UBC, CPSC 445: ALGORITHMS FOR BIOINFORMATICS (WINTER 2 2014/2015)

Assignment 1

Due: January 29, 2015 (at the beginning of the class)

1. Compute similarity matrix for RNA based on the following alignment of two Ribosomal RNA sequences:

```
-----CCACCCGGCGAUAGUGAGCGGGCAACACCCGGACUCAUCUCGAACCCGGAAGUAAAG-UCCC
GAUGGGUGCACGGUCAUAGCGGUGGAGUU-UACCCGGUCUCAUCCCGAACCCGGAAGUCAAGCCCUC

CUACGUUGGUAAG-GCA--GUGGGAUCCGCAAGGGCCUGCAGCCUUGCCAAGCUGGGAUGGACAUU
CUGCGUC-UGUCCC-AAUACUGUGGUACGAGAGUCCACGGGAACGGCGGUCACUGUG-C-------
```

   Note: Ignore aligned positions that contain a gap when calculating p_{ab} probabilities, but consider them when calculating q_a probabilities.

2. What is the number of alignments of a sequence x of length n and a sequence y of length m that do not contained any matched characters (that is: every character in x is a deletion and every character in y is an insertion). Note: all patterns (insertion-deletion, deletion-insertion) are allowed.

3. Develop a recurrence for counting the number of alignments in which a gap in one sequence is not allowed to be immediately followed by a gap in the other sequence (i.e., not insertion-deletion and deletion-insertion pattern). Use it to compute the number of such alignments for two sequences of the same length n, where $n = 1, \ldots, 7$ (you can write the program for that if you prefer or do it on paper with a calculator). Based on these numbers, guess whether this number is growing exponentially or polynomially.

4. (a) Implement a DP algorithm for global pairwise alignment with the affine gap penalty function. For scoring use the similarity matrix BLOSUM62 available at BLOSUM62.txt and the affine gap penalty function with constants $d = 8$ and $e = 3$. Use a programming language of your choice. Submit your code through Connect.

   Make sure that your code works correctly before proceeding to part (b).

   (b) Use your program to find the best 3 matches between the unknown sequence unknown.txt and selected 1000 human proteins available at uniprot list. This file is in FASTA format: for each protein there is line starting with `>` which contains identifying information about the protein, followed by several lines of lengths at most 60 with amino acids of the protein.

   Your answer should consists of 3 lines (top ranking matches), each line containing the name of the protein and the alignment score. The name of the protein is the third record in the identification line, the one ending with `_HUMAN`. For example, the name of the first protein in the list is `1A1L1_HUMAN`.

5. (a) Find the best global pairwise alignment of sequences XX and YYY using the following scoring model: the similarity score between X and Y is $s(X, Y) = -4$ and use the affine gap penalty function $\gamma(L) = -d - e(L - 1)$ with $d = 4$ and $e = 1$. What is the score of this alignment? (You can use any method you want, either recurrences from the lecture, or enumerate and compare all possible alignments, or some other method you come up with.)

(b) In the textbook (on page 30) it's claimed that *"if $-d-e$ is less than the smallest mismatch score (which we denoted by* MINs *at the lecture) then the optimal alignment does not contain a gap in sequence followed by a gap in the other sequence".* Use part (a) to argue that this statement is incorrect. What score would be computed for these two sequences by the simplified version of the algorithm for global pairwise alignment with affine gap penalty:

$$M(i,j) = s(x_i, y_j) + \max \begin{cases} M(i-1, j-1) \\ G_x(i-1, j-1) \\ G_y(i-1, j-1) \end{cases} \tag{1}$$

$$G_x(i,j) = \max \begin{cases} M(i-1, j) - d \\ G_x(i-1, j) - e \end{cases} \tag{2}$$

$$G_y(i,j) = \max \begin{cases} M(i, j-1) - d \\ G_y(i, j-1) - e \end{cases} \tag{3}$$

Hint. You can expect the score computed in part (b) to be smaller than the optimal score found in part (a).

6. Develop an algorithm for local pairwise alignment with a general gap penalty function $\gamma(L)$. You can assume that γ satisfies the "superadditivity" property: $\gamma(a) + \gamma(b) \leq \gamma(a+b)$ for every $a, b \geq 1$.