# **Computational Microbiome Analysis Workshop – Final Project Instructions**

## Project description

The objective of this project is to extrapolate microbiome composition based on dense time series using data from the Amboseli baboon project.

The project will be done in groups of 2 students. If you are looking for a partner, you can use the relevant forum on Moodle.

## Milestones

Throughout the course, there will be three milestones to ensure that progress is being made and to provide a formal opportunity for group discussions. Prior to each checkpoint meeting, you need to prepare a concise presentation (~20 minutes) that summarizes your progress and future plans. Coordinating these checkpoint meetings will be done via email. Please send your PowerPoint presentation at least 24 hours in advance.

Minimum progression expected and deadlines for each milestone is as follows:

|  |  |  |
| --- | --- | --- |
|  | Deadline | Minimum progression expected |
| Milestone 1 | 22/7/24 | Presenting preliminary exploration results and the main ideas that will be implemented in the pipeline. See below guiding questions for data exploration |
| milestone 2 | 15/9/24 | Presenting preliminary results. To present preliminary results, it is recommended to start with an overview of the current pipeline. Then, to test the model's performance, on two scenarios: 1. Extrapolate future compositions based on long measurements (n>=10). 2. Iterative extrapolation based on the first two timepoints. |
| Milestone 3 | 1/11/24 | Presenting the final results, including full pipeline, validation methods, their main results and possible future directions. **Model** **predictions should be sent a week before Milestone 3 meeting.** During the meeting we will present to you the model performance on the test set. |

Guiding questions for data explorations (milestone 1)

1. Basic overview – How many samples do we have from each subject? What is the time difference between them? Visualize.
2. Microbiome ordination – Calculate Bray-Curtis distances between all microbiome samples and plot PCoA and calculate average distance between subsequent samples. What are the auto correlations between a subject time series?
3. Explore FFQ data – plot the first two PCs and visualize samples. are the samples clustered according to social group or seasonality?
4. Naïve modeling – Try to predict the microbiome composition according to previous timepoints. Estimate the prediction accuracy based on Bray-Curtis dissimilarity. Does adding FFQ data improve the prediction?

## Final submission guidelines

Final submission includes: 1. Report 2. GitHub with your code 3. CSV with predictions on test set.

### Report

A detailed pdf document describing the problem, the pipeline, validation and validation results and possible future directions. The paper be max 3 pages long with up to 4 figures (not including bibliography). It should have the following section:

* Introduction – describing the problem, and a general description of the solution's approach, pipeline and a summary of the findings.
* Methods – detailed description of the pipeline and validation.
* Results – detailed description of your findings and the validation results. You may include the results of experiments you have done even if the results were not so good, and you excluded them later.
* Discussion – summary of the results, their importance and future possible directions.
* Bibliography – we recommend using either Mendeley or Zotero to insert bibliography.

Creating figures to visualize the results, pipelines, etc. is strongly encouraged. To easily create figures in Python we suggest using either seaborn, plotly or other similar Python libraries. **The report should be sent a week after milestone 3 meeting.**

### Code

Code should be submitted using a GitHub. Keep the code clean, well-documented and easy to read.

## Predictions on test data

Before submission you will be given metadata and FFQ in a similar structure to these of the training set. **Predictions should be sent a week before Milestone 3 meeting**. During the meeting we will present to you the model performance on the test set.