

CS 124 Final exam Winter 2006. I think the exam is hard, because it has a high concentration of conceptual questions rather than computation questions. But I think you are up to it.

1. (10 pts) Briefly explain how to use XPARAL in order to find the longest common substring between two strings.

2. (15pts) Suppose you have just sequenced a new gene and have translated it to an amino acid sequence. You want to know what the protein does. Briefly name and describe (at most one line per) the programs, services and databases we have learned about in this class that you could use to try to figure out what this protein does.

3. (20) Suppose you have a database D of protein sequences, and when you have a new query protein sequence P, a database program tells you which protein sequences in D it thinks P would strongly bind to. For each protein sequence in D that it reports, it gives two numbers, the expected binding energy (i.e., a measure of how well the two proteins are predicted to bind), and an E-value. Now you want to use this database but you haven't read the manual or tutorial for it. What would you expect the E-value to mean, and how do you think they determine it?

This question has nothing to do with binding energies - they are just used to make the problem concrete. Remember we have seen E-values in Blast and in Pfam.

4. (10) Both Clustal and star.pl build a multiple alignment that is consistent with a tree, but different trees. Explain how the trees differ. Then explain how the pairwise distances used by Clustal and star.pl differ.

5. Consider the Markov Model shown below. It is designed to distinguish "poly-T" islands in a sequence from "poly-G" islands (I just made those up). A poly-T island is a substring between two C's that has a high frequency of T's compared to G's, and a poly-G island is a substring between two C's with a high frequency of G's compared to T's. Also, poly-T islands tend to be longer than poly-G islands.

In this Markov Model when a Markov chain enters state 1, symbol C is emitted; when a chain enters state 2, T is emitted with probability $4/5$ and G is emitted with probability $1/5$; when a chain enters state 3, G is emitted with probability $4/5$ and T is emitted with probability $1/5$. The state transition probabilities are written on the edges of the Markov Model.

5a) (5) Is this model a simple Markov Model, or a Hidden Markov Model? Explain.

5b) (5) In a given Markov chain, which state identifies putative poly-T islands, and which state identifies putative poly-G islands?

5c) (10) Consider the sequence C G C T T T G T T C. What is the Markov chain that the Viterbi algorithm would find for this sequence? Justify your answer. You should be able to answer this question without actually working through the Viterbi algorithm (which would be doable but tedious).

5d) (5) Consider the sequence C T G C. What probability would the Forward algorithm give for this sequence? Don't actually calculate it, but just write out the expression you would need for that calculation.

6. (10) Is the following matrix ultrametric?

	A	B	C
A	0	3	5
B	3	0	5
C	5	5	0

If Yes, display the ultrametric tree for it. If No, explain why not.

7. In order to build an phylogenetic tree for a set of sequences from different species, we would ideally like to know the time-since-divergence for each pair of the sequences. Of course, we cannot know this, but instead try to deduce the time-size-divergence for each pair by aligning the pair and using the resulting alignment score. Suppose we have equal-length sequences that have only changed over time by mutations, with no insertions and deletions, so that the we can just line up two sequences and count the number of places where the sequences differ. For a pair S_i, S_j , let $D(i, j)$ denote the number of places where they differ, and let $T(i, j)$ denote the time-since-divergence of sequences of species i and j .

7a) (10) Generally we think that $T(i, j)$ is an *increasing* function of $D(i, j)$. What is a justification for this? Can you propose a plausible biological situation that would cause $D(i, j)$ to not be an increasing function of $T(i, j)$, where maybe it fluctuates over time?

7b) (5) In addition to thinking that $T(i, j)$ is an increasing function of $D(i, j)$, is it generally thought to be concave (see the figure). Why?

7c) (5) Let $S_{i,j}$ be the (unknown) ancestral sequence of S_i and S_j ; that is, $S_{i,j}$ is the sequence of the most recent common ancestor of S_i and S_j , and those two sequences derive from $S_{i,j}$. What theory or thesis justifies the expectation that (roughly) $D(S_{i,j}, S_i) = D(S_{i,j}, S_j) = D(S_i, S_j)/2$?