CS 5263 & 4593 Final Project
Draft version – March 4, 2015

Due Date – Submission opens early April, closes on May 2nd.

The final project is optional for 4593 – if you would like to get a boost on your grade - and required for 5263 (choose one from activity type 2-4).

Choose one or more activities below. The points that you will receive from this project will be determined by your instructor based on his opinion of the significance of the results and effort that you have put into the project, as compared to other students’ projects in the class. You are strongly encouraged to talk to the instructor before you choose an activity.

Your final grade in the course will be determined by the sum of points you received from all activities (homeworks, in-class quizzes, tests, etc) in the course.

For activity type 2 and 3, I expect some level of experimentation and a brief project report (2-4 pages) to describe your experimental procedure and results, demonstrating that you did achieve what you were attempting to achieve or explaining why you have failed to achieve your goals.

1. (Max 100 points) Redo one or more of the homework implementation problems and get partial credit for the homework(s).
2. (150 – 400 points) Substantially improve one of the programs that you wrote for the homeworks to make it (likely) useful in practice. Examples include but are not limited to:
   a. Implement Hirschberg’s idea for linear space global alignment.
   b. Given a pair of sequences, return up to k non-redundant highest-scoring local alignments. Note that one strand (but not both stands) of an already reported alignment may be reused in another alignment.
   c. Implement a global multiple sequence alignment algorithm with some optimization tricks. (For example, start with a progressive alignment, then iteratively take out sequences and re-align.)
   d. Improve the combinatorial motif finding algorithm to achieve a better balance between efficiency and accuracy by, say, using extended sample-based methods, or using results returned from your probabilistic motif finding algorithm.
3. (150 – 400 points) Empirical study of data mining algorithms in bioinformatics. Examples include but are not limited to:
   a. Take a large microarray dataset, cluster the samples or genes, and analyze the resulting samples / genes for biological significance using enrichment analysis, or apply multiple clustering algorithms and get the consensus clusters (e.g., which genes are consistently clustered together).
   b. Take a genome-wide ChIP-seq dataset, align the sequences back to genome to get chromosome coordinates of the reads, run peak finding, retrieve sequences around the
peak, and perform motif finding (using your program or existing software). Analyze the enrichment of the found motifs in the peak region compared to other regions.

c. Take a cancer microarray dataset, use classification algorithms to predict disease outcomes. Repeat multiple algorithms and compare their performances. Try some feature selection algorithms to see if results can be improved.

4. (Max 300 points) Literature study and report – read two-three papers in a selected topic of bioinformatics and write a critique review of the papers (8-12 pages, single column, double space).