BIOINFORMATICS
(BIOT 3102)

Time Allotted : 2½ hrs **Full Marks : 60**

Figures out of the right margin indicate full marks.

Candidates are required to answer Group A and any 4 (four) from Group B to E, taking one from each group.

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

Group – A

1. Answer any twelve: **12 × 1 = 12**

Choose the correct alternative for the following

(i) A Hidden Markov model satisfies which of the following criteria?
(a) It is a statistical model composed of a number of interconnected Markov chains
(b) It is a statistical model composed of a number of un-connected Markov chains
(c) It is a statistical method that separates true signals from the background
(d) None of the above.

(ii) Which of the following is not correct about coils and loops?
(a) They are irregular structures
(b) The connecting regions are completely irregular
(c) They are regular structures
(d) The loops are often characterized by sharp turns or hairpin-like structures.

(iii) Alignment method suitable for aligning closely related sequence is
(a) Local alignment (b) Multiple sequence alignment
(c) Global alignment (d) Pairwise sequence alignment.

(iv) Sequence alignment helps scientists to
(a) Trace out evolutionary relationships
(b) Infer the functions to newly synthesized genes
(c) Predict new members of gene families
(d) All the above.

(v) Which of the following is untrue regarding the gap penalty used in dynamic programming?
(a) Gap penalty is subtracted for each gap that has been introduced
(b) Gap penalty is added for each gap that has been introduced
(c) The gap score defines a penalty given to alignment when we have insertion or deletion
(d) Gap open and gap extension has been introduced when there are continuous gaps (five or more).

(vi) Which of the following does not describe local alignment algorithm?
(a) Score can be negative
(b) Negative score is set to 0
(c) First row and first column are set to 0 in initialization step
(d) In traceback step, beginning is with the highest score, it ends when 0 is encountered.

(vii) PRINTS is a secondary resource that provides a bridge between
(a) multiple and pairwise sequence search methods
(b) different profile based search methods
(c) single motif search methods and domain alignment search methods
(d) partial and full PROSITE sequence search.

(x) Which statement is used to enable strict mode in Perl?
(a) Strict mode (b) Use strict
(c) Use strict mode (d) None of (a), (b) & (c).

(ix) Which one of the following steps represents the fourth step of the pairwise energy approach of the threading algorithm?
(a) Query sequence selection (b) Fold library selection
(c) Scoring and ranking (d) Energy calculations.

(x) A database has which of the following characteristics?
(a) It is an archive of information
(b) It has a logical organization/structure
(c) It has tools to access the information
(d) All of the above.

Fill in the blanks with the correct word

(xi) Significant matches in alignment are indicated by a p value of _____.

(xii) _____ is an example for local sequence alignment algorithm.

(xiii) The value of the correlation coefficient CC for gene finding varies between _____ and _____.

(xiv) _____ is an example for a global sequence alignment algorithm.

(xv) In sequence alignment by BLAST, each word from query sequence is typically _____ residues for protein sequences and _____ residues for DNA sequences.

Group - B

2. (a) What are the two major defining characteristics of a biological database? Give examples of a relational database based on amino acid properties. What are the characteristics of two protein classification databases that present hierarchies of protein structures? Name those two databases. *[(CO1)(Analyse/IOCQ)]*

(b) What is implied by annotation of a database? Name three annotation characteristics of a biological database with specific examples in each characteristic. *[(CO1)(remember/LOCQ)]*

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

3. (a) Define machine learning in the context of bioinformatics data analysis. What are the two complementary aspects of machine learning? Cite with detail three types of numerical methods that are commonly applied to data analysis.
[(CO1)(Analyse/IOCQ)]

(b) Using two examples from biological databases each, explain the importance of annotation and quality control in biological databases. *[(CO4)(Remember/LOCQ)]*

(1 + 2 + 3) + (3 × 2) = 12

Group - C

4. (a) Itemize the distinguishing features of ab initio based approaches of gene prediction. *[(CO3)(Justify/IOCQ)]*

(b) For prokaryotic gene prediction, instead of identification of the initiation codon, other features associated with translation initiation may be used-explain how such an approach is helpful for accurate gene prediction Pointwise explain the use of the TESTCODE method for this purpose. *[(CO3)(Justify/IOCQ)]*

(c) Briefly and graphically represent the three states in a Hidden Markov Model (HMM) using a labeled diagram. *[(CO3)(memorize/IOCQ)]*

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

5. (a) Suppose the following sequences are of divergent groups. Justify the algorithm will be chosen for sequence alignment in this case- explain it. Find out the optimal alignment and score in the following pair of sequences Show the quantitative evaluation of the whole process of sequence alignment in a step wise manner following that algorithm. The sequences are as follows:
Seq 1. AGGCCTTGAATTCAGCGTAT
Seq2. AGCCTGAATAGGTT
[Given: opening gap penalty = (-10), End gap penalty= (-4), Gap extension penalty = (-1), Similarity = (+1), Identity = (+2), Mismatch= (-0.5)]
[(CO2)(Evaluate/HOCQ)]

(b) Make a comparative analysis between the following scoring matrices: PAM and BLOSUM. *[(CO2)(Analyse/IOCQ)]*

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

Group - D

6. (a) Describe the types of variables used in PERL programming with the help of suitable examples. *[(CO4)(Describe/LOCQ)]*

(b) Write a PERL program where user is asked to give a nucleotide sequence as an input then store it in a suitable file. Then check whether the input is having the standard reading frame within that stretch if it is present then only check if the stop codon is present there or not otherwise ask to give a fresh input.
[(CO4)(Design/HOCQ)]

6 + 6 = 12

7. (a) A DNA sequence is stored in a specific file. Write a PERL program to generate 10 successive mutation of the sequence. *[(CO3)(Analyse/HOCQ)]*

(b) Write a program to calculate the reverse complement of DNA. *[(CO4)(Describe/LOCQ)]*

8 + 4 = 12

Group - E

8. (a) Why are protein secondary structure prediction algorithms considered more accurate when combined with multiple sequence alignment (MSA) and neural networks (NN). Limit your answer to two specific points. How much is the accuracy percentage improved? Use a flowchart explaining the steps to represent the operation of a neural network that is applied only to the secondary structure of a protein. *[(CO5)(understand-Analyse/IOCQ)]*

(b) What are the common sources of errors that are prevalent in secondary structure prediction of proteins? Itemize them quantitatively in terms of occurrence. Name the analytical methods you would use to develop the robustness of secondary structure prediction methods.

[(CO5)(Remember-understand/LOCQ)]

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

9. (a) How can methods of bioinformatics be used to help select protein targets for experimental structure determination? List five such goals for target selection. How does one of these goals relate to computationally measuring the binding of a protein target P to a druggable ligand L? *[(CO6)(Understand-explain/IOCQ)]*

(a) What are the requirements for both structure based homology modelling and threading? "In doing successful fold recognition, there is necessity for calibrating the scores for the various models". Itemize in detail the methods in which such calibration of scores is obtained. *[(CO4)(Remember/HOCQ)]*

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

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	20.83	50	29.17

Course Outcome (CO):

After the completion of the course students will be able to

1. Gain and analyze knowledge about genes and proteins obtained through primary, secondary and specialized databases (e.g. NCBI, PDB).
2. Learn and apply principles and methodologies of pairwise and multiple sequence alignment towards biological problems (e.g. Smith Waterman, Needleman and Wunsch, CLUSTAL algorithm).
3. Learn and apply principles of gene prediction algorithms with respect to prokaryotic gene systems (e.g. Hidden Markov Model based gene annotation).
4. Learn and apply PERL for bioinformatics data interpretation (e.g. sequence analysis, protein to DNA translation).
5. Learn and apply principles and algorithms for secondary and tertiary structure prediction of globular and fibrous proteins (e.g. homology modeling, fold recognition methodologies).
6. Use introductory applications of bioinformatics procedures and protein structure prediction techniques to molecular modeling, molecular docking and virtual screening using representative examples.

**LOCQ: Lower Order Cognitive Question; IOCQ: Intermediate Order Cognitive Question; HOCQ: Higher Order Cognitive Question.*