Méthodes post hoc pour la génomique exploratoire

Stage de Master

Guillermo Durand, Mélina Gallopin

1 Description

Les méthodes d'analyse de données d'expression de gènes RNA-seq sont maintenant bien développées pour les études se focalisant sur une seule espèce. Lorsque l'objectif est de mettre en comparaison les données RNA-seq collectées chez plusieurs espèces, les méthodes d'analyse statistique ne sont pas aussi bien établies, notamment à cause de la difficulté de détection des relations d'orthologie entre les gènes des différentes espèces, de la qualité de l'assemblage des données transcriptomiques, de l'absence de génome de référence pour certaines espèces étudiées, et des relations évolutives entre populations(Bastide, Soneson, et al., 2022). Dans le cas où l'échantillonnage des espèces le permet, les approches de comparaison d'expressions de gènes par paires d'espèces proches sont prometteuses, mais accroissent la multiplicité des tests effectués.

Si des procédures de tests multiples bien connues comme celle de Bonferroni ou celle de Benjamini-Hochberg (Benjamini and Hochberg, 1995) renvoient des ensembles de rejet qui contrôlent des critères d'erreur classiques comme le Family-Wise Error Rate ou le False Discovery Rate, en recherche exploratoire (Goeman and Solari, 2011), on peut s'intéresser à l'approche inverse, c'est-à-dire, laisser l'utilisateur sélectionner un ensemble d'hypothèses qui lui semblent intéressantes, et lui fournir en retour une garantie (sous la forme d'une borne supérieure) sur les faux positifs dans cet ensemble. C'est ce que l'on entend par approche post hoc. Mathématiquement, cela se traduit par la construction d'une fonction aléatoire $\hat{V}$ agissant sur les sous-parties de $\{1, \dots, m\}$, où m est le nombre d'hypothèses testées, vérifiant le contrôle suivant :

$$\mathbb{P}\left(\forall S \in \{1, \dots, m\}, |S \cap \mathcal{H}_0| \leq \hat{V}(S)\right) \geq 1 - \alpha$$

où $\alpha \in]0, 1[$ est un niveau de confiance donné, typiquement 5%, et $\mathcal{H}_0$ désigne l'ensemble des vraies hypothèses nulles (autrement dit des faux positifs si on sélectionne ces hypothèses). $\hat{V}$ fournit donc un contrôle uniforme sur le nombre de faux positifs ou, de façon équivalente, sur le False Discovery Proportion (FDP), de tous les sous-ensembles d'hypothèses possibles. On appellera $\hat{V}$ une borne post hoc.

L'objectif de ce stage est d'appliquer de nouvelles approches de bornes post hoc Blanchard, Neuvial, et al. (2020) and Durand, Blanchard, et al. (2020) sur le jeu de données réels produits par García de la Torre, Majorel-Loulergue, et al. (2021). Ces nouvelles procédures post hoc prendront en compte la structure des données, notamment celle induite par l'arbre phylogénétique des espèces. On s'intéressera en particulier à la comparaison de la performance de différentes méthodes de construction pour les bornes post hoc. De multiples extensions sont possibles, comme le développement de nouvelles bornes à partir de l'hybridation de bornes existantes, ou la création de procédures du contrôle de FDP à partir de ces bornes. Ces extensions seront testées sur le jeu de données et également sur des données simulées. On utilisera le package R sansSouci (Neuvial, Durand, et al., 2021) qui implémente déjà les procédures publiées, le ou la stagiaire sera par ailleurs amené à y contribuer.

2 Conditions

Le stage sera co-encadré par Guillermo Durand, maître de conférences au LMO - Laboratoire de Mathématiques d'Orsay, et Mélina Gallopin, maîtresse de conférences à l'I2BC - Institut de Biologie Intégrative de la Cellule. Il durera de 4 à 6 mois et sera financé par l'Institut DATAIA. Pour toute question et/ou pour candidater, contacter Guillermo Durand (guillermo.durand@universite-paris-saclay.fr) et Mélina Gallopin (melina.gallopin@universite-paris-saclay.fr). Pour la candidature, un CV faisant ressortir la pratique des statistiques et de la programmation (R de préférence) sont attendus, et toute formation en biologie sera considérée comme un plus.

References

Bastide, Paul et al. (2022). "Benchmark of Differential Gene Expression Analysis Methods for Inter-species RNA-Seq Data using a Phylogenetic Simulation Framework". In: *bioRxiv*. DOI: 10.1101/2022.01.21.476612. eprint: https://www.biorxiv.org/content/early/2022/01/23/2022.01.21.476612.full.pdf. URL: https://www.biorxiv.org/content/early/2022/01/23/2022.01.21.476612.

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *J. Roy. Statist. Soc. Ser. B* 57.1, pp. 289–300. ISSN: 0035-9246. URL: https://www.jstor.org/stable/2346101.

Blanchard, Gilles, Pierre Neuvial, and Etienne Roquain (2020). "Post hoc confidence bounds on false positives using reference families". In: *Ann. Statist.* 48.3, pp. 1281–1303. ISSN: 0090-5364. DOI: 10.1214/19-AOS1847. URL: https://doi.org/10.1214/19-AOS1847.

Durand, Guillermo et al. (2020). "Post hoc false positive control for structured hypotheses". In: *Scand. J. Stat.* 47.4, pp. 1114–1148. ISSN: 0303-6898. DOI: 10.1111/sjos.12453. URL: https://doi.org/10.1111/sjos.12453.

García de la Torre, Vanesa S et al. (2021). "Wide cross-species RNA-Seq comparison reveals convergent molecular mechanisms involved in nickel hyperaccumulation across dicotyledons". In: *New Phytologist* 229.2, pp. 994–1006.

Goeman, Jelle J. and Aldo Solari (2011). "Multiple testing for exploratory research". In: *Statist. Sci.* 26.4, pp. 584–597. ISSN: 0883-4237. DOI: 10.1214/11-STS356. URL: https://doi.org/10.1214/11-STS356.

Neuvial, Pierre et al. (2021). *sansSouci: Post Hoc Multiple Testing Inference*. R package version 0.12.3-9999. URL: https://github.com/sanssouci-org/sanssouci.