Cascade : un package R pour étudier la dispersion d'un signal dans un réseau de gènes.

F. Bertrand[a], **N. Jung**[a,b], **M. Maumy-Bertrand**[a], **S. Bahram**[b] and **L. Vallat**[b]

[a] Institut de Recherche en Mathématique Avancées (IRMA)
Laboratoire d'Excellence IRMIA
Université de Strasbourg, 67084 Strasbourg Cedex, France

[b] Laboratoire d'Immunogénétique Moléculaire Humaine,
Institut National de la Santé et de la Recherche Médicale, Unité Mixte de Recherche S1109
Laboratoire d'Excellence Transplantex
Université de Strasbourg, 67085 Strasbourg Cedex, France

Correspondance : njung@math.unistra.fr

Mots clefs : Statistique, Biologie, Lasso, Réseau de régulation génique.

Un réseau de régulation est un outil de modélisation de systèmes complexes particulièrement bien adapté pour étudier les interactions entre des gènes. En effet, certains gènes activés ont la capacité de moduler l'expression (ARN messager) d'autres gènes, formant ainsi un système complexe qui peut-être modélisé par un réseau (orienté ou non) dans lequel les nœuds correspondent aux gènes et les flèches correspondent à l'action d'un gène sur un autre.

Pour inférer le réseau de gènes, il est possible d'étudier le niveau d'expression de ces derniers grâce à des microarrays qui permettent de mesurer la quantité d'ARN messager produite par chaque gène activé. Afin de pouvoir déterminer un lien de causalité, il est important de mesurer l'expression des gènes au cours du temps. Ces réseaux peuvent alors être modélisés sous forme d'interactions en cascade [1] (Figure 1), mais peu d'outils ont été développés à ce jour pour appréhender ces phénomènes.

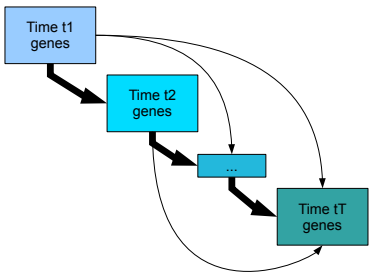

Figure 1: Réseau de gènes en cascade.

Le package R **Cascade** permet la sélection de gènes, l'inférence d'un tel réseau en cascade et la prédiction des effets d'une perturbation d'un ensemble de gènes dans le réseau (adapté de [1]). Une attention particulière a été portée dans la construction de ce package à la réalisation de graphiques facilement compréhensibles et interprétables par les biologistes. Les packages **animation** et **igraph** de R permettent ainsi de visualiser la propagation d'un signal dans le réseau. Vous pouvez trouver à cette adresse[1] un exemple.

[1] `http://www-irma.u-strasbg.fr/~njung/network/network.html`

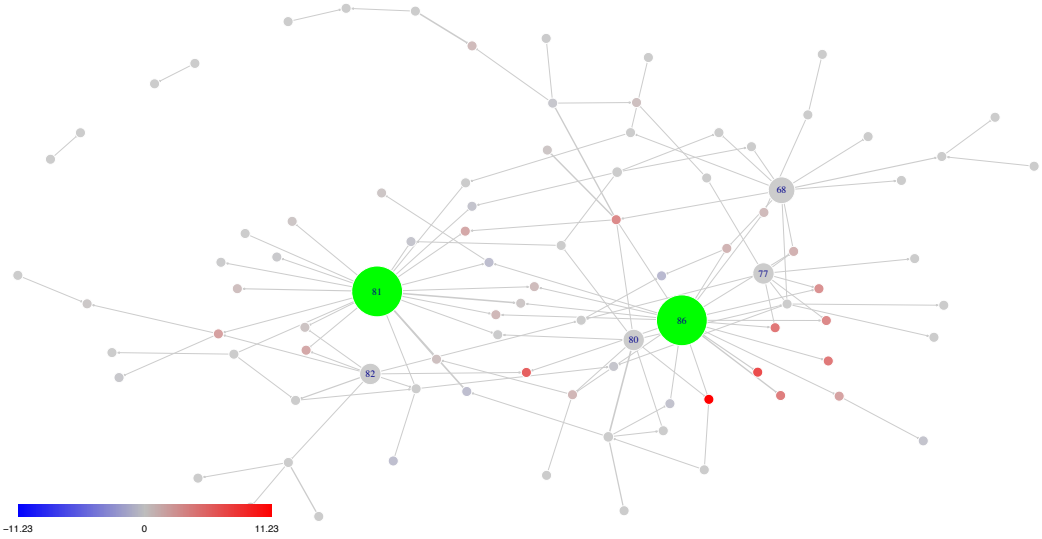

Figure 2: Prédiction de l'impact d'une intervention sur deux gènes du réseau (en vert).

L'étape de sélection des gènes permet de retenir des gènes fortement exprimés et des gènes présentant des cinétiques d'expressions spécifiques. Ceci permet notemment d'enrichir la sélection avec des gènes faiblement exprimés, mais présentant des cinétiques remarquables, en particulier aux temps précoces de la cascade [1]. Pour ce faire, nous avons construit des fonctions qui font une pleine utilisation des possibilités du package R de bioconductor **limma**.

Notre méthode d'inférence est basée sur celle développée dans [1]. C'est un modèle de régression linéaire pénalisé par une contrainte Lasso. De plus, la structure de réseau en cascade (Figure 1) permet d'assigner les gènes à des groupes temporels qui sont ensuite utilisés dans le modèle pour décrire l'action d'un gène sur un autre via les matrices $\boldsymbol{F}$:

$$\boldsymbol{Y} = \sum_{i=1}^{N} \boldsymbol{F}_{m(X_i)m(Y)} \omega_i \boldsymbol{X}_i + \lambda \sum_{i=1}^{N} |\omega_i| + \boldsymbol{\eta}, \tag{1}$$

où $\boldsymbol{Y}$ est le gène régulé et les $\boldsymbol{X}_i$ sont les gènes potentiellement régulateurs, les ω_i déterminent la puissance du lien entre X_i et $\boldsymbol{Y}$, et $m(.)$ est la fonction qui à un gène associe son groupe temporel ; λ est un cœfficient réel estimé par validation croisée déterminant la parcimonie du modèle et $\boldsymbol{\eta}$ est une erreur aléatoire. Plusieurs contraintes sont apportées afin d'assurer une évolution temporelle en cascade (voir [1] pour plus de détails).

A la fin du processus d'inférence, dans lequel chaque gène devient tour à tour gène régulé, un réseau en cascade est obtenu, matérialisé par les cœfficients ω. Cependant, cette méthode infère un grand nombre de lien. Or, il est parfois souhaitable pour le biologiste de réduire ce nombre de lien afin d'obtenir une information plus lisible. Pour cela nous proposons d'appliquer un seuillage sur les cœfficients ω. Notre package permet de voir l'évolution de la topologie du réseau en fonction de ce seuillage (voir exemple à cette adresse)[2]. En considérant la distribution du nombre de liens sortant (voir [2] par exemple), un test a été mis en place pour choisir le seuillage optimal. Des simulations ont par ailleurs confirmé l'intérêt d'un tel choix en montrant que cela permet de réduire le nombre de faux positifs parmi les liens inférés.

[2] `http://www-irma.u-strasbg.fr/~njung/evolution/evol.html`

Une fois le réseau inféré et le seuillage choisi, le résultat obtenu peut être visualisé sous plusieurs formes, dont la plus explicite est sans doute l'animation qui permet de voir un signal se propager dans le réseau [lien]. Pour finir, il est possible de prédire les effets d'une perturbation dans le réseau grâce au modèle donné dans l'équation (1), et de visualiser les changements prédits (voir Figure 2).

En conclusion, le package R **Cascade** a été conçu pour pouvoir être utilisé simplement et il apporte des représentations graphiques qui permettent une interprétation aisée des résultats. Il est disponible sur demande.

Références

[1]Vallat, L., Kemper, C. A., Jung, N., Maumy-Bertrand, M., Bertrand, F., Meyer, N., ... & Bahram, S. (2013). Reverse-engineering the genetic circuitry of a cancer cell with predicted intervention in chronic lymphocytic leukemia. *Proceedings of the National Academy of Sciences*, 110(2), 459-464.

[2]Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439), 509-512.