FICHE DESCRIPTIVE DU STAGE

> L'ENTREPRISE (nom, coordonnées…)

INRAE, Domaine du Vilvert, 78350 Jouy-en-Josas.

> INTITULE DU STAGE

Apprentissage par processus Gaussiens pour l'analyse intégrative de données omiques.

Objectif(s) du stage (problématique, missions, méthodologie...) :

Le développement récent des technologies à haut-débit donne accès à de nombreuses données complexes et de nature hétérogène, telles que les données –omiques. Nous nous plaçons dans ce projet dans le cadre de l'analyse conjointe de données d'expression de gènes (transcriptomique) et de données épigénétique de méthylation de l'ADN. L'objectif est de réaliser une analyse intégrée de ces deux types de données afin de mieux comprendre les liens de régulation sous-jacents. Pour cela, on souhaite construire des réseaux de corrélation incluant les gènes et les sites CpG au regard de leurs niveaux d'expression dans différentes conditions, mais également en prenant en compte leur structure de corrélation spatiale. Cette analyse intégrative sera basée sur des méthodes d'apprentissage par processus Gaussiens (GPs), à travers des structures de corrélation adaptées. Les développements méthodologiques porteront sur l'introduction de noyaux de covariance spécifiques pour traiter simultanément des relations spatiales et temporelles, entre le génome et données épigénétiques. Ces développements devront s'inscrire dans le cadre général de l'apprentissage GP multi-tâches, tel que proposé dans [1] et [2] pour la modélisation, prédiction et clustering de données fonctionnelles. Une implémentation logicielle de la méthodologie pourra être envisagée pour intégrer de nouvelles fonctionnalités au package R associé, MagmaClustR. La méthode proposée sera appliquée à un jeu de données généré à l'INRAE.

[1] Arthur Leroy, Pierre Latouche, Benjamin Guedj, Servane Gey (2022). MAGMA: inference and prediction using multi-task Gaussian processes with common mean. Machine Learning.
[2] Arthur Leroy, Pierre Latouche, Benjamin Guedj, Servane Gey (2023). Cluster-Specific Predictions with Multi-Task Gaussian Processes. JMLR.

Connaissances et aptitudes recherchées chez le stagiaire :

Le ou la candidat.e aura un goût pour la modélisation de données biologiques, et de bonnes aptitudes de programmation en R. Idéalement, le ou la candidat.e aura des compétences en apprentissage statistique (machine learning), en modélisation probabiliste, et/ou en statistiques bayésiennes. Le stage se déroulera dans un environnement stimulant de recherche, à l'interface entre la statistique et les applications biologiques.

> ENVIRONNEMENT DE LA MISSION

Le stage se déroulera à l'INRAE de Jouy-en-Josas dans l'Unité GABI (Génétique Animale et Biologie Intégrative)

Ressources mises à la disposition du stagiaire (informatiques, bureautiques, logiciels statistiques, matérielles...) :

Le ou la candidat.e disposera de toute l'infrastructure informatique nécessaire, avec un accès en particulier au logiciel R, de puissance de calcul et d'un espace mémoire suffisant pour le stockage et l'analyse des données –omiques et aux ressources bibliographiques d'INRAE.

> PERSONNE(S) A CONTACTER

Arthur LEROY, arthur.leroy@inrae.fr
Florence JAFFREZIC, florence.jaffrezic@inrae.fr

Durée du stage : 5 à 6 mois.

Gratification : environ 600 euros par mois.