

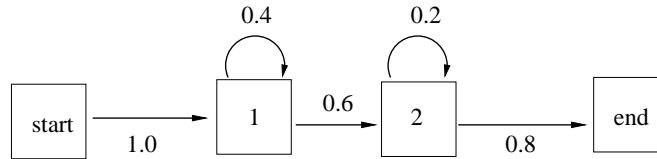
Biostatistics & Medical Informatics / Computer Sciences 576
Introduction to Bioinformatics
Fall 2002 Final Exam

Name: _____

Write your answers on these pages and show your work. You may use the back sides of pages as necessary. Before starting, write your name on every page of the exam. Also, make sure your exam has every page (numbered 1 through 11).

Problem	Score	Max Score
1.	_____	20
2.	_____	16
3.	_____	12
4.	_____	12
5.	_____	12
6.	_____	12
7.	_____	16
Total	_____	100

1. Hidden Markov Models: Consider the hidden Markov Model shown below. The transition probabilities are shown in the figure, and the emission probabilities are shown in the table. The *start* and *end* states are silent.

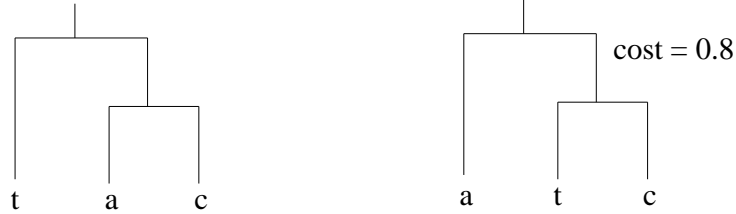


state 1	state 2
$e_1(A) = 0.4$	$e_2(A) = 0.1$
$e_1(C) = 0.1$	$e_2(C) = 0.4$
$e_1(G) = 0.1$	$e_2(G) = 0.2$
$e_1(T) = 0.4$	$e_2(T) = 0.3$

1a. (10 points) Show the values calculated by the Forward and Backward algorithms for the sequence **ACT**.

1b. (10 points) Using the Forward and Backward values you calculated in **1a.**, show how the Forward-Backward (Baum-Welch) method would update the transition probabilities going out of state 1. Assume that the training set consists only of the sequence **ACT**, and show the updates only for the first iteration of the method.

2. Maximum Parsimony Phylogeny: Consider the two simple phylogenetic trees shown below, and the *symmetric* cost matrix for assessing nucleotide changes. The tree on the right has a cost of 0.8.



	a	c	g	t
a	0	0.8	0.2	0.9
c	0.8	0	0.7	0.5
g	0.2	0.7	0	0.1
t	0.9	0.5	0.1	0

2a. (12 points): Show how the weighted version of Fitch's algorithm would determine the cost of the tree on the left.

December 17, 2002

Name _____

2b. (2 points): What are the minimal cost characters for the internal nodes in the tree on the left?

2c. (2 points): Which of the two trees would the maximum parsimony approach prefer? Briefly explain your choice.

3. Hierarchical Clustering (12 points): Given five genes $\{x_1, x_2, x_3, x_4, x_5\}$, and the following **similarity** matrix, show how complete-link hierarchical clustering would cluster the genes. In the matrix, larger numbers represent higher levels of similarity. Show each step of the algorithm as well as the final clustering returned.

	x_1	x_2	x_3	x_4	x_5
x_1		4	8	5	1
x_2			5	3	1
x_3				7	3
x_4					6
x_5					

4. *K*-Means Clustering: Suppose we want to run *k-means clustering* on the following genes represented as vectors:

$$x_1 = \langle 2, 0 \rangle$$

$$x_2 = \langle 2, 2 \rangle$$

$$x_3 = \langle 4, 0 \rangle$$

$$x_4 = \langle 6, 6 \rangle$$

Assume that $k = 2$, the initial coordinates of the cluster centers are $\langle 1, 1 \rangle$ and $\langle 5, 2 \rangle$, and we are using the similarity function:

$$s(x_i, x_j) = -|x_i^1 - x_j^1| - |x_i^2 - x_j^2|.$$

Here x_i^1 represents the first component and x_i^2 represents the second component of the vector x_i .

4a. (10 points): Show how *k-means clustering* moves the cluster centers on its first iteration.

December 17, 2002

Name _____

4b. (2 points) In this case, has k -means converged after one iteration? Briefly, explain your answer.

5. Protein Threading (12 points): Consider a simple threading problem in which we have a template with three segments (i, j, k) . We are given a sequence for which there are two possible starting positions for each segment. Given the following values for the scores of the individual segments and the scores for segment interactions, show how the branch and bound method would find the optimal threading.

$$\begin{aligned} g_1(i, 2) &= 4 \\ g_1(i, 3) &= 3 \end{aligned}$$

$$\begin{aligned} g_1(j, 8) &= 2 \\ g_1(j, 9) &= 5 \end{aligned}$$

$$\begin{aligned} g_1(k, 13) &= 1 \\ g_1(k, 14) &= 10 \end{aligned}$$

$$\begin{aligned} g_2(i, j, 2, 8) &= 6 \\ g_2(i, j, 2, 9) &= 0 \\ g_2(i, j, 3, 8) &= 1 \\ g_2(i, j, 3, 9) &= 0 \end{aligned}$$

$$\begin{aligned} g_2(i, k, 2, 13) &= 1 \\ g_2(i, k, 2, 14) &= 0 \\ g_2(i, k, 3, 13) &= 9 \\ g_2(i, k, 3, 14) &= 0 \end{aligned}$$

$$\begin{aligned} g_2(j, k, 8, 13) &= 3 \\ g_2(j, k, 8, 14) &= 12 \\ g_2(j, k, 9, 13) &= 5 \\ g_2(j, k, 9, 14) &= 11 \end{aligned}$$

Use the “simple lower bound” presented in class. When splitting a threading, split the segment having the minimal g_1 value for some position (e.g. split on k first since $g_1(k, 13) = 1$). To split a selected segment, divide it into two intervals of length one. Show the following:

- a) the threading sets considered in the search process,
- b) the lower bound calculated for each threading set,
- c) the threading that is returned by the method.

December 17, 2002

Name _____

6. EM Methods (12 points): We talked about two Expectation Maximization (EM) algorithms: (1) the Forward-Backward (Baum-Welch) method for learning parameters in HMMs, and (2) the clustering method with Gaussian clusters. For both of these algorithms, briefly describe the following.

- a) What is the hidden part of the problem (i.e. the hidden state) for each method?
- b) What are the key values computed in the expectation step (E-step)?
- c) What is adjusted in the maximization step (M-step)?

7. Branch and Bound Search (16 points): We talked about using branch-and-bound search in two contexts: (1) maximum parsimony inference of phylogenetic trees, and (2) protein threading. For both of these tasks, briefly describe the following.

- a) What do states (i.e. the things being considered in the search) represent?
- b) How are states expanded into other states (i.e. what do the operators do)?
- c) How are lower bounds calculated (i.e. what do they take into account)?
- d) What are the stopping criteria (i.e. how do we know when the search is done)?