INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

End-Spring Semester 2023-24

Date of Examination: 23-04-2024 Session: AN Duration: 3 hrs Full Marks: 80
Subject No: CS61060 Subject: Computational Biophysics: Algorithms to Applications
Department/Center/School: Department of Computer Science and Engineering
Specific charts, graph paper, log book etc., required: None
Special Instructions (if any): *(1) Answer all the questions. (2) In case of reasonable doubt, make practical assumptions and write that on your answer script. (3) The parts of each question must answered be together. (4) Non-programming calculator is allowed.*

---

1. A research study aims to understand the gene expression patterns across different tissues. They collected data on the expression levels of two RNA molecules (RNA1 and RNA2) across four different tissues. The expression levels are measured in transcript per million (TPMs). The data for the four tissues are as follows:

   Tissue 1: RNA1 expression = 3000, RNA2 expression = 2000
   Tissue 2: RNA1 expression = 7000, RNA2 expression = 1500
   Tissue 3: RNA1 expression = 2000, RNA2 expression = 2500
   Tissue 4: RNA1 expression = 6000, RNA2 expression = 1000

   (a) Normalize both the RNA expressions and calculate the covariance matrix from the normalized expressions.

   (b) Determine the eigenvalues and corresponding eigenvectors of the covariance matrix.

   (c) Compute the first principal components for all four tissues using the obtained eigenvectors.

   (d) Project the first principal component along the direction of the expression levels of RNA2 and comment on the RNA2 expression pattern across the tissues.

   Marks: 3+4+3+2=12

---

2. (a) Write down the steps for a de-novo protein design using Metropolis Monte-Carlo (MMC) Simulation. Clearly mention the input and output of the protein design method. A de-novo protein design starts with a random sequence.

   (b) Extend your MMC simulation to a replica-exchange MMC scheme to avoid local trap.

   Marks: 12+8=20

3.

```
>PS1
MVLSPADKTNVK
>PS2
KKVADALTNAVA
>PS3
MVLSGEDKSNIK
>PS4
KKVADALASAAG
>PS5
HASLDKFLASVS
>PS6
MSLTRTERTIIL
>PS7
SKVVAAVGDAVK
```

(a) The multiple sequence alignment (MSA) of a list of protein sequences in the FASTA format is given above. Compute the Position-Specific Scoring Matrix (PSSM) for this MSA.

(b) Write down the pseudo-code for the computation of the value of PI using Monte Carlo simulation technique, where side of a square is of unit length (1 cm). Iterate your process till the accuracy reaches upto six decimal places (You may consider the theoretical value of PI=3.14285714).

Marks: 10+10=20

4.

Element	S1	S2	S3	S4
0	0	1	0	1
1	0	1	0	0
2	1	0	0	1
3	0	0	1	0
4	0	0	1	1
5	1	0	0	0

Given the above matrix with six rows answer the following questions.

(a) Compute the minhash signature for each column if we use the following three hash functions:
$h1(x) = 2x + 1 \bmod 6$
$h2(x) = 3x + 2 \bmod 6$
$h3(x) = 5x + 2 \bmod 6$

(b) Which of these hash functions are true permutations? How close are the estimated Jaccard similarities for the six pairs of columns to the true Jaccard similarities?

Marks: 4+(1+3)=8

5. (a) What is the complexity of the Protein folding problem?

   (b) What is the protein energy landscape? Elaborate with an example. In the energy landscape, mark the locally trapped region, and native state.

   (c) What are the basic differences between a template-based protein folding and ab-initio protein folding problem?

   (d) Define a protein threading problem. What are the major steps of protein threading problem? Elaborate each step.

Marks: 2+4+4+(2+8)=20