Recherche automatique de réseaux de gènes co-exprimés

Ali Janbain [1] & Christelle Reynès [1] & Zainab Assaghir [2] & Hassan Zeineddine [2] & Robert Sabatier [1] & Laurent Journot [3]

[1] *Laboratoire de Biostatistique, Informatique et physique Pharmaceutique, UFR Sciences Pharmaceutiques, Université de Montpellier, Institut de Génomique Fonctionnelle - alijanbain.aj@gmail.com, christelle.reynes@umontpellier.fr, robert.sabatier@umontpellier.fr*

[2] *Département Mathématiques Appliquées-Faculté des Sciences-Université Libanaise- zainab.assaghir@ul.edu.lb, hassan.zeineddine@ul.edu.lb*

[3] *Institut de Génomique Fonctionnelle 141, rue de la Cardonille 34094 Montpellier- laurent.journot@igf.cnrs.fr*

Résumé. L'objectif de ce travail est de mettre au point une nouvelle approche automatique pour identifier les réseaux de gènes concourant à une même fonction biologique. Différentes stratégies ont été développées pour essayer de regrouper les gènes d'un organisme selon leurs relations fonctionnelles: génétique classique et génétique moléculaire. Ici, nous allons utiliser une propriété connue des réseaux de gènes fonctionnellement liés à savoir que ces gènes sont généralement co-régulés et donc coexprimés. Cette co-régulation peut être mise en évidence par des méta-analyses de données de puces à ADN (microarrays) telles que Gemma ou COXPRESdb. Pour cela, nous avons tout d'abord recherché les descripteurs de réseaux discriminant au mieux les ensembles de gènes fonctionnellement liés. Ensuite, nous combinons des outils de classification supervisée et des algorithmes génétiques pour séparer un ensemble de gènes fonctionnellement liés d'un ensemble de gènes sans lien connu.
Nous appliquerons cette méthode à l'étude de co-régulations en partant d'une annotation fonctionnelle connue de certains gènes.

Mots-clés. Réseaux, gènes, algorithme génétique.

Abstract. The aim of this work is to develop a new automatic approach to identify networks of genes involved in the same biological function. Various strategies have been developed to try to cluster genes of an organism according to their functional relationships: classical genetics and molecular genetics. We will use a known property of functionally related genes namely that these genes are generally co-regulated. This co-regulation can be detected by microarray meta-analyzes databases such as Gemma or COXPRESdb. Hence, we identified network descriptors discriminating at best the sets of functionally related genes. Then we combine supervised classification tools and genetic algorithms to separate one set of functionally related genes from genes with no known link.

We apply this method to the study of co-regulation starting from a known functional annotation of certain genes.

Keywords. Networks, genes, genetic algorithm.

1 Introduction

Différentes stratégies ont été développées pour essayer de regrouper les gènes d'un organisme selon leurs relations fonctionnelles (Kanehisa *et al* (2011), Ashburner *et al.* (2000)). Dans le présent travail, nous allons utiliser une propriété connue des réseaux de gènes fonctionnellement liés à savoir que ces gènes sont généralement co-régulés. Cette co-régulation peut être mise en évidence par des méta-analyses de données de puces à ADN (micro-arrays) telles que Gemma introduite par Zoubarev *et al.* (2012) ou COXPRESdb présentée par Obayashi *et al.* (2013). Un précédent travail réalisé par Al Adhami *et al.* (2015) a montré qu'en utilisant deux paramètres décrivant la topologie des réseaux de gènes co-régulés (Fig.1), on pouvait distinguer des groupes de gènes choisis au hasard de groupes de gènes ayant des liens fonctionnels. Nous avons choisi d'exploiter cette idée pour aider à l'amélioration des annotations fonctionnelles des gènes.

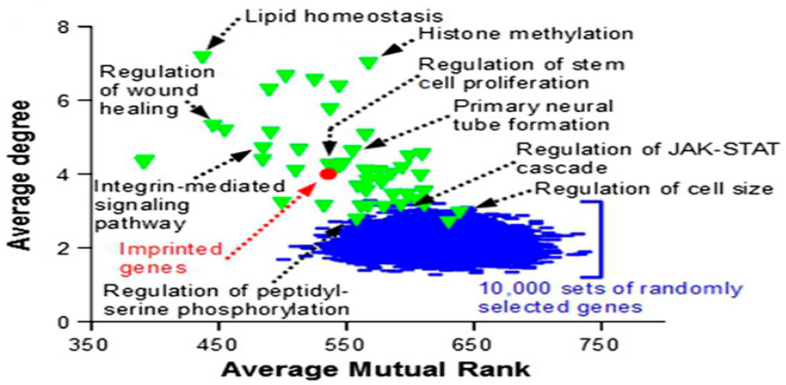

Figure 1: Le degré moyen (Average Degree) et le rang mutuel moyen (Average Mutual Rank) permettent de distinguer les réseaux obtenus à partir des gènes choisis au hasard (∘) ou appartenant à un même KEGG pathway (▲) ou au réseau des gènes soumis à empreinte génomique parentale (IG, ●).

2 Description générale des données

Nous disposons d'informations concernant 20959 gènes. Pour chaque couple de gènes, on connaît le coefficient de corrélation de leurs niveaux d'expression à travers les centaines ou milliers d'expériences de puces à ADN de la base de données. On quantifie alors le lien entre deux gènes par leur rang mutuel MR. En effet, dans COXPRESdb, les réseaux

de gènes coexprimés sont établis sur la base des rangs de corrélation : pour chaque gène, on donne le rang 1 au gène qui lui est le plus corrélé, puis le rang 2 au suivant. COXPRESdb, comme quantification du lien entre deux gènes, utilise le rang mutuel (MR) qui est la moyenne géométrique entre les deux rangs.
En accord avec les biologistes, nous avons décidé d'utiliser le MR pour décrire les interactions. Par ailleurs, pour chaque gène on connaît ses annotations fonctionnelles disponibles.

3 Étude des liens entre gènes

3.1 Construction d'un réseau de gènes

Un réseau (en mathématiques, un graphe) se compose de sommets reliés par des arêtes qui peuvent être dirigées ou non. Les sommets et les arêtes peuvent avoir des valeurs numériques.
Dans le cadre de cette étude, les sommets des réseaux sont les gènes et les arêtes représentent les interactions entre les gènes. Une relation entre deux gènes A et B dans un réseau, signifie simplement l'idée qu'un changement dans l'expression du gène A est susceptible de causer un changement dans l'expression du gène B. Si la valeur de MR est inférieure à un certain seuil alors on décide qu'il existera dans le réseau une relation donc une arête entre les deux gènes.

3.2 Types de réseaux

On définit deux types de réseaux:

- Réseau potentiel (PN): c'est un réseau dont les gènes sont connus pour participer à la même fonction.
- Réseau aléatoire (RN): c'est un réseau constitué d'un sous-ensemble de gènes choisis aléatoirement et dont on analyse les liens.

Pour chercher des réseaux potentiels, nous avons utilisé les KEGG pathways présentés par Kanehisa *et al* (2011) qui permettent d'obtenir des listes de gènes ayant été précédemment associés à une même fonction. Nous avons également utilisé les annotations "processus biologique" du projet Gene Ontology (GO) réalisé par Ashburner *et al.* (2000) qui associent à chaque gène des mots-clés liés à leur fonction.

3.3 Descripteurs pertinents

Il existe un certain nombre de paramètres usuels pour décrire un réseau (Diestel, R (2005)). Suivant le travail de Al Adhami *et al.* (2015), nous avons cherché parmi eux, des descripteurs dits pertinents, c'est-à-dire des descripteurs dont la distribution est significativement

différente entre réseaux potentiels PN et réseaux aléatoires RN. Par exemple, d'après la Fig.1, le *degré moyen (Average degree)* semble pertinent parce que sa distribution est disjointe pour de réseaux potentiels (● et ▲) et pour des réseaux aléatoires (○). Par contre le rang mutuel moyen ne semble pas pertinent.

Nous avons défini une centaine de réseaux potentiels en fixant le seuil de MR à 1200. Douze descripteurs parmi les descripteurs testés sont significativement discriminants.

4 Construction d'un modèle de discrimination

Nous allons maintenant utiliser ces descripteurs pour chercher à prédire le type de réseau. Par conséquent, nous avons construit une matrice composée de centaines de lignes et de 14 colonnes. Chaque ligne représente un réseau. Les douze premières colonnes représentent les valeurs des descripteurs pertinents de ce réseau, la treizième est la taille du réseau et la dernière colonne représente son type déjà connu (PN ou RN). Par validation croisée en deux blocs, le pourcentage moyen de bonne classification sur 1000 essais est de 95.18 %. Cette matrice est utilisée pour apprendre un modèle permettant de prédire le type d'un réseau quelconque. Pour cela, on applique la discrimination linéaire de Fisher (= LDA : Linear Discriminant Analysis) qui cherche à discriminer les réseaux en groupes PN et RN. On note $Score.LDA$ la coordonnée de chaque réseau sur l'axe discriminant de LDA.

5 Extraction d'un PN dans un ensemble de gènes

Supposons qu'on a un sous-ensemble de gène noté A formant un PN dans un espace de gènes noté Z. Notre but est de chercher dans Z les gènes qui constituent un réseau co-régulé contenant des gènes de PN. En d'autre terme, nous allons chercher des gènes qui peuvent participer à la même fonction biologique de PN. La nouvelle méthode impose l'ensemble Z comme un réseau contenant tous les gènes de la base de donnée. On repère alors l'ensemble PN des gènes annotés par une même annotation fonctionnelle, ils serviront d'initialisation pour la recherche des gènes co-régulés participant à cette fonction. L'objectif est alors de conserver les gènes participant à cette fonction et co-régulés entre eux, d'ajouter des gènes non annotés par cette fonction mais significativement co-régulé et d'éliminer des gènes annotés par cette fonction mais qui ne sont pas liés par une co-régulation.

5.1 Initialisation

L'une des propriétés des PN est l'existence de sous-graphes bien connectés qui contiennent la majorité des gènes. Nous cherchons donc, dans le réseau PN choisi, la plus grosse clique. En effet, une clique est un sous-ensemble de gènes dont deux gènes quelconques sont liés par une arête. La plus grosse clique est la clique possédant le plus grand nombre

de gènes. Cette grosse clique, notée GC est nommée "Cœur du réseau", elle constituera un point d'encrage pour l'algorithme.

5.2 Algorithme génétique (AG)

Nous utilisons un AG (Golberg, 1989) qui a pour but d'obtenir une solution approchée à un problème d'optimisation complexe. Ici, les solutions de l'AG sont un réseau candidat de la forme GC + autres gènes, dont on cherche à optimiser un critère pour obtenir une solution. Nous avons voulu nous servir de cette idée en définissant un critère (fitness) dont le fait de maximiser ce critère donne une solution contenant GC ainsi que des gènes co-régulés et donc participant potentiellement à la même fonction biologique que celle choisie pour construire PN.

Présenter le critère:
Une exploration supervisée, nous a permis de définir le critère (fitness) suivant comme étant un bon candidat pour quantifier la qualité d'une solution, comme nous le montrerons lors de l'exposé : $ScoreLDA \times Taille$ où $ScoreLDA$ est la valeur de $Score.LDA$ présentée dans la section 4 et $Taille$ est la taille de la solution candidate.

6 Conclusion

Le but initial de ce travail est de chercher des gènes qui peuvent participer à la même fonction biologique d'un PN.

Premièrement, nous avons étudié plusieurs descripteurs afin de choisir parmi eux ceux qui permettent de distinguer réseaux potentiels PN et réseaux aléatoires RN. Ensuite, en utilisant ces descripteurs, nous avons constitué une base de données de PN et construit un modèle LDA qui parvient à prédire le type d'un réseau quelconque(PN ou RN). Enfin, nous avons élaboré une stratégie d'exploration de l'espace des gènes, basée sur l'utilisation d'un algorithme génétique, pour permettre d'affiner les annotations fonctionnelles de gènes disponibles dans les bases.
Les applications de cette méthode sont nombreuses et divers exemples seront présentés.

Bibliographie

[1] Al Adhami, H. *et al.* (2015), *A systems-level approach to parental genomic imprinting: the imprinted gene network includes extracellular matrix genes and regulates cell cycle exit and differentiation*, Genome research, 25(3), 353-367.
[2] Ashburner, M. *et al.* (2000), *Gene Ontology: tool for the unification of biology*, Nature Genetics. 25: 25-29.

[3] Diestel, R. (2005), *Graph theory*, Springer-Verlag, Heidelberg, ISBN 3-540-26182-6.
[4] Kanehisa, M. et al (2011), *KEGG for integration and interpretation of large-scale molecular data sets*, Oxford University Press.
[5] Obayashi, T. *et al.* (2013), *COXPRESdb: a database of comparative gene coexpression networks of eleven species for mammals*, Nucleic Acids Research 41 :D1014-20.
[6] Goldberg, D.E. (1989), *Genetic algorithms in Search, Optimization and Machine Learning*, Addison-Wesley, Reading, Massachussetts.
[7] Zoubarev, A. *et al.* (2012), *Gemma: a resource for the reuse, sharing and meta-analysis of expression profiling data*, Bioinformatics. 28:2272-3, Canada.