Apprentissage parcimonieux pour pan-génomes

Antoine Villié*[1]

[1]Laboratoire de Biométrie et Biologie Evolutive - UMR 5558 – Université Claude Bernard Lyon 1, Centre National de la Recherche Scientifique : UMR5558 – France

Résumé

Les études d'association pan-génomiques visent à identifier des corrélations entre des variants génétiques et des phénotypes. Les approches classiques se concentrent sur les SNPs (Single-Nucleotide Polymorphisms) ou les petits indels, et ne sont donc pas appropriées lorsqu'on travaille avec des espèces présentant des génomes accessoires, métagénomes, translocations, ou encore des régions répétées. Plus récemment, des méthodes reposant sur l'utilisation de k-mers, c'est-à-dire des sous-séquences de longueur k, ont été développées. Les nouveaux variants, définis par la présence ou l'absence de certains k-mers dans une séquence biologique, permettent de mieux représenter les différentes variations génétiques.
Malgré cela, les k-mers ne sont pas assez expressifs pour représenter les régions polymorphiques, dont la présence est alors diluée sur plusieurs k-mers, représentant toutes les versions possibles de cette région. Une alternative plus flexible est alors d'utiliser des motifs de séquence probabilistes. Ces motifs peuvent être vus comme un récapitulatif de plusieurs k-mers, car ils modélisent la présence d'un mélange de nucléotides à chaque position.

Pour quantifier la présence d'un motif dans une séquence donnée, nous calculons son activation moyenne le long de cette séquence. Notre objectif est alors de trouver des motifs dont l'activation est significativement associée avec un phénotype donné. Cette tâche est rendue difficile par la présence d'une infinité de motifs possibles, puisque les proportions des nucléotides à chaque position sont continues.

Actuellement, nous développons une procédure pas-à-pas, permettant de sélectionner un petit nombre de motifs associés avec un phénotype. Puis nous utilisons les avancées récentes en inférence post-sélection pour aboutir à une procédure de test calibrée pour l'association entre les motifs ainsi sélectionnés et le phénotype, qui tient compte de notre procédure de sélection.

Nous faisons aussi un lien formel entre cette procédure de sélection et les CNN (Convolutional Neural Networks) utilisés en biologie, qui peuvent être vus comme un score de sélection particulier. Notre procédure pourrait donc être utilisée pour réaliser de l'inférence post-sélection sur les filtres des CNN à une couche.

*Intervenant