Homework 1
Due: Saturday Feb 8, before class starts

Note: the official language for this HW assignment is MATLAB. However, if you have some trouble with MATLAB, you can use another language (c/c++/java/python). Those who use MATLAB for implementation will receive 5 additional points for each of the two implementation assignments.

Problem 1 (25 points): Biology

You are given the following DNA sequence, which is believed to contain a small protein-coding gene.

GGAGGCGTAA AATGCGTACT GGTAATGCAA ACTAATGG

- If this sequence is fully transcribed (used as a coding strand), what is the corresponding mRNA sequence?
- Which region of the mRNA do you think can be translated into a protein (hint: Can you identify the start codon and stop codon from the mRNA sequence?)
- What is the protein sequence encoded by this gene?
- If the reverse-complementary strand of this DNA sequence is also transcribed, what will be the mRNA sequence?
- Do you think the reverse-complementary strand of this DNA sequence can encode a protein by itself (see hint above) ?

Problem 2 (25 points): Open reading frame finder.

Write a MATLAB function to identify open reading frames(ORF) from an input DNA sequence. ORF is a continuous stretch of codons that contain a start codon (i.e., AUG) at the beginning and a stop codon (i.e., UAA, UAG or UGA) at the end only (https://en.wikipedia.org/wiki/Open_reading_frame). Note that an ORF may be found on both strand of the DNA sequence.

For each of the DNA sequences in the input files hw1q2x.txt (one sequence per file), use your program to report up to 3 longest ORFs. For each ORF, report the strand (+ or -), start and end positions of the ORF, and the total number of amino acids encoded by the ORF. Use the input sequence as + strand and index of DNA start from 1.

Problem 3 (25 points): Global and local alignment - manually without coding

Consider the sequences v = TACGGGTAT and w = GGACGTACG. Assume that the match score is +1, and the mismatch and gap penalties are -1.

- Fill out a dynamic programming table for a global alignment between v and w. Draw arrows in the cells to store traceback information. What is the score of the optimal global alignment and what alignment(s) achieves this score?
- Fill out the dynamic programming table for a local alignment between v and w. Draw arrows in the cells to store traceback information. What is the score of the optimal local alignment in this case and what alignment(s) achieves this score?

Problem 4 (25 points): Global alignment

Use MATLAB to implement the Needleman-Wunsch algorithm with m = 1, s = -1, d = -1. The input and output of your programs should be as follows.

Input: two sequence files. Each file contains one sequence, which can be recorded in multiple lines. Discard spaces if there is any.

Output: the optimal global alignment between the two sequences and the alignment score. The output alignment should have three lines as shown in the example below, where matching characters are shown by a — character, mismatches by a dot (.), and gaps by a dash (-). For longer sequences, break the alignment into lengths of 50.

```
ACGTACGTAG--GACGTAAGCAGAGAACGAGAACCCGGGAAC-ACGAGGC
||.||. ||| |||.|||||..||||.||.||||| ||||| |||||||
ACCTAG-TAGCGGACTTAAGCGTAGAAGGACAACCC-GGAACGACGAGGC
TGGTCGGCTT
|.||||.|||
TGGTCGTCTT
```

First try your algorithm on the sequences used in Problem 3 to make sure it works correctly. Then download hw1prob4.fa (google: fasta format) from the course website and copy the three sequences into separate files; then use your program to align each pair of sequences in the file.

FYI, the sequences encode the hemagglutinin (HA) protein for different strains of the influenza viruses. From the sequences, can you manually identify the start codon and end codon? From the alignment, is there any particular pattern to where or how the gaps occur? If you are allowed to manually adjust the alignment, what might you do to improve the biological relevance of the alignment and why? Alternatively, how can you improve your alignment algorithm to achieve this?

Bonus (5 points): feedback

How much time did you spend on this homework? Who did you discuss with and what was the discussion about? How is the difficulty level? Do you have any comments about the course?