« Méthode bayésienne de sélection d'effets aléatoires : application en génétique »

Que ce soit dans les domaines de la médecine ou de l'agronomie, l'identification des régions du génome impliquées dans la variation des caractères d'intérêt est essentielle. Par exemple chez l'humain, une meilleure identification contribue à développer des stratégies thérapeutiques plus efficaces et, chez les plantes, cela améliore la compréhension des mécanismes d'adaptation face aux changements climatiques.

Les approches statistiques actuellement utilisées dans les études de génétique sont des approches de sélection d'effets fixes (méthodes stepwise backward/forward, méthodes de vraisemblance pénalisées, etc). Ces approches peuvent présenter certaines limites notamment à l'heure du « big data » : l'information génétique, sous forme de marqueurs moléculaires ou issue du séquençage, peut être extrêmement dense et les populations étudiées sont de plus en plus complexes car composées de cohortes d'origines génétiques différentes. Il s'avère donc particulièrement intéressant de se tourner vers des méthodes permettant de résumer l'information génétique disponible de façon pertinente. Récemment des approches appelées Regional Heritability Mapping (Resende et al., 2017) ont été proposées. Leur originalité repose sur l'utilisation de l'information moléculaire (les marqueurs) pour décomposer la variabilité génétique totale en une multitude de variabilités induites par des relations locales estimées en chaque position du génome. Dans ces approches, un modèle linéaire mixte est utilisé afin d'associer à chaque position du génome un effet aléatoire ayant pour matrice de covariance une matrice de similarité génétique calculée à partir des marqueurs. L'identification des positions pertinentes se traduit alors par la sélection d'effets aléatoires dans un modèle linéaire mixte.

Dans ce contexte, des approches de sélection bayésienne d'effets aléatoires ont été proposées (Lu et al., 2015, Heuclin et al., 2023). Ces approches, bien que prometteuses, présentent deux voies d'amélioration : d'une part la prise en compte de la structure de dépendance entre les positions génomiques le long du génome et d'autre part une meilleure estimation des matrices de similarité génétique.

Un premier objectif de ce travail de recherche sera d'utiliser les approches introduites dans la sélection bayésienne d'effets fixes avec une structure de dépendance comme le Bayesian Fused Lasso (Kyung et al., 2010) aux effets aléatoires. Une extension consistera à adapter ce type d'approches dans le modèle de Heuclin et al. (2023) afin de pénaliser les différences successives des composantes de la variance des effets aléatoires. Un autre objectif est de mieux comprendre l'influence/l'impact du nombre de marqueurs considérés pour le calcul des matrices, d'explorer la structure des liens d'apparentement le long du génome et d'optimiser le calcul des matrices en fonction du génome. Une possibilité envisagée est d'estimer des fonctions de similarité continues (Besnier and Carlborg, 2007).

Une application sur le palmier à huile sera faite sur un dispositif simple pour étudier un caractère d'intérêt pour lequel le gène responsable de la variabilité phénotypique est connu.

Mots clés : Statistique bayésienne, sélection de variables, modèles linéaire mixte, génétique

Profil recherché :

- École d'ingénieur ou Master 1 ou 2 en (bio)statistique ou mathématiques appliquée ou agronomie avec dominante en analyse de données
- Compétences en R
- Bonne maîtrise des modèles linéaires mixtes et des concepts bayésiens
- Intérêt pour la génétique
- Capacité rédactionnelle (français/anglais)

Dates du stage : 4 à 6 mois entre Janvier et Août 2024

Lieu du stage : Montpellier, CIRAD Lavalette, UMR AGAP

Encadrants principals : Marie Denis, PhD en biostatistique, Sébastien Tisné, PhD en biologie

Indemnité mensuelle : 611 euros

Modalités de candidature : CV et lettre de motivation à envoyer par email à marie.denis@cirad.fr

Besnier, F., & Carlborg, Ö. (2007). A general and efficient method for estimating continuous IBD functions for use in genome scans for QTL. BMC bioinformatics, 8(1), 1-9.

Lu, Z.-H., Zhu, H., Knickmeyer, R. C., Sullivan, P. F., Williams, S. N., Zou, F. and for the Alzheimer's Disease Neuroimaging Initiative, Multiple SNP Set Analysis for Genome-Wide Association Studies Through Bayesian Latent Variable Selection. Genet. Epidemiol., 39: 664–677, 2015

Heuclin, B., Denis, M., Trottier, T., Tisné, S., and Mortier, F. Continuous shrinkage priors for fixed and random effects selection in linear mixed models: application to genetic mapping. 2023. https://ifip.hal.science/I3M_UMR5149/hal-04238536v1

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ et al. (2009). Finding the missing heritability of complex diseases. Nature 461:747–53

Resende, R. T., Resende, M. D. V., Silva, F. F., Azevedo, C. F., Takahashi, E. K., Silva-Junior, O. B., & Grattapaglia, D. (2017). Regional heritability mapping and genome-wide association identify loci for complex growth, wood and disease resistance traits in Eucalyptus. New Phytologist, 213(3), 1287-1300