**CSE4305 Assignment 1**

**Pairwise Sequence Alignment using Dynamic Programming**

**Due 11:55PM, April 25, 2018**

**Introduction**

The first programming assignment is to implement the pairwise sequence algorithm for sequence alignment as discussed in class, parameterized in such a way that it can be used for either global or local sequence alignment. For this assignment, we expect your program to be able to trace **all the alignments** that give the best alignment score. Also, we expect **no end-gap penalties** to be imposed.

Remember, the differences between the local alignment and global alignments are:

1. Local alignment requires negative scores in the match matrix
2. Local alignment never records a score in the score matrix below 0.0
3. When looking for best match, local alignment looks for best score anywhere in matrix (global alignment looks only in last row and last col for best score).

**Program Specifications**

The input to the program is a file with lines containing the following information: (See sample files **align\_example\_input** and **align\_example\_input\_annotated**):

1. A sequence of letters indicating sequence A
2. A sequence of letters indicating sequence B
3. An indication of whether local (1) or global (0) alignment is sought.
4. A set of gap penalties for introducing gaps into A or B.
5. The symbol alphabets to be used (e.g. ATGC for DNA strands and 20 single-letter abbreviations for the proteins, but there could be other symbols)
6. A matrix which shows the score for a match between an element in B and one in A.

For simplicity, you could assume that the input sequences will never be longer than 300 letters and the symbol alphabet size is never larger than 26.

The output of the program is a file with lines containing the following information: (See **align\_example\_output**)

1. score for the alignment.
2. sequence A (with necessary gaps)
3. sequence B (with necessary gaps) such that aligned characters from A and B are in the same column.

When you need to report more than one equivalent alignment, you can leave a blank line and then print the alignments in a similar manner. (See **align\_all\_example\_output**) There are 7 more align examples for your convenience.

**(Part 1)** Submit your source code and a readme file to e-class.

* 1. Source code with the comments and explanations of what is going on. Especially if you expect partial credit for "almost working" code.
  2. Readme file that describes the followings:
     + Your name and student ID.
     + Programming language used and instruction on how to run your program

**(Part 2)** Run your program on the 10 test cases provided (see attached files) and answer the following questions.

**Question 1**

Test file: test1\_dat.txt

Description: Global alignment of two DNA sequences.

Alignment Score: 5.6

**Question 2**

Test file: modify test1\_dat.txt gap penalty to (1 0.5 1 0.5) -> test2\_dat.txt

Description: Global alignment of the same two DNA sequences with different gap penalties.

Alignment Score: 5.4

**Question 3**

Based on test 2, what is a first aligned subsequence (a region aligned between gaps) in sequence 1? (Pick ONLY one alignment if you have multiple answers.)

Answer: TTGACACC

**Question 4**

Test file: test3\_dat.txt

Description: Global alignment of two protein sequences with gap penalty (1 0 1 0).

Alignment Score: 589.0

**Question 5**

Based on Test 3, on average, how many gaps does the first sequence from the data file have? (5+4+5+4+4+3)/6 = 4.17 -> 4 gaps on average

**Question 6**

Test file: modify test3\_dat.txt gap penalty to (1 0.5 1 0.5) -> test4\_dat.txt

Description: Global alignment of the same two protein sequences with different gap penalties.

Alignment Score: 586.0

**Question 7**

Based on Test 4, on average, how many gaps (up to one digit after the decimal point) does the first sequence from the data file have? (3+2)/2 = 2.5 gaps on average

**Question 8**

Test file: modify test3\_dat.txt gap penalty to (3 1.5 3 1.5) -> test5\_dat.txt

Description: Global alignment of the same two protein sequences with different gap penalties.

Alignment Score: 580.0

**Question 9**

Based on Test 5, on average, how many gaps (up to one digit after the decimal point) does the first sequence from the data file have? 0 gaps

**Question 10**

Test file: test6\_dat.txt

Description: Local alignment of two protein sequences with gap penalty (2 1 2 1).

Alignment Score: 155.0

**Question 11**

Based on Test 6, the first aligned letter in sequence 1 is: I

**Question 12**

Based on Test 6, the last aligned letter in sequence 1 is: K

**Question 13**

Based on Test 6, the first aligned letter in sequence 2 is: V

**Question 14**

Based on Test 6, the last aligned letter in sequence 2 is; K

**Question 15**

Test file: modify test6.dat gap penalty to (4 1 4 1) -> test7\_dat.txt

Description: Local alignment of the same two protein sequences with different gap penalties.

Alignment Score: 137.0

**Question 16**

Test file: test8\_dat.txt

Description: Local alignment of two DNA sequences.

Alignment Score: 21.4

**Question 17**

Test file: test9\_dat.txt

Description: Local alignment of the above (DNA translated into protein) sequences.

Alignment Score: 27.0

**Question 18**

Test file: modify test9.dat to global alignment -> test10\_dat.txt

Description: Global alignment of the above (DNA translated into protein) sequences.

Alignment Score: 26.0

**Question 19**

Compare Test 9 and Test 10. Which gives you a higher alignment score?

a. Local alignment

b. Global alignment

**a. Local alignment**

**Question 20**

Compare Test 3, 4, 5 or Test 6, 7.

When you increase the gap penalty, does your alignment score change (general trend)?

a. Increase

b. Decrease

c. Unchange

**b. Decrease**

**Question 21**

When you increase the gap penalty, does the number of gaps in the alignment change (general trend)?

a. Increase

b. Decrease

c. Unchange

**b. Decrease**