6.047 / 6.878 Computational Biology: Genomes, Networks, Evolution
Fall 2008

# 6.047/6.878 Fall 2007 Midterm Exam

October 25, 2007

Name:

No books, notes or electronic aids (such as calculators) are permitted. Both 6.047 and 6.878 have the same exam and scoring rubric, but they will be considered separately in determining final grades.

## True/False and Multiple Choice (20 points)

Read each statement or question carefully, and circle the correct answer.

1. **True / False** Given the exponential number of possible alignments, the Needleman-Wunsch algorithm requires exponential space to recover the best alignment, even though it requires only polynomial time to determine its score.

2. **True / False** While dynamic programming sequence alignment algorithms are guaranteed to find the optimal alignment in quadratic time $O(N^2)$, **bounded** dynamic programming is guaranteed to find the optimal alignment in linear time $O(kN)$, where $k$ is the width searched around the diagonal of the matrix.

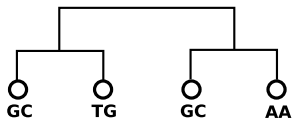
3. **True / False** Two sequences of length 10 that have 70% identity (no gaps) are guaranteed to have a stretch of four identical nucleotides in a row.

4. **True / False** Posterior decoding finds the most probable hidden state path, $\arg\max_\pi P(X, \pi)$.

5. Which hierarchical linkage method is used by the UPGMA algorithm?

   (a) Single link
   (b) Average link
   (c) Complete link
   (d) None of the above

6. **True / False** The minimum number of substitutions for this tree is 4.




7. **True / False** If we use a given phylogenetic tree to compute a distance matrix for all leaves based on the branch lengths of the tree, the resulting distance matrix will always be additive.

8. **True / False** In a finite population, an allele with frequency $0 < p < 1$ that is selectively neutral will never reach fixation.

9. **True / False** If we compare the sequences of two orthologous genes from different species and find that $dN/dS > 1$, we have evidence that the gene has evolved under positive selection.

10. **True / False** In the absence of selection, older alleles tend to be associated with longer haplotypes than newer alleles.

## Short Answer (48 points)

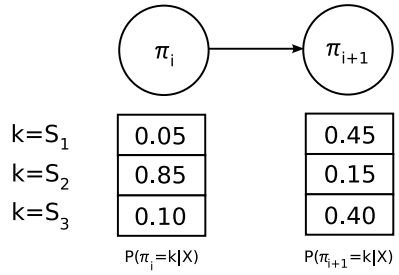
11. In the four-nucleotide DNA code, the 20 amino acids are encoded in codons of length three. Suppose Martians have 40 different amino acids and a five-nucleotide code (A, C, G, T, Z). What is the minimum codon length required to encode Martian proteins? Justify your answer.

12. Given a BLOSUM amino acid alignment score matrix $s(X, Y)$ derived from a database of protein alignments containing one million letters, how can you determine the overall frequency of amino acid $A$ in the database?

13. Write a comb of length 7 that would be well-suited for nucleotide BLAST when a large fraction of the genome is protein-coding. You can use 1 for "match" and 0 for "don't care" positions. Justify your answer.

14. Your have two HMMs, one trained for gene finding in the human genome and one for gene finding in the fruit-fly genome. You are given an unknown, unannotated genome sequence, and you'd like to decide whether it is more likely to have come from the human or the fruit-fly. Explain how you could use your two HMMs to make this decision. Which algorithm(s) would you use, and how would you use the output?

15. We have performed posterior decoding on a hidden Markov model with three states, $S_1$, $S_2$, and $S_3$. The table below shows the posterior probability distribution for two adjacent positions in the sequence, as determined by posterior decoding.




Which states will posterior decoding assign to $\pi_i$ and $\pi_{i+1}$?

Fill in the blanks in a transition matrix for this HMM such that any subsequent Viterbi decoding of the same emission sequence will give a different path than **this** posterior decoding. (There are many correct answers.)

| from \ to | $S_1$ | $S_2$ | $S_3$ |
|---|---|---|---|
| $S_1$ | 0.90 | | |
| $S_2$ | | 0.75 | |
| $S_3$ | | | 0.80 |

16. We have an HMM-based gene finder with known transition probabilities $a(k, l)$ and emission probabilities $e_l(x)$. We would now like to "upgrade" this HMM to a linear chain conditional random field. Define a feature function $F(k, l, i, X)$ such that, if the CRF contains only this feature, it is equivalent to our HMM. ($k$ is the previous label/hidden state assignment, $l$ is the current label, $i$ is the current position, and $X = x_1 x_2 \ldots x_N$ is the genome.)

17. What objective function does the $k$-means algorithm try to optimize?

18. Does the following feature set for gene finding in the human genome satisfy the naive Bayes assumption? Explain your answer.

    {GC content, protein BLAST hit to mouse, nucleotide BLAST hit to chicken}

19. State one advantage and one disadvantage of the naive Bayes classifier compared to the support vector machine.

20. Is neighbor joining more or less robust to fast-evolving lineages than UPGMA? Why?

21. Under what conditions is neighbor joining guaranteed to find the correct tree?

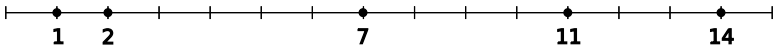
22. We have seen several algorithms based on the principle of expectation-maximization, a general probabilistic framework for updating parameters when there is some unknown hidden data. What are the parameters and hidden data in each application?

| | parameters | hidden data |
|---|---|---|
| HMMs (Baum-Welch) | | hidden state assignments ($\pi_i$) |
| motif finding (MEME) | motif position weight matrix | |
| clustering (fuzzy $k$-means) | | |

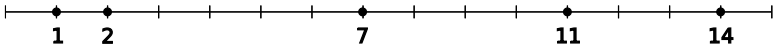
## Practical Problems (16 points)

23. Draw the hierarchical trees produced from clustering the following points by *single* linkage and *complete* linkage on their Euclidean distance.

(single linkage)




(complete linkage)




24. Consider one iteration of the EM algorithm for a motif of length $3$. Below we have provided you with a set of sequences and a $Z$ matrix ($Z_{ij}$ gives the probability that position $j$ in sequence $i$ is the start of the motif). Compute the next $M$ matrix (a position weight matrix representing the motif at the next iteration). Assume that all pseudocounts are $0$ and that the background nucleotide distribution is uniform.

| Position | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Sequence 1 | T | A | G | C | A | A |
| Sequence 2 | T | G | A | G | A | C |
| Sequence 3 | G | C | T | A | C | A |

| $Z$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Sequence 1 | 1.0 | 0.0 | 0.0 | 0.0 |
| Sequence 2 | 0.0 | 0.5 | 0.0 | 0.5 |
| Sequence 3 | 0.0 | 0.0 | 1.0 | 0.0 |

| $M$ | 1 | 2 | 3 |
|---|---|---|---|
| A | | | |
| G | | | |
| C | | | |
| T | | | |

## Design Problem (16 points)

25. The single best-scoring alignment produced by the Needleman-Wunsch algorithm might not inform us of interesting, different, and only slightly sub-optimal alignments. Design a pairwise global sequence alignment algorithm similar to Needleman-Wunsch that produces the $K$ best-scoring alignments between two sequences $X = x_1 x_2 \dots x_M$ and $Y = y_1 y_2 \dots y_N$ given a score matrix $s(x, y)$ and linear gap penalty $d$. Describe how your algorithm stores the information necessary to recover the $K$ best paths, specify the dynamic programming update rule for the score matrix, describe the basic idea of the traceback procedure, and state the time and space requirements.