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. You should have picked something by Thursday, Feb 24. 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 **required** to email me an update of your progress on the 1st and 15th of every month until end of semester. 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 to be mailed to me. Your project should be completed and submitted by **April 8**. Your presentations will start from April 12. Contact me for detailed information on the projects.

Genome-specific databases

1. Create *PseudoNEXUS*, a genome-specific DB for the bacterium *P. aeruginosa*. It will need workflow, visualization, BLAST, genome browser, data analysis tools. This requires coordination with 9 undergraduates doing their senior project. [Daniel Medvin]

Transcription Factor Binding Sites

2. *Detecting Transcription Factor Binding Sites*: The idea is to use a tool called **IEM** to improve your predictions by incorporating comparative genomics data. The first step is to try multiple sequence alignments of promoter regions of orthologous genes. The genomes of interest are multiple strains of the bacterium *Pseudomonas aeruginosa*. (Collaborators: Prof. Kalai Mathee, Prof. Lisa Schneper)

Protein Structure Analysis

3. *Finding common substructures in proteins*: You will implement a tool for detecting all common substructures between 2 or more protein structures. The suggestion is to use geometric hashing to achieve the goal.

Degenerate Primer Design

4. *Improved DePiCt*: The current version of DePiCt is unable to handle inputs of size more than 50 because it uses hierarchical clustering with time complexity O(n^3). This

project is to find a more efficient implementation of clustering that is well-suited for degenerate primer design. [Melita Jaric]

RNA Secondary Structure Prediction

5. *RNA Structure Prediction*: Your task is to understand the best algorithm (and code) for RNA structure prediction, and to implement a version that predicts the structure of two interacting RNA strands.

Genome Assembly

6. Assemble two new genomes for which next-generation sequence data is available. The genomes are for a particularly virulent strain of *Pseudomonas aeruginosa.* The assembled genomes need to be analyzed for comparative genomic information. (Collaborators: Drs. Kalai Mathee and Lisa Schneper) [Tram Ta]

Comparative Genomics

7. *Study recently sequenced genome*: The BROAD Institute sequenced the bacterial genome *Burkholderia dolosa* (AU0158). We have an improved draft assembly for this genome. A number of very closely related genomes have been previously sequenced and are also available. Your job is to compare the new genome with previously sequenced genomes in a number of different ways. A local BLAST server may be needed for this project. More specific goals are:
    - Identify new genes and its orthologs within the updated regions.
    - Study differences in the order of genes and functional annotations.

    (Collaborators: Profs. Kalai Mathee & Stephen Lory, & Broad Institute)

Next Generation Sequencing (NGS) Tools

8. Implement a next generation sequencing tool that takes into account comparative information from multiple closely-related genomes. The genome of interest is *Pseudomonas aeruginosa.* (Collaborators: Drs. Kalai Mathee and Lisa Schneper)

Comparative Genome Visualization Tools

9. Take an existing tool such as CGView or GenomeAtlas and implement as a genome visualization webservice.

Finding Novel Genes in Plants using RNA-Seq

10. Plants of the genus Piper (black pepper) are diverse across the tropics, with several rare species limited to Jamaica. Low coverage RNA-seq was performed on 7 individuals of 3 species with Illumina sequencing to find EST-microsatellite or SNP loci for further population characterization of remaining populations of the rarest species. Secondly, the dataset provides an opportunity to uncover novel leaf-expressed genes in these plants of widespread ethnobotanical importance, such as to the maroon culture of rural Jamaica. (Collaborator: Dr. Eric von Wettberg)