CMSC423: Midterm 2 Recap

Fall 2014

The midterm exam will consist of: ~6-10 quick questions (multiple choice, true/false), ~5 short questions, 1 or 2 longer questions.

It will cover the following material (this is not an exhaustive list, there may be material, especially from the motif finding chapter, that is not listed here, that may be included in the test):

1. Biological sequence comparison: Why do we need algorithms that find inexact alignments between DNA or aminoacid sequences? Why are exact matching algorithms not sufficient for biological questions?

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

3. Inexact Alignment. Dynamic progamming 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?

4. 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.