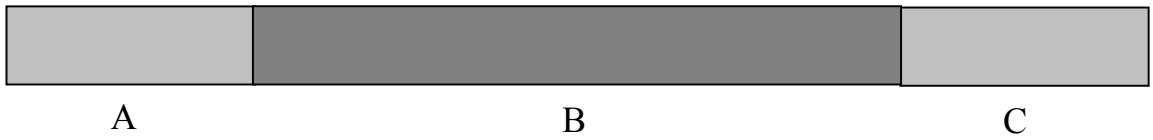
Due Date: Monday, March 28, 2011, 11:59PM

CENG 465
Spring 2010-2011

Assignment #1
(Programming Assignment)

Pairwise Sequence Alignment using Location Specific Scores

In this assignment you are going to implement a modified version of the Needleman-Wunsch pairwise sequence alignment algorithm. Needleman-Wunsch algorithm is the dynamic programming algorithm for global pairwise sequence alignment. You will use your program to align DNA sequences. The modification is that you will have different gap penalties, match scores, and mismatch penalties in different regions of the sequences while performing the alignment. For this assignment, a sequence has three regions as shown below:



The regions A and C are called flanking regions and the region B is called the center region. The lengths of the left and right flanking regions are the same. In other words, the regions A and C contain the same number of base pairs. For a given sequence of length *L*, if the length of region A is given as *M*, the lengths of regions C and B will be *M* and *L*-2**M*, respectively.

You will use a scoring scheme which varies with respect to regions while aligning a pair of DNA sequences..

Below is the detailed description of the scoring scheme:

- If there is a match between two center region nucleotides, the match score is 7
- If there is a mismatch between two center region nucleotides, the mismatch penalty is -5
- If there is a match between a center region nucleotide and a flanking region nucleotide, the match score is 2
- If there is a match between two flanking region nucleotides, the match score is 4
- If there is a mismatch between a center region nucleotide and a flanking region nucleotide, the mismatch penalty is -8
- If there is a mismatch between two flanking region nucleotides, the mismatch penalty is -3
- Linear gap penalty is used
- If a center region nucleotide is aligned against a gap, the gap penalty is -3
- If a flanking region nucleotide is aligned against a gap, the gap penalty is -1

Below is an example alignment (not necessarily the optimum alignment), in which the lengths of the flanking region in both sequences are given as 4 nucleotides. The flanking regions are marked.

```
ATCG-TTACACGGTATCACAACAAG-
-TGAC-T-CA-GGTCTCAGAATA-GA
```

The corresponding alignment score is:
-1 + 4 + -3 + -3 + -1 + -3 + 7 + -3 + 7 + 7 + -3 + 7 +7 + 7 + -5 + 7 + 7 + 7 + -5 + 7 + 7 + -3 + 4 + -1 + 4 + -1 = 57

Your program will get two DNA sequences as input and a single integer for the length of the flanking regions in both sequences (the lengths of the flanking regions in both sequences are equal). Your program will find the optimum alignment between the two sequences based on the described scoring scheme.

Test your program with the test data provided at:

http://www.ceng.metu.edu.tr/~tcan/ceng465_s1011/Assignments/hw1_test_data.zip

In the zip file provided at the above URL, there are five pairs of sequences in separate files. The first line in the file is the length of an individual flanking region given as a single integer, *M*. The following two lines contain two DNA sequences each. The length of the sequence is guaranteed to be greater than 2**M*.

Run your program for the test sequences and write a report answering these questions.

Question 1 (20 pts): Describe your modification to the original NW algorithm.

Question 2 (10 pts): Analyze the running time complexity of your program theoretically. Did the modifications to the original algorithm cause an increase in the running time? You may ignore constant time increases.

Question 3 (25 pts): What are the alignment scores for the 5 test sequence pairs?

Question 4 (25 pts): What are the alignments for the 5 test sequence pairs?

Question 5 (20 pts): For test5, i.e., the fifth test sequence pair, if the length of the flanking region was different, could you get a better alignment score? In other words, ignore the length provided in the test file and then find the best length of the flanking region which results in the maximum alignment score. There are $\lceil L/2 \rceil - 1$ different lengths you should try if the length of the shorter sequence is *L* and any region has at least on nucleotide.

Deliverables:

1. The source code of your program. You may use any programming language of your choice.
2. Your report which contains the answers to the questions above.

Submission:
Submit the deliverables as a zip bundle or as a tarball using the COW system.

Late Submission Policy:
Late submission is allowed. Your final assignment grade will be penalized 20 points per late day.

USE THE NEWSGROUP IF YOU HAVE QUESTIONS ABOUT THE ASSIGNMENT.

CHECK THE NEWSGROUP REGULARLY FOR POSSIBLE UPDATES ON THE ASSIGNMENT.