Ph.D. Scholarship Application

February 2022

Title: “Applying Hybrid Evolutionary Machine Learning techniques on Environmental DNA data to predict biodiversity in marine environment”

Summary

Evaluating the level of biodiversity in marine and terrestrial locations is currently a universal practice for estimating the impact of human activities and climate change on the ecosystems of these environments. An innovative and modern approach to perform this observation is the high-throughput sequencing of DNA taken from the environment ("environmental DNA)

Environmental DNA (eDNA) analysis consists mainly of the identification of species from the DNA they leave in their environment. Many studies show that the use of eDNA allows a good estimation of the different species (or taxa) present in different types of environments. eDNA analysis is based on classical molecular biology techniques (PCR, sequencing...). The identified taxa are assigned to ecological weights which are used to calculate the biotic indices (BI). These indices are used to determine the ecological quality of the site in question (generally classified in five categories from "very poor" to "very good"). However, these studies relied on reference sequence databases (Silva[1] Greengenes[2] LTP) for taxonomic assignment, in order to retrieve taxon-specific ecological weights and to calculate BI values. This phase may fail if some sequences in the environment to be analysed are incomplete or do not appear in the reference databases.

Recent works [1,2,5] has demonstrated that machine learning can be used to predict accurate values of biotic indices from eDNA metabarcoding, regardless of sequence affiliation. The idea is to generate a clustering model to group closely related sequences (high percentage of identity) belonging to a well-defined taxon. The approach then consists in using a clustering technique on the analysis data of the marine samples applicable even if the determined sequences are incomplete. These sequences are then affiliated automatically to the taxon of the central sequence of the cluster identified in a reference database. The validity of the model can be checked by searching for the exact affiliation of the cluster centres in reference databases.

Objectives:

The subject of this thesis has a double objective. The first objective is the implementation of conventional techniques for clustering eDNA data of a marine environment allowing to evaluate the BI of this environment. The generated models in this phase may be non-generalizable solutions and very dependent on the sequencing methods used in the analysed samples. Hence the second objective of the thesis which aims at the generalization of prediction models to data from other samples of the same marine environment using evolutionary methods. The idea is to develop a hybrid evolutionary algorithm that evolves clustering models as a function of sample parameters (i.e. sequencing markers) and clustering parameters (i.e. distance measurement indicators). The model scheme can be based on the "Genetic K-means algorithm" (GKA) of Krishna and Murty.

[1] https://www.arb-silva.de/

[2] https://greengenes.secondgenome.com/

The first models are to be tested on marine samples studied by Cordier [1,2] and available online. Other samples will be studied after validation of the algorithm. The validation of the method and its application on other marine data may lead to a possible exchange with OFB (French Office of Biodiversity).

We seek a student with:

- Good knowledge in Machine Learning
- Good knowledge in Computer Science and Mathematics.
- Good programming skills, especially in Python programming.
- Some knowledge in Bio-informatic would be good but it is not mandatory.
- Good command of written and spoken English.
- Master's degree in computer science or equivalent degree giving access to PhD studies.

Advisors and contact:

Sana Ben Hamida (Sana.mrabet@dauphine.psl.eu) and Marta Rukoz
(marta.rukoz@lamsade.dauphine.fr)

Offer:

- Starting date: September 2022
- Fixed-term contract of 3 years.
- Job location: LAMSADE CNRS UMR 7243, University of Paris Dauphine, France.

To apply:

Please send the following material before April 15th, 2022 to Sana.mrabet@dauphine.psl.eu and marta.rukoz@lamsade.dauphine.fr:

- fully detailed CV,
- academic records (Master degree or equivalent),
- cover letter,
- recommendation(s) and supporting letter(s).

References

[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