HAL
archives-ouvertes.fr

Inférence sur réseaux géniques par Analyse en Facteurs

Yuna Blum, Chloé Friguet, Sandrine Lagarrigue, David Causeur

To cite this version:

Yuna Blum, Chloé Friguet, Sandrine Lagarrigue, David Causeur. Inférence sur réseaux géniques par Analyse en Facteurs. 42èmes Journées de Statistique, 2010, Marseille, France, France. 2010. <inria-00494802>

HAL Id: inria-00494802
https://hal.inria.fr/inria-00494802

Submitted on 24 Jun 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Inférence sur réseaux géniques par Analyse en Facteurs

Yuna Blum[1,2,3], Chloé Friguet[1], Sandrine Lagarrigue[2,3] & David Causeur[1]

[1] *Laboratoire de Mathématiques Appliquées - Agrocampus Ouest, Rennes*
[2] *INRA, UMR598 Génétique Animale, Rennes*
[3] *Agrocampus Ouest, UMR598 Génétique Animale, Rennes*

Résumé : La technologie des puces à ADN permet l'analyse simultanée du niveau d'expression de plusieurs milliers de gènes. Un des enjeux de l'analyse de ce type de données est de comprendre la structure de dépendance, qui rend compte des relations biologiques entre les gènes. En particulier, on s'intéresse ici à la modélisation du réseau de régulation des gènes impliqués dans le contrôle d'un caractère phénotypique. Dans un premier temps, on définit un cadre général pour la prise en compte de la dépendance par l'identification de facteurs latents, modélisant la variation commune à l'ensemble des gènes. On montre que l'introduction de ces facteurs dans les procédures d'analyse différentielle en améliore la puissance ainsi que la stabilité des taux d'erreurs. De plus, dans le contexte des modèles graphiques gaussiens pour la modélisation des réseaux d'interactions entre gènes, on présente une méthode d'estimation des corrélations partielles s'appuyant sur la réduction de la dimension des données par les variables latentes. La méthode est illustrée par son application à une étude visant à identifier les gènes impliqués dans le métabolisme des lipides chez le poulet (UMR INRA Génétique Animale de Rennes).

Mots-clés : Procédures de tests multiples, Analyse en facteurs, Modélisation de la dépendance, Réseau génique, Modèle graphique.

Abstract : Microarray technology allows the simultaneous analysis of thousands of genes within a single experiment. The analysis of such data aims at understanding the dependence structure due to the genes biological functions, by gene network reconstruction. We focus here on modeling the network of genes implied in the biological control of a given phenotypic character. A general framework for multiple testing dependence is defined, considering a factor model for the correlation matrix through latent variables modeling the common variability. It is shown to improve power and error rates stability in multiple testing procedures. Moreover, in the context of gaussian graphical models, we use the estimated correlation matrix obtained from the factor model to infer a gene regulatory network. This method is illustrated thanks to a study that aims at identifying the genes implied in the lipid metabolism (UMR INRA Génétique Animale de Rennes).

Key words : Multiple testing procedures, Factor Analysis, Dependence modeling, Gene networks, Graphical models.

Sélection de gènes et caractérisation fonctionnelle

La première étape de l'analyse consiste à sélectionner les gènes liés au caractère d'intérêt. La réponse statistique à cette question biologique passe par la mise en oeuvre d'une procédure de tests multiples contrôlant le taux moyen de gènes déclarés à tort liés au caractère phénotypique (Benjamini & Hochberg, 1995). De nombreuses études récentes ont montré que la dépendance entre gènes pouvaient conduire à une très forte variabilité des taux d'erreurs de ces procédures. Leek et Storey (2008) proposent un cadre général d'étude de la dépendance basé sur la proposition suivante :

Proposition [voir Leek et Storey (2008)]

Soit $\varepsilon^{(k)} = Y^{(k)} - \mathbb{E}(Y^{(k)}|x)$ l'erreur résiduelle du modèle de régression de l'expression du $k$ème gène sur la covariable x mesurant le caractère d'intérêt. S'il n'existe aucune fonction g mesurable telle que $\varepsilon^{(k)} = g(\varepsilon^{(1)}, \ldots, \varepsilon^{(k-1)}, \varepsilon^{(k+1)}, \ldots, \varepsilon^{(m)})$ presque sûrement, alors il existe un vecteur aléatoire $Z = (Z_1, Z_2, \ldots, Z_q)$, avec $0 \leq q \leq m$ et, pour tout $k = 1, \ldots, m$, il existe des vecteurs b_k tels que $\varepsilon^{(k)} = b_k' Z + e^{(k)}$, où $e = (e^{(1)}, \ldots, e^{(m)})'$ est un vecteur aléatoire dont les composantes sont indépendantes.

Les variables Z introduites dans la proposition ci-dessus peuvent être vues comme des facteurs de variations communes à l'ensemble des données. De ce point de vue, la décomposition de la variabilité proposée par Leek and Storey (2008) s'apparente à un modèle d'analyse en facteurs. Connue des psychométriciens et des sociologues comme une technique de réduction de la dimension, l'Analyse en Facteurs est par ailleurs apparue récemment comme une technique d'analyse de la dépendance des données en grande dimension issues des expérimentations à haut-débit comme les biopuces (Pournara and Wernisch, 2007, Kustra et al., 2006). Friguet *et al.* (2009) proposent une méthode d'estimation de ce modèle en grande dimension et présentent une procédure de tests multiples basée sur une estimation conditionnelle du taux de faux positifs, ce qui améliore la performance par rapport aux méthodes classiques.

Cette méthode, dénommée FAMT (Factor Analysis for Multiple Testing) est ici appliquée à des données générées pour la détection de QTL contrôlant la variabilité du gras abdominal chez deux lignées divergentes de poulets (Blum *et al.*, 2010). Des mesures de $m = 11213$ expressions hépatiques ont été réalisées par biopuces sur $n = 45$ poulets. La masse de gras abdominal est la variable explicative X de l'étude, il s'agit ici d'une variable quantitative. L'étude des facteurs de variabilité commune permet d'expliquer une forte dépendance par certaines caractéristiques du dispositif expérimental. En effet, l'échantillon est constitué de poulets ayant le même père et pour certains la même mère. Par ailleurs, une structuration de la variabilité par lots d'éclosion est également révélée par l'interprétation des facteurs. La méthode FAMT permet ici de détecter deux fois plus de gènes corrélés au caractère d'intérêt par rapport à une approche classique.

Dans un premier temps, on confirme par l'analyse exploratoire des données d'expressions la pertinence d'un point de vue biologique de la liste de gènes corrélés au caractère gras

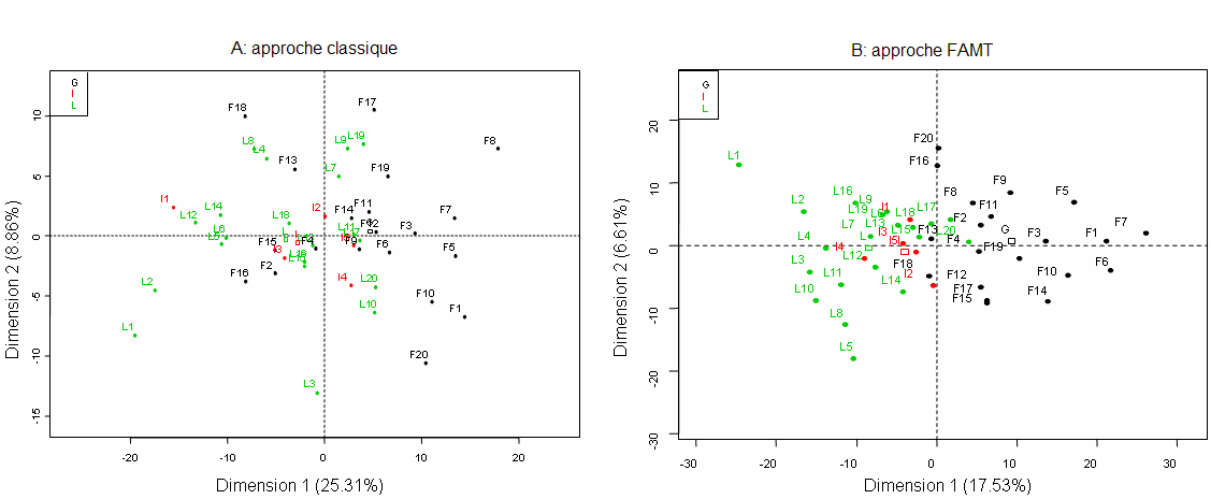

Figure 1: ACP menées respectivement sur les gènes détectés par l'approche classique (A) et par FAMT (B) (nuages des individus) Vert: individus maigres - Rouge: individus intermédiaires - Noir: individus gras

trouvée par FAMT (voir figure 1). La comparaison des premiers plans factoriels issus des ACP menées respectivement sur les gènes détectés par l'approche classique (A) et par FAMT (B) montre nettement une meilleure séparation des individus selon leur quantité de gras avec la liste de gènes (B).

Par ailleurs, des tests d'enrichissement, basés sur un test exact de Fisher comme proposé par Man *et al.* (2000), sont utilisés pour caractériser les propriétés fonctionnelles de l'ensemble des gènes sélectionnés, en utilisant les informations disponibles dans les bases de données de termes fonctionnels (Gene Ontology, KEGG). Ici, seule la liste (B) permet de mettre en évidence des termes biologiques lié au métabolisme des lipides (processus de biosynthèse des hormones stéroïdiennes).

La pertinence de la méthode FAMT pour la détection de gènes différentiellement exprimés dans un contexte biologique donné est confirmée.

On considère maintenant le sous-ensemble de $p = 688$ gènes parmi les m gènes initiaux, qui sont liés au caractère d'intérêt.

Inférence sur réseaux géniques

Un des enjeux majeurs en génomique est la recherche de liens fonctionnels entre les gènes : on cherche à comprendre la logique des régulations entre les gènes déclarés comme étant en lien avec le caractère d'intérêt, et par extension, celle des processus biologiques associés. On appelle réseau de gènes un réseau dont les noeuds représentent les gènes et les arêtes les régulations ayant lieu entre les gènes. Les modèles graphiques permettent la représentation

des relations de dépendances conditionnelles entre différentes variables aléatoires. On va donc utiliser cette méthode pour modéliser les réseaux de gènes, en inférant à partir de l'information apportée par les données transcriptomiques. Nous choisissons d'inférer à partir de modèles graphiques gaussiens. Cette méthode nécessite l'estimation de la matrice des corrélations partielles, puisqu'on s'intéresse aux interactions directes entre les gènes. Or, en grande dimension, cette étape est sujette à une forte instabilité. Nous proposons d'utiliser l'espace de dimension réduite décrit par les facteurs de variabilité commune pour améliorer l'estimation de cette matrice. Nous étudions les propriétés de cette méthode par comparaison avec celle introduite par Schäfer et Strimmer, inspirée d'une procédure de "shrinkage" (Schäfer et Strimmer, 2005). Pour chacune des méthodes d'estimation de Σ, nous construisons le réseau de gènes associé à $\hat{\Sigma}^{-1}$ puis nous analysons la pertinence des régulations trouvées à l'aide des annotations géniques disponibles.

Bibliographie

[1] Benjamini, Y. et Hochberg, Y. (1995). Controlling the false discovery rate: a pratical and powerful approach to multiple testing, Journal of the Royal Statistical Society, B 57, 289-300.

[2] Blum, Y., Le Mignon, G., Lagarrigue, S. et Causeur, D. (2010). A factor model to analyze heterogeneity in gene expression, BMC Bioinformatics, submitted.

[3] Friguet, C., Kloareg, M. et Causeur, D. (2009). A Factor Model Approach to Multiple Testing Under Dependence, Journal of the American Statistical Association, 104:488, 1406-1415.

[4] Kustra, R., Shioda, R. et Zhu, M. (2006). A factor analysis model for functional genomics, BMC Bioinformatics, 7, 216.

[5] Leek, J. T. et Storey, J. (2008). A general framework for multiple testing dependence, Proceedings of the National Academy of Sciences, 105, 18718-18723.

[6] Man, M, Wang, X et Wang, Y (2000). POWER-SAGE: comparing statistical tests for SAGE experiments, Bioinformatics, 16, 953-959.

[7] Pournara, I. et Wernish, L. (2007). Factor analysis for gene regulatory networks and transcription factor activity profiles, BMC Bioinformatics, 8, 61.

[8] Rubin, D. B. et Thayer, D. T. (1982). EM Algorithms for ML Factor Analysis,Psychometrika, 47, 69-76.

[9] Schäfer, J. et Strimmer, K. (2005). A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics, Statistical Applications in Genetics and Molecular Biology, 4:32.