INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR
Mid-Spring Semester 2023-24

Date of Examination: __-02-2024 **Session:** FN/AN **Duration:** 2 hrs **Full Marks:** 60
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.*

1. (a) Assume one DNA sequence is of the length 1000 of which a stretch is represented as follows:

..... AAGCGAATAATATATTTATACTCAGATTATTGCGCG .....

You are asked to perform a Knuth–Morris–Pratt (KMP) algorithm based substring search method to check the existence of a motif (pattern) given as: TAATATATTTAT. Draw the state-transition diagram that will be generated in your preprocessing stage.

(a) What will be the time and space complexity for the preprocessing stage? What will be the searching time complexity?

Marks: 8+4=12

2. (a) Consider two DNA sequences, Sequence A and Sequence B, each represented as strings of length 1000 nucleotides. Implement a locality-sensitive hashing (LSH) scheme with a hash function that maps similar DNA sequences to the same bucket with a probability of 0.9. If the LSH scheme uses 20 hash functions, calculate the expected number of common buckets for Sequence A and Sequence B. Assume a uniform distribution of hash values.

(b) When a false positive case will occur in the case of LSH?

(c) Assume we have collected 5000 such sequences of DNA samples. Calculate the expected number of false positives. All parameters are the same as given in part a.

Marks: 4+1+3=8

3. (a) Define protein-protein docking problem.

(b) Write down the major steps of a hashing-based protein docking decoy generation method.

Marks: 2+8=10

4.

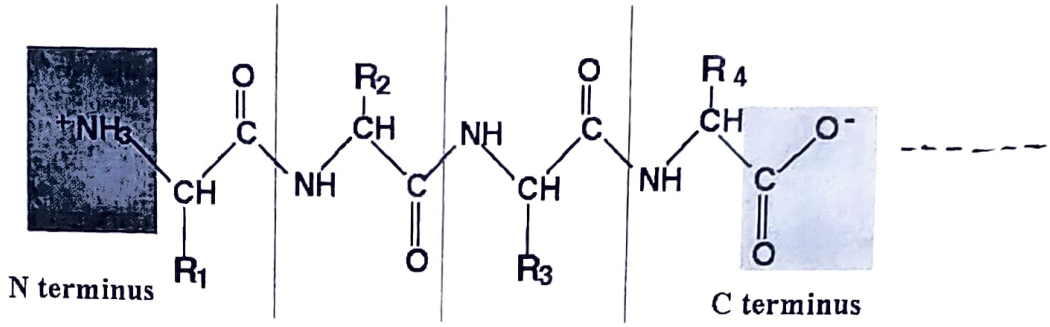

Figure 1: Sample polypeptide chain (assume it is extended on the right-hand side.). Phi and Psi angles are formed among the main chain atoms (…CA-C-N-CA-C-N-CA…., where CA indicates alpha carbon, CH in the figure)

```
ATOM      1  N   GLU A  20      12.985 -50.707  16.620  1.00 50.43      N
ATOM      2  CA  GLU A  20      13.697 -50.232  15.443  1.00 45.68      C
ATOM      3  C   GLU A  20      14.889 -49.408  15.879  1.00 45.56      C
ATOM      4  O   GLU A  20      14.785 -48.578  16.782  1.00 41.87      O
......
ATOM     10  N   VAL A  21      16.015 -49.627  15.200  1.00 48.57      N
ATOM     11  CA  VAL A  21      17.228 -48.856  15.421  1.00 42.52      C
ATOM     12  C   VAL A  21      17.297 -47.819  14.327  1.00 41.73      C
ATOM     13  O   VAL A  21      17.096 -48.127  13.146  1.00 47.15      O
.....
```

Figure 2: PDB format of a protein structure file. Main chain data of only two amino acids are presented.

(a) Compute the number of main-chain hydrogen bonds for a given protein (hydrogen bond formed between NH of ith amino acid and O of jth amino acid, i≠j) (Figure 1). The input to your program is Protein Data Bank (PDB) format file as downloaded from RCSB website (Figure 2). Hence, include the PDB parsing process in your algorithmic steps. Assume that the hydrogen atom is not present in the PDB file.

(b) Next write down a separate algorithm that will print the phi and psi angle for each of the amino acid from the Protein Data Bank (PDB) format file as downloaded from RCSB website (Figure 1 and 2). You may use the same PDB parsing process as you have done in the previous question (by mentioning the step number).

NOTE: Use vector algebra or Euclidean geometry for the above questions. Make the necessary diagram and use the three dimensional coordinate information in your algorithm. No marking (not even partial) for only mentioning "used dot product", "used cross product", "computed distance/angle", etc.

Marks: 8+7=15

5. (a) Apply dynamic programming algorithm to align following two sequences. Draw the scoring matrix, and the trace-back matrix for the optimum alignment. Report your alignment score, and show the alignment. Given, match/mismatch score as per BLOSUM62 matrix (Figure 3, provided at the end), gap opening penalty (including one gap extension) is -10 and gap extension penalty is -1.

Input:

```
>Design Protein Sequence| C terminus:
AGVWGAPVRESVPSL

>11AS_1|Chains A|C terminus:
VQAGVWPAAVRESVPSLL
```

Marks: 6+3+2+4=15

Figure 3: BLOSUM62 Matrix

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9	-1	-1	-3	0	-3	-3	-3	-4	-3	-3	-3	-3	-1	-1	-1	-1	-2	-2	-2
S	-1	4	1	-1	1	0	1	0	0	0	-1	-1	0	-1	-2	-2	-2	-2	-2	-3
T	-1	1	4	1	-1	1	0	1	0	0	0	-1	0	-1	-2	-2	-2	-2	-2	-3
P	-3	-1	1	7	-1	-2	-1	-1	-1	-1	-2	-2	-1	-2	-3	-3	-2	-4	-3	-4
A	0	1	-1	-1	4	0	-1	-2	-1	-1	-2	-1	-1	-1	-1	-1	-2	-2	-2	-3
G	-3	0	1	-2	0	6	-2	-1	-2	-2	-2	-2	-2	-3	-4	-4	0	-3	-3	-2
N	-3	1	0	-2	-2	0	6	1	0	0	-1	0	0	-2	-3	-3	-3	-3	-2	-4
D	-3	0	1	-1	-2	-1	1	6	2	0	-1	-2	-1	-3	-3	-4	-3	-3	-3	-4
E	-4	0	0	-1	-1	-2	0	2	5	2	0	0	1	-2	-3	-3	-3	-3	-2	-3
Q	-3	0	0	-1	-1	-2	0	0	2	5	0	1	1	0	-3	-2	-2	-3	-1	-2
H	-3	-1	0	-2	-2	-2	1	1	0	0	8	0	-1	-2	-3	-3	-2	-1	2	-2
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5	2	-1	-3	-2	-3	-3	-2	-3
K	-3	0	0	-1	-1	-2	0	-1	1	1	-1	2	5	-1	-3	-2	-3	-3	-2	-3
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5	1	2	-2	0	-1	-1
I	-1	-2	-2	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4	2	1	0	-1	-3
L	-1	-2	-2	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4	3	0	-1	-2
V	-1	-2	-2	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4	-1	-1	-3
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6	3	1
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	2
W	-2	-3	-3	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11