

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 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 (i.e. the gap penalty function is linear). Show the filled-in scoring matrix, the traceback pointers in the matrix, and all of the optimal alignments.

x: TGT
y: TACT

		T	A	C	T
T	0	-2	-4	-6	-8
G	-2	1	-1	-3	-5
T	-4	-1	0	-2	-4
T	-6	-3	-2	-1	-1

x: T A C T x: T A C T
y: T - G T y: T G - T

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 need to be 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.

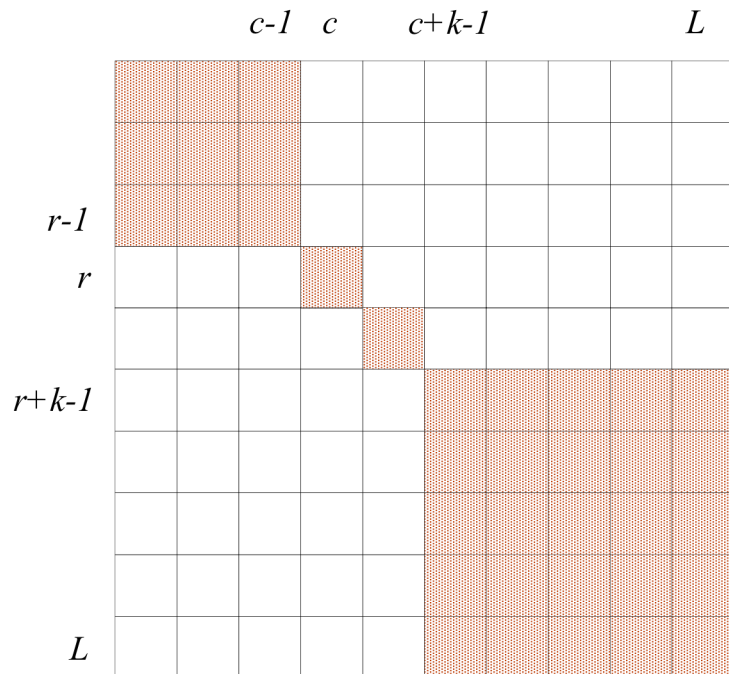
Not counting initialization, this requires filling in

$$(r-1) \times (c-1) +$$

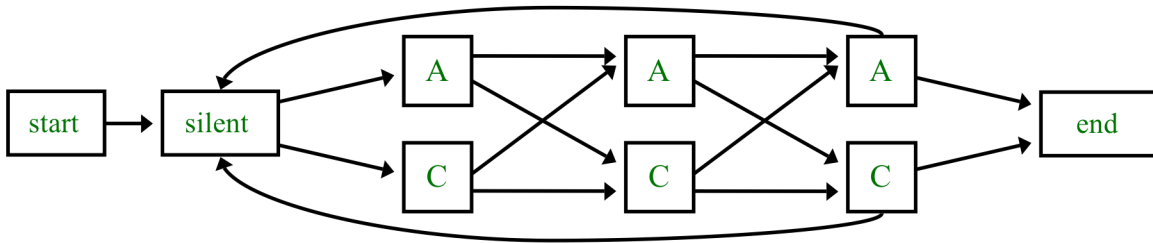
$$(k-1) +$$

$$(L - (r+k-1) + 1) \times (L - (c+k-1) + 1)$$

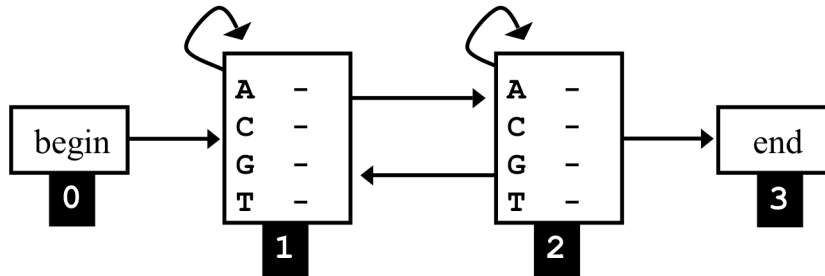
cells.



3. Markov Chain Models (16 points): Suppose we want to use an inhomogeneous 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 order states to represent the characters that appear in the first position of each triplet, and 1st order 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 k^{th} character must be emitted by *state 1*. Describe how we should adjust standard Baum-Welch learning to handle this situation. Be specific in your description.



We can adjust the Forward and Backward steps as indicated below, and then do the rest of the E-step and the M-step as it is usually done.

$$f_0(0) = 1$$

compute $f_1(i), f_2(i)$ for $1 \leq i < k$ as usual

$$f_1(k) = e_1(x_k)(f_1(k-1) + f_2(k-2))$$

$$f_2(k) = 0$$

compute $f_1(i), f_2(i)$ for $k < i \leq L$ as usual

$$b_3(L) = 1$$

compute $b_1(i), b_2(i)$ for $k \leq i < L$ as usual

$$b_1(k-1) = a_{11}e_1(x_k)b_1(k)$$

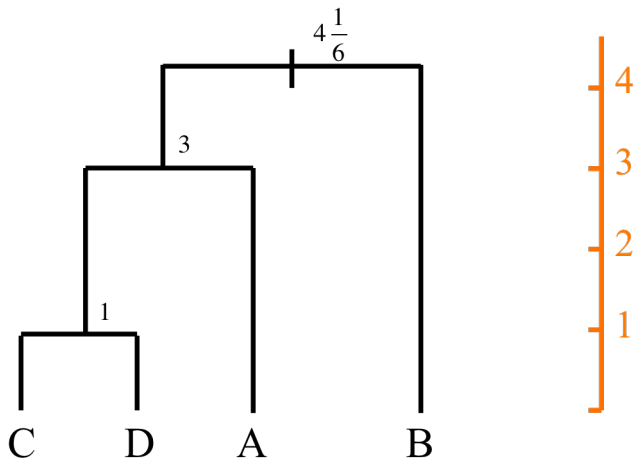
$$b_2(k-1) = a_{21}e_1(x_k)b_1(k)$$

compute $b_1(i), b_2(i)$ for $0 \leq i < k-1$ as usual

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

5a (12 points): Given the distance data below, show 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? Briefly justify your answer.

No, because the data is not ultrametric. Neighbor joining would be a better choice.

6. Multiple Sequence Alignment (14 points): Given n sequences to align, how many pairwise alignments and how many merging operations do the tree-based and star-based 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.

Tree-based: $n-1$ pairwise alignments (one at each internal node of tree) and 0 merging operations.

Star-based with “maximal similarity” selection of the center: first compute all $\frac{n(n-1)}{2}$ pairwise alignments then do $n-2$ merges.

Star-based with “try each sequence” as the center: first compute all $\frac{n(n-1)}{2}$ pairwise alignments, then do $n(n-2)$ merges.

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