Stage M2 ou 3A Ingénieur, domaines bioinformatique, écologie ou évolution.
Durée : 5 à 6 mois, débutant mars ou avril 2027.

Sujet : Extension d'une plateforme de benchmarking dans le cadre de la détection et l'identification de virus émergents

Le suivi des émergence virales est une activité importante dans le contexte de la santé animale. Il permet de réagir rapidement lorsque qu'une potentielle nouvelle épidémie débute, limitant ainsi les impacts économiques et sociaux associés. Aujourd'hui, le séquençage d'ADN via les technologies de reads longs Nanopore offre la capacité de séquencer très rapidement des échantillons animaux et de repérer des fragments d'ADN associés à un nouveau virus ou un nouveau variant [1].

Cette identification se base se base sur une analyse taxonomique du virus échantillonné. Elle se fait souvent par des outils bio-informatiques basés sur des k-mers et des algorithmes de classification ; on peut citer les outils Kraken [2], très utilisé en métagénomique virale. L'analyse fine et phylogénétique du génome échantillonné arrive dans un second temps, souvent après assemblage des reads et plusieurs étapes de contrôle qualité. Les séquences candidates sont introduites dans un arbre de génomes viraux connus, potentiellement datés ou géographiquement localisés. Ces informations supplémentaires permettent d'affiner le scénario d'émergence, et donc les actions à prendre pour limiter son impact.

Cette inférence phylogénétique n'est pas applicable directement aux milliers de reads séquencés par le Nanopore, car le coût de calcul associé est trop important. Le « placement » phylogénétique est une méthode alternative compatible avec l'analyse de millions de reads pour un coût de calcul bien inférieur. Elle présente le potentiel d'une analyse bioinformatique fine directement depuis la production des reads viraux, permettant un gain de temps et de réactivité. Les méthodes basées sur les phylo-k-mer, développées dans nos équipes, portent fortement ce potentiel de mise à l'échelle [3].

Le stage s'intéressera au benchmarking des outils de « placement » phylogénétique dans le contexte de la classification taxonomiques des séquences virales. Il s'appuiera sur un pipeline d'évaluation existant, et cherchera à l'étendre sur deux aspects :

- La collecte et l'évaluation de la qualité de nombreux arbres phylogénétiques viraux, afin d'établir une liste de genres et espèces virales candidates pour étendre les données d'entrée du benchmark.
- L'extension d'une plateforme de benchmarking PEWO [4], pour pourvoir évaluer la qualité des identification taxonomiques atteignables sur la base de ces données génomiques. Cette étape impliquera le développement de nouvelles métriques pour la comparaison des outils de classification taxonomique testés et l'évaluation de la qualité de l'identification.

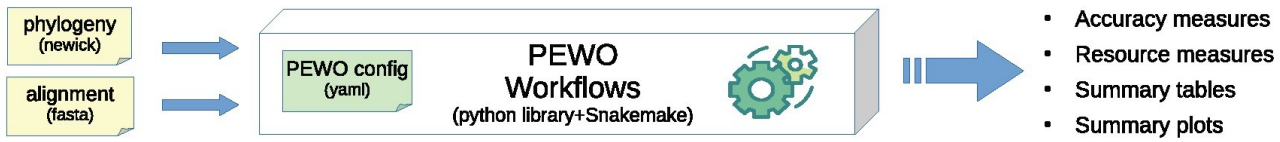

Figure 1. La plateforme de benchmark PEWO permet d'évaluer la qualité des placements phylogénétiques produits à partir d'un arbre de référence donné. Il permet d'estimer la qualité de la prédiction taxonomique d'un nouveau génome par rapport aux espèces et variants connus. Il sera étendu pour une application à de larges panels d'espèces virales, dans un contexte de détection de virus émergents.

Le but général du stage est d'évaluer la précision du placement phylogénétique dans un contexte de diversité virale bien plus large. Le stage s'intéressera donc à la fois a) à explorer, compiler et évaluer des données génomiques virales complexes, bruitées et souvent incomplètes du point de vue phylogénétique, et b) impliquera une forte composante de développement basée sur Python et le gestionnaire de WorkFlow Snakemake.

Il sera hébergé au MIAT (INRAE Occitanie-Toulouse), dans le cadre d'un projet nationale (ANR Pro-k-mer) il se co-encadré par des collègues de l'ENVT (compétences de virologie vétérinaire) et impliquera des interactions avec l'équipe MAB du LIRMM à Montpellier (Fabio Pardi, Eric Rivals, compétences phylogénétiques). Il impliquera des échanges réguliers, et de potentiels déplacements chez ces partenaires.

REFERENCES :
[1] https://doi.org/10.3389/fcimb.2024.1257586 [2] https://doi.org/10.1038/s41596-022-00738-y [3] https://doi.org/10.1093/bioinformatics/btad692
[4] https://doi.org/10.1093/bioinformatics/btad692

Objectifs du stage :

- Produire une collection de jeu de données virales pour amorcer le benchmark (arbre phylogénétiques issus de bases de données.
- Étendre les pipelines d'évaluation pour évaluer la qualité des identifications taxonomiques basées sur le placement phylogénétique et des méthodes de classification supervisées (type Kraken).
- Développer de nouveaux critères d'évaluation de la qualité des identification phylogénétiques produites.
- Implémenter les extensions du pipeline via du code python et le gestionnaire de workflow snakemake.

Profil de candidat souhaité :

- Une expérience du langage python est obligatoire.
- Des bases en analyse des séquences biologique et en phylogénie.
- Une connaissance des outils Snakemake ou NextFlow sera bienvenue, mais non obligatoire.
- Intérêt pour les contextes fortement multidisciplinaires et appliqués.
- Autonomie et capacité de travail en équipe, de rédaction, de synthèse.

Encadrement :

- Le stage sera co-encadré par Benjamin Linard (MIAT, INRAE, Toulouse), spécialisé dans le développement d'algorithmes et d'outils pour l'analyse des génomes et métagénomes, Guillaume Croville (ENVT, INRAE, Toulouse), et Fabio Pardi (LIRMM, CNRS, Montpellier).
- Le ou la stagiaire sera hébergé au sein de l'unité MIAT (Mathématique et Informatique Appliqués de Toulouse), à l'INRAE Occitanie-Toulouse (24, Chemin de Borde Rouge 31320 Auzeville-Tolosane).
- Des déplacements ponctuels seront possibles en interaction avec les équipes co-encadrantes à Toulouse et à Montpellier.

Contact :

benjamin.linard@inrae.fr
Envoyer CV + lettre de motivation.