Proposition Sujet de thèse :

Janvier 2022

"Applying Hybrid Evolutionary Machine Learning techniques on Environmental DNA data to predict biodiversity in marine environment"

« Prédiction du niveau de biodiversité des lieux marins basée sur des techniques d'apprentissage évolutionnaires hybrides appliquées aux ADN Environnemental »

Cadre

— **Encadrants de la thèse :**

Sana Ben Hamida (Maître de Conférences) LAMSADE — UMR 7243

Sana.mrabet@dauphine.psl.eu

Marta Rukoz (Professeur), LAMSADE — UMR 7243

marta.rukoz@lamsade.dauphine.fr

— **École doctorale** : École doctorale de Dauphine - Université Paris Dauphine-PSL

— **Laboratoire d'accueil** : LAMSADE - Université Paris Dauphine-PSL. Pôle Sciences de Données.

Contexte :

L'évaluation du niveau de la biodiversité dans les lieux marins et terrestres est actuellement la pratique universelle pour estimer l'impact des activités humaines et le changement climatique sur les écosystèmes de ces milieux. Une démarche innovante et moderne pour effectuer cette observation est le séquençage haut-débit d'ADN prélevé dans l'environnement (« ADN environnemental ») [3].

L'analyse de l'ADN environnemental (ADNe) consiste principalement en l'identification d'espèces à partir de l'ADN qu'elles laissent dans leur environnement. De nombreuses études montrent que l'utilisation de l'ADNe permet une bonne estimation des différentes espèces (ou taxons) présents dans différents types d'environnements.

L'analyse de l'ADNe s'appuie sur des techniques classiques de biologie moléculaire (PCR, séquençage…). Les taxons identifiés sont attribués à des poids écologiques qui sont utilisés pour calculer les indices biotiques (BI). Ces indices permettent de déterminer la qualité écologique du lieu en question (généralement classée dans cinq catégories de « très mauvais » à « très bon »). Cependant, ces études se sont appuyées sur des bases de séquences de référence (Silva[1], Greengenes[2], LTP) pour l'assignation taxonomique, afin de récupérer des poids écologiques spécifiques aux taxons et de calculer les valeurs BI. Cette phase peut échouer si quelques séquences du milieu à analyser sont incomplètes ou n'apparaissent pas dans les bases de références.

Des travaux récents [1,2,5] ont démontré que l'apprentissage automatique peut être utilisé pour prédire des valeurs précises des indices biotiques à partir de *métabarcoding* de l'ADNe, quel que soit l'affiliation des séquences. L'idée est de générer un modèle de clustering permettant de grouper les séquences proches (un fort pourcentage d'identité) appartenant à un taxon bien défini. La démarche consiste alors à utiliser une technique de clustering sur les données d'analyse des échantillons marins

[1] https://www.arb-silva.de/
[2] https://greengenes.secondgenome.com/

applicable même si les séquences déterminées sont incomplètes. Ces séquences sont alors affiliées automatiquement au taxon de la séquence centrale du cluster identifiée dans une base de référence. La validité du modèle peut être vérifiée en cherchant l'affiliation exacte des centres des clusters dans les bases de référence

Objectif de la thèse :

Le sujet de cette thèse a un double objectif. Le premier objectif est la mise en place de techniques d'apprentissage conventionnelles pour le clustering des données relatives aux ADNe d'un milieu marin permettant d'évaluer le BI de ce milieu. Les modèles générés à cette phase peuvent être des solutions non généralisables et très dépendantes des méthodes de séquençage utilisées dans les échantillons analysés. D'où le deuxième objectif de la thèse qui vise la généralisation des modèles de prédiction aux données d'autres échantillons du même milieu marin en utilisant des méthodes évolutionnaires. L'idée est de développer un algorithme évolutionnaire hybride qui fait évoluer des modèles de clustering en fonction des paramètres des échantillons (i.e. marqueurs de séquençage) et les paramètres de clustering (i.e. indicateurs de mesure de distance). Le schéma du modèle peut se baser sur celui du « Genetic K-means algorithm » (GKA) de Krishna et Murty [6].

Les premiers modèles sont à tester sur des échantillons marins étudiées par Cordier [1,2] et disponibles en ligne. D'autres échantillons seront étudiés après validation de l'algorithme. La validation de la méthode et son application sur d'autres données marines peuvent amener à un éventuel échange avec OFB (Office Français de la Biodiversité).

Prérequis : Des bonnes compétences en Machine Learning et programmation Python avec, de préférence (mais pas obligatoirement), des connaissances en Bio-informatique.

Références :

[1] Cordier T, Forster D, Dufresne Y, Martins CIM, Stoeck T, Pawlowski J. Supervised machine learning outperforms taxonomy based environmental DNA metabarcoding applied to biomonitoring. *Molecular Ecology Resources*. November 2018; Vol. 18, Issue 6: 1381-1391. https://doi.org/10.1111/1755-0998.12926

[2] Tristan Cordier, Philippe Esling, Franck Lejzerowicz, Joana Visco, Amine Ouadahi, Catarina Martins, Tomas Cedhagen and Jan Pawlowski, Predicting the Ecological Quality Status of Marine Environments from eDNA Metabarcoding Data Using Supervised Machine Learning. June 2017 - Environmental Science and Technology 51(16) DOI: 10.1021/acs.est.7b01518

[3] Actes de la journée ADN Environnemental – Tristan LEFEBURE, Université Lyon 1 – Barcoding, Métabarcoding : petite entrée en matière - http://www.graie.org/zabr/zabrdoc/Actes_Adne_web.pdf

[4] Guilhem Sommeria-Klein. From models to Data: understanding biodiversity patterns from environmental DNA data. Doctoral dissertation, Université Paul Sabatier-Toulouse III. 2017.

[5] Dully, V., Wilding, T. A., Mühlhaus, T., & Stoeck, T. (2021). Identifying the minimum amplicon sequence depth to adequately predict classes in eDNA-based marine biomonitoring using supervised machine learning. Computational and structural biotechnology journal, 19, 2256-2268.

[6] Krishna K, Narasimha Murty M. Genetic K-means algorithm. IEEE Trans. Systems Man Cybernetics Society, Part B (Cybernetics). June 1999; 29(3):433-439. doi: 10.1109/3477.764879. PMID: 18252317