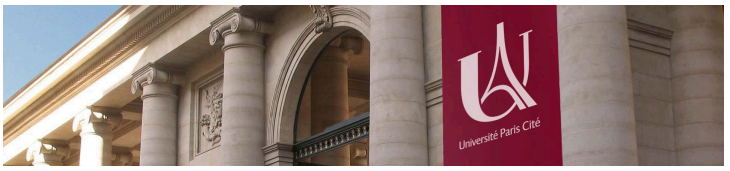
Université Paris Cité



Master Degree in Bioinformatics

Speciality : Bioinformatics and Computational Science

2024 - 2025

Supervised Project

Protein Language Model pour l'Alignement Multiple de Séquences Phylogénétiquement Informé

Supervisor :

Name: GELLY

First Name: Jean-Christophe

Le projet sera co-supervisé par Ragou RADJASANDIRANE et Gabriel CRETIN, doctorants dans l'équipe DSIMB

Professional Address:

INSERM UMR_S1134 - Université Paris Cité

Hôpital Necker APHP | Bâtiment Lavoisier 7e étage - Bureau 704

149 Rue de Sèvres, 75015 Paris

Position: Professor

Laboratory:

Institution: Academic - INSERM / Université Paris Cité

email: jean-christophe.gelly@u-paris.fr

Title: Protein Language Model pour l'Alignement Multiple de Séquences Phylogénétiquement Informé

Short Project Summary :

Scientific Background

L'alignement multiple de séquences (MSA) est une méthode fondamentale en bio-informatique, utilisée pour étudier les relations évolutives entre protéines. Les modèles de langage protéique basés sur les Transformers, tels que le MSA

Transformer développé par Meta, ont démontré une bonne capacité à capturer des informations structurales et fonctionnelles à partir d'alignements de séquences. Cependant, les alignements de séquences ne prennent pas en compte les relations évolutives et la distance évolutive entre séquences. Le projet vise à intégrer cette information phylogénétique dans un modèle de langage, en l'occurrence MSA Transformers.

Project Objective

Ce projet vise à adapter et à appliquer le modèle MSA Transformer pour l'alignement multiple de séquences de protéines paralogues, en intégrant des informations phylogénétiques afin d'améliorer la précision et la pertinence des alignements.

Working packages

1. Revue bibliographique : Étudier les approches actuelles permettant d'intégrer les alignement multiple de séquences dans des modèles de langage basés sur les Transformers, en se concentrant sur celles qui permettrait d'intégrer simplement des informations phylogénétiques.

2. Collecte et préparation des données : Assembler des ensembles de données de protéines paralogues à partir de bases de données publiques, en veillant à la qualité et à la représentativité des séquences sélectionnées.

3. Adaptation du modèle MSA Transformer : Configurer et entraîner le modèle MSA Transformer sur les ensembles de paralogues, en intégrant des informations phylogénétiques.

4. Évaluation des performances : Comparer les MSA Transformers au MSA Transformers avec information phylogénétique sur un ensemble de tâches classiques.

5. Analyse et interprétation : Examiner les résultats pour identifier les améliorations apportées par l'intégration des informations phylogénétiques et proposer des pistes pour de futures recherches.

Relevant Publications linked to the project :

· Rao, R., Liu, J., Verkuil, R., Meier, J., Canny, J. F., Abbeel, P., Sercu, T., & Rives, A. (2021). MSA Transformer. bioRxiv. https://doi.org/10.1101/2021.02.12.430858

· Ahdritz, G., Bouatta, N., Kadyan, S., Xia, Q., Gerecke, W., Meier, J., ... & AlQuraishi, M. (2022). OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. bioRxiv. https://doi.org/10.1101/2022.11.20.517210

· Ishikawa, S. A., Zhukova, A., Iwasaki, W., & Gascuel, O. (2019). A fast likelihood method to reconstruct and visualize ancestral scenarios. Molecular Biology and Evolution, 36(9), 2069–2085. https://doi.org/10.1093/molbev/msz131

Keywords:

Bioinformatics, multiple sequence alignment, Phylogeny, Deep Learning, Transformers

Supervision and Work Environment

The student will work under the supervision of **Pr. Jean-Christophe Gelly, Ragou**

Radjansandirane and Gabriel Cretin within the **BIGR** team (Integrated Biology of the Red Blood Cell, INSERM UMR_S1134). The scientific supervision will rely on the team's expertise in bioinformatics and deep learning.