Indian Statistical Institute

M.Tech. (CS), Second Year, Mid-Sem of First Semester Examination, 2024-25

Computational Molecular Biology and Bioinformatics

Full Marks: 30 Date: 13-09-2024 Time: 2 Hours

Answer any *three* of the following questions

$3 \times 10 = 30$

1. (a) Why do mutations in a DNA sequence not always have an effect in the translation (protein generation) process? Justify your answer.
   (b) Can a poly-A tail include CpG islands? Justify your answer.
   (c) Suppose a DNA sequence is given to you that appears between the location 101 and 1000 (both inclusive) in the +Ve strand of chromosome 1 in a mammal. The following additional details are provided to you.
      (i) The GC content is 60%.
      (ii) The occurrences of the neucleotides C and G are 20% and 30%, respectively.
      (iii) The occurrences of the 2-mer CG is 50%.

   Is the said DNA segment a CpG island?

   2+2+6

2. Derive the global alignment between the DNA sequences "**TATAG**" and "**TAGAT**" in a tabular form. Show the optimal alignment and find out the best global alignment score between them if the following scoring strategy is adopted.
      (i) Match: If a column has two identical characters, it will receive value +2.
      (ii) Mismatch: Different characters will give the column value −2 (a mismatch).
      (iii) Gap: A space in a column drops down its value to −3.

   10

3. (a) Given the protein sequence "**HNCNHH**", apply the Burrows-Wheeler transform to convert it to a sequence that is better compressible. Show the steps of performing the transformation.
   (b) What is the best data structure to be used for applying the inverse Burrows–Wheeler transform? Justify your answer.

   8+2

4. (a) Let there be a fixed but unknown protein sequence M (the motif) of length l mentioned to you. The problem is to determine M, given t sequences each of length n, and each containing a planted variant of M. The variants of interest are sequences that are at a maximum Hamming distance of d from M (i.e., they have at most d point-substitutions). Assuming that the background sequences are i.i.d. and no overlapping motifs does exist, what is the probability that a given l-mer is an (l, d)-motif?
   (b) What are the various approaches of performing significance analysis on network motifs found within a gene-gene interaction network with respect to background random networks?

   6+4
