CSI 5126 Algorithms in Bioinformatics — Fall 2010
Computer Science — University of Ottawa

Homework Assignment #1 (100 points, weight 15%)
Due: Tuesday October 5, at 11:30 a.m. (in lecture)

Molecular Biology and Sequence Similarity

1. (10 marks) Exercise 1.9-1 (note: collect info for 10 representative protein sequences (protein name, first amino-acids), provide info on gene bank used)
   Visit the gene bank and check a number of protein sequences. What are the first amino acids in those protein sequences? Any explanation for your observations?

2. (10 marks) Exercise 1.9-4 (note: explain how you obtained your answer and give appropriate references)
   Different species encode the same amino acid using different codons. For the amino acid Q, check the most frequently used codons in Drosophila melanogaster, human, Saccharomyces cerevisiae. Is there any codon bias for different species?

3. (15 marks) Exercise 1.9-5 (note: briefly explain how you got your answer.)
   For the following mRNA sequence, extract its 5'UTR, 3'UTR and the protein sequence.

   ACTTGTCATGGTAACTCCGTCGTACCAGTAGGTCATG

4. (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 $h = -4$, $s = -1$.

5. (20 marks) Exercise 2.7-10
   Consider two sequences $S[1..n]$ and $T[1..m]$. Assume match, mismatch and indel scores 1, -1, and -2, respectively. Give an algorithm that computes the maximum score of the best alignment (S', T') of S and T, forgiving spaces at the beginning of S' and spaces at the end of T'.

6. (25 marks) Exercise 2.7-9
   Given two DNA sequences S_1 and S_2 of length n and m, respectively, we would like to compute the minimum number of operations required to transform S_1 to S_2, where the allowed operations include: (1) insertion of one base (2) deletion of one base (3) replacement of one base, and (4) reversal of a DNA substring. In addition, for operation (4), once it is applied for one segment of the DNA sequence, the bases in the segment cannot be further transformed using any operation. Give an efficient algorithm that returns the minimum number of operations to transform S_1 to S_2. Give the time complexity.