Indian Statistical Institute

M.Tech. (CS), Second Year, Mid-Sem of First Semester Examination, 2025-26

Computational Molecular Biology and Bioinformatics

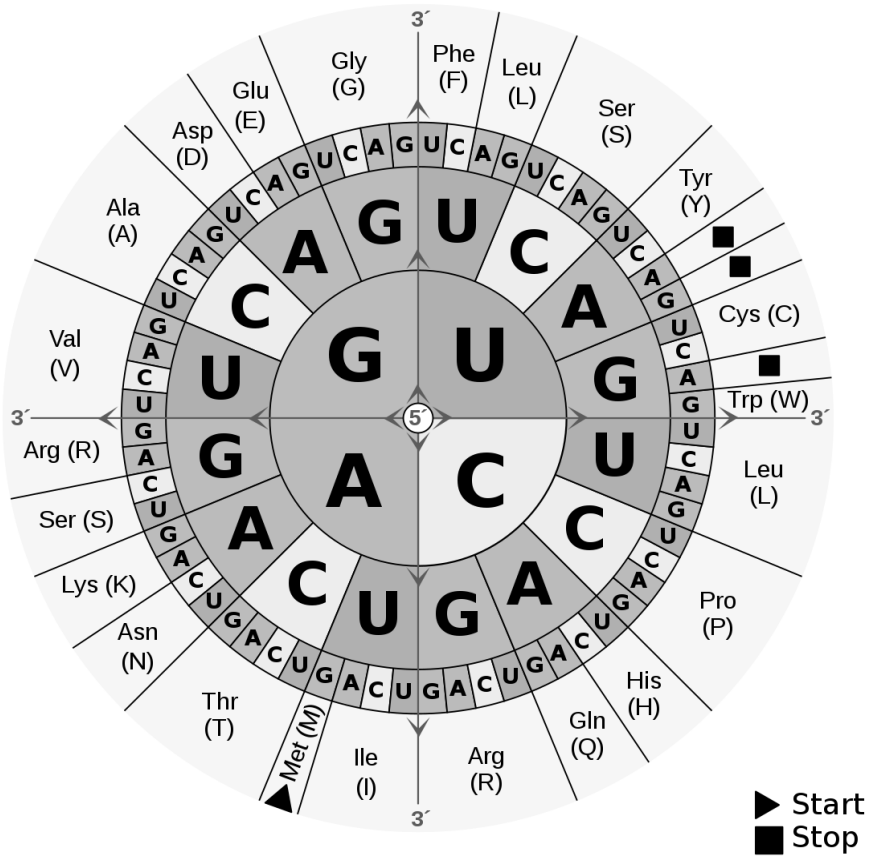
Full Marks: 30 Date: 09-09-2025 Time: 2 Hours

Answer any *three* of the following questions

3 × 10 = 30

1. (a) Let Σ = {A, C, G, T, N} denotes the DNA alphabet with 4 nucleotides as bases and an unknown base N. Suppose we generate a random DNA sequence of length 10 where each character is chosen uniformly at random from Σ. What is the expected number of distinct 3-mers that appear in the sequence?

   (b) Consider the codon table given below.



   Given the first 5 amino acids of the α-globin chain of human hemoglobin protein as **MVLSP**, what is the probability that its source DNA sequence is **AUG GUG UUG AGU CCG**?

   (c) Can the complementary strand of a Poly-A tail include CpG islands? Justify your answer.

4+4+2

2. Given the two DNA sequences **AATCG** and **ATG**, align them globally with gap penalty functions such that a maximal series of consecutive characters in one sequence can be aligned with spaces (gaps) in the other. Mention your scoring scheme based on which you perform the alignment and derive the best alignment score.

7+3

3. (a) Given the DNA sequence **TGCAAA**, apply the X-Mapper method to construct a pyramid of all possible x-mers. Derive the hashcode of each of the x-mers generated directly from the pyramid.
   (b) What is the utility of a lazy Needleman-Wunsch algorithm in the X-Mapper method?

(5+3)+2

4. (a) Cite an example to explain how the dependencies between datasets from different biological sources can be modeled using the concept of canonical correlation.
   (b) Cite an example to explain how a network flow problem can be formulated as a Mixed-integer Linear Programming (MILP) problem.

5+5
