BE7-R3: APPLIED BIO-INFORMATICS

NOTE:

> 1. **Answer question 1 and any FOUR questions from 2 to 7.**
> 2. **Parts of the same question should be answered together and in the same sequence.**

Time: 3 Hours **Total Marks: 100**

1.

a) Define open reading frame. Write three forward frames for the following sequence
TATACGTAGTATTCGAATGGG

b) How is a typical prokaryotic gene structure different from a eukaryote?

c) List at least four distinct substrings of "AATAG".

d) Compare and contrast the two scoring matrices PAM and BLOSUM.

e) What types of errors do occur in DNA fragment assembly problems?

f) Why are profile models better than consensus model?

g) Write four applications of Hidden Markov model framework in Bioinformatics.

(7x4)

2.

a) Using dynamic programming method for a pairwise alignment problem (global and local), write the initialization conditions and recurrence relations.

b) Given a score of 1 for a match, 0 for a mismatch and a linear gap cost 0.5, use the global alignment algorithm to score the following scoring matrix.

0	---	c	g	c	a	t	G
---							
a							
c							
g							
a							
g							

c) Why is dynamic programming based optimal alignment not suitable for multiple sequence alignment? Briefly mention any one method for the multiple sequence alignment.

(4+6+8)

3.

a) Mention four versions of Blast programs. Explain how Blast algorithm finds similar sequences from a database and how is the alignment quality evaluated?

b) Briefly write about the structure of a typical "GenBank" record.

c) What are low complexity sequences and why should they be masked?

(10+4+4)

4.

a) EcoRI is a restriction Enzyme that cuts DNA wherever the sequence GAATTC is found. Cuts are made between the G and the first A. If so, consider the sequence ATCCATTGAATTCTCGGACC and write down the resulting fragment cut by EcoRI.

b) When some binding site sequence from four different species were alignment was obtained. Write the consensus sequence which can represent as a signature for the binding sites.

G T A G A C

G	T	A	G	A	C
G	T	G	G	A	T
G	T	G	C	A	A
G	T	A	G	T	C

c) Briefly state the steps taken to sequence a whole genome sequence? Also mention how these one dimensional symbolic sequences are annotated in terms of biological functions?

(4+4+10)

5.

a) Define first order Markov chain? Why are higher order models not generally used in Bioinformatics?

b) Let the state symbols for the positive model be given as A+, T+, G+ and C+ and similarly for the negative model be given as A-, T-, G- and C-. In the hidden Markov model transition can take place in two ways i.e. within and between the states. Draw the state-transition diagram separately for these transitions.

c) How is hidden Markov model framework applied to multiple sequence alignment?

(4+6+8)

6.

a) Compute the log odds ratio using the transition probability matrix given below for “+” model and “-“ model:

+	A	C	G	T
A	0.180	0.274	0.426	0.120
C	0.171	0.368	0.274	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182

-	A	C	G	T
A	0.300	0.205	0.285	0.210
C	0.322	0.298	0.078	0.302
G	0.248	0.246	0.298	0.208
T	0.177	0.239	0.292	0.292

b) Distinguish between optimal and heuristic methods.

c) Explain briefly “sum of pairs (SP)” measure in the context of multiple sequence alignment.

(8+4+6)

7.

a) Discuss algorithmic complexity in a typical pairwise alignment problem.

b) How do PSI-BLAST and PHJ-BLAST algorithms fetch similar sequence from the database?

c) Consider the two amino acid sequences

1) CAEFDDH

2) CDAEFPDDH

Suppose their respective paths through a protein model HMM of length 10 are

$m_0m_1m_2m_3m_4d_5d_6m_7m_8m_9m_{10}$ and

$m_0m_1i_1m_2m_3m_4d_5m_6m_7m_8m_9m_{10}$,

respectively. Find the alignment induced by the above path.

(4+6+8)