Computational Biology II (populations, interactions, evolution) Microbiome project (2022)

antoine.frenoy@univ-grenoble-alpes.fr

Guidelines

This project will be realized by teams of one or two students.

Email me (or upload on Teide if it comes back to life) your work **before 2023-01-15 at 20:00** (strict deadline). You should provide your code, a synthetic report (including figures when relevant), and a file containing your prediction as explained below.

Your report should not only present your results or paraphrase your code, but should also explain and defend your choice of methods and offer a critical look at your results.

I must be able to run your code to reproduce your results with minimal efforts, which implies that your organize and document it.

Each team is expected to produce individual work: sharing of code or answers with other teams is strictly forbidden, as well as copying code or material found online or written by someone else. However, your are encouraged to:

- use any documentary source you want, with appropriate references (this does *not* allow you to copy-paste text without quotation marks and proper references).
- use the standard Python data science packages (scipy, numpy, matplotlib, pandas, scikit-learn). If you would like to use other existing software package, ask the lecturer beforehand. Same if you would like to use anoother programming language than Python.

Important news 2022-12-16:

- As several of you told me they had problems with the `hdf` files, I replaced them by `csv` versions of the exact same datasets. I further noticed that the column names (bacterial taxons) were previously lost in the `hdf` export, which means it was not possible for you to use the information provided in `bacterial_species.csv`. The `csv` files now contain this information.
- Some of you asked me about the RAM requirements: this is clearly a large dataset, and finding how to work with it efficiently is part of the challenge. It is up to you to do any relevant dimensionality reduction. Note that many reductions are possible without loading the full dataset in memory.

Context and expected work

A consortium of microbiology labs is working to catalog the abundance of all known bacterial species in a large number of different environments (terrestrial, aquatic, associated with humans, animals or plants, etc...).

The objective of this project is to use their data to build a machine learning model predicting the environment in which a sample (described by a vector of abundances of the different bacterial species) has been collected.

The training dataset is provided in the files `training_descriptors.hdf`, which indicates the number of occurrences of each of the ∼ 122000 known bacterial species (recognized from their 16S ribosomal RNA) in each of the 20983 samples of the training set; associated with the file `training_environments.csv`, indicating the environment in which each of these samples was taken. These environments are described at three different levels of precision (`empo_1` is the most generic, `empo_3` the most precise).

The file `bacterial_species.csv` further describes the different bacterial species in the dataset at different taxonomic levels: it may help to bring more biological knowledge about the descriptors into the model.

A challenge dataset is provided in the file `challenge_descriptors.hdf`, containing 2332 samples described by the respective abundances of the different known bacterial species. Your task is to infer as precisely as possible the environments in which each these samples have been collected. Your will store your predictions in a file `predictions.csv` that will be uploaded with your work. Each line of this file will indicate the name of the sample and the infered environment, in the same format than the file `training_environments.csv`.

Note that before performing this challenge on the test dataset, you are expected to assess the performance of your predictive model using cross-validation on the training dataset. A critical assessment of your model and predictions is expected.

In addition to statistical predictions, we would like to get biological information from the model: for example which species or taxa are discriminating for the detection of the environment? Any relevant information can be included in your report.