BE7-R3: APPLIED BIOINFORMATICS

NOTE:

1.	**Answer question 1 and any FOUR from questions 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 Gene expression and Gene regulation.

b) How can you find all open reading frames in a given sequence?

c) Suppose you wish to train a predictor using machine learning techniques to determine whether a given residue will be on the inside or the outside of a protein when folded. What data would we need to gather in order to train the predictor? Give an example of different types of problem which might occur with your data.

d) Consider partial digest L = {2, 2, 3, 3, 4, 5, 6, 7, 8, 10}. Solve the Partial Digest problem for L (i.e., find X such that ΔX = L).

e) How will you search motifs using the PSI-BLAST approach?

f) Explain the reasons why protein searches are faster than DNA searches.

g) What is the Hamming distance between these two strings?

i) BIOINFORMATICS_IS_THE

ii) BEST_FOR_STRUCTURE_PREDICTION

(7x4)

2.

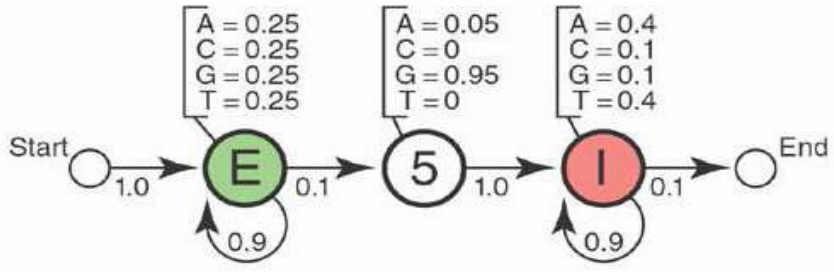
a) Explain the biological motivation of sequence analysis.

b) Describe Karlin-Altschul statistics for local sequence alignment.

c) Discuss PAM approach in brief. What are the problems related to PAM?

(6+6+6)

3.



Given the above state diagram of the Markov Chain, answer the following questions about the sequence ACGGT.

a) What is the probability of the sequence ACGGT if we assume the sequence of states to be EE5II?

b) What is the probability of the sequence ACGGT?

c) What is the confidence that the penultimate letter (ACG**G**T) is the splice site?

(4+8+6)

4.

a) Discuss in detail the dynamic programming approach for global alignment. How can this algorithm be modified for local alignment?

b) Draw a dot-plot to produce a likely global alignment for these two sequences: DILVDEQ and IVQDEQ. Show on the above dot-plot how you produced this alignment.

(10+8)

5. The following Markov chain model governs the intracellular (I) and extracellular (E) placement of amino acid residues in trans membrane proteins.

Transmission probabilities:

	E	I
E	0.7	0.3
I	0.2	0.8

Emission Probabilities:

E		I	
L	0.00	L	0.45
V	0.16	V	0.25
D	0.50	D	0.25
K	0.34	K	0.05

Assume that the probability of a sequence starting with an intracellular (I) placement of amino acid residues is about 40%.

a) Construct (Draw) the HMM for this scenario showing all the possible states, all transition and emission probabilities. Just consider Leucine (L), Valine (V), Aspartate (D) and Lysine (K) in your protein sequence.

b) Determine the sequence of states most likely to generate the amino acid sequence KLD from this model.

c) Explain the fundamental difference between Viterbi and Forward-Backward Algorithm in terms what probability they calculate.

(6+6+6)

6.

a) What is DNA sequencing? Briefly describe the methodologies of clone based and shotgun sequencing.

b) Briefly describe biological applications of multiple sequence alignment.

(10+8)

7.

a) What is polymerase chain reaction? Write down the procedure and purpose of PCR.

b) What are the major differences between BLAST and FASTA algorithms?

c) Briefly describe the problem related with PSI-BLAST for homology searching.

(3x6)