

**BCB 567/CprE 548**  
**Fall 2007**  
**Homework 5**  
**Due Tuesday, Dec 4**

**Note:** Use constant gap penalty and the following parameters, unless otherwise specified  $\alpha$  denotes the score of a match,  $\beta$  denotes mismatch penalty, and  $\gamma$  denotes the gap penalty.

1. (4 points) Let  $s_1, s_2, \dots, s_k$  be  $k$  sequences, each of length  $n$ . What is the maximum possible SP-score of a multiple sequence alignment between them? What is the worst possible SP-score?
2. (4 points) Devise an example of three (short) sequences  $s_1, s_2$  and  $s_3$  such that the induced alignment of  $s_1$  and  $s_2$  has a lower score than the optimal pairwise alignment of  $s_1$  and  $s_2$ .
3. (5 points) Consider generalizing the  $k$ -band algorithm for the multiple sequence alignment of  $r$  sequences, of length  $n$  each. Derive a stopping condition for the same.
4. (5 points) In the center star method, the multiple alignment is constructed by successively aligning each new string to the center string  $s_c$ . The *order* in which the strings are aligned to  $s_c$  is not specified. Argue that the same multiple alignment is built, regardless of the order used.
5. (7 points) Let  $S$  be a set of amino acid sequences. A *profile* for a multiple alignment of  $S$  is a table  $P$  such that  $P(x, j)$  gives the frequency with which character  $x$  appears in column  $j$  of the alignment, where  $x$  is either an amino acid or the gap character. Let  $P$  be a profile for a multiple alignment of  $S$  and let  $s$  be some given sequence; let  $m$  be the number of columns of  $P$  and  $n$  be the length of  $s$ . An *alignment* of  $s$  to  $P$  is obtained by inserting spaces into  $s$  and into the sequence  $\langle 1, \dots, m \rangle$  (that is the sequence of column indices of  $P$ ) so that the resulting sequences have the same number of characters and that no position of the alignment has two gaps. For example, suppose that  $s = ACTAG$  and that profile  $P$  has 5 columns. Shown below is an alignment of  $s$  and  $P$ :

$A$	$C$	$T$	$-$	$A$	$G$
1	$-$	2	3	4	5

Suppose that we have a fixed gap penalty  $\gamma$  and that we are given a weight matrix  $W = [w_{ab}]$ , where  $w_{ab}$  is the score for aligning amino acids  $a$  and  $b$ . To score a sequence to profile alignment  $A$ , we score each position separately and compute the total. A position  $i$  of  $A$  is scored as follows: If position  $i$  has a gap, then its score is  $-\gamma$ . Otherwise, if the position corresponds to character  $a$  from  $t$  and column  $j$  from  $P$ , then its score is  $\sum_b(w_{ab} \cdot P(b, j))$ .

Design a dynamic programming algorithm that finds a maximum-score sequence to profile alignment and analyze its running time.