Apprentissage de signatures de facteurs de transcription à partir de données d'expression

Mohamed Elati*, Céline Rouveirol*[1], François Radvanyi**

* LRI, UMR CNRS 8623 ,
Université Paris Sud, bât 490
91405 ORSAY cedex
elati, celine@lri.fr
** Institut Curie, UMR CNRS 144,
26 rue d'Ulm
75248 Paris cedex 05
francois.radvanyi@curie.fr

Résumé. L'inférence de signatures de facteurs de transcription à partir des données puces à ADN a déjà été étudié dans la communauté bioinformatique. La principale difficulté à résoudre est de trouver un ensemble d'heuristiques pertinentes, afin de contrôler la complexité de résolution de ce problème NP-difficile. Nous proposons dans cet article une solution heuristique alternative à celles utilisées dans les approches bayésiennes, fondée sur la recherche de motifs fréquents maximaux dans une matrice discrétisée issue des données numériques de puces ADN. Notre méthode est appliquée sur des données de cancer de vessie de l'Institut Curie et de l'Hôpital Henri Mondor de Créteil.

1 Introduction

Un des principaux objectifs de la biologie moléculaire consiste à comprendre la régulation des gènes d'un organisme vivant dans des contextes biologiques spécifiques. Les facteurs de transcription (notés Tfs dans la suite) sont les régulateurs de la transcription qui vont réagir avec les promoteurs de la transcription des gènes cibles. Ils ont deux modes d'action : ils peuvent *activer* ou *inhiber* l'expression d'un gène. Les mécanismes d'interaction facteurs de transcription/gènes cibles sont complexes. Plusieurs facteurs de transcription peuvent être nécessaires pour l'induction (resp. la répression) d'un gène cible et, dualement, un facteur de transcription peut induire ou réprimer plusieurs gènes. Les techniques récentes d'analyse du transcriptome, telles que les puces à ADN permettent de mesurer simultanément les niveaux d'expression de plusieurs milliers de gènes. Un ensemble de puces permet donc de connaître l'expression de ces milliers de gènes dans plusieurs conditions expérimentales d'intérêt. En général, ces mesures (appelés *données d'expression* dans la suite) sont représentées dans une matrice dont les lignes représentent les gènes et les colonnes représentent les différentes puces disponibles. Certains travaux d'analyse de puces font l'hypothèse que l'observation de corrélations dans les données d'expression va permettre d'inférer des relations

[1]Ce travail a été effectué pendant la délégation CNRS de Céline Rouveirol à l'Institut Curie.

de co-régulation entre ces gènes. En effet, la plupart des mécanismes biologiques sont régulés au niveau transcriptionnel par des cascades d'activateurs ou d'inhibiteurs.

Les approches qui extraient des régulations géniques à partir de données d'expression exploitent cette dépendance. (Qian et al. 2003) présentent une méthode d'apprentissage supervisée qui utilise les *Support Vector Machines*. Les exemples en entrée du système sont les profils d'expression de couples Tf/gène cible pour des relations de régulations confirmées ou infirmées par des méthodes biologiques expérimentales. L'algorithme fournit un classifieur capable de prédire, pour de nouveaux couples Tf/gène cible, s'ils sont en relation de régulation. Cette méthode est peu adaptée si le nombre d'exemples positifs, i.e. les relations de régulation connues, est insuffisant, ce qui est le cas chez l'homme. Par ailleurs, et surtout dans le cas des eucaryotes, un gène est généralement régulé par un complexe de facteurs de transcription, ce qui n'est pas pris en compte dans cette méthode. Dans (Pe'er et al. 2002), les auteurs utilisent l'approche bayésienne introduite par (Friedman et al. 2000) pour modéliser le réseau de régulation génique par un graphe[2] orienté sans cycle tel que l'expression d'un gène est une fonction de probabilité conditionnelle de l'expression de ses régulateurs. Malheureusement, la recherche du réseau de régulation optimal à partir des données d'expression est un problème NP-complet[3] (Friedman et al. 2000). (Pe'er et al. 2002) propose un algorithme *Minreg*, qui recherche une approximation de l'ensemble minimal de régulateurs (Tfs) actifs R pour un ensemble des gènes cibles donné. *Minreg* est guidé par un score *local* (l'information mutuelle) qui évalue le degré de dépendance entre un sous-ensemble des régulateurs et un gène cible. *Minreg* calcule R itérativement et de manière gloutonne : à chaque étape, il ajoute à R le facteur de transcription Tf qui permet de maximiser le score global de corrélation entre chacun des gènes cibles et tous les sous-ensembles de $R \cup Tf$ de taille $\leq d$, contrainte sur la taille des régulateurs d'un gène. Hors, comme indiqué par les auteurs, bien que l'information mutuelle soit bien fondée du point de vue statistique, rien n'assure que nous obtiendrons ainsi pour chaque gène cible le meilleur ensemble de régulateurs de taille $\leq d$. Par exemple, si $R \cup Tf_1$ et $R \cup Tf_2$ ont chacun un faible score mais que $R \cup (Tf1 \cup Tf2)$ a un score élevé, ce dernier ne sera pas détecté.

Pour résoudre ce problème, nous proposons dans ce travail une méthode originale qui fait l'hypothèse qu'une relation de régulation (activation ou inhibition) entre un sous-ensemble de Tfs et son gène cible s'observe dans les données d'expression sous la forme d'un *motif fréquent*. Nous utilisons donc la recherche des motifs fréquents maximaux pour approximer une topologie de régulation, autrement dit pour calculer un ensemble de régulateurs candidats réduit pour chaque gène cible. Puis, nous effectuons une recherche exhaustive, guidée par une fonction de score locale, du sous-ensemble de régulateurs candidats qui explique au mieux sa variation d'expression.

La suite de cet article est organisée comme suit. Dans la section 2, nous introduisons la technique de recherche des motifs fréquents ainsi que le pré-traitement effectué sur les données de puces à ADN. Puis, nous présentons le détail de notre approche. Dans la section 3, nous exposons des résultats préliminaires concernant l'application de cette approche à partir des données de cancers de vessie. Enfin, nous concluons sur les perspectives ouvertes de ce travail.

[2]Les nœuds sont les gènes et les arcs représentent des liens de régulation entre les gènes.
[3]L'énumération de réseaux possibles est exponentielle en fonction du nombre de gènes.

2 Méthodologie

2.1 Recherche des motifs fréquents

La recherche de motifs fréquents a été introduite la première fois par (Agrawal et al. 1993), dans le cadre du problème d'"analyse du panier de la ménagère", où chaque transaction est constituée d'une liste d'articles achetés afin d'identifier les articles fréquemment achetés ensemble. Cette technique d'apprentissage non supervisé a été utilisée dans plusieurs domaines où une grande masse de données est disponible, en particulier l'analyse du transcriptome (Morishita et al. 2001). Nous allons présenter brièvement les notations indispensables à la suite de l'article. On définit un *contexte d'extraction* comme un triplet (O, A, M) tel que O est un ensemble fini d'observations, A est un ensemble fini d'attributs et M est une matrice booléenne dont les colonnes sont les attributs de A et dont les lignes sont les descriptions booléennes des observations de O en fonction de A. M sera également appelé contexte d'extraction, par abus de langage. On appelle *motif* tout sous-ensemble d'attributs de A. Le *support* (fréquence) d'un motif est le nombre d'observations dans lesquelles le motif apparaît. Un motif est dit *fréquent* si son support est supérieur ou égal à un support minimum S_{min} fixé par l'utilisateur. Un motif est dit fréquent *maximal*, s'il est fréquent et que tous ses sur-ensembles stricts sont non fréquents. L'ensemble des motifs fréquents maximaux est une représentation très compacte de l'ensemble des motifs fréquents, puisqu'il représente la frontière qui sépare l'ensemble des motifs fréquents des motifs non fréquents. Plusieurs algorithmes ont été mis en place pour calculer cet ensemble, nous pouvons citer entre autres Eclat (Borgelt 2003).

2.2 Pré-traitement des données de puces à ADN

L'expression brute des gènes est représentée par des variables continues (réelles) dans les données d'expression. Pour pouvoir les traiter dans un algorithme d'apprentissage booléen, il nous faut les discrétiser. Il est possible d'utiliser des algorithmes de discrétisation classiques présentés dans (Dougherty et al. 1995), mais comme nous étudions des échantillons *tumoraux* humains et que nous disposons d'un ensemble d'échantillons *normaux*, nous avons préféré définir un algorithme de discrétisation par rapport aux échantillons normaux, afin que le profil d'expression caractérise au mieux le processus tumoral étudié. Le niveau d'expression sera discrétisé en 3 valeurs : (1) représente l'état sur-exprimé des gènes, (0) l'état stable et (-1) l'état sous-exprimé. Puis chaque gène cible g sera représenté par deux variables booléennes : $g\uparrow$ qui représente l'état *sur-exprimé* et $g\downarrow$ qui représente l'état *sous-exprimé*. Etant donné qu'un facteur de transcription n'a une action de régulation sur ses gènes cibles que lorsqu'il est lui même exprimé, nous associons à chaque Tf une seule variable booléenne, qui est à 1 si Tf est sur-exprimé, et à 0 sinon. A l'issue de cette discrétisation, nous obtenons une matrice booléenne des niveaux d'expression des Tfs et gènes cibles (attributs) dans un ensemble d'échantillons tumoraux (observations).

2.3 Estimation de la topologie de régulation

La première étape de notre algorithme approxime pour chaque gène cible g un ensemble de régulateurs candidats $F(Ra(g), Ri(g))$, avec $Ra(g)$ les activateurs et $Ri(g)$ les inhibiteurs, de taille très réduite par rapport à l'ensemble initial de facteurs de transcription. Nous faisons l'hypothèse que chaque facteur de transcription fréquemment exprimé lorsqu'un gène cible est sur-exprimé (resp. sous-exprimé) peut être un activateur (resp. inhibiteur) de ce gène cible. À partir du *contexte booléen* d'extraction construit à l'étape précédente, nous générons tous les motifs fréquents maximaux. Un post-traitement de ces motifs nous permet de ne conserver que l'ensemble M des motifs *mixtes*. Un motif est *mixte* s'il contient au moins un gène cible et au moins un facteur de transcription. Nous pouvons donc calculer, à partir de M, une *topologie* de régulation approximée entre un ensemble T de Tfs et un ensemble V de gènes cibles, représentée comme un graphe bipartie $G = (T, V, S)$. L'ensemble S de relations de régulation (active ou inhibe) qui relient les Tfs à leurs gènes cibles, est construit de la façon suivante : un Tf est lié à un gène cible g par la relation active (resp. inhibe) s'il existe un motif $m \in M$ tel que $\{Tf, g\uparrow\}$ (resp. $\{Tf, g\downarrow\}$) est inclus dans m. Ainsi pour chaque $g \in V$, $Ra(g)$ est l'ensemble de Tfs adjacents à g avec la relation active et $Ri(g)$ est l'ensemble des Tfs adjacents à g avec la relation inhibe (voir figure 1).

O/A	$\mathbf{Tf_1}$	$\mathbf{Tf_2}$	$\mathbf{Tf_3}$	$\mathbf{g_{1\uparrow}}$	$\mathbf{g_{1\downarrow}}$
$\mathbf{o_1}$	0	1	1	1	0
$\mathbf{o_2}$	1	1	0	0	1
$\mathbf{o_3}$	1	1	0	0	1
$\mathbf{o_4}$	1	1	1	1	0

Soit $S_{min} = 20\%$
$M = \{\{\mathbf{Tf_1}, \mathbf{Tf_2}, \mathbf{g_{1\downarrow}}\}, \{\mathbf{Tf_2}, \mathbf{Tf_3}, \mathbf{g_{1\uparrow}}\}\}$
$Ra(g_1) = \{\mathbf{Tf_2}, \mathbf{Tf_3}\}$
$Ri(g_1) = \{\mathbf{Tf_1}, \mathbf{Tf_2}\}$

FIG. 1 – Approximation d'une topologie de régulation entre 3 Tfs et un gène cible.

Notons que nous choisissons un S_{min} faible puisque l'objectif de cette phase est de réduire *sans perte d'information* l'ensemble des facteurs de transcription candidats pour la régulation d'un gène cible.

2.4 Extraction d'interactions entre Tfs et gènes cibles

La deuxième étape consiste à extraire, pour chaque gène, le sous-ensemble des régulateurs candidats calculé à l'étape précédente qui maximise un *score local*. Ce score évalue le degré de dépendance entre un régulé et ses régulateurs candidats. Plusieurs mesures statistiques peuvent être utilisées à ce titre (voir (Azé 2003) pour une synthèse). Dans le cas général, ces fonctions d'évaluation ne sont ni *monotones* ni *anti-monotones*, ce qui ne nous permet pas d'effectuer une recherche efficace du meilleur sous-ensemble de régulateurs. Au contraire de *Minreg*, et parce que nous disposons d'un ensemble de régulateurs candidats réduit, nous pouvons calculer exhaustivement le score de tous les sous-ensembles de $Ra(g)$ (resp. $Ri(g)$) et sélectionner celui ayant le meilleur score comme activateur (resp. inhibiteur) du gène cible g.

3 Premiers résultats expérimentaux

Nous disposons de 85 expériences effectuées sur des puces ADN de type $Affymetrix$ (U95A). Cinq puces ont été obtenues par le grattage de l'épithélium vésical d'individus sains et sont utilisées pour discrétiser les 80 puces résultats d'une biopsie effectuée sur des patients atteints de cancer de vessie à différents stades de la maladie. Une liste de 469 facteurs de transcription a été extraite de *Gene Ontology* (http ://www.geneontology.com). Pour évaluer notre méthode, nous avons sélectionné dans ces données d'expression une liste des gènes cibles récemment publiée (Stratton et al. 2004) de 239 gènes connus comme étant liés à l'initiation et au développement tumoral. Pour un S_{min}=20%, l'algorithme génère un ensemble M de 150995 motifs fréquents maximaux mixtes. À partir de M, 202 états fréquents des gènes cibles (127 sous-exprimés et 85 sur-exprimés) sont détectés avec un ensemble de régulateurs candidats pour chaque état de taille moyenne égale à 11. Puis, l'algorithme effectue une recherche exhaustive guidée par un score local de corrélation (Lerman 1981) pour extraire les régulateurs potentiels (taille moyenne = 2) de chaque régulation génique. Nous donnons un extrait des résultats obtenus dans le tableau 1.

Gène cible	**Régulateurs (Tfs)**	**Relation**	**Score**	**Méth. expérimentales**
PRCC	MYC	Active	0,46	oui (Zeller et al. 2003)
FGFR3	TWIST MEF2 TCF3	Inhibe	0,71	oui (Bonaventure et al. 2003)
PDGFRB	NR2F1 TAF11 PMX1	Active	0,61	non confirmé

Tab. 1 – Extrait des résultats obtenus

Pour une première évaluation de ces relations de régulation inférées, nous avons recherché si ces relations ont été observées à partir de méthodes expérimentales dans la littérature et à l'aide des experts du domaine. En particulier, la régulation négative du gène FGFR3 a été confirmée[4]. En effet des résultats biologiques récents indiquent que TWIST exerce un rôle inhibiteur sur la différenciation ostéoblastique associée à une modulation de l'expression des gènes de la famille FGFR, et que cette régulation nécessite la formation d'un complexe qui lie TWIST à d'autres partenaires, comme TCF3 et MEF2. Ces résultats positifs nous autorisent à penser que notre système, bien que la tâche soit complexe, est prometteur.

4 Conclusion et perspectives

Nous avons conçu et implanté une méthode générale, modulaire et rapide pour l'identification des signatures de facteurs de transcription. Théoriquement, notre méthode permet de résoudre les limites de *Minreg* (Pe'er et al. 2002) et fournit des résultats optimaux vis-à-vis de la fonction d'évaluation choisie et de l'ensemble des facteurs de transcription sélectionnés par l'étape de calcul de motifs fréquents. Une extension envisageable de notre méthode est de combiner les données de *site de fixation* sur l'ADN

[4]Communication personnelle de Jacky Bonaventure (INSERM U93, Hôpital Necker).

des facteurs de transcription dans le but de vérifier une dépendance physique entre le niveau d'expression des gènes cibles et leurs régulateurs.

Références

Agrawal R., Imielinski T. et Swami A. (1993), Mining Association Rules between sets of items in large databases, Proc. of the ACM SIGMOD, 207-216, 1993.

Azé J. (2003), Extraction de Connaissances dans des Données Numériques et Textuelles, Thèse de l'Université Paris 11, `http://www.lri.fr/~aze/`, 2003.

Bonaventure J. et El Ghouzzi V. (2003), Molecular and cellular bases of syndromic craniosynostoses, Exp. Rev. Mol. Med., Vol 5, 2003.

Borgelt C. (2003), Efficient Implementations of Apriori and Eclat, Workshop Frequent Item Set Mining Implementations, FIMI 2003, Melbourne, FL, USA, 2003.

Dougherty J., Kohavi R. et Sahami A. (1995), Supervised and Unsupervised Discretization of Continuous Features, Proc. of Int. Conf. on Machine Learning,487-499, 1995.

Friedman N., Linial M., Nachman I. et Pe'er D. (2000), Using Bayesian Networks to Analyze Expression Data, J. Computational Biology, 7, 601-620, 2000.

Futreal P., Hubbard T., Wooster R., Rahman N. et Stratton M. (2004), A census of Human Cancer Genes, Nature Reviews Cancer, 4, 177-183, 2004.

Lerman C. (1981), Classification et analyse ordinale des données, Dunod, 32-41, 1981.

Morishita S., Hishiki T. et Okubo K. (2001), Towards Mining Gene Expression Database, Bioinformatics, 15 255, 2001.

Pe'er D., Regev A. et Tanay A. (2002), Minreg : inferring an active regulator set, Bioinformatics, 18, S258-S267, 2002.

Qian J., Lin J., Yu H. et Gerstein M. (2003), Prediction of regulatory networks : genome-wide identification of transcription factor targets from gene expression data, Bioinformatics, 19(15), 1917-26, 2003.

Zeller K, Jegga A., Aronow B, Donnell K. et Dang V. (2003), An integrated database of genes responsive to the MYC oncogenic transcription factor : identification of direct genomic targets, Genome Biology, 4, 2003.

Summary

DNA microarrays provide a way to learn relationships between transcription factors and their target genes by observing the simultaneous values of expression of thousands of genes and transcription factors. We propose an alternative heuristic solution to those implemented in bayesian approaches for inferring transcription factors' signatures from DNA chip expression values. We describe in this paper our inference method, which relies on the computation of maximal frequent itemsets from a boolean matrix obtained by discretizing DNA chips numerical gene expression values. We provide first evaluation results on bladder cancer data from Institut Curie and Hospital Henri Mondor.