Project Ideas

Given below is a list of possible projects for you to work on. Some projects are better defined than others. But they are all interesting, and the only limitations are the amount of effort you put in and your creativity. If you wish to pick a project outside this list, please contact me as soon as possible. It is optimal for you to have picked something by Monday, Feb 23. Your choice has to be approved by me, since I have to make sure that there is no conflict with another group. Depending on the project, you will work in groups of size 1 or 2. Lot of the work is research-oriented and also result-oriented. I want to see some good results by the end of the semester. So start early. You are **advised** to email me an update of your progress on the 1st and 15th of every month until the end of semester and seek some feedback from me. Maintain a log file (or journal) with your activities on this project including: updates on your reading, progress on implementations and partial testing, ideas for future work, ideas that you may not be able to follow up, bug fixes, known current bugs in your code, organization of your program files and data files, etc.

At the end of the project, you will need to write a report (in docx or LaTeX format). It must include a short summary of your project. State clearly the following: your name, e-mail addresses, date, title of project, goals, hypotheses or assumptions, background with references and URLs, methods used (with references), what was implemented or achieved, summary of results, conclusions, possible future work.

Finally, prepare: (1) a 25-minute PowerPoint presentation of your work, (2) a short handout to distribute to your classmates, (3) web page describing your project, and (4) a zip-compressed file containing your (commented) source code, data, results, report, and webpage to be mailed to me. Your project should be completed and submitted by **April 9** because your presentations will start from Monday, April 13. Contact me for detailed information on individual projects.

Genome Assembly

Many assembly programs exist for specific technologies. But very few exist for hybrid technologies. For two particularly virulent strains of *Pseudomonas aeruginosa,* we have next generation sequence data from 2 different technologies – Illumina and PacBio. Illumina produces short sequences with very low error rates, while PacBio produces long sequences with high error rates. Additionally, we have 6 closely related genome sequences that could be used as reference genomes.

1. Use Ragout & BOWTIE to obtain two different genome assemblies for the genomes mentioned above. (Collaborators: Drs. Kalai Mathee and Vanessa Aguiar-Pulido)
2. Implement in Java, the CloG algorithm published in a paper by Yang et al. (http://users.cis.fiu.edu/~giri/papers/CloG_Yang.pdf). This algorithm focuses on closing gaps using next generation sequence data from two different technologies.

Degenerate Primer Design

3. *Implement DePiCt*: Implement in Java or C++ the algorithm published in a paper by Jaric et al. (*BMC Proceedings* 2013, **7**(Suppl 7):S4).

Epigenetics

4. It is well known that even identical twins have different reactions to the environment, partly because of epigenetics. These are traits we inherit from our parents that are not part of the genetic code. One such set of features is the data on methylation sites. Experiment with tools and techniques to perform differential analysis with such data sets. (See Human Epigenome Project website)

Improving the Metagenomics Pipeline with Intelligent Classifier

5. We have a metagenomics pipeline set up which takes NGS data on samples and converts them to abundance matrices. One of the steps is a classifier that labels reads and assigns them to bacterial taxa. Unfortunately, it does not remember classes across independent runs. Modify the pipeline with a classifier that uses a smart database to label the classes. (Collaborators: Camilo Valdes and Dr. Vanessa Aguiar-Pulido)

Evaluating Community Structure in Networks

6. There are many kinds of networks that are of interest for Systems Biology. In order to analyze these biological networks, we often use tools from the field of Social Networks. Newman and Girvan [http://arxiv.org/pdf/cond-mat/0308217.pdf] devised techniques to find “central” nodes in a social network. How can we generalize thei techniques for directed and weighted networks? What kinds of networks can be evaluated using these methods?

Visualization

7. *Visualize a collection of interactions using Cytoscape tools*: The end product must be a flexible tool that allows drawing networks with the following properties. (a) You should be able to draw nodes of different shape, color, size; (b) Edges may be directed or not, may have varying thickness, color, length; (c) node placement must be determined by some standard placement algorithm (e.g., Fruchterman Reingold) (d) nodes must be labeled as needed; (e) there must be reasonable ways to group nodes; (f) show how to deal with arbitrary cases, (g) add any other extra properties you can think of.

Protein Structure Analysis

8. *Finding common substructures in proteins*: In a previous project, students have implemented a tool in Java for finding common substructures in protein structures. This work needs to be debugged, then tested extensively with known results, and then benchmarked and improved.

Computational Metabolomics

9. Microbial organisms produce metabolites, which then may interact with other microbes or with the host they infect. This is a new field that has not received much attention. Survey existing literature, databases, tools & techniques and build an appropriate database of interest. Some papers are at: http://users.cis.fiu.edu/~giri/teach/Bioinf/S15/Papers/Metabolomics/

RNA Secondary Structure Prediction

10. *RNA Structure Prediction*: Survey existing literature, tools and techniques, and suggest a problem.