COMP 571: Homework #1
Spring 2017

Assigned on January 26, 2017

Due at the beginning of class on February 7, 2017.

The Honor Code applies to all homework sets. Sign the pledge on your solutions.

1. Let two sequences be $q = CDAA$ and $d = AEECA$, and a scoring matrix:

	A	C	D	E
A	2	-2	-2	-1
C		1	0	0
D			2	-2
E				2

   (a) Find the highest score by aligning q and d when the gap penalty is $g(\ell) = -2\ell$. Then, find the best alignment.

   (b) Now use gap penalty $g(\ell) = -1.8 - 0.4\ell$. The dynamic programming table will be partly filled as below, using the general DP procedure:

q/d			A	E	E	C	A
		0	1	2	3	4	5
	0	0.0	-2.2	-2.6	-3.0	-3.4	-3.8
C	1	-2.2	-2.0	-2.2	-2.6	-2.0	-4.2
D	2	-2.6	-4.2	-4.0	-4.2	-2.6	-4.0
A	3	-3.0	-0.6	-2.8	-3.2		
A	4	-3.4	-1.0	-1.6			

   Fill in the rest of the table, and find the best alignment.

   (c) Compare the alignments found under (a) and (b), and find for each of them the minimum number of mutations which might have occurred, when we suppose that only one residue is included in a substitution.

2. If we use a gap penalty $g(k) = -10k$, we observe that all optimal alignments of strings A and B are identical for all values of the mismatch score μ when $\mu < -20$. Explain why this is true.

3. In some cases one wants to score gaps at the ends of an alignment as 0.

   (a) In what cases is this reasonable (what is the relation between the two sequences)?

   (b) Modify the global alignment algorithm to take into account end gaps with zero score.

   (c) We have sequences $q = ART$ and $d = AARRTRT$. Use score 1 for a match, -1 for a mismatch, and a linear gap penalty of -1, and find the best alignments when a score of 0 is used for the end gaps.

4. At times it is desirable to force a pairwise global alignment to align certain sections of the two sequences—regardless of whether the regions would be aligned in the optimal global alignment. The positions where the algorithm is forced to align certain nucleotides are called *anchors*.

   Describe an algorithm for globally aligning sequences A and B with a set of anchors $C = \{(i_1, j_1), ..., (i_n, j_n)\}$ where (i_k, j_k) is a pair of positions in A and B, respectively, that must be paired with one another in the final global alignment. Argue that your solution is correct.

5. Show that the number of pairwise alignments of two sequences of lengths m and n, respectively, is

$$\sum_{k=0}^{\min(m,n)} \binom{n+m-k}{k, m-k, n-k},$$

   where

$$\binom{p}{k_1, k_2, k_3} = \frac{p!}{k_1!k_2!k_3!}.$$

   Hint: Think of each alignment as a path in the alignment matrix (that one that the dynamic programming algorithm builds).