Structure d’accueil : Unité mixte de recherche ASTRE

Localisation : Campus international de Baillarguet. 34398 Montpellier

Contexte du stage

L’évolution récente des approches métagénomiques permet aujourd’hui l’étude de communautés virales. Ces avancées permettent une meilleure compression de ces communautés et, notamment, de leur rôle dans les infections par des agents pathogènes et dans les maladies associées.

Notre équipe se concentre sur l’étude des communautés virales retrouvées chez des moustiques vecteurs de maladies humaines et vétérinaires en Afrique et Europe. L’objectif à court terme est de mieux appréhender la diversité des virus eucaryotes ainsi que les différents facteurs qui peuvent structurer cette diversité (espèces de moustiques, géographie, anthropisation du milieu, etc.).

La caractérisation de communautés virales par métagénomique requiert du séquençage shotgun, c’est-à-dire, le séquençage aléatoire des acides nucléiques de l’échantillon. Notre équipe a mis au point un pipeline bioinformatique en snakemake, ‘Snakevir’, optimisé pour l’analyse des données de ce type de séquençage. Cependant, le clustering de séquences obtenues en unités taxonomiques reste en défi. Contrairement au clustering de séquences à partir du séquençage de barcodes génétiques (i.e. métabarcoding), nous manquons des méthodes de clustering performantes et adaptées à la métagénomique shotgun. Par exemple, un grand nombre de nouvelles espèces virales sont présentes dans nos jeux de données et les méthodes actuelles, souvent basées sur l’alignement, ne permettent pas de discriminer entre des séquences des virus proches connus et inconnus.

Nous cherchons une méthode de clustering adaptée pour l'analyse des données de séquençage métagénomique viral. Nous avons déjà développée une méthode basée sur les réseaux de neurones artificielles. Dans ce contexte, ce projet de stage vise à comparer cette méthode avec des méthodes existantes et des nouvelles méthodes basées sur l’intelligence artificielle. Plus précisément, les objectifs du stage sont :

(1) Identifier des nouvelles méthodes de clustering s’appuyant notamment sur le machine learning

(2) Produire un protocole de benchmarking adapté pour la comparaison des méthodes de clustering (production de jeux de données in silico, études de données de terrain...)

(3) Tester les méthodes et documenter les résultats

(4) Intégrer la/les méthodes sélectionnées au pipeline d’analyse

Compétences demandées :

Formation recherchée : M1 ou première année d’école d’ingénieur avec parcours bio-informatique.

• Maîtrise du langage bash

• Bonnes pratiques de développement de code.

• Des connaissances en analyses métagénomiques serait appréciées

• Des connaissances en langages de programmation tels que python ou snakemake seraient un

plus.

Contact :

\* Antoni EXBRAYAT : antoni.exbrayat@cirad.fr

\* Serafin GUTIERREZ : serafin.gutierrez@cirad.fr