L'analyse d'un réseau de co-expression génique met en valeur des groupes fonctionnels homogènes et des gènes importants relatifs à un phénotype d'intérêt

Nathalie Villa-Vialaneix[1,2] & Laurence Liaubet[3] & Thibault Laurent[4] & Adrien Gamot[5] & Pierre Cherel[6] & Magali SanCristobal[3]

[1] *Institut de Mathématiques de Toulouse (UMR5219), Université de Toulouse, 118 route de Narbonne, F-31062 Toulouse cedex 9, France*

[2] *Université de Perpignan Via Domitia, IUT, Département STID, Domaine Universitaire d'Auriac, Avenue du Dr Suzanne Noël, F-11000 Carcassonne, France*

[3] *INRA de Toulouse, Laboratoire de Génétique Cellulaire (UMR444), BP 52627, F-31326 Castanet Tolosan cedex, France*

[4] *Toulouse School of Economics, Université Toulouse 1, Manufacture des Tabacs, 21 allées de Birenne, F-31000 Toulouse, France*

[5] *Laboratoire de Biologie Moléculaire Eucaryote (UMR5099), Université Toulouse III, Bâtiment IBCG, 118 route de Narbonne, F-31062 Toulouse cedex 9, France*

[6] *Hendrix Genetics RTC, 100 avenue Denis Papin, F-45808 St Jean en Braye Cedex, France*

Résumé

Cet article présente l'analyse d'un réseau de co-expression entre gènes dont la particularité est d'être régulés génétiquement. Cette étude est menée selon deux axes : une classification des gènes impliqués dans le réseau permet de mettre en valeur des groupes fonctionnels homogènes. Par ailleurs, une analyse conjointe du réseau et d'un phénotype d'intérêt permet de mettre en évidence des gènes candidats importants.

Mots clé : réseau de co-expression génique ; eQTL ; phénotype ; classification ; modularité ; recuit simulé ; diagramme de Moran ; points influents

Abstract

Focusing on genes that are genetically regulated, a gene co-expression network is studied following two purposes: first, the genes are clustered into dense groups that appear to have a great functional homogeneity. Then, jointly studying the network structure and a phenotype of interest, candidate key genes are extracted.

Keywords: co-expression network; eQTL; phenotype; clustering; modularity; simulated annealing; Moran's plot; influential observations

1 Introduction

Nous présentons, dans cet article, une étude de l'expression d'un groupe de gènes contrôlés génétiquement. Cette étude est menée par le biais de la définition puis de l'analyse d'un réseau de co-expression génique et structurée selon deux axes d'intérêt. D'un côté, nous montrons qu'une méthode de classification de sommets dans un graphe permet d'isoler des groupes de gènes homogènes d'un point de vue fonctionnel. L'utilisation d'une telle approche donne des informations aux biologistes sur le rôle de gènes encore inconnus. Par ailleurs, nous expliquons comment l'utilisation d'outils issus de la statistique spatiale permet d'intégrer des informations sur la structure de corrélation des gènes et d'autres relatives à un phénotype d'intérêt. Cela nous conduit à mettre en valeur des gènes potentiellement importants pour ce phénotype.

L'article est structuré de la manière suivante : dans la Section 2, nous présentons les données et le réseau utilisé dans cette étude. La Section 3 présente les résultats de l'analyse du réseau et la Section 4 donne les conclusions de l'étude.

2 Données et définition du réseau de co-expression génique

Les données utilisées dans l'étude présentée ici ont été obtenues à partir de 56 cochons d'une même famille F2. L'expression de gènes a été extraite sur le muscle de la longe des animaux (*longissimus dorsi*) à partir d'une biopuce 9K (GEO numéro d'accession GPL3729). Le protocole d'hybridation et de traitement des données transcriptome (Ferré et al. (2007) ; Lobjois et al. (2008)) a permis l'identification de 2 464 gènes exprimés. A partir des mêmes animaux, l'ADN génomique a été extrait et 170 microsatellites couvrant les 18 autosomes avec un espacement moyen de 17 cM. Les analyses ont été effectuées par SAGA LICOR logiciel. Les animaux F2, leurs parents et grands-parents, ont tous été génotypés et la ségrégation mendélienne a été vérifiée. Les cartes génétiques ont été reconstruites avec le logiciel CRIMAP (Green, 1992). Les matrices de relation IBD (Identity By Descent) ont été estimées tous les 2 cM à l'aide du logiciel LOKI 2,5 (Heath, 1997) et la variance a été estimée à l'aide du maximum de vraisemblance résiduelle (REML) avec la version 2.0 du logiciel ASREML (Gilmour et al., 2006). Ainsi les variations d'expression de 272 gènes ont été identifiées comme étant régulées génétiquement par au moins un locus (ou eQTL, expression Quantitative Trait Locus).

À partir de l'expression des 272 gènes ainsi sélectionnés, un réseau de co-expression génique a été défini en utilisant un modèle graphique Gaussien (Schäfer and Strimmer (2005), parmi les nombreuses méthodes d'inférence de réseau qui existent : voir De Smet and Marchal (2010) pour une revue récente sur la question). 4 000 échantillons bootstrap, chacun de taille 20, ont été utilisés pour estimer les corrélations partielles entre l'expression de toutes les paires de gènes. Le réseau finalement obtenu est modélisé par un graphe de :

- 272 sommets, chacun représentant un gène ;
- 4 690 arêtes entre paires de sommets pour lesquelles la corrélation partielle de l'expression était significativement non nulle (test basé sur une approche bayésienne : Schäfer and Strimmer (2005)) ;
- les arêtes ont été pondérées par la valeur absolue de l'estimation de la corrélation partielle. Les poids des arêtes du graphe sont symétriques et positifs.

3 Analyse du réseau de co-expression génique

3.1 Classification des gènes

Utilisant la structure de corrélation entre expressions, les gènes ont été classés à partir d'un algorithme de classification de sommets (voir Fortunato (2010) pour une revue des méthodes de classification de sommets dans un graphe). L'objectif est d'obtenir des groupes de gènes fortement connectés (c'est-à-dire pour lesquels les corrélations entre expressions sont fortes). De manière plus précise, une mesure de la qualité de la classification de sommets dans un graphe, la modularité, a été optimisée par un algorithme de recuit simulé comme suggéré dans Reichardt and Bornholdt (2006) ou Villa-Vialaneix et al. (2011). La classification obtenue contient 7 classes et conduit à la représentation simplifiée du graphe donnée dans la Figure 1. Cette classification a été confrontée au logiciel

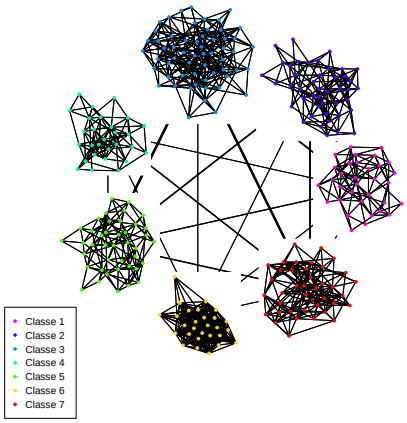

FIGURE 1 – Représentation des classes ainsi que de leurs relations : l'épaisseur des arêtes entre deux classes est proportionnelle à la somme des poids entre les sommets des classes.

"Ingenuity Pathways Analysis" (IPA, https://analysis.ingenuity.com/pa/) pour étudier la pertinence biologique des réseaux de gènes extraits. Les résultats sont donnés dans

Classe	Nombre de gènes dans la classe	Nombre de gènes éligibles pour Ingenuity	Proportion de gènes présents dans un même réseau biologique	Fonctions biologiques correspondantes
1	33	21	94%	Métabolisme des acides gras et des acides nucléiques
2	44	31	96%	Tissus de connexion et morphologie cellulaire
3	58	34	59%	Synthèse protéique et développement musculaire
4	28	27	95%	Prolifération et mort cellulaire
5	41	36	89%	Biochimie et transport moléculaire
6	28	44	63%	Développement et fonction musculaire
7	40	30	75%	Métabolisme des acides aminés et des carbohydrates

TABLE 1 – Résumé de la confrontation des classes trouvées dans le réseau de co-expression génique avec un logiciel d'analyse fonctionnelle.

la Table 1. On remarque une très grande homogénéité fonctionnelle des groupes mis en évidence (le pourcentage de gènes impliqués dans le même réseau fonctionnel est toujours très élevé, sauf pour la classe 3 qui est la plus grande et la moins dense des classes mises en évidence). Ceci plaide en faveur de la pertinence biologique des groupes et peut permettre, par analogie, au biologiste de formuler des hypothèses sur la fonction biologique de gènes inconnus. Dans un contexte où une bonne partie de l'information génétique n'est pas connue, ces hypothèses sont précieuses.

3.2 Lien entre co-expression génique et phénotype d'intérêt

Cette section présente un travail reliant la structure du réseau de co-expression génique à un phénotype d'intérêt impliqué dans la qualité de la viande : le pH. Pour cela, les corrélations partielles entre expression des gènes et valeur du pH de la viande ont été estimées. Les valeurs de ces corrélations partielles sont représentées sur la Figure 2 (gauche) avec des niveaux de couleurs correspondant à la corrélation partielle entre l'expression du gène représenté par le sommet et le pH. L'ajout de cette information supplémentaire peut être modélisée par un réseau dont les sommets sont étiquetées (ici, par la valeur de la corrélation partielle entre l'expression du gène et le pH) : Laurent and Villa-Vialaneix (2010) proposent l'utilisation d'outils issus de la statistique spatiale pour analyser les relations entre la structure d'un graphe et les valeurs des étiquettes sur ses sommets. En particulier, le diagramme de Moran (voir Figure 2, à droite) représente la valeur moyenne des étiquettes des voisins d'un sommet en fonction de la valeur de l'étiquette de ce sommet.

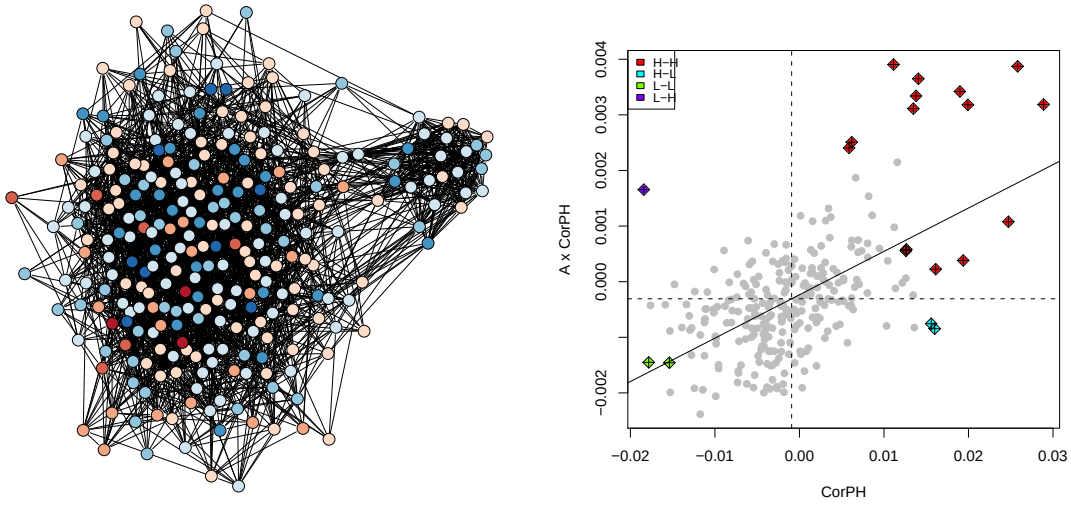

FIGURE 2 – *À gauche* : Niveau des corrélations partielles entre expression des gènes et pH : les sommets bleus sont des gènes dont l'expression est corrélée négativement avec le pH, les sommets rouges des gènes dont l'expression est corrélée positivement. L'intensité de la couleur correspond à la force de la corrélation. *À droite* : Diagramme de Moran des corrélations partielles avec le pH dans le réseau de co-expression génique.

Dans notre cas, ce diagramme présente une tendance linéaire et permet de mettre en valeur des points *influents* : ces points sont ceux qui influencent fortement la valeur de la droite de tendance du diagramme de Moran et dont le comportement peut donc être considéré comme atypique, au sein du réseau, pour la corrélation partielle au pH (voir Belsley et al. (1980) ou Cook and Weisberg (1982) pour plus de détails). La plupart des gènes en rouge, par exemple, correspondent à des gènes dont la corrélation partielle au pH est forte et positive et qui sont entourés de voisins ayant aussi des corrélations partielles avec des valeurs fortes et positives. Beaucoup de ces gènes (10 gènes rouge et le gène violet) se retrouvent dans la classe 4 alors que les autres classes contiennent, au maximum, 5 gènes repérés comme influents. Ainsi, la classe 4 apparaît comme une classe dont la corrélation avec le pH est singulière. Les gènes de cette classe sont des gènes candidats importants pour expliquer le phénotype d'intérêt : ils n'ont pas été sélectionnés par l'approche habituelle consistant à rechercher des différences d'expression selon la valeur du pH car, travaillant sur des gènes dont la particularité est d'être régulés génétiquement, les différences d'expression relatives à un phénotype restent très faibles. De plus, les gènes mis en évidence ne sont pas simplement singuliers par leur relation individuelle au phénotype d'intérêt mais aussi parce que les gènes avec lesquels ils interagissent sont également singuliers.

4 Conclusion

L'utilisation de réseaux est naturelle en biologie puisqu'ils permettent de modéliser des phénomènes de dépendances complexes entre un grand nombre d'objets (ici, des gènes) et donc, de mieux comprendre le système biologique dans son ensemble. Nous avons montré ici que l'utilisation de méthodes statistiques dédiées aux graphes permet de formuler des hypothèses biologiques sur la fonction de certains gènes et de proposer des gènes candidats impliqués dans un phénomène biologique d'intérêt.

Références

Belsley, D., Kuh, E., and Welsch, R. (1980). *Regression Diagnostics*. Wiley, New York.

Cook, R. and Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman and Hall, London.

De Smet, R. and Marchal, K. (2010). Advantages and limitations of current network inference methods. *Nature Reviews Microbiology*, 8:717–729.

Ferré, P., Liaubet, L., Cocordet, D., SanCristobal, M., Uro-Coste, E., Tosser-Klopp, G., Bonnet, A., Toutain, P., Hatey, F., and Lefebvre, H. (2007). Longitudinal analysis of gene expression in porcine skeletal muscle after post-injection local injury. *Pharmaceutical Research*, 24:1480–1489.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486:75–174.

Gilmour, A., Gogel, B., Cullis, P., and Thompson, R. (2006). *ASReml User Guide Release 2.0*. VSN International Ltd, Hemel Hempstead, HP1 1ES, UK.

Green, P. (1992). Construction and comparison of chromosome 21 radiation hybrid and linkage maps using CRI-MAP. *Cytogenetics and Cell Genetics*, 59:122–124.

Heath, S. (1997). Markov chain monte carlo segregation and linkage analysis for oligogenic models. *American Journal of Human Genetics*, 61(3):748–760.

Laurent, T. and Villa-Vialaneix, N. (2010). Analysis of the influence of a network on the values of its nodes: the use of spatial indexes. In *1ère Conférence Modèles et Analyse des Réseaux : Approches Mathématiques et Informatique (MARAMI)*, Toulouse, France.

Lobjois, V., Liaubet, L., SanCristobal, M., Glenisson, J., Feve, K., Rallieres, J., Le Roy, P., Milan, D., Cherel, P., and Hatey, F. (2008). A muscle transcriptome analysis identifies positional candidate genes for a complex trait in pig. *Animal Genetics*, 39(2):147–162.

Reichardt, J. and Bornholdt, S. (2006). Statistical mechanics of community detection. *Physical Review E*, 74(016110).

Schäfer, J. and Strimmer, K. (2005). An empirical bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6):754–764.

Villa-Vialaneix, N., Dkaki, T., Gadat, S., Inglebert, J., and Truong, Q. (2011). Recherche et représentation de communautés dans un grand graphe : une approche combinée. *Document Numérique*, 14(1):59–80.