Python for biologists

Final project guidelines

The goal of the final project is to present a small research project implemented with computational methods and tools taught throughout the course.

You may choose any subject or question, preferably one which is connected to your research or for which you have available data.

The project should contain the following sections:

Abstract

General background and overview of the project. The abstracts you submitted must be edited to answer the following:

1. Overview: what is done in the lab? What is the background and context of your project?
2. What is the research question? What is its benefit to the comprehensive research?
3. What are the expected results?

Data files

Full and elaborate description of each of your data sources.

Methods

Describe the analysis performed, the algorithm used to implement the analysis and statistical tests applied to the data. List the python packages used for the implementation, and fully describe how you organized your data and processed it. Divide the process to sections and sketch or provide a workflow of your process (for instance, by numbering the titles of the sections).

Results

Summary of the results, statistical analyses etc. Attach here the plots and describe them.

Summary

Conclusions drawn from the results, answers to the biological question and comparison to the expected results.

- All sections may, and should, be accompanied by appropriate figures and tables, especially to clarify the results and their biological meaning.
- **Write your text as clear as possible to a reader that is not familiar with the field.**

Code

You should use python code to perform the processing and analysis of the data. You may use any python module you see fit or any external code, as long as the main body of the code is original.

External software and tools such as Microsoft Excel, text editors or analytic software may be used in processing the data, and if possible embedded within the python code. The core analysis should be performed with python code, including statistical analysis, plotting etc. If you are not sure, please ask in the forum.

Project submission

The final submission will include the following as a zip file:

Paper

Containing the sections mentioned above (abstract, methods, results, summary) in doc / docx / PDF file.

Code

All the code used for the processing and analysis of the data should be attached. The code should be readable and clear, including comments (the comments within the code must be in English) and function documentation. It is important that the code can be run and the results are reproducible. Attach a text/word file called "readme": a file containing explanations and usage instructions of the code, i.e., how to run your code.

Data

The original data on which analyses were performed. Make sure that by using the code on the provided data, the results will be reproduced.

Final submissions will be done through the Moodle website. **Due date is June 30th at 23:55**.

Grading

Proper use of python code, code readability and clarity, usage of the right tools and modules. Correct and comfortable usage of data structures: strings, lists, sets, dictionaries, numpy arrays, pandas, biopython etc.

The project will be evaluated based on the following criteria.

1. Accessing and parsing the data files (CSV, txt, online data): your data should be easily read and processed. You are encouraged to read or load data directly to data structure

(denoted in the next section) through the module's functions (for example, numpy.load()).

2. Generating and organizing the data in a convenient pythonic data structure (Pandas, Numpy, BioPython etc.). Make sure to use dictionaries, lists, sets and strings in an elegant way that fits your work.
3. Data processing: process the data using the modules taught in class. Make use of **two** procedures, for example:
    - Statistical analysis using Numpy and Scipy (regression, optimization, hypothesis testing, linear algebra, normalization etc.)
    - Simulations
    - Sequence analysis with Biopython and/or regular expressions (regex)
    - Query online databases (blast, entrez, etc.)
    - Comparison to other method or data: you can compare your method to an existing method that does something similar in terms of running time (complexity), goals achieved, additive features of your method and more, or to another published dataset, for example, if you analyzed phenotypic values of a mutant, you can compare it to the wild-type data published in another paper, or if you analyzed mutations in a sequence, download the wild-type and compare them, or use blast to check if it's a known phenomenon. **These are only examples, be creative!**
4. Visualization of the data: prepare 2 different plots of your data that summarize the process: histograms, scatter plots, etc. The plots should reflect the idea and contain suitable titles, axis labels, legends (if needed) etc.

Help

Questions regarding the project are welcome. Use the course forum or consult one of the instructors for questions about the instructions or the programming involved. You are also encouraged to consult your adviser or anyone else; however, **the actual code should be written by you alone**. Please ask your questions early, so we can attend all issues.