BMI/CS 576
Introduction to Bioinformatics
Fall 2007 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, make sure your exam has every page (numbered **1** through **8**).

Problem	Score	Max Score
1.	_____	16
2.	_____	12
3.	_____	16
4.	_____	14
5.	_____	16
6.	_____	14
7.	_____	12
Total		100

1. Pairwise Global Alignment (16 points): Show how the dynamic programming approach would be used to determine a <u>global</u> 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 (i.e. the gap penalty function is <u>linear</u>). Show the filled-in scoring matrix, the traceback pointers in the matrix, and <u>all</u> of the optimal alignments.

x: **TGT**
y: **TACT**

2. Constrained DP for Pairwise Alignment (12 points): Suppose you are given two sequences of length L, x and y, for which you want to find a global alignment. Moreover, suppose you are given the constraint that k characters starting position r in sequence x must be aligned with the k characters starting at position c in sequence y. Otherwise, the alignment is unconstrained. If we are using a linear gap-penalty function, how many cells in the dynamic programming matrix <u>need to be</u> computed? Justify your answer. You may include or exclude the cells that are filled in as part of the initialization, but be sure to indicate whether you are counting them or not.

3. Markov Chain Models (16 points): Suppose we want to use an <u>inhomogeneous</u> Markov chain model to represent sequences that consist of triplets of characters (like codons in an exon). Each sequence may contain an arbitrary number of triplets. Suppose we want to use 0th <u>order</u> states to represent the characters that appear in the first position of each triplet, and 1st <u>order</u> states to represent the characters that appear in the second and third positions of each triplet. Draw the states and transitions for such a Markov model. Be sure to include *start* and end *states*. For simplicity, assume that our alphabet consists only of the characters **A** and **C**.

4. Parameter Learning in Hidden Markov Models (14 points): Consider a situation in which we are learning the parameters of the HMM shown below. We are given training sequences for which the correct state is unknown (i.e. hidden) for most characters in the sequence. However, for some sequences we know that the kth character must be emitted by *state 1*. Describe how we should adjust standard <u>Baum-Welch</u> learning to handle this situation. Be specific in your description.

5. Distance-Based Phylogenetic Tree Inference (16 points)

5a (12 points): Given the distance data below, show how how UPGMA would produce a phylogenetic tree. 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	7	6	6
B		0	9	9
C			0	2
D				0

5b (4 points): Do you think UPGMA is the most appropriate phylogenetic-tree algorithm for this data? <u>Briefly</u> justify your answer.

6. Multiple Sequence Alignment (14 points): Given n sequences to align, how many pairwise alignments and how many <u>merging operations</u> do the <u>tree-based</u> and <u>star-based</u> methods perform in computing a multiple alignment of these sequences. In your answer for the star-based approach, state which method you are assuming for determining the center of the star.

7. Short Answer (12 points): Briefly define each of the following terms.

affine gap-penalty function

maximum likelihood estimation

nearest neighbor interchange

outgroup

profile HMM