CMSC423: Midterm 1 Recap

Spring 2017

The final exam will consist of: ~10 quick questions (multiple choice, true/false), ~3 short questions, 2 long questions. It is cumulative, but will be skewed towards material covered in the last third of the class (num. 13 below). Questions from previous midterms may reappear.

It will cover the following material:

1. Molecular Biology concepts. Most questions will be about term/concept identification, possibly a short question to test your understanding of basic molecular biology processess: e.g., replication, transcription, translation.
2. Programming for Bioinformatics. What are the types of entities commonly encapsulated by data structures in bioinformatics libraries? What is the importance of reproducibility in the analysis of genomics data?
3. Bioinformatics Resources. Quick questions of to check your ability to identify resources containing specific types of data. E.g., genomic sequences may be found in Refseq, sequencing experiments in the Short Read Archive
4. Motif finding. What are transcription factors? What are motifs? What is the motif finding problem (biologically and computationally)? What is a "profile"? What are entropy and relative entropy, and why should we use it to score motifs? Why should we score motifs? How is motif finding an optimization problem? In Gibbs sampling, what is the benefit of randomly selecting a starting point for a given sequence instead of choosing the starting point that maximizes probability?
5. Clustering. What is the general clustering problem (assuming gene expression, or other continuous measurements/features). What is the k-means algorithm? What is the objective function minimized by the k-means algorithm? What is the fuzzy k-means formulation and what is its' relationship with the randomized algorithms used in motif finding.
6. Peptide sequencing. What are peptides? What is a peptide's mass spectrum? What is the cyclospectrum peptide sequencing problem? Given an observed experimental spectrum, and a peptide, what is the difference between scoring as a linear vs. cyclic peptide?
7. General algorithmics. What is a branch and bound algorithm? What is a heuristic algorithm? What is a randomized algorithm? For each of these, what are appropriate ways of analyzing the correctness and performance (running time) of these algorithms.

8. Biological sequence comparison: Why do we need algorithms that find inexact alignments between DNA or amino-acid sequences? **Why are exact matching algorithms not sufficient for biological questions?**

9. Sequence Assembly. The Hamiltonian and Eulerian approaches to sequence assembly. High-level understanding of Lander-Waterman statistic for sequencing coverage requirements of genome assembly.

10. Inexact Alignment. Dynamic programming algorithms: Global alignment (Needleman-Wunsch), Local alignment (Smith-Waterman). Linear gap penalties, affine gap penalties. The probabilistic interpretation of scoring matrices (as log odds of two probabilistic models). Formulating inexact alignment algorithms as finite state machines. How to do global alignment with linear space complexity?

11. Clustering. What is the general clustering problem (assuming gene expression, or other continuous measurements/features). What is the k-means algorithm? What is the objective function minimized by the k-means algorithm?

12. EM algorithm. How is fuzzy k-means an instance of the EM algorithm. How is the EM algorithm related to randomized algorithms used in motif finding.

13. Exact Matching methods. I will ask you questions about properties of algorithms (time and/or space complexity), details about their implementation, and their application to specific string problems that can be solved using exact matching. The following are fair game: fundamental preprocessing (z-algorithm), KMP algorithm, suffix tries, suffix trees, suffix arrays, the Burrows-Wheeler transform.