Biological Data Analysis (CSE 182) Final project

1 Logistics

The final lectures of the class will be devoted to final presentation of the project. This project is on a team annotation of the genome of *Acinobacter baumanni.* We have two projects. Project 1 is on the core annotation of the genome, and Project 2 is on serving up the annotations via web browser.

Project 1 deliverables

1. Select a collection of 100 sequences to annotate. However, your code should run on any collection of sequences. It is not sufficient to copy and paste your sequences in an online annotation tool.

2. Set up protein annotation tools, either by querying online databases, or by downloading the databases and the tools for querying them.

3. At the minimum, run Blast against an annotated protein database (UniProt full is a good choice), Pfam, and Prosite. However, you can choose to run other tools as well including tools for GO annotation. Blocks, Prints, tools for assigning sub-cellular localization of the protein, tools for predicting 3D structure. A good strategy would be to choose the required tools and one other annotation.

4. For each entry in the Biological database (Proteins, Pfam, Prosite), write a script to mine the entries and collect functional information in the form of keywords.

5. Produce a table of 'hits'. Each query sequence is a row. For each query sequence, and each biological database, enter a comma separated (or semi-colon separated) list of keywords extracted from searching your query against the database. For example, if your Blast search hits E. coli protein ANK03648.1 with the annotation Dihydropteroate synthase (plasmid), you should use that. *For extra credit*, store (in an indexed file) the raw results that led to a keyword being extracted, including for example, Blast alignments, and so on.

6. At the very end, add your best explanation of the function of the query sequence by writing in the comments field. You can earn *extra-credit*, by assigning function to a sequence that does not have a Blast hit. Note that your automated script must produce the entire table with the exception of the comment field.

7. In trying to compute statistics on the functions, it is good to use standard terms. *For extra credit*, identify a GO-slim term (see link on web-site) for each prtoein sequence, and add it to the comments field. For example, one such term could be 'GO:0003676-nucleic-acid-binding'.

8. Negotiate with a project 2 team on the data exchange format, and periodically, give them examples of the files that you are generating. If you save the raw files of your searches, break them into multiple files, and provide that to the Project 2 team to hyperlink to your keywords.

Project 2

1. Decide on a web-framework. As the project does not require a fancy interface with complex queries, it is OK to use any framework, or work with static html/javascript pages.
2. Discuss with some project 1 teams and fix a data exchange format, and build your own test cases in that format. You should engage with as many teams as possible as points will be awarded for having as extensive a data-set as possible. Each project 1 team must engage with at least one project 2 team.
3. Write the code to display results.
4. Upload as much data as possible, and develop the following tools for computing statistics.
   (a) **search.** You should be able to search each column and filter rows using keywords in the column.
   (b) **Standardizing the vocabulary of keywords.** You must simplify the keywords and putting them in a standard vocabulary. Preferably, use standard GO-terms from the available database to define and merge the keywords. For example, if you see a keyword called 'transcription factor', you can replace it with 'GO:0003700-transcription factor activity, sequence-specific-DNA-binding'. This will allow you to gather similar terms together. Add your own column called 'GO', that contains a standardized term that best represents the function of the sequence.
   (c) **distribution of functions**. Draw a pie-chart of the 20 most frequent keywords in the GO field. Each slide of the pie should be annotated with a keyword, and its frequency in your data. Use GO terms for your keywords.
   (d) **Statistics**. For each column (Biological database), compute the fraction of sequences annotated, and display it at the bottom of the table.