Nom de l'entreprise ou du laboratoire :

Centre Inria de l'université de Bordeaux, équipe Pleiade

Adresse où se déroulera le stage :

200 Avenue de la Vieille Tour

33405 Talence

France

Responsables du stage (personnes qui seront contactées par les candidates et candidats):

- Nom, Prénom : Guilhem Sommeria-Klein (Chargé de Recherche Inria), Frioux Clémence (Chargée de Recherche Inria)

- Coordonnées : Guilhem.sommeria-klein@inria.fr; clemence.frioux@inria.fr

Titre du stage : Dimensionality Reduction Techniques for Human Gut Microbiome data

Mots clés résumant les méthodes et techniques à utiliser au cours du stage :

Non-Negative Matrix Factorization (NMF) – Latent Dirichlet Allocation (LDA) – Dimensionality reduction – Metagenomics – Microbial communities – DNA sequences – Machine learning.

Début du stage souhaité : janvier - février 2026

Résumé du projet de stage [*French version on the last page of this document*] :

The study of microbial communities, known as *microbiomes*, generates high-dimensional data comprised of the DNA sequences (*genomes*) of microorganisms present in various samples. These genomes each contain thousands of *genes*, which encode cellular effectors. Identifying and annotating these genes provides insights into the *functions* and roles of microbes, such as the molecules they consume or produce.

Among microbial ecosystems, the gut microbiome is one of the most extensively studied. Over the past two decades, the microbial DNA from thousands of samples has been sequenced, leading to the development of comprehensive databases[1,2] and a deeper understanding of the diversity and functions of microbial populations[3] . Nearly 300,000 genomes have been reconstructed and grouped into approximately 5,000 distinct microbial species. These genomes encompass several million genes, which can be clustered based on sequence similarity. This represents hundreds of times the number of genes present in our own genome. Over evolutionary history, our organisms appear to have evolved a dependence on some of these microbial genes to function, making them critical to human health.

This internship aims to investigate the distribution and co-occurrence of genes across the microbial genomes found in human gut by applying dimensionality reduction to the vast collection of genes and genomes described above. Our team's previous work has focused on

applying dimensionality reduction to the taxonomic diversity of microbiomes in human gut and in the environment[4,5]. In this project, the emphasis will be on microbial functions. The goal is to establish a base for decomposing the gut microbial gene content of individuals into "functional profiles" that will make it more interpretable and will ease downstream analyses. The core dataset to obtain these functional profiles consists of matrices showing the presence or absence of millions of genes in thousands of bacterial genomes.

Objectives of the internship include:

- Reviewing common application domains of dimensionality reduction (e.g., image processing, text mining) to identify techniques that can scale up to the size of our data.
- In particular, exploring the applicability of Non-negative Matrix Factorisation (NMF), Latent Dirichlet Allocation (LDA), and Variational Autoencoder (VAE) techniques for the dimensionality reduction of our data.
- Implementing and comparing various algorithms.
- Investigating bi-cross validation approaches for large datasets to select for the optimal number of dimensions.
- Using the obtained functional profiles to characterise human gut microbiome samples.

If you are interested in computational biology and data science, and want to apply advanced machine learning techniques to real-world biological data, we encourage you to apply!

Expected skills and profile

- Expected:
    - Proficiency in Python
    - Good level in English (written and spoken)
- Appreciated:
    - High-performance computing
    - Bash programming language
    - Interest in microbiology applications

Working language: French or English

We are seeking a student with one of the following profiles:

– A Master's degree in computer science or artificial intelligence,

– or a Master's degree in computational biology with significant coursework in computer science.

Montant des indemnités de stage : gratification au taux en vigueur.

Références

1. Richardson, L. *et al.* MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res* (2022) doi:10.1093/nar/gkac1080.
2. Gurbich, T. A. *et al.* MGnify Genomes: a resource for biome-specific microbial genome catalogues. *J Mol Biol* 168016 (2023) doi:10.1016/j.jmb.2023.168016.
3. Fan, Y. & Pedersen, O. Gut microbiota in human metabolic health and disease. *Nat Rev Microbiol* 1–17 (2020) doi:10.1038/s41579-020-0433-9.
4. Sommeria-Klein, G. *et al.* Global drivers of eukaryotic plankton biogeography in the sunlit ocean. *Science* **374**, 594–599 (2021).
5. Frioux, C. *et al.* Enterosignatures define common bacterial guilds in the human gut microbiome. *Cell Host Microbe* (2023) doi:10.1016/j.chom.2023.05.024.

[French version of the internship project]

Techniques de réduction de dimension pour des données de microbiome intestinal humain.

L'étude des communautés microbiennes, appelées *microbiomes*, génère des données de haute dimension composées des séquences d'ADN (*génomes*) des micro-organismes présents dans divers échantillons. Chaque génome contient des milliers de *gènes* qui codent des effecteurs cellulaires. L'identification et l'annotation de ces gènes apportent des informations sur les *fonctions* et les rôles des microbes, comme les molécules qu'ils peuvent consommer ou produire.

Parmi les écosystèmes microbiens, le microbiome intestinal est l'un des plus étudiés. Au cours des deux dernières décennies, l'ADN microbien de milliers d'échantillons a été séquencé, ce qui a permis la création de bases de données [1,2] et une compréhension approfondie de la diversité et des fonctions des populations microbiennes[3]. Près de 300 000 génomes ont été reconstruits et regroupés en environ 5 000 espèces microbiennes distinctes. Ces génomes recensent plusieurs millions de gènes, qui peuvent être regroupés selon leur similarité de séquence. Cela représente des centaines de fois le nombre de gènes présents dans notre propre génome. Au fil de l'évolution, nos organismes semblent avoir développé une dépendance à certains de ces gènes microbiens pour fonctionner, ce qui les rend essentiels à la santé humaine.

Ce stage vise à étudier la distribution et la co-occurrence des gènes au sein des génomes microbiens du microbiote intestinal humain en appliquant des techniques de réduction de dimension à l'immense collection de gènes et de génomes décrite ci-dessus. Les travaux antérieurs de l'équipe se sont concentrés sur l'application de la réduction de dimension à la diversité taxonomique des microbiomes humains et environnementaux[4,5]. Dans ce projet, l'accent sera mis sur les fonctions microbiennes. L'objectif est d'établir une base permettant de décomposer le contenu génétique microbien intestinal des individus en « profils fonctionnels » afin de le rendre plus interprétable et de faciliter les analyses en aval. Le jeu de données central pour obtenir ces profils fonctionnels se compose de matrices présentant la présence ou l'absence de millions de gènes dans des milliers de génomes bactériens.

Les objectifs du stage incluent :

- Effectuer une revue des grands domaines d'application de la réduction de dimension (par exemple : traitement d'image, exploration de texte) afin d'identifier les techniques capables de passer à l'échelle de nos données.
- Explorer en particulier l'applicabilité des méthodes de factorisation en matrices non négatives (NMF), d'allocation de Dirichlet latente (LDA) et d'auto-encodeur variationnel (VAE) pour la réduction de dimension de nos données.
- Implémenter et comparer différents algorithmes.
- Étudier l'approche de bi-cross validation sur des grands jeux de données afin de sélectionner le nombre optimal de dimensions.
- Utiliser les profils fonctionnels obtenus pour caractériser des échantillons du microbiome intestinal humain.

Si vous vous intéressez à la biologie computationnelle et à la science des données, et souhaitez appliquer des méthodes avancées d'apprentissage automatique à des données biologiques réelles, nous vous invitons à postuler !

Compétences et profil attendus :

Compétences attendues :

- Maîtrise de Python

- Bon niveau d’anglais (écrit et oral)

Compétences appréciées :

- Calcul haute performance (HPC)
- Programmation Bash
- Intérêt pour les applications en microbiologie

Langue de travail : français ou anglais

Nous recherchons un étudiant ou une étudiante avec l’un des profils suivants :

– Master en informatique ou en intelligence artificielle

– ou Master en biologie computationnelle avec une part significative des cours en informatique