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)]
