Homework 1
Due: Wed, Feb 4 before class starts

Problem 1 (15 points)
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 the gene?

• If the reverse-complementary strand of the DNA sequence is also transcribed, what will be the mRNA sequence?

• Do you think this reverse-complementary strand can encode a protein by itself?

Problem 2 (15 points)
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 the 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 3 (20 points)
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 and may contain spaces. Discard all spaces before performing the alignment.

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-GGAACGAGGC

TGGTCGGCTT

.||| ||| ||

TGGTCGTCTT
```

Download three sequence files from the course website and use your program to align each pair of sequences. FYI, the sequences encode the hemagglutinin (HA) protein for different strains of the influenza viruses. From the sequences, can you identify the start 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?

**Problem 4 (20 points)**

Implement the Smith Waterman algorithm and use it to perform plagiarism detection. The input should be similar as specified in Problem 3 but spaces are not discarded. For output, report the alignments and scores for the highest-scoring aligned segments. (Bonus will be given for reporting three non-overlapping highest-scoring alignments.)

Test data files will be posted online soon.

Discuss what is the limitation of this plagiarism detection tool and what you might do to improve the detection accuracy for different situations.

**Bonus (5 points)**

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?