Proposition de stage de master 2 pour 2024

Meta-analyse exploratoire de jeux de données publics de différents écosystèmes microbiens de la chaîne alimentaire

Laboratoire d'accueil : Unité MaIAGE, Université Paris-Saclay, INRAE, Domaine de Vilvert, 78352 Jouy-en-Josas cedex

Gratification : environ 560 euros/mois (grille de rémunération INRAE)

Encadrants :

- Christelle Hennequet-Antier (christelle.hennequet-antier@inrae.fr), Cédric Midoux (cedric.midoux@inrae.fr) et Hélène Chiapello (helene.chiapello@inrae.fr), UR MaIAGE
- Eric Dugat-Bony (eric.dugat-bony@inrae.fr) UMR SAYFOOD et Julien Tap (Julien.tap@inrae.fr), Institut MICALIS

Contexte du projet

L'ouverture des données (Open Data) fait désormais partie intégrante de la recherche scientifique, révolutionnant la manière dont les connaissances sont générées, partagées et ré-utilisées. Dans le cadre de la politique OpenScience d'INRAE (Institut national de recherche pour l'agriculture, l'alimentation et l'environnement) et de la volonté de (ré-)utiliser des jeux de données publiés, le département MICA (Microbiologie et Chaîne alimentaire) a mis en place une étude pilote transversale d'analyse comparative, *in silico*, des écosystèmes microbiens présents dans la chaîne alimentaire de l'Homme et de l'animal, jusqu'aux systèmes anaérobies de dépollution à partir de données de type métagénomique par amplicon ciblant le gène codant pour l'ARNr 16S. Il s'agit du projet pilote "Mica Open16S".

Données disponibles

Les données de séquences de type Illumina ou Ion Torrent provenant de 17 études publiques, couvrant 3 grands types d'environnements (aliments, microbiotes humains et animaux, digesteurs anaérobies) et plus de 2000 échantillons disponibles publiquement dans les bases de données internationales, ont été téléchargées puis ré-analysées avec le même pipeline bioinformatique afin de produire une table d'abondance comparable entre les différents échantillons analysés. Cette table d'abondance ainsi que les métadonnées disponibles sur les échantillons constitueront le jeu de données à exploiter.

Travail demandé

L'objectif principal est d'évaluer des méthodes de statistiques de type "data-driven" (analyses sans a priori, guidées par la structure des données) pour réaliser des méta-analyses en écologie microbienne.

La première partie du stage sera donc consacrée à une revue des méthodes statistiques les plus utilisées dans le cadre de méta-analyses (analyse simultanée de nombreux jeux de données provenant d'études indépendantes). Un travail autour des méthodes d'inférence de réseaux microbiens sera aussi réalisé pour détecter et visualiser les associations à partir d'une ou plusieurs tables d'abondance. Ceci permettra de sélectionner les outils les plus adaptés en fonction des besoins spécifiques du projet et en accord avec les encadrants.

La deuxième partie du stage sera consacrée à la mise en œuvre des outils sélectionnés sur les données du projet MICA Open16S. Une comparaison des résultats obtenus sera effectuée en prenant en compte différents aspects, tels que les métadonnées et les considérations écologiques. On pourra en particulier rechercher s'il est possible d'identifier des facteurs biotiques et abiotiques clés dans la structuration des communautés microbiennes de différentes communautés. On pourra aussi rechercher si des drivers (facteurs) peuvent expliquer la distribution d'un taxon ou d'un ensemble de taxons d'intérêt dans différents biotopes.

Démarche et outils utilisés

Une démarche basée sur des bonnes pratiques (principes FAIR) de gestion de données et de scripts R reproductibles sera utilisée. Les développements et analyses seront effectués sur l'infrastructure migale, en utilisant l'environnement R/Rstudio pour les analyses statistiques (notebooks, packages R), des outils de gestion de version du code (e.g. git) et des outils de management de projet (e.g. gantt chart, kanban, etc).

Ce projet constitue une opportunité unique d'acquérir des compétences de science des données appliquées à l'écologie microbienne**.** A l'issue du projet, en fonction de l'avancée du stage, il pourra être envisagé d'intégrer les données d'études supplémentaires dans la méta-analyse et de participer à une publication de valorisation des résultats.

Compétences recherchées

- Connaissances en statistiques pour l'analyse de données biologie à haut-débit, en bioinformatique et en biologie (génomique & métagénomique, microbiologie).
- Maîtrise du langage de programmation R, éventuellement Python
- Travail de recherche et approche scientifique
- Maîtrise des techniques de communication orale et écrite
- Maîtrise de l'anglais (lu exigée, écrit si possible).