

Stage M2 – Annotation de méta-transcriptomes de sols forestiers.

Mots-clés :

Méta-transcriptomique – annotation – pipeline – champignons – bases de données

Contact : lucas.auer@inrae.fr

Description

Les champignons sont des acteurs majeurs des écosystèmes forestiers où ils sont impliqués dans la croissance des arbres (mycorhiziens symbiotiques), leur santé (parasites pathogènes) et le recyclage de matière morte (dégradeurs saprotrophes). Ils sont ainsi un des acteurs clés du cycle du carbone des forêts et l'adaptation des arbres face au changement climatique. Mais la connaissance des liens entre diversité fongique et potentiel fonctionnel des espèces reste encore limitée. Pour dépasser la seule description des taxa présents dans un écosystème, notre laboratoire est précurseur dans le développement d'approches de méta-transcriptomique des sols qui se sont déjà révélées très prometteuses.

Le pipeline d'annotation fonctionnelle des ARNm fongiques développé dans l'équipe repose sur l'utilisation de bases de données génomiques, principalement MycoCosm, issue du 1 000 Fungal Genomes Project (actuellement près de 1500 génomes disponibles répartis dans 800 familles distinctes). Si cette ressource est un atout précieux, elle est à relativiser au regard des 1,5 million d'espèces de champignons présents sur Terre. Il est donc essentiel de pouvoir améliorer la qualité des résultats d'annotation obtenus à partir de ces ressources génomiques sur des échantillons de sol comportant des centaines d'espèces au génome en majorité inconnu, et de pouvoir estimer leur précision et leur fiabilité. Dans le contexte de ce stage, le stagiaire mènera une approche comparative entre annotations attendues et prédites, en utilisant des génomes connus mais non contenus dans MycoCosm, et en retirant des génomes ciblés de MycoCosm avant de les interroger contre le reste de la base de données. Des approches d'annotation par intelligence artificielle et de données simulées pourront également être mise en œuvre.

Objectifs

- Designer et générer des sous-échantillons de la base de données Mycocosm permettant de tester différents scénarios de proximité phylogénétique
- Mettre en œuvre les outils d'annotation fonctionnelle sur les données générées
- Designer et développer des métriques et méthodes d'analyse des résultats obtenus
- Analyser les résultats générés
- Participer à la préparation d'une publication à partir des résultats obtenus. Présenter les résultats obtenus en conférence (JOBIM 2024 à Toulouse fin juin)
- (*) mettre en œuvre des outils d'annotation de séquence par intelligence artificielle et en comparer les performances d'autres outils d'annotation sur les métriques choisies
- (*) simuler des données de séquençage à partir de données génomiques pour tester l'ensemble du pipeline (en plus de l'étape d'annotation)

(*) : optionnel et en fonction de l'avancement du projet et de l'intérêt et l'affinité de l'étudiant

Compétences requises

Les candidats devront être inscrits en Master2 bio-informatique ou équivalent, et répondre à minima aux critères suivants :

- Maîtrise d'Unix/Bash
- Maîtrise d'au moins 1 langage parmi python / perl / R
- Compétences en phylogénie et connaissances minimales en biologie moléculaire et cellulaire (relatives aux annotations fonctionnelles)
- Bonnes compétences en statistiques

Encadrement et conditions d'accueil

L'étudiant ou l'étudiante sera accueilli au sein de l'unité Interactions Arbres-Microorganismes (<https://mycor.nancy.inra.fr/> @UMR_IAM), dans l'équipe Ecogénomique des Interactions, sur le Centre INRAE Grand Est Nancy à Champenoux.

Il sera encadré et dirigé par Lucas Auer, ingénieur de recherche en bio-informatique et statistiques spécialisé dans les approches méta-omiques, et s'intégrera dans l'équipe qui compte 11 chercheurs et ingénieurs permanents (dont 2 ingénieurs en bio-informatique) et une dizaine de doctorants, postdoctorants, chercheurs étrangers accueillis et contractuels (dont 1 doctorante et une postdoctorante en bio-informatique), ainsi que 4 ou 5 stagiaires M2 sur la période. L'équipe héberge la « plateforme Ecogénomique », ce qui lui permet de disposer de son propre cluster de calcul (200 CPU, 2To DDR).

Ce stage s'inscrit dans le cadre du projet FunGAAL (Fungal Genes Annotation Accuracy and Limits), financé par le Laboratoire d'Excellence ARBRE.

La gratification de stage est au taux horaire pour les organismes publics (~575€/mois). Le Centre INRAE Grand Est Nancy dispose d'un restaurant d'entreprise (tarif étudiant) et fournit gratuitement une carte de transports en commun pour rejoindre le Centre depuis Nancy.

Le stage est prévu pour une durée de 6 mois et devra débuter entre janvier et mars 2024, à la convenance du stagiaire. Il inclut un déplacement à Toulouse fin juin pour y présenter les résultats en congrès (JOBIM).