Computer Sciences 838 – Bioinformatics
Spring 2001 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.	_____	18
2.	_____	20
3.	_____	16
4.	_____	26
5.	_____	8
6.	_____	12
Total	_____	100

May 8, 2001 Name _____

1. Pairwise Alignment:

1a. (8 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: `ACCC`

1b. (3 points) For these two sequences and the scoring scheme used above, how many distinct *optimal* alignments are there? Explain how you arrived at your answer.

1c. (3 points) Show the *lowroad* and the *highroad* alignments given by the filled-in matrix above.

1d. (4 points) Briefly describe how <u>two</u> elements of the dynamic programming procedure would differ if you were computing a *local alignment* of *protein* sequences with an *affine gap penalty function* instead of computing an alignment as you did in part 1a.

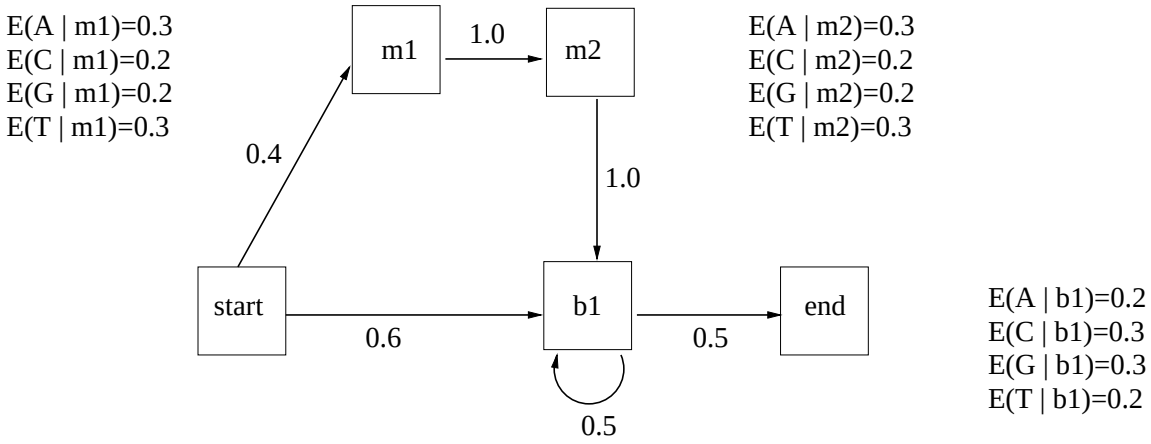
2. Markov Chain Models:

2a. (10 points) Suppose that we wanted to construct a *first-order Markov chain model* for representing a particular class of DNA sequences. Draw a picture of that shows the states and transitions of such a model. Your model should include a *start* state, but need not include an *end* state. With the two sequences, `ACTACGT` and `ACG`, estimate the transition probabilities for this model using Laplace estimates (i.e., pseudocounts of 1). Estimate the probabilities for transitions from the *start* state just like transitions from any other state. It is fine to express all of your probabilities as fractions.

2b. (10 points) Draw a picture of a *second-order Markov chain model* for these sequences. Your picture should depict all of the states. Do not show all of the transitions, but instead show only those that are used by the two sequences considered previously: `ACTACGT` and `ACG`. For just these transitions, show the probabilities as determined by maximum likelihood estimates.

3. Hidden Markov Models

(16 points) Suppose we wanted to use the simple HMM below to detect sequences that start with a two-base motif of interest. The states $m1$ and $m2$ provide a probabilistic representation of the bases in the motif, and the state $b1$ represents other bases in the sequence. Show how you would use one of the dynamic-programming algorithms for HMMs to decide if the sequence `ATT` probably starts with the motif or not. Show the calculations made by this algorithm as well as your final prediction.



4. Clustering and Gene Expression Analysis

4a. (10 points) Show how an *average-link hierarchical clustering* method would cluster the following four vectors

$$x_1 = \langle 4, 4 \rangle$$
$$x_2 = \langle 4, 3 \rangle$$
$$x_3 = \langle 1, 2 \rangle$$
$$x_4 = \langle 2, 1 \rangle$$

using the similarity function:

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

Here, x_i^1 represents the first component and x_i^2 represents the second component of the vector x_i. Show each step of the algorithm as well as the final clustering returned.

4b. (8 points) Using the same four vectors, show how *k-means clustering* would move its cluster centers on its first iteration. Assume that $k = 2$ and the initial coordinates of the cluster centers are $\langle 0, 3 \rangle$ and $\langle 5, 2 \rangle$.

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

4d. (6 points) Briefly compare and contrast clustering approaches, like those considered above, to the approach of Friedman *et al.* which learns a *Bayes network* to characterize expression data.

5. Protein Structure Prediction

(8 points) Briefly compare and contrast the problems of *pairwise sequence alignment* and *protein threading*, and the methods used for these two tasks. Focus on the most essential similarities and differences.

6. Stochastic Context Free Grammars

6a. (6 points) Briefly describe the relationship of hidden Markov models (HMMs) and stochastic context free grammars (SCFGs) to the Chomsky hierarchy, and explain why SCFGs are well suited RNA modeling.

6b. (6 points) Briefly compare and contrast the basic EM approach for learning sequence motifs to the inside-outside method for SCFGs. Focus on the most essential similarities and differences.