

Inférence de réseaux de régulation de gènes à partir de données d'expression et de polymorphisme

Jimmy Vandel, Simon de Givry

Laboratoire de mathématiques et informatique appliquées, INRA, Toulouse
BP 52627, 31326 Castanet Tolosan Cedex France
{jimmy.vandel,degivry}@toulouse.inra.fr

Résumé : *La reconstruction de réseaux de régulation est un problème complexe, en particulier du fait que les données d'expression mesurent un très grand nombre de gènes seulement à travers un petit nombre d'échantillons. Parmi les nombreuses approches proposées, nous choisissons d'explorer l'approche fondée sur l'apprentissage de la structure d'un modèle probabiliste, en l'occurrence un réseau Bayésien. Nous évaluons différents algorithmes d'apprentissage sur des données simulées d'expression de gènes pour un échantillon d'individus apparentés et génotypés. Nous proposons d'améliorer la qualité de la reconstruction en exploitant l'information des génotypes dans le modèle et dans les algorithmes.*

Mots-clefs : bio-informatique, données d'expression de gènes, données de polymorphisme, réseaux Bayésiens, apprentissage automatique de la structure.

1 Introduction

Un gène est une unité fonctionnelle de l'hérédité, porteur d'une information d'une génération à l'autre. L'expression d'un gène dans une cellule se traduit par la production d'enzymes, chacune pouvant réguler l'expression d'autres gènes. Une meilleure connaissance du réseau de régulation d'un ensemble de gènes est une aide précieuse pour les biologistes dans leur analyse des caractères complexes comme par exemple la susceptibilité à certaines maladies ou la résistance à des stress hydriques et thermiques dans le cas des plantes.

La reconstruction de réseaux de régulation est un problème complexe [7], en particulier du fait que les données d'expression mesurent un très grand nombre de gènes seulement à travers un petit échantillon. Ce cadre méthodologique découle des données sur les transcrits des gènes. En effet, les technologies de type micro-array permettent d'observer sur une unique puce un très grand nombre de transcrits mais le coût de cette technologie, bien que diminuant chaque mois, est encore trop élevé pour envisager un échantillon (de puces) de grande taille.

Les premières approches permettant de prédire la topologie du réseau se sont attachées à décrire et quantifier les relations locales entre deux noeuds quelconques du réseau, le réseau global étant directement construit par seuillage de l'estimateur choisi pour quantifier la relation locale [4]. Les approches globales de reconstruction de réseaux sont basées soit sur des modèles d'équations structurelles et le cadre de la vraisemblance pénalisée [8], soit sur des réseaux Bayésiens ou des modèles graphiques et le cadre de l'inférence Bayésienne [3]. Nous avons choisi d'explorer l'approche fondée sur l'apprentissage de la structure d'un modèle probabiliste, en l'occurrence un réseau Bayésien.

Ces dernières années, plusieurs travaux [6,10,8] ont montré l'intérêt d'exploiter des données de polymorphisme pour mieux reconstruire un réseau de régulation de gènes. Ces données proviennent de la variabilité génétique entre individus et correspondent à l'observation de différents génotypes sur des marqueurs moléculaires pour un ensemble d'individus. Notre objectif est de proposer et d'évaluer une méthode d'apprentissage qui exploite données d'expression et données de polymorphisme.

2 Travaux en cours

Actuellement, il n'y a pas à notre connaissance de données d'expression combinées à des données de polymorphisme disponibles sur le Web. Pour évaluer des modèles et algorithmes d'apprentissage, un premier travail a consisté à développer un simulateur de données d'expression à partir d'individus apparentés génotypés en reprenant le processus décrit dans [8]. Pour cela, nous générons aléatoirement des données de génotype pour N marqueurs répartis uniformément le long d'un chromosome pour un ensemble P d'individus obtenus par croisements de type *lignées recombinantes auto-fécondées*. Dans un premier temps, une hypothèse simplificatrice consiste à associer un gène par marqueur. Un système d'équations différentielles ordinaires modélise la cinétique des niveaux d'expression des gènes en fonction de leurs polymorphismes [8] et d'un réseau de régulation (activation ou inhibition) généré lui aussi aléatoirement [9]. Le simulateur de réactions biochimiques Copasi [5] fournit les niveaux d'expression pour chaque individu à partir du système d'équations. La figure 1 montre un exemple de réseau et une distribution des niveaux d'expression pour un gène donné. Une étape importante avant l'apprentissage du réseau Bayésien est la discrétisation des niveaux d'expression. Nous avons choisi une discrétisation classique tri-valuée (sous, normal et sur-exprimé) [3].

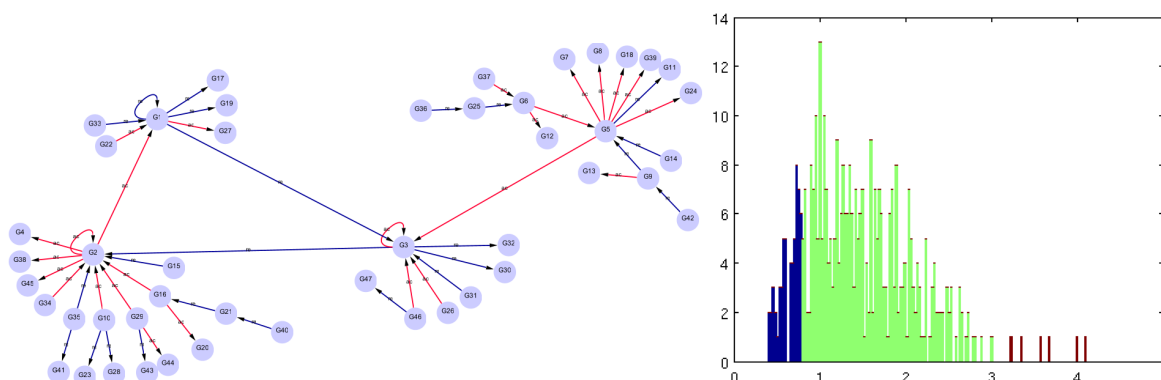


Fig. 1. Réseau de régulation artificiel de type *scale-free* (Web50 [9], $N = 50$) et distribution des niveaux d'expression d'un gène pour $P = 400$ individus (en rouge, niveaux sur-exprimés, en vert normaux et en bleu sous-exprimés).

Dans un premier temps, nous comparons des méthodes d'apprentissage de la structure d'un réseau Bayésien composé de N variables aléatoires (correspondant aux gènes) ayant un domaine à 3 valeurs en exploitant uniquement les données d'expression. Les méthodes d'apprentissage à partir de données complètes sans variable latente se divisent en deux grandes familles. Les méthodes fondées sur la recherche de causalité (effectuant des tests d'indépendance conditionnelle) et celles parcourant l'espace des structures possibles en optimisant un score local. Ce score combine la vraisemblance des données à une pénalité liée à la complexité de la structure apprise. En suivant la démarche de [1], nous avons choisi le score *Bayesian Information Criterion* et nous utilisons la librairie d'algorithmes d'appren-

tissage *Bayes Net Toolbox Structure Learning Package*¹ [2]. Pour faire face à la taille des instances (N pouvant être de l'ordre du millier de gènes), des modifications dans la librairie ont été apportées (en particulier, limite sur le nombre maximum de régulateurs, exploration du voisinage d'un réseau de type *sélection du premier voisin améliorant le score*). L'évaluation des résultats en fonction de la taille des données, de la topologie du réseau à reconstruire et des paramètres sur les lois cinétiques contrôlant les niveaux d'expression est en cours.

Par la suite, les données de polymorphisme seront intégrées dans le réseau Bayésien (ajout de nouvelles variables associées aux génotypes). L'influence des génotypes sur l'expression des gènes peut être analysée par des méthodes de cartographie de QTL (Quantitative Trait Loci) appliquées aux données quantitatives que sont les niveaux d'expression. Le résultat de cette analyse est de déterminer pour chaque gène les marqueurs dont le polymorphisme contribue à expliquer la variation du niveau d'expression du gène cible. Les gènes localisés aux marqueurs sont des régulateurs potentiels du gène cible. Cette information permettra de contraindre l'espace des structures possibles de manière à *guider* l'algorithme d'apprentissage vers de meilleures structures [10].

Références

- [1] Cedric Auliac, Vincent Frouin, Xavier Gidrol, and Florence D'alche-Buc. Evolutionary approaches for the reverse-engineering of gene regulatory networks : a study on a biologically realistic dataset. *BMC Bioinformatics*, 9 :91–104, 2008.
- [2] Olivier François and Philippe Leray. Étude comparative d'algorithmes d'apprentissage de structure dans les réseaux bayésiens. *Journal électronique d'intelligence artificielle*, 5(39), 2006.
- [3] Nir Friedman. Inferring Cellular Networks Using Probabilistic Graphical Models. *Science*, 303(5659) :799–805, 2004.
- [4] A. Ghazalpour, S. Doss, B. Zhang, S. Wang, C. Plaisier, R. Castellanos, A. Brozell, E. Schadt, T. Drake, A. Lusic, and S. Horvath. Integrating genetic and network analysis to characterize genes related to mouse weight. *PLOS GENETICS*, 2(8) :1182–1192, 2006.
- [5] S. Hoops, S. Sahle, R. Gauges, C. Lee, J. Pahle, N. Simus, M. Singhal, L. Xu, P. Mendes, and U. Kummer. COPASI—a COMplex PATHway SIMulator. *Bioinformatics*, 22(24) :3067–3074, 2006.
- [6] Ritsert C. Jansen and Jan-Peter Nap. Genetical genomics : the added value from segregation. *Trends in Genetics*, 17(7) :388 – 391, 2001.
- [7] Huai Li, Jianhua Xuan, Yue Wang, and Ming Zhan. Inferring regulatory networks. *Frontiers in Bioscience*, 13 :263–275, 2008.
- [8] Bing Liu, Alberto de la Fuente, and Ina Hoeschele. Gene network inference via structural equation modeling in genetical genomics experiments. *GENETICS*, 178(3) :1763–1776, 2008.
- [9] Pedro Mendes, Wei Sha, and Keying Ye. Artificial gene networks for objective comparison of analysis algorithms. *Bioinformatics*, 19 :ii122–ii129, 2003.
- [10] Jun Zhu, Matthew C. Wiener, Chunsheng Zhang, Arthur Fridman, Eric Minch, Pek Y. Lum, Jeffrey R. Sachs, and Eric E. Schadt. Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. *PLOS COMPUTATIONAL BIOLOGY*, 3(4) :692–703, 2007.

¹ <http://www.cs.ubc.ca/~murphyk/Software/BNT/bnt.html> et <http://bnt.insa-rouen.fr/>.