FINAL PROJECT

Please select from the appropriate section and read the general note at the end of the page. **Final projects are due by 5 PM EDT on Friday, May 7, 2010.**

Please email final projects to **cbb752@gersteinlab.org**

MBB452/752 and MCDB452/752:
Pick one of the topics below and write a research paper. Research papers should be approximately ten pages in length (double spaced). Carefully cite your references at the end of the paper (not included in the 10 pages). Please note that the papers are research proposals and therefore should contain both a) a review of the literature and b) a proposal for something new or an analysis of existing programs. (i.e. Propose a simple project for a new algorithm or improving an existing algorithm, or select an algorithm that you feel best fits a particular research task and explain why you selected that algorithm).

CHOOSE ONE:

1. As discussed in the data mining section of the course, spectral methods are often used to reduce the dimensionality of large data sets and to find patterns within the data. Write a paper detailing the different algorithms/techniques currently applied, examples of how they are applied to biological data (i.e. microarrays, etc)., and compare/contrast the different methods (i.e. what are the strengths and weaknesses of each method and when/on what types of data would one use each).

2. As covered in class, structural alignments are less straightforward than sequence alignments. Write a research proposal dealing with aligning the structures of two macromolecules. Assuming that both structures are known, what issues arise in creating the alignment? Is it possible to produce a verifiably "best" alignment? What technique would you use to align two macromolecules? Why did you choose this method over the others?

3. Multiple sequence alignments cannot be efficiently handled using purely dynamic programming. Write a research proposal dealing with these alignments. How do existing methods approach this problem? Can other information be determined from these alignments (i.e. phylogenetic trees or motif finding)? What technique would you use to align a family of sequences? Why did you choose this method over the others?

CPSC752 and CBB752:
Using the language of your choice, please select one of the following options. Please submit source code and a brief (one to three pages) write-up explaining: the task your program is used for, the algorithm you implemented, and instructions for compiling and using the program (include the language version you used). A test-run showing the output put from your program along with your training/test data must also be included as well as a written description of how your algorithm works. Programs that do not run will not receive credit.

A significant portion of each program should be implemented from scratch and not simply rely upon calling existing libraries or packages!

CHOOSE ONE:

1) Implement GOR IV.

2) Implement a simple structural alignment algorithm.

3) Analyze the following microarray data set using principal component analysis (PCA): http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS1096 In detail, discuss the process and rationale behind each step (for example, normalization of the data, the PCA, etc). Discuss in detail what the output of the PCA means and the information that can be derived (i.e. interpret the loadings, what do the loadings mean, relationship of the number of principal components to variability, etc). Interpret your results and draw biological conclusions based on YOUR analysis. The goal of this project choice is to allow students to drill down into the data, design and implement an analysis from scratch, and draw their own conclusions based upon these analyses.

NOTE TO ALL:
If you have another topic that you would like to write about or implement, please talk with Mark Gerstein before **Tuesday, April 20, 2010**.