# RNAseqNet : un package pour l'inférence de réseaux à partir de données RNA-seq

**A. Imbert**[a] et **N. Villa-Vialaneix**[a]
[a]MIAT, Université de Toulouse, INRA
31326 Castanet-Tolosan cedex, France
{alyssa.imbert, nathalie.villa-vialaneix}@inra.fr

**Mots clefs** : RNA-seq, inférence de réseau, individus manquants, imputation multiple hot-deck

La technologie RNA-seq est une approche haut débit permettant de mesurer simultanément l'expression de milliers de gènes dans un échantillon biologique donné. Ces données se présentent sous la forme d'une matrice, $X$, à $n$ lignes ($n$ échantillons biologiques) et $p$ colonnes ($p$ gènes) dans laquelle l'entrée $x_{ij}$ est une donnée de comptage mesurant le nombre de copies d'ARNs messagers correspondant à un gène, $j$, donné, dans l'échantillon $i$. Ces données sont d'un grand intérêt pratique en biologie car elles permettent d'obtenir un instantané du fonctionnement des cellules des échantillons considérés. Parmi les analyses statistiques utilisées pour extraire de l'information biologique de ce type de données, l'inférence de réseau d'interaction de gènes permet de représenter les relations de dépendance entre gènes qui peuvent modéliser les phénomènes de régulation, de co-régulation ou de co-expression au sein de la cellule. Ces approches sont plus difficiles à mettre en œuvre pour les données RNA-seq du fait du caractère discret des données (contrairement au cas gaussien [5], il n'y a pas de cadre formel consensuel pour représenter les relations de dépendances conditionnelles entre gènes) et du faible nombre d'échantillons de la plupart des expériences réalisées (ces données étant très coûteuses à acquérir).
Cette proposition se focalisera sur le package R **RNAseqNet** qui permet l'inférence de réseau à partir de données RNA-seq. En particulier, le package contient l'implémentation (dans la fonction `GLMnetwork`) du modèle graphique log-linéaire de Poisson (LLGM) [1]. Cette approche est fondée sur un modèle linéaire généralisé avec une pénalité type Lasso. **RNAseqNet** inclut le choix du paramètre de pénalisation Lasso par un critère de stabilité, StARs [7], dans la fonction `stabilitySelection`.
Ce modèle (ainsi que d'autres modèles d'inférence de réseaux pour les données issues du séquençage haut-débit) était déjà proposé dans le package **XMRF** [11] mais le package **RNAseqNet** permet aussi d'utiliser de l'information externe pour améliorer la qualité de l'inférence lorsque les données RNA-seq ont été mesurées simultanément avec d'autres données biologiques plus complètes et liées au phénomène étudié. De manière formelle, l'approche, décrite dans [6], suppose que les données d'expression RNA-seq sont contenues dans une matrice $X$ de dimension $n_1 \times p$. Des données auxiliaires ont été obtenues sur les mêmes $n_1$ échantillons ainsi que sur d'autres échantillons. On note $Y$ la matrice de dimension $n \times q$ avec $n > n_1$ contenant ces données, les $n_1$ individus communs avec $X$ correspondant aux $n_1$ premières lignes. Ce problème peut donc être vu comme des valeurs manquantes dans une matrice $\left[\tilde{X}, Y\right]$ de dimension $n \times (p+q)$ dans laquelle, $\tilde{x}_i = \begin{cases} x_i & \forall i = 1, \ldots, n_1 \\ \tilde{x}_i \text{ est manquant} & \forall i \geq n_1 + 1 \end{cases}$. De manière similaire à [10], une approche de type « hot-deck » [2, 3] est utilisée pour imputer des lignes entières dans $X$ en utilisant l'information de proximité entre individus venant de $Y$. Cette méthode a l'avantage de respecter les caractéristiques initiales des données (caractère discret et positivité) ainsi que de conserver la structure de corrélation entre les variables imputées, ce qui est particulièrement important pour l'inférence de réseau. La méthode est mise en œuvre dans un cadre d'imputation multiple [8, 9] qui permet d'observer la stabilité des arêtes inférées.

En résumé, les étapes de la méthode, appelée **hd-MI**, sont les suivantes :

- pour tout individu $i > n_1$ dans $\tilde{X}$, on définit le groupe de donneurs $\mathcal{D}(i)$ comme l'ensemble des individus $i' \leq n_1$ qui sont « similaires » à $i$. Les similarités entre les individus sont calculées sur le jeu de données auxiliaire $Y$ ;
- un individu $i'$ est alors choisi aléatoirement dans $\mathcal{D}(i)$. La ligne entière $i$ de $\tilde{X}$ est imputée avec la ligne $i'$ de $\tilde{X}'$ Cette étape est répétée pour tout $i = n_1 + 1, \ldots, n$ afin d'obtenir un tableau de données complet, $X^*$.

Dans le cadre de l'imputation multiple, cette procédure est répétée $M$ fois pour obtenir $M$ tableaux de données complets, $X^{*,m}$ pour $m = 1, \ldots, M$, à partir lesquels $M$ réseaux sont inférés en utilisant le modèle LLGM. Enfin, une étape d'agrégation combine les différents réseaux en un seul réseau défini de manière à ne conserver que les arêtes les plus stables. De manière plus précise, la fréquence d'apparition d'une arête donnée, $e$, dans les $M$ réseaux est calculée :

$$r(e) = \frac{\text{nombre de fois où l'arête } e \text{ est prédite}}{M}$$

et les arêtes telles que $r(e) \geq r_0$ où $r_0$ est un seuil fixé (de manière typique 90% sont conservées). Dans **RNAseqNet**, la fonction `imputeHD` permet de mettre en œuvre **hd-MI** et la fonction `imputedGLMnetwork` permet d'appliquer **hd-MI** pour l'inférence de réseau avec le modèle LLGM. Enfin, un objet de type `igraph` [4] peut être créé à partir des résultats par la fonction `GLMnetToGraph` : il permet de manipuler et d'analyser facilement le réseau inféré.
Disponible sur le CRAN, le package **RNAseqNet** est accompagné d'une vignette complète illustrant son utilisation sur des données réelles issues du projet GTEx `https://gtexportal.org`.

# Références


[1] G. Allen and Z. Liu. A log-linear graphical model for inferring genetic networks from high-throughput sequencing data. In *Proceedings of IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2012.

[2] R. Andridge and R.J.A. Little. A review of hot deck imputation for survey non-response. *International Statistical Review*, 78(1):40–64, 2010.

[3] S.J. Cranmer and J. Gill. We have to be discrete about this: a non-parametric imputation technique for missing categorical data. *British Journal of Political Science*, 43:425–449, 2012.

[4] G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems, 2006.

[5] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[6] A. Imbert, A. Valsesia, C. Le Gall, C. Armenise, G. Lefebvre, P. Gourraud, N. Viguerie, and N. Villa-Vialaneix. Multiple hot-deck imputation for network inference from RNA sequencing data. *Bioinformatics*, 2018. Forthcoming.

[7] H. Liu, K. Roeber, and L. Wasserman. Stability approach to regularization selection (StARS) for high dimensional graphical models. In *Proceedings of Neural Information Processing Systems (NIPS 2010)*, volume 23, pages 1432–1440, Vancouver, Canada, 2010.

[8] D.B. Rubin. *Multlipe Imputation for Nonresponse in Surveys*. Wiley, 1987.

[9] J.L. Schafer. Multiple imputation: a primer. *Statistical Methods in Medical Research*, 8(1):3–15, 1999.

[10] V. Voillet, P. Besse, L. Liaubet, M. San Cristobal, and I. Gonzáles. Handling missing rows in multi-omics data integration: multiple imputation in multiple factor analysis framework. *BMC Bioinformatics*, 17:402, 2016.

[11] Y. Wan, G. Allen, E. Yang, P. Ravikumar, M. Anderson, and Z. Liu. XMRF: an R package to fit Markov networks to high-throughput genetic data. *BMC Systems Biology*, 10(Suppl 3):69, 2016.