Proposition de stage de M2
Modèles à blocs latents pour la caractérisation de la biodiversité

Contexte finalisé et problématique scientifique

La biodiversité est classiquement caractérisée par un ensemble d'indices, comme l'indice de Shannon ou de Rao, qui sont des informations globales et ne décrivent pas l'organisation des espèces entre elles. Ils ont été conçus en un temps où la diversité était essentiellement décrite à partir d'observations des traits phénotypiques des espèces, et où le frein était la production des données, peu nombreuses et longues à collectionner. En parallèle avec les phylogénies moléculaires, il est possible d'envisager une caractérisation plus riche de la biodiversité, grâce à l'accès à des données moléculaires (des séquences d'ADN). En effet, la notion de biodiversité peut se refléter dans un pattern de dissimilarités entre séquences, quantifiées par des distances d'alignement. Il s'agit alors d'identifier à partir de ces données, les groupes de séquences correspondant à une même espèce, puis le pattern d'organisation des espèces entre elles. Des premiers travaux [1] ont permis de représenter les séquences sous la forme d'un nuage de points en petite dimension et d'en extraire des zones de fortes densités de points. L'étape suivante consiste à revenir à l'espace de grande dimension initial et à analyser ces zones dans cet espace : est-ce qu'une zone correspond vraiment à une seule espèce, ou découvre-t'on des sous-groupes? Les "espèces" sont-elles connectées, ou isolées? Comment se positionnent les zones les unes par rapport aux autres ? Pour répondre à ces questions, il est possible par seuillage des distances, de se ramener à une représentation des zones par un graphe où un sommet est une séquence, et un lien existe si la distance est inférieure au seuil puis d'analyser la structure de ce graphe pour en déduire une description de la biodiversité associée. Le stage proposé porte sur l'analyse de ce graphe basée sur l'utilisation des modèles à blocs stochastiques (Stochastic Block Model, SBM) qui permettent non seulement d'identifier des groupes de sommets connectés mais aussi de caractériser les interactions entre ces groupes [2].

Projet de stage

Il s'agira dans un premier temps de s'approprier le cadre des SBM et l'algorithme VEM (Variational Expectation Maximization) utilisé classiquement pour l'estimation du modèle [2]. Puis l'étudiant(e) travaillera sur la mise en œuvre en pratique de cette approche sur des données de distances d'alignement (quelle(s) initialisation(s), quel nombre de blocs, quel découpage du jeu de données s'il dépasse les capacités du VEM?). Une des premières sorties attendue du stage sera la mise en place du pipeline complet de traitement depuis la matrice des distances d'alignement entre séquences jusqu'à l'identification des groupes et de leurs liens.

Ensuite le pipeline sera mis en œuvre sur des données de biodiversité de la forêt guyanaise dans le but de caractériser l'organisation des distances entre espèces à partir d'un pool de séquences. Différents niveaux de distance de seuillage correspondant aux espèces, genres, familles seront considérés puis il s'agira d'analyser la structure des SBM ainsi obtenus et des variations de cette structure selon le seuil.

Enfin, selon l'avancé du stage et la formation de l'étudiant(e) deux pistes sont envisagées: l'exploration du formalisme des produits tensoriels pour proposer une alternative à l'approximation variationnelle pour l'étape E du EM qui permette de traiter de plus grand problèmes, ou la définition de descripteurs mathématiques de la biodiversité à partir de l'objet SBM.

Compétences attendues

- bon niveau en statistique ou machine learning
- pratique du logiciel R
- goût pour les applications

Bibliographie

[1] Blanchard P., Chaumeil P., Frigerio JM, Rimet F, Salin F., Thérond S., Coulaud O, Franc A. A geometric view of biodiversity: scaling to metagenomics. Research report INRIA n 9144, 2018

[2] Daudin J.-J., Picard F, Robin S. A mixture model for random graphs. Statistics and Computing, 18, 173-183, 2008.

Lieu et Encadrement

Le stage se déroulera au sein de l'unité MIAT de l'INRA de Toulouse. Il sera co-encadré par Nathalie Peyrard (nathalie.peyrard@inra.fr, MIAT), Alain Franc (alain.franc@inra.fr, BioGeCo INRA et Pleiade INRIA Bordeaux) et Olivier Coulaud (olivier.coulaud@inria.fr, Hiepacs, INRIA Bordeaux).

Durée

6 mois, début possible à partir de mars 2019 (stage rémunéré environ 560 /mois)