LABORATOIRE DE BIOMÉTRIE ET BIOLOGIE ÉVOLUTIVE

Proposition de stage de Master en Bioinformatique

Venez explorer le côté obscur du vivant à partir des données massives de séquençage métagénomique

Lieu: Université Claude Bernard Lyon1, PRABI-AMSB (Villeurbanne)
Durée : 6 mois, dès février 2023
Encadrement : **Vincent Navratil** (PRABI, Univ Lyon 1) vincent.navratil@univ-lyon1.fr ;
Damien de Vienne (CNRS, LBBE, Univ Lyon 1) damien.de-vienne@cnrs.fr

1. Contexte scientifique du projet :

Depuis deux décennies et avec les avancées de la métagénomique, les chercheurs se sont rendus compte que le séquençage des génomes des organismes présents dans un milieu donné (sol, eau, sang, intestin, etc.) révélait systématiquement une diversité énorme et jusqu'alors inconnue. L'étude de cette **"matière noire biologique"** [1] (biological dark matter), composée de séquences appartenant à des espèces non identifiées et généralement non cultivables en laboratoire (bactéries, virus, champignons, etc.) a permis des avancées scientifiques importantes notamment en santé humaine. Par exemple, l'étude du microbiome intestinal humain a permis un inventaire des séquences trouvées chez de très nombreux patients et à la mise au point de méthodes diagnostiques pour prédire des prédispositions à certaines maladies.

Mais l'étude de la matière noire peut avoir plusieurs autres intérêts, notamment en biologie évolutive et en écologie: quantifier le plus finement possible la biodiversité actuelle des différents écosystèmes, son évolution au cours du temps, ou en fonction de facteurs environnementaux (effet du réchauffement climatique, effet de l'anthropisation, etc.). En effet, le nombre d'espèces vivant sur terre aujourd'hui est encore inconnu, et les estimations sont très variables, de quelques millions [2] à plusieurs milliards [3], voire millions de yota (10^{31}) si l'on intègre la biodiversité du monde viral [4]. Or les projets de séquençage de métagénomes et de l'ADN environnemental (eDNA) sont de plus en plus nombreux [5], et couvrent une grande diversité de localisation géographiques et des écosystèmes. Il est donc désormais possible de quantifier directement depuis les données de séquençage le nombre et l'abondance des espèces déjà répertoriées sur terre et d'estimer la matière noire biologique encore inexplorée. Cette entreprise est de plus facilitée par l'arrivée récente de nouveaux algorithmes bioinformatiques et d'infrastructures de calculs extrêmement performants [6,7] permettant d'effectuer de façon automatique et massive, à savoir à l'échelle de Pétabases de données, l'assignation taxonomique de l'intégralité des séquences présentes dans les archives publiques de séquences tel que Short Read Archive (SRA, https://www.ncbi.nlm.nih.gov/sra). Le projet de master proposé est une première étape dans la réalisation d'une méta-analyse de l'ensemble des métagénomes disponibles sous SRA dans l'optique d'explorer, de quantifier et de mieux caractériser cette matière noire biologique par une approche biogéographique.

2. Objectifs du projet :

L'objectif du stage est **d'évaluer par une approche bioinformatique la qualité des assignations taxonomiques proposées massivement par l'algorithme STAT** en les comparant, pour un sous-ensemble de métagénomes préalablement sélectionnés, aux assignations taxonomiques inférées par d'autres méthodes référentes dans le domaine (kraken2, kraken unique, centrifuge, Metaphlan4). Compte tenu de leur rôle prépondérant dans le fonctionnement de la biosphère, un intérêt particulier sera porté sur l'exploration du microbiome et plus particulièrement du virome et des bactériophages (virus infectant les bactéries) au sein des fractions de séquences assignées 1) aux génomes de références, 2) aux Métagenome-Assembled Genomes disponibles (MAG) et 3) aux lectures non-assignées restantes. Après une revue la de bibliographie (peu abondante) sur les méthodes de quantification de la matière noire biologique, l'étudiant·e 1) proposera et et implémentera un ou plusieurs indices permettant de quantifier et caractériser cette matière noire biologique pour les échantillons sélectionnés, et 2) étudiera l'influence du choix des bases de données de références (génomes de références, MAG) et des biais liés à l'acquisition des séquences (profondeur de séquençage, librairies, protocole d'extraction etc.) sur ses résultats. Ces indices permettront de proposer un contour rationnel de cette matière noire biologique en explorant notamment les communautés biologiques et leur interactions ainsi que les ecosystèmes associées. Si l'étudiant·e est intéressé·e, ce projet pourra s'ouvrir sur un sujet de thèse innovant et transdisciplinaire avec une dimension biogéographique en collaboration avec le laboratoire EVS de l'Université Lyon 2.

Le·la candidat·e travaillera - en collaboration étroite avec les responsables scientifiques du stage et nos collaborateurs du laboratoire EVS Lyon2 pour la composante géographique - pourra bénéficier d'un support technique de la plateforme PRABI-amsb (http://amsb.prabi.fr) pour être guidé au mieux dans les choix méthodologiques et leur mise en œuvre (bonnes pratiques de développement, infrastructure de calcul). Les bases de données, les jeux de données, l'infrastructure informatique et les outils bioinformatiques nécessaires à la bonne réalisation des tâches listées ci-dessus seront disponibles dès le début du projet.

3. Compétences recherchées :

Nous recherchons un·e candidat·e de niveau bac **+3 ou + 4 en bioinformatique**. Il·elle devra être capable d'interagir avec différents chercheurs et ingénieurs du projet pour développer et mettre en place des méthodes innonvantes en sciences de la vie.

Compétences requises :

- Python, R, algorithmes bioinformatiques

- versionning sous git

- Fortes capacités relationnelles et d'organisation, autonomie

- Esprit d'analyse et de synthèse

- Capacités rédactionnelles

4. Informations pratiques :

Ce stage bénéficiera d'une gratification de 6 mois à partir de février 2023 (552 euros/mois).
Pour tout renseignement ou candidature, merci d'adresser un CV et une lettre de motivation le plus rapidement possible pour démarrage le 01/02/2023 à l'attention de : **Vincent Navratil** (PRABI-AMSB, Univ Lyon 1) vincent.navratil@univ-lyon1.fr ; **Damien de Vienne** (CNRS, LBBE, Univ Lyon 1) damien.de-vienne@cnrs.fr

Bibliographie

1. Marcy, Y. *et al.* Dissecting biological 'dark matter' with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 11889–11894 (2007).
2. Chapman, A. D. *Numbers of living species in Australia and the world*. (Department of the Environment Water Heritage and the Arts, 2009).
3. Locey, K. J. & Lennon, J. T. Scaling laws predict global microbial diversity. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 5970–5975 (2016).
4. Cobián Güemes, A. G. *et al.* Viruses as Winners in the Game of Life. *Annu. Rev. Virol.* **3**, 197–214 (2016).
5. Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457–463 (2017).
6. Katz, K. *et al.* The Sequence Read Archive: a decade more of explosive growth. *Nucleic Acids Res.* **50**, D387–D390 (2022).
7. Katz, K. S. *et al.* STAT: a fast, scalable, MinHash-based k-mer tool to assess Sequence Read Archive next-generation sequence submissions. *Genome Biol.* **22**, 270 (2021).