Estimation de liens épidémiologiques par apprentissage statistique sur données génomiques

Samuel Soubeyrand[1] & Maryam Alamil[2]

[1] *BioSP, INRA, 84914, Avignon, France ; samuel.soubeyrand@inra.fr*
[2] *BioSP, INRA, 84914, Avignon, France ; maryam.alamil@inra.fr*

Résumé. Qui a infecté qui au cours d'une épidémie causée par une maladie infectieuse ? Ou plus généralement, quels individus hôtes sont liés épidémiologiquement ? Sous l'angle de la statistique, répondre à ces questions revient à estimer les liens dans un réseau dont les noeuds sont les individus hôtes, que ceux-ci soient des humains, des animaux, des plantes, des foyers, des troupeaux ou encore des champs. Diverses approches permettent de répondre à ces questions, dont des approches exploitant des données génomiques caractérisant, au niveau individuel, le pathogène causant la maladie (des individus portant des variants identiques ou proches du pathogène étant vraisemblablement liés épidémiologiquement). Nous avons récemment développé une telle approche, permettant d'estimer les liens épidémiologiques au sein d'une population d'hôtes, fondée sur un modèle semi-paramétrique dans l'espace des génomes du pathogène et sur une technique d'apprentissage exploitant des données de contact pouvant être collectées par exemple dans le cadre d'une enquête épidémiologique.

Au cours de la présentation, nous détaillerons la construction du modèle semi-paramétrique et l'implémentation de la méthode d'estimation, puis nous illustrerons l'application de notre approche à des données simulées et des données réelles portant sur Ebola, la grippe porcine et un potyvirus du salsifis sauvage.

Mots-clés. apprentissage, données génomiques, maladie infectieuse, modèle semi-paramétrique, pseudo-vraisemblance