Trichet F. (ed.). Actes de la conférence IC2007 : journées francophones d'ingéniérie des connaissances. Toulouse : Cépaduès, 2007, p. 337-338. ISBN 978-2-85428-790-5

Conférence IC2007 : Journées Francophones d'Ingénièrie des Connaissances, 18., 2007/07/04-06, Grenoble