Indian Statistical Institute

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

Computational Molecular Biology and Bioinformatics

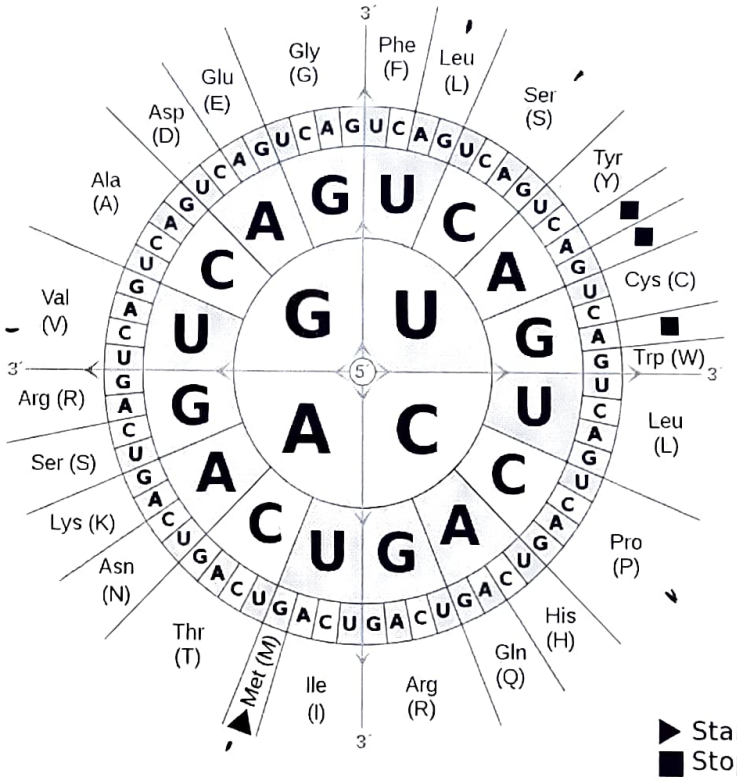
Full Marks: 30 Date: 09-09-2025 Time: 2 Hours

Answer any *three* of the following questions

$3 \times 10 = 30$

1. (a) Let $\Sigma = \{A, C, G, T, N\}$ 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 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 through 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

A A T C G
G

Indian Statistical Institute

M.Tech. (CS), First Semester Examination, 2025-26

Computational Molecular Biology and Bioinformatics

Full Marks: 50 Date: 18-11-2025 Time: 3 Hours

Answer any *five* of the following questions

$5 \times 10 = 50$

1. Consider a dataset $T = \{T_1, \ldots, T_m\}$ of paired healthy and diseased samples, where each element T_i is a triplet $\langle \mathbf{v}^h, \mathbf{U}, \mathbf{v}^d \rangle$ with normalized gene expression values of healthy cell line $\mathbf{v}^h \in [0, 1]^N$, disease-causing gene set $\mathbf{U}$, and gene expression values of diseased cell line $\mathbf{v}^d \in [0, 1]^N$, where N is the number of genes. Suppose that the goal is to find, for each sample $T_i = \langle \mathbf{v}^h, \mathbf{U}, \mathbf{v}^d \rangle$, the variable set $\mathbf{U}'$ with the highest likelihood of shifting gene states from diseased $\mathbf{v}^d$ to healthy $\mathbf{v}^h$ state. Formulate this as a representation learning problem. [10]

2. (a) How is Kolmogorov-Arnold representation theorem useful for function approximation in neural networks? How is this theorem useful in prediction problems in molecular biology?
   (b) How can a $(n, 2n + 1, 1)$-Kolmogorov-Arnold network be made deep? [(4+2)+4]

3. (a) What are the triple-effects that often occur in Cell Painting data?
   (b) How can the triple-effects be corrected by the cpDistiller model? [3+7]

4. (a) How can a cross-entropy loss function be modified to a more conservative form to control the influence on the penalty term in EvoGradient method?
   (b) Given an input peptide, how can an iterative gradient descent approach be applied to discover potential antimicrobial peptides? [4+6]

5. (a) How does Evo 2 train a DNA language model by separately prioritizing repeating (low priority) and non-repeating (high priority) regions of the sequence?
   (b) How are the rotary embeddings used by Evo 2 for context extension toward adapting to longer DNA sequences during the training phase? [4+(3+3)]

6. (a) How can a causal Knowledge Graph, trained on multiple evidences of the same fact, be used for performing probabilistic inference?
   (b) Consider a query peptide sequence MGYIN and a target peptide sequence MDPKI. How can you design an interaction language model based on biochemical properties of individual amino acids for predicting unknown protein-protein interactions? [5+5]

7. (a) Design an agentic AI model for predicting antimicrobial peptides.
   (b) State two limitations of each of the following types of sequence alignment methods.
   (i) Dynamic programming based. (ii) Hashcode based (iii) de Bruijn graph based. [4+(2+2+2)]
