Outils pour l'analyse et la simulation de données RNA-seq

A. Imbert[a] and **N. Villa-Vialaneix**[a]
[a]MIAT, Université de Toulouse, INRA
31326 Castanet-Tolosan cedex, France
nathalie.villa@toulouse.inra.fr

Mots clefs : RNA-seq, inférence de réseaux, simulation

Dans cette proposition de communication, nous présentons les packages R qui permettent d'analyser et de simuler des données issues des technologies de séquençage du transcriptome (RNA-seq). Dans un premier temps, nous ferons un tour rapide des packages qui permettent d'effectuer la normalisation et l'analyse différentielle. Puis, nous discuterons des méthodes pour inférer un réseau de gènes à partir de telles données et les packages R associés. Enfin, nous présenterons quels packages permettent de simuler des réseaux et des données transcriptomiques à partir de ces réseaux.

Données RNA-seq

Les données RNA-seq sont issues d'une technique récente de séquençage à haut débit qui mesure l'abondance de séquences d'ARN pour des milliers de gènes simultanément et permet ainsi de quantifier l'expression de ces gènes. Ces données se présentent sous la forme d'un tableau de comptage de taille $p \times n$, où chaque élément (i, j) correspond au nombre de copies du transcrit du gène j, $j \in \{1, ..., p\}$, pour l'échantillon i, $i \in \{1, ..., n\}$. Les difficultés de modélisation de ces données sont liées à leur caractère discret et au faible nombre d'échantillons disponibles. Plusieurs approches ont été proposées pour les modéliser : des lois de Poisson, des lois de Poisson sur-dispersées ou des lois binomiales négatives.

La première étape pour analyser ce type de données est de corriger les biais techniques en normalisant les données. Cette étape est utile pour répondre à une question souvent importante pour le biologiste, qui est de trouver des gènes différentiellement exprimés entre différentes conditions expérimentales testées. Deux packages sont couramment utilisés pour normaliser les données RNA-seq et réaliser une analyse différentielle. Il s'agit des packages **DESeq2** et **edgeR**, disponibles à partir de Bioconductor. Un autre package, également utilisé pour l'analyse différentielle, est le package **limma** qui utilise des données RNA-seq transformées (voir la section suivante pour une discussions sur la transformation de données RNA-seq).

L'analyse exploratoire de ces données peut également être poursuivie en effectuant une classification des gènes. Une approche fructueuse dans ce domaine utilise un modèle de mélange de lois de Poisson et est implémentée dans le package **HTSCluster**.

Inférence de réseaux RNA-seq

Les biologistes s'intéressent aussi aux liens de régulation entre les gènes. L'inférence de réseau permet de formaliser ces interactions et de les visualiser. Deux stratégies sont possibles pour inférer un réseau de gènes à partir de données RNA-seq.

La première consiste à transformer les données pour que leur distribution approche une loi normale et à utiliser des modèles graphiques gaussiens (GGM). Une des transformations possibles est la transformation box-cox, présente dans le package **MASS**. Une fois les données transformées, il est alors possible d'utiliser des GGM [2, 5, 6]. De nombreux packages proposent des implémentations de ces méthodes pour inférer les réseaux : **GeneNet**, **glasso**, **simone**, **huge**. Une fois le graphe construit, le package **igraph** permet de l'analyser et de le manipuler.

La deuxième stratégie consiste à prendre en compte le caractère discret des données RNA-seq en utilisant des modèles basés sur des lois de Poisson. Il existe plusieurs modèles :

— un modèle graphique log-linéaire de Poisson proposé par [1] : pour prendre en compte

la surdispersion des données RNA-seq, elles sont transformées avec une tansformation puissance $y_{ij} \to g(y_{ij}) = y_{ij}^{\alpha}$, avec $\alpha \in]0, 1]$, implémentée dans le package R **PoiClaClu**. Soit z_j le vecteur transformé des valeurs d'expression pour le gène j des n échantillons, le modèle est le suivant : $p(Z_{ij}|z_{i(-j)}) \sim P(\mu_j)$ avec $\log(\mu_{ij}) = \sum_{j' \neq j} \beta_{jj'} \widetilde{z}_{ij}$ où $\widetilde{z}$ correspond aux données log-transformées et standardisées et $\beta_{jj'}$ à une arête entre j et j' pour les valeurs non nulles du coefficient. Il faut alors estimer les paramètres β_j en utilisant l'algorithme de type LASSO implémenté dans le package R **glmnet** qui permet d'obtenir une solution parcimonieuse. Afin de sélectionner un nombre pertinent d'arêtes, il est possible de rajouter une pénalité de type LASSO dans le modèle et d'utiliser le critère StaRs [4] pour sélectionner le paramètre de régularisation λ optimal. Ce critère, implémenté dans **huge**, n'est pas présent dans **glmnet**.
Enfin, le package **XMRF** propose également d'inférer directement des réseaux en suivant ce modèle.

— un modèle graphique hiérarchique log-normal de Poisson proposé par [3] : l'expression des comptages pour le gène j et l'échantillon $i \in \{1, ..., n\}$ est modélisé par $Y_{ij} \sim \mathcal{P}(\theta_{ij})$ avec $\log(\theta_{ij}) = \sum_{j' \neq j} \beta_{jj'} \widetilde{y}_{ij} + \epsilon_{ij}$ où $\widetilde{y}$ correspond aux données log-transformées et standardisées et $\epsilon_j \sim \mathcal{N}(0, \sigma_j^2 I_n)$. Comme pour le modèle précédent, il faut alors estimer les paramètres β_j en utilisant le package **glmmixedlasso**.

Simulations

Le package **igraph** permet de simuler différents types de réseaux : réseaux aléatoires (Erdös-Renyi), réseaux « scale-free »et les réseaux « small world ». Des packages comme **GeneNet**, ou **simone**, proposent des fonctions pour simuler des réseaux et des données d'expression suivant un GGM associé à un réseau fixé.

La technologie RNA-seq, étant relativement récente, peu de packages proposent de simuler directement des données de comptages à partir d'un réseau donné. Il est possible néanmoins de simuler un réseau et de simuler soi-même avec R des données RNA-seq [1]. Deux packages récents proposent de simuler des données de comptage à partir d'un réseau « scale-free » : **XMRF** et **SynRNASeqNet**.

En conclusion, bien que la technologie RNA-seq soit récente, il existe déjà des packages R permettant d'analyser ces données. Ces packages sont encore en cours de développement. Lors de la présentation, nous montrerons et comparerons quelques exemples de résultats obtenus avec ceux-ci.

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

[3] M. Gallopin, A. Rau, and F. Jaffrézic. A hierarchical poisson log-normal model for network inference from rna-seqrna-seq data. *PLoS ONE*, 2013.

[4] 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.

[5] N. Meinshausen and P. Bühlmann. High di-mensional graphs and variable selection with the lasso. *Annals of Statistic*, 34(3) :1436 :1462, 2006.

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