Homework 2: Multiple Sequence Alignment

BCH4300B, Winter 2014
Assigned: March 13, 2014
Due: March 20, 2014

1. In this question we will practice several elements of progressive multiple sequence alignment, as outlined in class. For this problem, we will use the same gap penality $P = -1$ and scoring matrix S (below) as in the first homework. Suppose we are trying to construct a multiple alignment of the four DNA sequences: $X_1 = CTGG$, $X_2 = CAGG$, $X_3 = CTGT$, and $X_4 = ACAC$. (Of course, this is an artificial example. Usually would would be aligning much longer sequences, and probably many more than four.

$S =$

	A	C	G	T
A	3	-3	-1	-3
C	-3	3	-3	-1
G	-1	-3	3	-3
T	-3	-1	-3	3

Recall that progressive alignment has three phases: (1) computing all pairwise alignment scores (or "similarities") by a pairwise sequence alignment approach; (2) computing the "guide tree", which groups together the most similar sequences; and (3) using the guide tree to perform pairwise alignments (between either individual sequences or groups of sequences), ultimately leading to our multiple alignment. You know well how to perform Step (1) by the Needleman-Wunsch dynamic program; we covered it in class and you practiced it in homework 1. Rather than make you repeat such tedious computations for the current example, I will simple provide you with the pairwise scores obtained:

Pairwise scores:

	X_1	X_2	X_3	X_4
X_1	·	6	7	-2
X_2	·	·	1	1
X_3	·	·	·	-1
X_4	·	·	·	·

1A) Use the method described in class to compute (by hand) the guide tree for these sequences based on the pairwise scores above. Recall that at each step of the construction, you (a) find the highest score in the matrix, (b) group together the sequences corresponding to that score, and (c) create a new (smaller) score matrix where each entry (e.g. row i and column j) is the average pairwise alignment score between the one or more sequences in group i and the one or more sequences in group j. Show the groups after each step of guide tree construction, along with the new score matrix computed after the grouping.

1B) Having constructed your guide tree, you could, in principle, proceed to construct the full multiple alignment by performing a several pairwise alignments. (Aside: How many would it be, exactly?) However, instead of doing all of them, let's just do one. (In the eventual homework solution, I can post all the steps, if you like.) In particular, if you constructed your guide tree correctly, then at one point you should have found that it recommends aligning the grouping $\{X_1, X_3\}$ with the sequence X_2. Let's do this one by hand, using the Needleman-Wunsch pairwise alignment procedure–but accounting for the fact that one of the things we're aligning is a pair of sequences. Suppose that the optimal alignment between X_1 and X_3 is:

$$\begin{array}{ccccc} C & T & G & G & - \\ | & | & | & | & | \\ C & T & - & G & T \end{array}$$

Therefore, if we want to fill in a dyanmic programming table aligning that with X_2, the table should like something like what is below. I've helped you out by filling in the first row and first column. These correspond, respectively, to aligning more and more of what's on top or what's down the left side, with gaps. (Do you understand why the gap penalties are higher along the first row than down the first column?)

		$\begin{array}{c} C \\ \vert \\ C \end{array}$	$\begin{array}{c} T \\ \vert \\ T \end{array}$	$\begin{array}{c} G \\ \vert \\ - \end{array}$	$\begin{array}{c} G \\ \vert \\ G \end{array}$	$\begin{array}{c} - \\ \vert \\ T \end{array}$
	0	-2	-4	-5	-7	-8
C	-1					
A	-2					
G	-3					
G	-4					

You should fill in the rest of this table in Needleman-Wunsch style. So for example, consider filling in row $i = 3$ and column $j = 4$. This means that we're trying to find the best alignment between the first two letters of X_2, namely CA, with the first three columns of the alignment of X_1 with X_3, namely $\begin{array}{ccc} C & T & G \\ | & | & | \\ C & T & - \end{array}$. Just as in classic Needleman-Wunsch, we need to consider three options: aligning A with $\begin{array}{c} G \\ | \\ - \end{array}$, aligning A with a gap, or aligning $\begin{array}{c} G \\ | \\ - \end{array}$ with a gap. Each option is evaluated by the score of the column it creates plus the score of the best alignment of what remains after that choice (which is in the box up & left, up, or left respectively). Remember, to score a column, we use the sum of pairs. So for example, if we decided to align A with $\begin{array}{c} G \\ | \\ - \end{array}$, we would get the column $\begin{array}{c} G \\ | \\ - \\ | \\ A \end{array}$, which has score -1 (G with gap) plus -1 (G with A) plus -1 (gap with A), or -3 total.