

SGN-6156 Computational Systems Biology II

Exam, May 14th 2008

You can use a (graphical/programmable) calculator in the exam.

1. a) Calculate the dynamic programming matrix and an optimal global alignment for DNA sequences GAATTC and GATAC using the Needleman-Wunsch algorithm. Score a match (i.e., AA, CC, GG or TT) by +2 and a mismatch (e.g., AC, AG, AT, CA, etc.) by -1. Use the linear gap penalty $\gamma(g) = -dg$ with $d = 2$. Hint: Recall that, in the Needleman-Wunsch algorithm, the recursion matrix $F(i, j)$ is initialized with $F(0, 0) = 0$, $F(i, 0) = -id$ (for all i) and $F(0, j) = -jd$ (for all j). (6p)
2. a) Explain how hidden Markov models (HMM) can be used for pairwise sequence alignment. You can do this, for example, by showing how a finite state automaton (FSA) can be converted and extended to a pair HMM. You can focus on global sequence alignment. (You do NOT need to explain any of the algorithms: Viterbi, forward algorithm, etc.) (3p)
b) The most probable (hidden) path of a HMM can be found efficiently using the Viterbi algorithm. Explain the Viterbi algorithm. Recursion equations can be shown, but they are not necessary. You can explain the algorithm in terms of a general HMM or using an application (e.g. pairwise sequence alignment, dishonest casino example from the course book, etc.) (3p)
3. a) Scoring methods for multiple sequence alignment are typically defined so that the score is defined separately for each column in an alignment. Explain the minimum entropy score and the sum of pairs score for multiple alignment. (4p)
b) What is the (algorithmic) computational complexity and memory requirements for the Needleman-Wunsch algorithm (for sequences of length n and m)? In addition to the formula, explain your answer. (2p)
4. A commonly used statistical model for DNA sequence motifs (such as transcription factor binding sites) is the position specific frequency matrix (PSFM). Explain how PSFM can be used to predict transcription factor binding sites. Also explain how the statistical significance (hypothesis testing) of predicted binding sites can be computed. (6p)
5. Briefly explain the following: (2p each)
 - a) The Gillespie simulation algorithm
 - b) The basics of reconstructing a metabolic network, including the stoichiometric matrix and fluxes.
 - c) General cross-validation method for model selection