BI296: Linux Programming For Bioinformatics

# Course Projects: Suggestions

2018 Spring

---

In this project, you need to write a package/module to take advantage of at least one of the following packages or modules by organizing into groups with 3-4 persons per-group.
You are required to submit a report with at least 4 pages of A4 paper specifying the following contents:

- Title of your report;
- Task of your project;
- Data set and analysis methods applied to the data;
- Results obtained through the analysis;
- Discussion of the results and perspectives;
- Conclusion;
- Contribution of the group members

Your python scripts should be no less than 500 lines.

## 1 Python Packages

### 1.1 Visualization

1. `matplotlib` (`https://matplotlib.org`): Python 2D/3D plotting library producing publication-quality figures in a variety of hardcopy formats and interactive environments.
2. `Seaborn` (`https://seaborn.pydata.org/`): High-level interface for drawing attractive statistical graphics.
3. `Bokeh` (`https://bokeh.pydata.org`): Interactive visualization library that targets modern web browsers for presentation.
4. `plotly` (`https://plot.ly/`): Leading open source tools for composing, editing, and sharing interactive data visualization via the Web.

### 1.2 Mathematics

5. `scipy` (`https://www.scipy.org`): Python-based ecosystem of open-source software for mathematics, science, and engineering.
6. `numpy` (`http://www.numpy.org`): Fundamental package for scientific computing with Python.
7. `pandas` (`https://pandas.pydata.org`): Python library for high-performance, easy-to-use data structures and data analysis tools for the Python programming.

### 1.3 Machine learning and Statistics

8. `scikit-learn` (`http://scikit-learn.org`): Simple and efficient tools for data mining and data analysis.
9. `statsmodels` (`https://www.statsmodels.org`): Python module that provides classes and functions for the estimation of many different statistical models, as well as for conducting statistical inference.
10. `pymcmc` (`http://pymcmc.readthedocs.io/`): Bayesian inference with MCMC methods in Python.

### 1.4 Bioinformatics

11. `pysam` (`https://pysam.readthedocs.io`): An interface for reading and writing SAM files
12. `khmer` (`https://khmer.readthedocs.io`): In-memory nucleotide sequence k-mer counting, filtering, graph traversal.
13. `biopython` (`http://biopython.org`): Python libraries and applications for biological computation.
14. `scikit-bio` (`http://scikit-bio.org`): Data structures, algorithms, and educational resources for bioinformatics.

### 1.5 Scrawler/Spiders

15. `BeautifulSoup` (`http://www.pythonsite.com/?p=211`): Python scrawler.
16. `Scrapy` (`https://scrapy.org`): Scrawler framework.
17. `Requests` (`http://docs.python-requests.org/`): HTTP requests and response.
18. `re` (`https://docs.python.org/2/library/re.html`): Regular expression

### 1.6 Game Development

19. `PyGame` (`https://pygame.org`)

## 2 Datasets

- Kaggle: `https://www.kaggle.com/datasets`
- UCI Dataset: `https://archive.ics.uci.edu/ml/datasets.html`
- Tianchi Dataset: `https://tianchi.aliyun.com/`
- RosaLind: `http://rosalind.info/problems`

# 3 Tutorials

- `http://www.edvancer.in/python-data-science-made-simple-step-step-guide/`
- `https://inventwithpython.com/pygame/`

# 4 Write a report

A formal report usually follows this basic structure:

1. Introduction
2. Main Body
3. Summary/Conclusions
4. Appendix
5. References.

## 4.1 INTRODUCTION

Even if everyone who will be interested in the report is deeply familiar with the problem, data, and questions, it is always important to provide the following, so that the report stands entirely on its own:

- Subject matter background what is broad scientific context and what are the challenges and unresolved issues? Here, the report should give a short description of the subject matter problem and why it is important in the domain science area.
- A brief summary of the study carried out to address the challenges.
- A statement of each of the specific scientific questions to that will be addressed.
- A "high-level" summary of the conclusions of the data analysis in the context of the subject matter area.
- A brief roadmap for the rest of the report indicating what can be found in each subsequent section.

## 4.2 MAIN BODY

The main portion of the report can be organized in whatever way makes the most sense to you given the nature of the study and questions. Here is one standard way.

- Detailed summary of the data and the source.
- Detailed description of the model or method applied to the data, but do not include lots of equations, formula, matrices, and mathematical symbols.
- Describe the method used to train the model and mention which package/module was used.
- Present relevant numerical results and interpret them in the context of the subject matter.

- For each analysis, point out any limitations or caveats.
- Do not include code or raw output! It is fine to summarize results in a table, but the table should not just be the raw output from software. All columns and entries should be explained in a table caption, with additional explanation if needed in the text. Code and output belong in an appendix to the report; see below. Never instruct or expect readers to go look at these; everything that investigators need to read should be in the main report.

### 4.3 SUMMARY/CONCLUSIONS

As with any report on any topic, there should be a section providing an overarching summary of the objectives and results. This final section should present the problems again, the conclusions of the analysis, and the interpretation of them in terms of the subject matter. Discussion of the implications of the results for the science; any additional observations or findings that, while not directly related to the questions, seem interesting; and possible future studies suggested by the analyses and results can be given here.

### 4.4 APPENDIX

An appendix to the report contains technical details and supporting information. Typical things that would be presented include:

- A detailed description of all statistical models, with all symbols precisely defined, and precise statements of the assumptions that were made.
- A precise, technical explanation of the methods that were used, including how they were implemented (e.g., software used, with any special options or assumptions noted).
- Code implementing the methods; it is prudent to document all code with extensive descriptive comments.
- Additional data summaries, tables, figures, that might be of interest but are not directly relevant to the results.

It is fine to refer readers to the appendix in the main part of the report for more information and details, but looking at the appendix should not be required. All necessary information for investigators to understand what was done and the results should be in the main report. As above, never refer investigators to output or code.

### 4.5 MISCELLANEOUS

Some additional items:

- If any literature (papers, books, software documentation) is cited, there should be a References section with full information on each, in a consistent format.
- There should be no misspellings or grammatical errors; always spell check any report before finalizing it!
- The entire report should be organized in a logical fashion, with section headings that make it easy for a reader interested in a particular result or question to locate that portion.

- Keep the writing straightforward and to the point, but not to the point that it is so brief that important information is obscured or missing. Check for run-on sentences and avoid language that is too flowery and wordy. Do not use terminology or words that are unfamiliar to a likely reader unless it is necessary, and, if you do, define them. If you use acroynms, present the entire term the first time you use it and define the acronym; e.g., "convolutional neural network (CNN)."
- The narrative should flow naturally and "tell a story", so should be easy to follow.

Good report writing, both the writing itself and the knack for organizing the information in a sensible, logical way, is a skill that some people are born with but most people must learn. Use each report you write as an opportunity to develop and hone your skill. Good report writing skills will serve you well in not only collaborative work but in writing research papers.