Proposition de stage de Master 2 en bioinformatique

Contact : Laurence Garczarek / Gregory Farrant

Courriel : laurence.garczarek@sb-roscoff.fr / gregory.farrant@sb-roscoff.fr

Structure : Station Biologique de Roscoff, Roscoff, France
UMR 7144 Adaptation et Diversité en Milieu Marin (AD2M)

Statut : M2 à pourvoir

Titre du stage :
Développement d'une base de données de marqueurs génétiques pour inférer les variations spatiales et temporelles des picocyanobactéries marines *Prochlorococcus* et *Synechococcus*

Mots clés résumant les méthodes et techniques à utiliser au cours du stage :
Python, bash, R, postgresql, utilisation des banques de données publiques, blast, alignements, placement phylogénétique et analyses de variants

Résumé du projet de stage :

En permettant d'acquérir d'énormes sets de données environnementales, l'avènement des nouvelles techniques de séquençage à haut débit a révolutionné nos connaissances sur les communautés microbiennes et notre compréhension des écosystèmes, notamment dans le milieu marin. Cependant, la détection de changements significatifs dans la structure des communautés est encore fortement limitée par i) la rareté des bases de référence expertes nécessaires à l'assignation taxonomique fiable des séquences, ii) le manque d'homogénéité entre les analyses génétiques réalisées sur différents sites et iii) la dispersion de l'expertise taxonomique sur différents groupes d'organismes.

A ce niveau, plusieurs marqueurs phylogénétiques sont couramment utilisés pour analyser la diversité génétique et la distribution des picocyanobactéries marines, *Prochlorococcus* et *Synechococcus*, qui constituent les deux organismes photosynthétiques les plus abondants dans les océans, et qui a ce titre jouent un rôle majeur dans tous les grands cycles biogéochimiques de l'océan [1] : i) L'ARNr 16S, codant pour la petite sous-unité des ribosomes, constitue le marqueur universel des procaryotes et est de ce fait disponible pour de nombreuses souches en culture [1]. Cependant ce marqueur ne permet d'assigner les séquences environnementales qu'au niveau du clade, un niveau taxonomique insuffisant pour identifier les écotypes, i.e. les groupes génétiques présentant différents patrons de distribution *in situ*, ii) *petB*, un gène photosynthétique (chloroplastique chez les eucaryotes), codant pour le cytochrome b_6, qui présente une résolution taxonomique beaucoup plus fine, de plus en plus utilisé par la communauté scientifique travaillant sur *Synechococcus* [2, 3], iii) *rpoC1*, un gène également chloroplastique chez les eucaryotes, codant pour la grande unité de l'ARN polymerase et qui présente une résolution taxonomique similaire à *petB* mais pour lequel la correspondance avec les clades identifiés par l'ARNr 16S est plus difficile à réaliser [4], et enfin plus récemment iv) *psbO*, codant pour une protéine de stabilisation du cluster manganèse du photosystème II, et qui peut également être utilisé pour étudier la diversité

des eucaryotes photosynthétiques puisqu'il est codé par le noyau des cellules eucaryotes et donc présent en simple copie dans la cellule.

L'objectif de ce stage sera d'élaborer à partir d'une base de référence existante développée par l'équipe d'accueil (base CyanoRefDB utilisant le gène *petB*), un base multi-marqueurs intégrant les trois autres marqueurs génétiques couramment utilisés (ARNr 16S, *rpoC1*, *psbO*), de tester leur résolution taxonomique respective et d'adapter les pipelines d'analyse de données omiques à ces différents marqueurs. Pour cela, l'étudiant(e) sélectionné(e) sera chargé(e) i) d'extraire à partir des bases de séquences publiques (NCBI, EMBL) les séquences disponibles correspondant à chacun de ces marqueurs, ii) de réaliser des analyses phylogénétiques afin d'assigner taxonomiquement ces séquences en utilisant les séquences de référence disponibles pour chacun de ces marqueurs et enfin iii) d'utiliser les organismes pour lesquels plusieurs marqueurs sont disponibles afin de réconcilier la taxonomie basée sur les différents marqueurs. Dans un deuxième temps, la résolution taxonomique de ces marqueurs sera testée en adaptant des scripts maison, préalablement développés pour le gène *petB*. Enfin, l'étudiant(e) pourra également adapter et développer des pipelines d'analyse de la diversité utilisant ces bases de références par placement phylogénétique et analyses de variants en utilisant pour cela les données de métabarcoding et métagénomiques de la campagne *Tara Océans* et/ou de la série à long terme SOMLIT ASTAN. Cette étude qui s'intègre dans le cadre du projet RoskoBaz (EU-H2020 Assemble Plus et SAD Région Bretagne), pourra donner lieu à la rédaction d'un article décrivant la base de référence CyanoRefDB.

Nous recherchons un(e) étudiant(e) de Master 2 motivé, ayant une solide formation en bioinformatique et génomique. Le projet nécessite notamment une expérience générale en programmation (Python, Bash, R), en analyse de données de séquences (blast, utilisation des banques de données publiques, alignements, analyses phylogénétiques) et si possible en développement logiciel. Des connaissances en bases de données relationnelles (PostgreSQL) et/ou en écologie constitueront également un atout majeur.

Références

[1] Scanlan DJ, et al. (2009) Ecological genomics of marine picocyanobacteria. Microbiol Mol Biol Rev 73:249–299.

[2] Mazard S, Ostrowski M, Partensky F, Scanlan DJ. Environ Microbiol. 2012 14: 372-86.

[3] Farrant GK, Doré H, Cornejo-Castillo FM, Partensky F, Ratin M, Ostrowski M, et al. Delineating ecologically significant taxonomic units from global patterns of marine picocyanobacteria. Proc Natl Acad Sci USA. 2016;113:E3365–74

[4] Kent AG, Baer SE, Mouginot C, Huang JS, Larkin AA, Lomas MW, Martiny AC. ISME J. 2019 13:430-441.