

Biostatistics & Medical Informatics / Computer Sciences 576
Introduction to Bioinformatics
Fall 2003 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 9).

Problem	Score	Max Score
1.	_____	25
2.	_____	10
3.	_____	15
4.	_____	15
5.	_____	10
6.	_____	9
7.	_____	16
Total	_____	100

1. Distance-Based Phylogeny

1a. (10 points): Given the following distance data for four species, show how UPGMA would produce a phylogenetic tree for these species. Show the partial tree at each step of the algorithm and indicate the distances represented by edges in the final tree.

	A	B	C	D
A	0	3	5	6
B		0	6	5
C			0	9
D				0

1b. (10 points): Show how neighbor-joining would produce a tree given this same distance matrix. Show the partial tree at each step of the algorithm and indicate the distances represented by edges in the final tree.

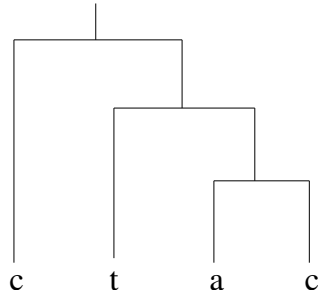
	A	B	C	D
A	0	3	5	6
B		0	6	5
C			0	9
D				0

December 18, 2003

Name _____

1c. (5 points): How do the trees produced by UPGMA and neighbor-joining differ in this case. Which one do think is the better tree, and why?

2. Maximum Parsimony Phylogeny (10 points): Show how the weighted version of Fitch's algorithm would determine the cost of the tree below using the following symmetric cost matrix. Show your calculations and the resulting minimal cost characters for the internal nodes.



	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

3. Hierarchical Clustering (15 points): Suppose we were using a hierarchical clustering method to arrive at a partitional clustering with two clusters. Given the following **distance** matrix, would single-link and complete-link clustering produce the same two clusters? Show your work to justify your answer.

	A	B	C	D	E
A	0	4	8	5	2
B		0	5	4	1
C			0	7	3
D				0	6
E					0

4. Protein Threading (15 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.

$$\begin{aligned} g_1(i, 2) &= 5 \\ g_1(i, 3) &= 4 \end{aligned}$$

$$\begin{aligned} g_1(j, 8) &= 3 \\ g_1(j, 9) &= 6 \end{aligned}$$

$$\begin{aligned} g_1(k, 13) &= 2 \\ g_1(k, 14) &= 10 \end{aligned}$$

$$\begin{aligned} g_2(i, j, 2, 8) &= 7 \\ g_2(i, j, 2, 9) &= 1 \\ g_2(i, j, 3, 8) &= 2 \\ g_2(i, j, 3, 9) &= 1 \end{aligned}$$

$$\begin{aligned} g_2(i, k, 2, 13) &= 2 \\ g_2(i, k, 2, 14) &= 1 \\ g_2(i, k, 3, 13) &= 10 \\ g_2(i, k, 3, 14) &= 5 \end{aligned}$$

$$\begin{aligned} g_2(j, k, 8, 13) &= 4 \\ g_2(j, k, 8, 14) &= 12 \\ g_2(j, k, 9, 13) &= 6 \\ g_2(j, k, 9, 14) &= 12 \end{aligned}$$

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 $g_1(k, 13) = 2$). 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.

5. Short Answer (10 points): Briefly define each of the following terms:

guide tree

homology modeling

molecular clock assumption

outgroup

progressive multiple sequence alignment

6. Tree-Based Representations (9 points): We talked about using tree-based representations in three different contexts: (1) phylogenetic inference, (2) hierarchical clustering and (3) decision-tree classification of gene-expression profiles. For each of these tree-based representations, briefly answer the following:

- a) What do the internal nodes represent?
- b) What do the leaves represent?
- c) Are the trees rooted?

7. EM Clustering of Sequences (16 points): We discussed an instantiation of EM clustering in which the clusters were represented by Gaussians. Consider a case in which we want to use EM to cluster protein sequences. Suppose that we will use one profile HMM (instead of a Gaussian) to represent each cluster. Briefly, describe how you would adapt EM clustering for this case. In particular, discuss:

- a) What is the hidden state in this problem?
- b) How will you compute the E (expectation) step of the clustering procedure?
- c) How will you compute the M (maximization) step of the clustering procedure?