amU Institut microbiologie, bioénergies et biotechnologie Aix Marseille Université

AFMB

Offre de stage Master 2 au sein d'une plateforme technologique
année 2025-2026

Sujet: CAZy 3D prédiction

Durée du stage: 6 mois

Laboratoire d'accueil:
Laboratoire AFMB – Architecture et Fonction des Macromolécules Biologiques
CNRS – Aix-Marseille Univ. UMR7257
Parc Scientifique et Technologique de Luminy – Case 932
163 avenue de Luminy
13288 Marseille CEDEX 09

Responsable(s) du stage :
Vincent Lombard, Ingénieur de recherche,
Plateforme : CAZy Bioinformatics

Site web : https://www.afmb.univ-mrs.fr/facility/bioinformatique-cazy/

Contexte :
La plateforme bioinformatique CAZy (Carbohydrate Active enZymes) analyse les données génomiques et métagénomiques afin d'identifier les séquences codantes pour des enzymes impliquées dans l'assemblage(glycosyltransférases) ou la déconstruction des sucres complexes (glycoside hydrolases, polysaccharide lyases, carbohydrate estérases, activités auxiliaires). Elle met à disposition de la communauté scientifique – qu'il s'agisse de structures académiques ou non-académiques, locales ou nationales – ses ressources techniques et l'expertise de ses personnels pour réaliser des analyses spécialisées dans le domaine des CAZymes.
S'appuyant sur les outils bioinformatiques développés[1] autour de la base de données CAZy[2], conçue et continuellement enrichie par l'équipe de Glycogénomique du laboratoire, la plateforme offre différents niveaux d'analyse, allant de l'annotation automatique à la curation manuelle des résultats.
Les services proposés comprennent :
L'annotation manuelle de génomes
L'annotation automatique de données -omiques
L'expertise et l'analyse comparative de génomes
L'expertise et la prédiction fonctionnelle
Dans une perspective d'amélioration continue, le développement d'outils de détection plus performants est nécessaire afin d'accroître la précision des analyses -omiques. L'intégration de nouvelles technologies, notamment l'intelligence artificielle, représente un enjeu majeur pour répondre à ces besoins et renforcer l'efficacité de la plateforme. Nous souhaitons ainsi développer cette dimension pour offrir des prestations toujours plus performantes et adaptées aux attentes de la communauté scientifique.

Objectif du stage :
L'étudiant.e intégrera le groupe Glycogénomique ainsi que la plateforme Bioinformatique CAZy de l'AFMB, une équipe composée actuellement de trois enseignants-chercheurs, d'un chercheur émérite, de deux ingénieurs, d'un post-doctorant et de deux doctorants. L'équipe bénéficie d'une infrastructure informatique, adaptée à l'analyse des enzymes agissant sur les sucres (CAZymes), comprenant notamment un serveur web, trois serveurs SMP dédiés au calcul génomique, ainsi qu'un serveur de bases de données pour la gestion des données. Pour les analyses nécessitant des ressources importantes, un cluster de calcul de plus de 650 cœurs est disponible, permettant l'exécution de logiciels parallélisés via un système de batch, particulièrement utile pour le traitement de grands volumes de données, comme en métagénomique.
Le projet vise à améliorer l'annotation automatique de données -omiques de la plateforme en intégrant les données structurales et fonctionnelles des protéines. Contrairement aux méthodes actuelles, qui reposent principalement sur la comparaison de séquences d'acides aminés, cette approche exploitera les récents développement du deep learning et de l'intelligence artificielle pour développer la prédiction fonctionnelle de CAZyme.
L'étudiant·e devra recueillir les informations structurales des CAZymes en interrogeant différentes bases de données spécialisées, telles que la Protein Data Bank ou l'ESM Metagenomic Atlas[3], qui regroupent des millions de structures protéiques prédites ou expérimentales. Ces données constitueront le socle pour développer un nouveau pipeline de prédiction des sites catalytiques, spécifiquement adapté aux particularités des CAZymes, en s'appuyant sur des outils utilisant le deep learning comme P2Rank[4] ou TopEC[5]. Les informations tridimensionnelles obtenues seront ensuite croisées avec des données expérimentales tabulaires sur les CAZymes et leurs activités, issues notamment de l'outil CAZac récemment développé[6]. L'ensemble de ces données sera exploité pour développer un nouvel outil basé sur le deep learning, capable d'intégrer à la fois les informations structurales et les données expérimentales. Il permettra d'améliorer significativement la précision des prédictions fonctionnelles des CAZymes. Une attention particulière sera portée à l'ergonomie et à la documentation des outils développés, afin d'en faciliter l'accès aux chercheurs non spécialistes et d'assurer une prise en main aisée des résultats obtenus.
Ce stage offrira à l'étudiant.e l'opportunité d'acquérir une vision approfondie des CAZymes et de rechercher des informations détaillées sur les structures protéiques. Il/Elle pourra également se familiariser avec différents modèles de deep learning et renforcer ses compétences en programmation Python, notamment à travers l'utilisation d'interfaces telles que Jupyter Notebook.

Profil attendu du candidat:
L'étudiant(e) recherché(e) doit faire preuve d'une forte motivation, alliée à des compétences en informatique.
Compétences techniques attendues :
- Maîtrise des bases de la programmation, des connaissances en SQL et Python seraient un atout.
- Compréhension des principes du deep learning et de l'intelligence artificielle.
- Familiarité avec les bases de données internationales telles que UniProt, PDB, AlphaFold, ainsi qu'avec les outils dédiés à l'analyse de la structure 3D des protéines (Pymol ou ChimeraX) (appréciée).
- Aptitude à rédiger des documentations fonctionnelles et techniques.
- Capacité à communiquer efficacement et à faire preuve de pédagogie.
Compétences comportementales :
- Autonomie dans le travail tout en sachant collaborer efficacement au sein d'une équipe.
- Motivation et engagement dans les missions confiées.

Modalités des candidatures : Merci de faire parvenir un CV, une lettre de motivation, une lettre de recommandation et les résultats de la licence 3 et du Master 1 (notes et classement), par mail à vincent.lombard@univ-amu.fr

Références bibliographiques :

1. Delannoy-Bruno, O. et al. Evaluating microbiome-directed fibre snacks in gnotobiotic mice and humans. Nature 595, 91–95 (2021).

2. Drula, E et al. The carbohydrate-active enzyme database: functions and literature. Nucleic Acids Res 50(D1):D571-D577 (2022).

3. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 379, 1123–1130 (2023).

4. Krivák, R. et al. D. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. J. Cheminformatics 10, 39 (2018).

5. van der Weg et al. TopEC: prediction of Enzyme Commission classes by 3D graph neural networks and localized 3D protein descriptor. Nat. Commun. 16, 2737 (2025).

6. Lombard, V. et al. CAZac: an activity descriptor for carbohydrate-active enzymes. Nucleic Acids Res. 53, D625–D633 (2025).