BE7-R3: APPLIED BIOINFORMATICS

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 the terms “Gene” and “Genome”.

b) Why is eukaryotic gene finding more difficult than prokaryotic gene finding?

c) Write atleast four distinct substrings of “TTAG”.

d) Why is blastp preferred over blastn in homology searches?

e) For any probabilistic model of sequences we can write the probability of the sequence as

$P(x) = P(x_L\ x_{L-1}\ \ldots,\ x_1)$

$= P(x_L|x_{L-1},\ \ldots,\ x_1)\ P(x_{L-1}|x_{L-2},\ \ldots,\ x_1)P(x_1)$

Using the key property of markov chain, write down the expression of P(x).

f) Alignments are models that reflect different biological perspectives. There are two approaches which consider similarity. Name the algorithms and why do these exploit dynamic programming?

g) Define the term “overlap” with an example.

(7x4)

2.

a) Restriction enzymes precisely recognize its site and cut the DNA into two fragments. One such enzyme BamHI cuts the DNA wherever the sequence GGATCC is found. Cuts are made between the first G and the second G.
Consider the DNA sequence
TAATTGGATCAACCGTACC
Write down the resulting fragment cut by the above mentioned enzyme.

b) Describe briefly any sequencing strategy or technique employed to decipher the genome.

c) How does the availability of database serve the purpose of prediction strategies in typical bioinformatics tasks?

(4+8+6)

3.

a) Describe the “STAR” alignment strategy for the multiple sequence alignment problem and describe how the resulting alignments are evaluated by Sum of Pairs method.

b) Find the optimal alignment for the DNA sequences AAAGTC and AGATTC, scoring +2 for a match, -1 for a mismatch and with a linear gap penalty of d=2.

c) Distinguish between a “profile” and a “motif”.

(8+6+4)

4.

a) Give reasons for employing “Heuristic” methods over “Optimal” methods for large sale database homology searches.

b) Briefly describe how does BLAST algorithm work in homology searches and how does one infer about the significance of the database hits?

c) What are the advantages of using “PSI-BLAST”?

(4+8+6)

5.

a) Write a recursive algorithm to sort n elements.

b) Give a pseudo code that accepts as input a DNA sequence of any size and returns its reverse complement.

c) How genome sequencing is modelled as a fragment assembly problem and explain the steps involved in it.

(5+5+8)

6.

a) Distinguish between Markov Chain and Hidden Markov Models.

b) What do you understand by memory less property in the context of Markov process? How will you apply this concept for problems dealing symbolic nucleotide sequence?

c) 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 models. Draw the state transition diagram separately for these transitions.

(6+4+8)

7.

a) In the context of bioinformatics applications, discuss the algorithms employed in hidden markov model framework.

b) Why are Markov Models successful in prediction problems? Mention any four applications where Hidden Markov Model (HMM) is found to be useful.

(10+8)