B.TECH/BT/5TH SEM/BIOT 3102/2018
BIOINFORMATICS
(BIOT 3102)

Time Allotted : 3 hrs **Full Marks : 70**

Figures out of the right margin indicate full marks.

Candidates are required to answer Group A and any 5 (five) from Group B to E, taking at least one from each group.

Candidates are required to give answer in their own words as far as practicable.

Group – A
(Multiple Choice Type Questions)

1. Choose the correct alternative for the following: **10 × 1 = 10**

(i) Which of the following bioinformatics tools specialize in DNA analysis?
(a) PORTER (b) PHYRE2 (c) SWISS MODEL (d) BPROM.

(ii) In a Hidden Markov Model (HMM) the probability of going from one state to another is known as
(a) emission probability
(b) transition probability
(c) true positive probability
(d) false negative probability .

(iii) Which of the following IS AN INCORRECT choice for description of PAM matrices?
(a) This family of matrices lists the likelihood of change from one amino acid to another in homologous protein sequences during evolution
(b) There is presently no other type of scoring matrix that is based on such sound evolutionary principles as are these matrices
(c) Even though they were originally based on a relatively small data set, the PAM matrices remain a useful tool for sequence alignment
(d) It stands for Percent Altered Mutation.

(iv) In regular expressions, which of the following pair of pattern is wrongly matched with its significance?
(a) [ ] – Or (b) { } – Not (c) ( ) – Repeats (d) Z – Any.

(v) Lead compound improvement involves which of the following steps
(a) Interactive computational modelling (b) combinatorial chemistry
(c) experimental library screening (d) all of the above.

(vi) In sequence alignment by BLAST, each word from query sequence is typically _____ residues for protein sequences and _____ residues for DNA sequences.
(a) ten, eleven (b) three, three (c) three, eleven (d) three, ten.

(vii) The UNIQUE mathematical step in the Metropolis Monte Carlo procedure for protein structure calculations involves which of the following choices?
(a) generation of a random set of values of x to provide a starting conformation
(b) perturbation of the variable from x to x,"
(c) calculation of the energy of the new conformation at x" if energy is decreased
(d) calculation of the energy of the new conformation at x" if the energy is increased.

(viii) In which of the following computations have neural networks been used?
(a) prediction of secondary structures of proteins
(b) interpretation of NMR structures of proteins
(c) generation of quantitative structure-activity relationships
(d) none of the above.

(ix) Which of the following DOES NOT describe local alignment?
(a) A local alignment aligns a substring of the query sequence to a substring of the target sequence
(b) A local alignment is defined by maximizing the alignment score, so that deleting a column from either end would reduce the score, and adding further columns at either end would also reduce the score
(c) Local alignments have terminal gaps
(d) The substrings to be examined may be all of one or both sequences; if all of both are included then the local alignment is also global.

(x) Which of the following classes of proteins are NOT specifically included in a MARCOIL database for structure prediction?
(a) Dyneins (b)tropomyosins (c) SNARE proteins (d) proteases.

Group – B

2. (a) How have bioinformatics methods played a role in the development of forensic DNA analysis? What areas of forensic DNA typing do methods development in bioinformatics address? Why is massively parallel sequencing *such an important* area of bioinformatics methods development?

(b) What makes protein structural bioinformatics an important part of the subject of bioinformatics? Cite two ways by which bioinformatics tools have been applied to agricultural research and development.

(c) How has *functional bioinformatics* been applied to mechanistically understand the varied medical applications of aspirin? Explain your answer pointwise.

(2 + 1 + 1) +(2 + 2) + 4 = 12

3. (a) What is data mining as it pertains to bioinformatics? Define machine learning in the context of such bioinformatics based data mining. What are the two complementary aspects of machine learning? Explain your answer with specific examples.

(b) Briefly describe three categories of numerical methods applied to bioinformatics data analysis. Distinguish between *supervised* and *unsupervised* learning in bioinformatics analysis.

(c) Using an example, define a *markup* language in the context of positional formatting and biological databases. Suggest two methods of reducing redundancy in a biological database.

(1+2+1) + (2+2) + (2+2) = 12

Group – C

4. (a) Discuss the relationship among homology , orthology and paralogy with suitable example.

(b) "The scoring systems in a dynamic programming alignment algorithm are referred to as substitution matrices." Why isthis and how is a substitution matrix derived? Briefly enumerate the calculation steps for amino acid substitution matrices.

(c) Explain the basis of your choice for BLOSUM45 and BLOSUM80 in BLAST.

3 + (2 +2 + 2) + 3 = 12

5. (a) Five amino acid sequence stretches of enolase were obtained from five different organisms viz. rice, wheat, rat, yeast and frog. What is the unique bioinformatics advantage if you were to align the above sequences? Enumerate five specific applications of multiple sequence alignment.

(b) Write out the steps needed to establish the multiple sequence alignment amongst the above sequences using a progressive alignment method (e.g like CLUSTAL W or ω). Why is the method referred to as a heuristic one?

(c) Answer the following questions with the help of your answers to (b) above

(i) briefly discuss how the final alignment is obtained.

(ii) what are the important features of this method?

(iii) disadvantages of the progressive alignment method compared to the iterative alignment method.

(1 + 2) + (4 + 1) +(1 + 1 + 2) = 12

Group – D

6. (a) Define and briefly describe the term PERL. Explain the role of meta symbols in PERL programming with suitable examples

(b) A pair of sequences of hexokinase expressed in *Staphylococcus aureus* and *Oryzasativa* is stored in two files in v.pep and p.pep. Write a PERL program in proper syntax where percentage of identity (including matches) calculation between themselves can be successfully done.

(c) Explain the function of the split and join operators in PERL using suitable examples.

(1 + 3) + 4 + (2 +2) = 12

7. (a) Compose a PERL program where the user is provided with the following constraints:(i) a nucleotide sequence is to be used as an input (ii)the sequence is stored as a suitable variable(iii) a check is performed whether the input has the standard reading frame within that stretch (iv) if standard reading frame presence is positive, THEN ONLY check for the presence/absence of a stop codon(v) last step is to OTHERWISE query for a FRESH input.

(b) Write a PERL program where a nucleotide sequence stretch by mistake is stored in a scalar variable but need to find the length of the stretch. Along with that find out its complimentary sequence

6 + (3 + 3) = 12

Group – E

8. (a) What are the steps involved in a protein tertiary structure prediction using homology modelling? Schematically represent the loop modelling step in the above procedure. What technical reasons make loop modelling a computationally difficult problem?

(b) What parametric range does a database protein have, to make it suitable as a template? How can the template selection process be improved?

(c) Using SWISS-MODEL as an example, explain how quality estimation is done for a modelled protein? What is the importance of the ligand modelling step in SWISS-MODEL?

(2+ 2 + 1) + (1 + 2) + (2 + 2) = 12

9. (a) An example of a QSAR equation to relate a possible lead molecule's biological activity to its electronic characteristics and hydrophobicity is given by

$\text{Log}\ (1/C) = k_1 \text{logP} - k_2\ (\text{logP})^2 + k_3\sigma + k4$

(i) Define and explain the different terms in the above equation.

(ii) Considerable effort has been expended to develop computer programs that can estimate log P values entirely on the basis of chemical structure. How would such a program be useful?

(b) Why are knowledge based potentials commonly used as threading scoring functions? Give an example of such a scoring function.

(c) What is a deductive method for prediction of the binding site on a target protein? What are the two essential criteria for a deductive method to work? Explain your answers.

(3 + 3) + 3 +(1.5 ×2=3) = 12