Introduction to Focus Areas in Bioinformatics – WS19/20

Prof. Dr. Tim Conrad

6 Project 6

- Deadline: 01.12.2019, 23:59
- All files need to be available through your GIT repository, in the directory "Project 6".
- You can work in teams up to 3 people. Please state in the report the group member names.
- If your code is in python, I must be able to run your code within a Google Colab notebook. If your code is not in Python or R, you must provide a manual how to compile and run it on a Linux machine.

6.1 Analyzing Microbiome time-series data

We will use the Prophet library (https://facebook.github.io/prophet/) to analyze and forecast time-series data and apply it to longitudinal data from Microbiome measurements. The data is described in the paper "Moving pictures of the human microbiome" by Caporaso et al. and can be downloaded from here: https://github.com/twbattaglia/MicrobeDS/tree/master/data-raw/MovingPictures .

Each of the following sub-tasks should be done on at least 2 selected species (OTUs) – which should come from the same person and the same body site. All resulting figures and answers to the questions must appear in your report.

(1) Visualize the time-series data.
(2) Compute, visualize and briefly discuss the components of the time-series (trend, etc.).
(3) Perform and visualize a forecasting analysis and include an uncertainty visualization.
(4) Compare the results from the additive seasonalities to multiplicative seasonalities.
(5) Perform a change-point analysis and visualize detected change-points. Can the change points be explained in a meaningful way?
(6) Check whether the data contains outliers that should be removed for the model forecast. If you think that this is the case, remove the outliers and repeat the forecasting analysis. Compare the results (with outliers included and outliers removed).

Hint: if you really do not like the Prophet library and the Microbiome data-set is really not working for you, here is an alternative: https://towardsdatascience.com/analyzing-time-series-data-in-pandas-be3887fdd621. However, if you use the alternative, you need to describe all the options you have tried with the Microbiome dataset and where it failed.

6.2 Analyzing snap-shot Microbiome data

We will analyze the data from "Reduced diversity and altered composition of the gut microbiome in individuals with myalgic encephalomyelitis/chronic fatigue syndrome" by Giloteaux et al. (Microbiome, 4 (30), 2016). In particular, you are supposed to use the provided Jupyter notebook and execute all steps (=cells) in there. Include the following in your report:

(1) Briefly describe all relevant steps in the report (Hint: Cells containing import or install commands are not relevant).
(2) In step "Get differentially abundant bacteria": report the name and Phylum of the top 3 differentially abundant bacteria. Describe how you got this information (e.g. which website / database you have used).
(3) Perform a classification with the goal to distinguish patient samples from control samples. Explain what you did and report the results including accuracy and confusion matrix. (Hint: this might help http://biocore.github.io/calour/notebooks/microbiome_machine_learning.html)

6.3 Deliverables

Your need to upload all source codes and a report to your GIT repository.

- The report should be about 600-1200 words in length (this is roughly 1-2 pages, depending on your layout).
- The report must be delivered in PDF format using the BMC template.
- The report must have an abstract as defined in project 5.
- The following sections must be present (you can add more if needed):
    - Analysing Microbiome time-series data
        - Background and goal of this analysis
        - Description of the data
        - Results
          (including the produced figures comparison results and answers to the questions)
    - Analysing snap-shot Microbiome data
        - Background and goal of this analysis
        - Description of the data
        - Results
          (including the produced figures and answers to the questions)
    - Discussion: why is this a typical project for a data-scientist? (Or why not?)