Homework Assignment #1 (100 points, weight 15%)
Due: Friday October 7, at 11:30 a.m. (in lecture)

Molecular Biology and Sequence Similarity

1. (10 marks) Explain with your own words the central dogma of molecular biology which describes the transformation from DNA to RNA to protein. Use at most half a page and provide a reference to any resource consulted.

2. (10 marks) For the following mRNA sequence, extract its 5'UTR, 3'UTR and the protein sequence. Briefly explain how you got your answer.

   CTTTTTGCTATGAATGCTGCGATTTAAGAGAATCCTTCTTCG

3. (20 marks) Use BLOSUM62 Score Matrix for aminoacids to calculate the score of the following protein sequence alignment, using the following types of alignment problems and gap penalty functions. Show the breakdown of your score calculation.

   ```
   QKKMIWGTCSYC----
   ----IWAGC--CFPST
   ```

   (a) global alignment with uniform gap penalty model (indel score -4).
   (b) global alignment with general gap penalty model $g(q) = \lfloor\sqrt{q}\rfloor$.
   (c) semi-global alignment with affine gap penalty model $g(q) = 3 + q$.

4. (20 marks) Simulate the steps of Hirschberg's global alignment algorithm, for the sequences given below and the following score function: match = +1; mismatch = -1 from purine to purine or from pyrimidine to pyrimidine and mismatch = -3 for other cases; indels = -1. Use a similar display as in the example given in class. Calculate the memory requirement for this example (in terms of number of matrix positions needed).

   S: `AAGTCGTT` and T: `CCGGCC`

5. (20 marks) Consider global alignment and note that the traceback paths in the dynamic programming table correspond one-to-one with the optimal alignments. Therefore the number of distinct optimal alignments can be obtained by computing the number of distinct traceback paths. Give an algorithm (pseudocode) to compute this number in $O(nm)$ time.

   Hint: Use dynamic programming.

6. (20 marks) Exercise 2.7-4
   Consider two DNA sequences of the same length n and let the scoring function be defined as follows: 1 for match, -1 for mismatch, -2 for indel. Let the score of the optimial global alignment be G and that of the optimal local alignment be L.

   (a) Prove that $L \geq G$ and construct an example such that $L = G$.
   (b) What is the maximum value of $L - G$? Construct an example with this maximum value.
   (c) If we want to find a pair of non-overlapping substrings within a given sequence S with the maximum global alignment score, can we simply compute the optimal local alignment between S and itself? Explain your answer.