Inférence conjointe de réseaux de gènes dans de multiples états

Nicolas Jung, Myriam Maumy-Bertrand, Laurent Vallat, Frédéric Bertrand

Inférence conjointe de réseaux de gènes dans de multiples états

Nicolas Jung [1,2], Myriam Maumy-Bertrand [1], Laurent Vallat [2] & Frédéric Bertrand [1]

[1] *Institut de Recherche en Mathématique Avancée (IRMA), Strasbourg* [2] *Institut d'Hématologie, Faculté de Médecine de Strasbourg*

Résumé. Quand une cellule est stimulée, le programme génique qu'elle contient est activé. Les gènes mis en action apportent alors une réponse concertée au stimulus. Cette réponse est modélisée statistiquement par un réseau dans lequel les noeuds correspondent aux gènes et les liens correspondent à leurs interactions. À partir des expressions de ces gènes, un nombre important de méthodes statistiques a été proposé pour inférer les réseaux de gènes sous-jacents.

Certaines maladies, comme le cancer par exemple, affectent le programme génique. Pour tenter de comprendre les perturbations qui en résultent, il est nécessaire d'estimer le réseau de gènes dans les différents états (sain/malade, par exemple). En supposant que seule une partie restreinte du réseau est affectée par la maladie, une estimation simultanée des différents réseaux (correspondant chacun à un état particulier) est nécessaire.

Cette estimation simultanée permet d'une part d'utiliser pleinement l'information commune dans les sous-parties du réseau inchangées d'un état à l'autre, et d'autre part, d'obtenir des réseaux plus facilement comparables. La méthode que nous proposons s'inscrit dans ce cadre.

Mots-clés. Réseau de régulation de gènes, Sélection de variables, Méthodes de classification.

Abstract. When a signal triggers a cell, the inherent genetic program is activated, leading to a concerted action of stimulated genes. This response is modeled statistically thanks to a network in which nodes correspond to genes and links correspond to potential interactions. Based on these gene expressions, lots of methods have been proposed to reverse-engineer underlying gene networks.

Some diseases, such as in cancer for example, modify the genetic program. In order to understand which perturbations are linked to these modifications, it is necessary to reverse-engineer the gene network in the different states (eg., healthy/ill). Assuming that the part of the network which is affected by the disease is restricted, a simultaneous reverse-engineering procedure might be necessary.

This would allow to use the common information contained in the part of the network that is not affected by the disease. Futherthemore, this estimation leads to comparable networks.

Keywords. Gene regulatory networks, Variable selection, Clustering.

1 Introduction et motivations

Quand une cellule est stimulée, le programme génique qu'elle contient est activé. Les gènes mis en action apportent alors une réponse concertée au stimulus. Cette réponse est modélisée statistiquement par un réseau dans lequel les noeuds correspondent aux gènes et les liens correspondent à leurs interactions. Depuis l'introduction de technologies à haut débit qui permettent de mesurer simultanément l'expression de milliers de gènes, beaucoup de méthodes statistiques ont été proposées pour l'inférence de ces réseaux de régulation.

Ces méthodes peuvent être regroupées en trois catégories principales. Il y a d'abord les méthodes dites d'interactions, dans lesquelles une mesure de proximité entre les gènes est définie, comme l'entropie dans la méthode ARACNe de Margolin et al. (2006). Parmi ces méthodes, se trouve la classe des GGMs (Graphical Gaussian Models), dans laquelle l'hypothèse de normalité permet de calculer le cœfficient de corrélation partiel linéaire (Chiquet (2011), par exemple). Ces méthodes sont relativement peu coûteuses en temps de calcul, mais elles ne permettent pas de décrire la dynamique des systèmes biologiques. Nous trouvons ensuite les méthodes dites d'optimisation dans lesquelles il convient de distinguer les réseaux booléens d'une part (Liang et al. (1998)), et les réseaux bayésiens d'autre part (Dondelinger et al. (2011)). Dans ces derniers, l'ensemble des gènes régulateurs d'un gène donné est appelé *parents*. Des probabilités *a priori* de chaque gène (sachant ses parents) sont alors définies. Par la formule de Bayes, nous cherchons alors la structure de réseau qui maximise la probabilité *a posteriori* (sachant les valeurs observées pour les expressions de gènes). Ces méthodes sont particulièrement efficaces dans l'inférence de réseaux de gènes et leur intérêt majeur est de pouvoir distinguer les interactions directes de celles qui sont indirectes (grâce au conditionnement par rapport aux parents). Ces méthodes, dans lesquelles un algorithme de recherche des réseaux possibles est souvent nécessaire, ne sont pas adaptées aux réseaux contenant plusieurs centaines de gènes. Enfin, nous avons les méthodes basées sur des équations différentielles ou des régressions, dans lesquelles des techniques spécifiques doivent être utilisées, du fait que le nombre d'observations est souvent largement inférieur au nombre de variables (les gènes). Vu sous cet angle, le problème peut se poser sous la forme d'un choix de variables. L'approche la plus courante consiste à pénaliser l'estimation des paramètres dans la régression linéaire, comme dans Gustafsson et al. (2009, 2010). Ces méthodes sont quant à elles particulièrement bien adaptées dans le cadre d'inférence de réseaux de grande taille.

Certaines maladies comme le cancer affectent le programme génique. Il est alors intéressant de chercher à inférer le programme génique des individus sains et des patients malades. Il est légitime de supposer que seule une partie du réseau de gènes soit altérée d'un état à l'autre ; par conséquent, l'estimation simultanée du réseau des individus sains et des patients malades permettrait de prendre en compte l'information commune entre

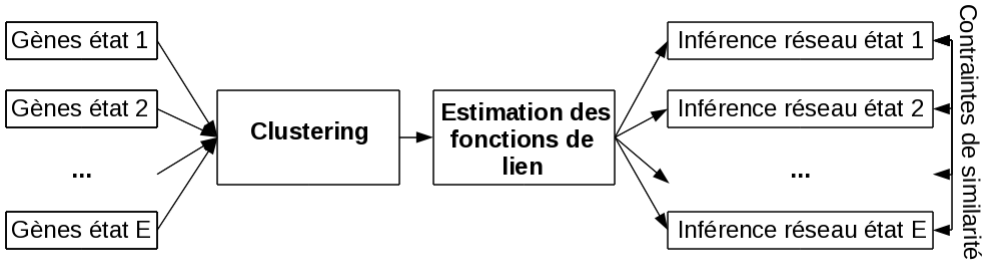

Figure 1: Principe de la méthode proposée pour inférer des réseaux de gènes provenant de plusieurs états

les différents états. Dans les méthodes présentées ci-dessus, seule Dondelinger et al. (2011) permet de prendre en compte cette problématique grâce à un réseau dynamique bayésien, dans le cadre de réseaux de taille limitée. Aussi, nous proposons ici une nouvelle méthode permettant l'estimation simultannée de larges réseaux de gènes issus de multiples états.

Cette méthode, développée ci-dessous, se base sur une régression linéaire précédée d'une étape de clustering ; elle s'inspire d'un premier modèle proposé dans Vallat et al. (in prep) et exposé lors du second colloque international BIO-SI en biostatistique à Rennes. Cette méthode se décompose, comme montré dans la Figure 1, en plusieurs étapes. Un clustering est d'abord réalisé afin de regrouper les gènes ayant une expression similaire au cours du temps. Nous chercherons ensuite des fonctions de liens entre les gènes de différents clusters. Ensuite, nous inférerons les différents réseaux de gènes, en les contraignant à rester similaires dans les parties de réseaux inchangées par la maladie.

2 Etape 1 : Clustering

Supposons que nous disposons d'observations provenant de N gènes, mesurés chacun dans E états, sur T temps de mesure et sur P patients (considérés ici comme des répétitions indépendantes). Ainsi, X_{netp} correspond à l'expression du gène $n \in 1, ..., N$ mesuré pour l'état $e \in 1, ..., E$, au temps $t \in 1, ..., T$ et sur le patient $p \in 1, ..., P$. Nous noterons $\boldsymbol{X}_{net.} = (X_{net1}, ..., X_{netP})'$ le vecteur pour le gène, l'état et le temps fixés ; cette notation est valable quelques soient la place et le nombre de coordonnées remplacées par des points.

Comme montré dans la Figure 1, les expressions de gènes issus des différents états sont mélangées pour le clustering ; autrement dit, notre liste de vecteurs sur laquelle sera appliqué le clustering sera :

$$\{\boldsymbol{X}_{ne..}\}_{n,e}\,.$$

Les deux méthodes les plus courantes pour faire du clustering d'expression de gènes sont les cartes auto-organisatrices (SOMs) et la méthode k-means. Supposons que nous cherchons à classer les gènes en C clusters. Pour prendre en compte les différents états nous pouvons modifier l'algorithme k-means (référence) en cherchant à minimiser la fonction suivante :

$$J(U,V) = \sum_{c=1}^{C} \sum_{e=1}^{E} \sum_{n=1}^{N} \sum_{p=1}^{P} \left(\alpha_n \mu_{nec} + (1-\alpha_n)\gamma_{nc}\right) \left(\boldsymbol{X}_{ne.p} - \boldsymbol{V}_c\right)^2 \tag{1}$$

où $U = \{\mu_{nec}, \gamma_{nc}\}_{n,e,c}$ et $V = \{\boldsymbol{V}_c\}_c$, contiennent les profils temporels (vecteurs de longueur T) pour chaque cluster. Nous rajoutons les contraintes $\sum_c \mu_{nec} = \sum_c \gamma_{nc} = 1$. L'ensemble $\{\alpha_n\}_n$ est constitué de constantes fixées *a priori*, comprises entre 0 et 1. Plus α_n est petit, et plus les différents états d'un même gène n seront contraints d'appartenir au même cluster. Une manière de fixer les α_n est de comparer la variabilité de l'expression des patients entre les différentes conditions et à l'intérieur d'une condition donnée :

$$\alpha_n = \max(1 - \alpha'_n, 0)$$

avec :

$$\alpha'_n = \frac{(E-1)\,P}{P-1} \times \frac{\displaystyle\sum_{e=1}^{E} \sum_{p=1}^{P} \sum_{\substack{p'=1 \\ p'>p}}^{P} \left(\boldsymbol{X}_{ne.p} - \boldsymbol{X}_{ne.p'}\right)^2}{\displaystyle\sum_{e=1}^{E} \sum_{\substack{e'=1 \\ e'>e}}^{E} \sum_{p=1}^{P} \sum_{p'=1}^{P} \left(\boldsymbol{X}_{ne.p} - \boldsymbol{X}_{ne'.p'}\right)^2}.$$

Pour minimiser la fonction J de l'équation (1), il faut procéder à un algorithme itératif. L'initialisation, qui détermine l'ensemble V des réprésentants de chaque cluster peut se faire en effectuant une analyse en composantes principales.

Etape 2 : Estimation des fonctions de liens

Nous cherchons maintenant un ensemble de fonctions $\{f_{ij}\}$, $1 \leqslant i, j \leqslant c$, $i \neq j$ qui décrit comment un élement du cluster c_1 (ie., un gène dans un état donné, noté $\boldsymbol{X}_{ne..}$) agit sur un élément du cluster c_2. Nous supposerons que l'état d'un gène au temps t est entièrement régulé par l'état d'autres gènes au temps $t-1$. Les clusters c_i et c_j étant fixés, nous allons chercher à minimiser :

$$\min_{f_{ij}\in\mathcal{F}}\left\{\sum_{\substack{n_1=1,...,N\\ e_1=1,...,E\\ \boldsymbol{X}_{n_1e_1..}\in c_i}}\sum_{\substack{n_2=1,...,N\\ e_2=1,...,E\\ \boldsymbol{X}_{n_2e_2..}\in c_j}}\sum_{p=1,...,P}\sum_{t=2,...,T}\left\|X_{n_1etp}-f_{ij}\left(X_{n_2e(t-1)p}\right)\right\|_2^2\right\}$$

avec $\|.\|_2$ la norme euclidienne, et $\mathcal{F}$ un espace de fonction de $\mathbb{R}$ dans $\mathbb{R}$ à définir. Cet espace de fonctions peut être général ou peut contenir un ensemble discret de fonctions choisies *a priori* comme dans Gustafsson et al. (2009).

Etape 3 : Inférer le réseau de gènes

À partir de maintenant, nous décomposons le problème en N problèmes indépendants. Supposons que nous voulons connaître les régulateurs du gène 1 *sachant qu'il appartient aux clusters* $c_1, ..., c_E$ *pour les états respectifs* $1, ..., E$.

Nous transformons d'abord tous les régulateurs potentiels du gène 1, par les fonctions estimées ci-dessus. Précisément :

$$\forall n \in 1, ..., N \quad \forall e \in 1, ..., E \quad \forall t \in 1, ..., T \quad \widetilde{X}_{netp} = f_{cl(n,e)c_e}\left(X_{npet}\right)$$

où $cl(.,.)$ est la fonction qui à un gène et à un état associe son cluster pour l'état en question. Ensuite, nous minimisons :

$$\min_{\{\beta_{n,e}\}\in\mathbb{R}}\left\{\sum_{e=1}^{E}\sum_{p=1}^{P}\sum_{t=2}^{T}\left\|\boldsymbol{X}_{1etp}-\sum_{n=1}^{N}\beta_{n,e}\widetilde{\boldsymbol{X}}_{ne(t-1)p}\right\|_2^2+\sum_{e=1}^{E}\sum_{n=1}^{N}\rho(|\beta_{n,e}|\,|\gamma,\lambda_1)\right.$$
$$\left.+\lambda_2\sum_{e=1}^{E}\sum_{n=1}^{N}\sum_{e'=1}^{E}\sum_{n'=1}^{N}|\theta_{nen'e'}|\left(\beta_{n,e}-sgn\left(\theta_{nen'e'}\right)\beta_{n',e'}\right)^2\right\}$$

- γ, λ_1 et λ_2 sont des paramètres à estimer par validation croisée,
- $$\rho(t\,|\gamma,\lambda_1) = \lambda_1\int_0^t\left(1-\frac{x}{\gamma\lambda_1}\right)_+dx$$
- $\theta_{nen'e'}$ reflète la proximité *a priori* du paramètre $\beta_{n,e}$ et $\beta_{n',e'}$. Nous pouvons par exemple choisir :
$$\theta_{nen'e'} = \text{corr}\left(\boldsymbol{X}_{ne..}, \boldsymbol{X}_{n'e'..}\right)$$
- $sgn(.)$ est la fonction qui à un nombre associe son signe.

La fonction ρ, présentée dans Zhang (2010), sert à sélectionner les variables dans la régression linéaire. Contrairement à la régression Lasso, dans laquelle tous les termes sont affectés également par la pénalité (introduisant ainsi du biais dans la méthode), les termes supérieurs à $\gamma\lambda_1$ ne seront pas affectés (en effet : $x > \gamma\lambda_1 \Rightarrow \rho'(x) = 0$). La contrainte L_2 a été étudiée dans Huang (2011).

L'ensemble $\{\beta_{n,e} \neq 0\}_n$, pour $e = 1, ..., E$ fixé, réprésente les régulateurs du gène 1 dans l'état e. Précisément, $\beta_{n,e}$ représente l'intensité de l'action du gène n sur le gène 1 dans l'état e.

Bibliographie

[1] Chiquet J. (2011), *Réseaux biologiques*, SMF Gazette, No. 130, 76–82.
[2] Dondelinger F., Husmeier D. et Lèbre S. (2011), *Dynamic Bayesian networks in molecular plant science: inferring gene regulatory networks from multiple gene expression time series*, Euphytica, 1–17.
[3] Gustafsson M., Hornquist M., Lundstrom J., Bjorkegren J., et J. Tegner J. (2009). *Reverse engineering of gene networks with LASSO and nonlinear basis functions*, Annals of the New York Academy of Sciences, Vol. 1158, No. 1, 265–275.
[4] Gustafsson M. et M. Hornquist (2010),*Gene Expression Prediction by Soft Integration and the Elastic Net-Best Performance of the DREAM3 Gene Expression Challenge*, PLoS One, Vol. 5, No. 2, e9134.
[5] Huang J., Ma S., Li H., et Zhang C. (2011), *The sparse Laplacian shrinkage estimator for highdimensional regression*, The Annals of Statistics, Vol. 39, No. 4, 2021–2046.
[6] Liang S., Fuhrman S., et Somogyi R. (1998),*Reveal, a general reverse engineering algorithm for inference of genetic network architectures*, Pacific symposium on biocomputing, Vol. 3, 18–29.
[7] Margolin A., Nemenman I.,. Basso K, Wiggins C.,. Stolovitzky G, Favera R., et Califano A. (2006), *ARACNE : an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context*, BMC Bioinformatics, Vol. 7, S7.
[8] Steinhaus H. (1956), *Sur la division des corps matériels en parties*, Bull. Acad. Polon. Sci., Vol. 1, 801–804.
[9] Vallat L., Kemper C., Jung N., Pocheville A, Maumy-Bertrand M, Bertrand F., Meyer N., Bahram S., Fisher J. et Gribben J. (in prep), *Predicted intervention in a cancer cell genetic program.*
[10] Zhang C. (2010), *Nearly unbiased variable selection under minimax concave penalty*, The Annals of Statistics, Vol. 38, No. 2, 894–942.