Etudes avec le modèle SBM de la diversité présente dans un tableau de distances moléculaires

Alain Franc*[1], Mohamed-Anwar Abouabdallah , and Nathalie Peyrard

[1]UMR BIODIVERSITE GENES COMMUNAUTES (UMR BIOGECO) – Institut national de la recherche agronomique (INRA), Université de Bordeaux (Bordeaux, France) – 69, route d'Arcachon 33610 Cestas, France

Résumé

Classification non supervisée et taxonomie ont coévolué depuis des décennies. Ce lien s'est cristallisé récemment autour de la notion d'OTU (Operational Taxonomical Unit), avec un grain proche de l'espèce, sur données moléculaires. Nous avons étudié l'adéquation de ces deux visions de la classification du vivant, à des niveaux taxonomiques plus grossiers, sur un jeu de données d'environ 1500 arbres d'une parcelle en Guyane (200 espèces). Nous avons comparé la classification botanique en taxons à différents grains (espèce, genre, famille, ordre) avec quatre méthodes de clustering : trois méthodes pour la CAH (simple lien, lien complet et Ward), et l'utilisation du Stochastic Block Model (SBM). Toutes ces méthodes sont en très bonne adéquation avec la botanique pour le grain espèces, et donnent des résultats de bons (CAH Ward ou lien complet, SBM) à dégradés (CAH simple lien) pour des niveaux taxonomiques de plus en plus grossiers. Une des causes de la dissonance entre botanique et méthodes numériques est que le tableau de distances entre individus n'est pas structuré en communautés bien établies. Cela nous amène à porter une attention particulière au modèle SBM qui ne fait pas l'hypothèse que les clusters trouvés doivent être des communautés. Nous présenterons pourquoi nous l'envisageons comme un bon candidat pour caractériser la diversité présente dans une communauté, avec trois axes de recherche en cours : la définition et l'identification automatique de différentes topologies d'OTU (clique, clique avec halo, structure en réseau de sous-communautés) ; la construction d'une caractérisation fine, multidimensionnelle, de la diversité présente dans un tableau de distances entre individus ; et enfin le passage à l'échelle des méthodes d'estimation du modèle SBM.

Mots-Clés: modèle SBM, biodiversité, distances moléculaires, OTU, taxonomie moléculaire

*Intervenant