

Biostatistics & Medical Informatics / Computer Sciences 576
Introduction to Bioinformatics
Fall 2002 Midterm 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 7).

Problem	Score	Max Score
1.	_____	25
2.	_____	10
3.	_____	35
4.	_____	30
Total	_____	100

1. Pairwise Sequence Alignment:

1a. (20 points) Show how the dynamic programming approach would be used to determine a *global* alignment for the two sequences below. Matching bases should be scored +1, mismatching bases should be scored -1, and each gap position should be penalized -2. Show the filled-in scoring matrix and the traceback pointers in the matrix.

x: TCG

y: TCCG

1b. (5 points) For these two sequences and the scoring scheme used above, show **all** of the optimal alignments. Indicate which one is the highroad alignment.

2. Multiple Sequence Alignment (10 points): Briefly describe one approach for calculating multiple sequence alignments. State one advantage and one disadvantage of this approach.

3. Markov Models:

3a. (25 points) Suppose we want to design a Markov model for the following problem. We are given DNA sequences that consist of alternating subsequences of type X and type Y . We want to use a second-order, homogeneous submodel to represent X subsequences and a first-order, homogeneous submodel to represent Y subsequences. We know the following conditions will hold:

- the sequences we process will always start with type Y segments,
- the X and Y subsequences can be of arbitrary length (≥ 2),
- subsequences of type X always start with **A** and end with **C**,
- subsequences of type Y always start with **G** and end with **T**.

Draw a picture showing the graphical structure of such a model. In particular show (a) the start state, (b) all of the states of the second-order X submodel, (c) all of the states of the first-order Y submodel, (d) all of the transitions *within* the Y submodel, (e) the transitions leading from the start state, and (f) the transitions between the X submodel and the Y submodel. You do *not* need to show the transitions that are just within the states of the X submodel, nor do you need to show an end state or transitions to it.

October 29, 2002

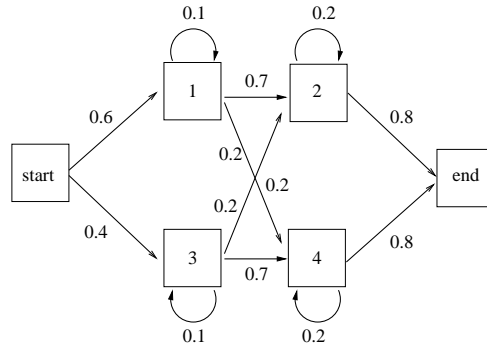
Name _____

3b. (5 points) Is this model a hidden Markov model? Explain why or why not.

3c. (5 points) Suppose that we have used a training set to determine the probability parameters of the model above, and now we want to use it to process a test sequence. In particular, we want to predict which parts of the test sequences represent X subsequences and which represent Y subsequences. Briefly describe how we could do this?

4. Markov Models:

4a. (20 points) 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. Show the values calculated by the Forward algorithm for the sequence **ATT**.



state 1	state 2	state 3	state 4
$e_1(A) = 0.5$	$e_2(A) = 0.2$	$e_3(A) = 0.3$	$e_4(A) = 0.1$
$e_1(C) = 0.2$	$e_2(C) = 0.2$	$e_3(C) = 0.3$	$e_4(C) = 0.1$
$e_1(G) = 0.1$	$e_2(G) = 0.4$	$e_3(G) = 0.2$	$e_4(G) = 0.1$
$e_1(T) = 0.2$	$e_2(T) = 0.2$	$e_3(T) = 0.2$	$e_4(T) = 0.7$

4b. (5 points) How many possible paths are there for this sequence through the HMM?

4c. (5 points) For this HMM, how many values would the Forward algorithm need to compute (i.e., how many elements in the F matrix would need to be filled in) if the sequence we were processing was 10 bases long? Explain your answer.