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 DOES NOT describe the outcome of a local alignment algorithm?
(a) Score can be negative
(b) Negative score is set to zero
(c) First row and first column are set to zero in initialization step
(d) In the traceback step, the beginning is with the highest score and ends when zero is encountered.

(ii) In order to explore equilibrium protein conformations, protein structure calculations need which of the following steps?
(a) Monte Carlo minimizations to find the most stable protein conformation
(b) An experimentally determined protein structure
(c) A Taylor series based dynamical protein conformation
(d) Use of a three dimensional profile to assess the quality of the structure.

(iii) A Support vector machine (SVM) is an algorithm that
(a) is a data structure and falls under supervised learning
(b) is an example of unsupervised machine learning
(c) has a lower accuracy than an artificial neural network
(d) has a dimensionality higher than 2.

(iv) Which of the following does not describe PAM matrices?
(a) These matrices are used in optimal alignment scoring
(b) It stands for Point Altered Mutations
(c) It stands for Point Accepted Mutations
(d) It was first developed by Margaret Dayhoff.

(v) 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.

(vi) An example of the homology and similarity tool is
(a) PROSPECT (b) EMBOSS (c) RASMOL (d) BLAST.

(vii) Which of the following is not a variant of BLAST?
(a) BLASTN (b) BLASTP (c) BLASTX (d) TBLASTNX.

(viii) Which of the following does not describe BLOSUM matrices?
(a) It stands for BLOCKs Substitution Matrix
(b) It was developed by Henikoff and Henikoff
(c) The year it was developed was 1992
(d) These matrices are logarithmic identity values.

(ix) 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).

(x) Among the following which one is not the approach to the local alignment?
(a) Smith-Waterman algorithm (b) K-tuple method
(c) Words method (d) Needleman-Wunsch algorithm.

Group- B

2. (a) "The nucleotide sequence databases contain more than $6X10^{11}$bases (>600GBp); the database of macromolecular structures (proteins) contains over 100,000 entries of typical length 400 residues". Cite two scientific-technical observations from the preceding statement about the nature of bioinformatics data that *has contributed to increasing sophistication of existing biological databases and development of new ones*. In order to make a database effective, modes of access must be provided. Cite four examples of KEY *database queries in bioinformatics.* [(CO1)(Understand-analyze/IOCQ)]

(b) What is the relevance of computed structure models (CSM) in PDB in methodological terms? Cite and define the "confidence" score parameter that is calculated for each computed structure model (CSM) in PDB. Use one example to highlight the utility of this parameter. Briefly cite three important anticipated applications of such computed structure models. [(CO1and CO5)(Remember-understand/IOCQ]

(c) For efficiency in archival and retrieval of biological information, the internal structure of a database must reflect the inter-relationships of its contents. Illustrate/explain the TWO types of database organization/structure in common use in biological databases with one example in each. [(CO1)(Reasoning-Analyze/IOCQ)]

4 + 4 + 4 = 12

3. (a) What are the categories of information that typically are included in a well annotated biological database? " Errors in biological databases are especially problematic because they tend to propagate through links into other databases". Elucidate on this statement by answering the following two questions <u>*briefly and specifically*</u> : (i) what are the two general approaches that can be adopted to improve database quality? and (ii) use an example/procedure from *either* UniPROT *or* PDB to highlight external/internal methods of improving database quality control. [(CO1)(Analyze-IOCQ)]

(b) What are the two main scientific reasons connected to increased productivity that has required the use of bioinformatics tools in agriculture? Elucidate your answer with two examples of bioinformatics tools that are used for this purpose.

[(CO1)(remember-understand/LOCQ)]

(2 + 5) + 5 = 12

Group - C

4. (a) There are two proteins with weak and remote similarity. Their sequences are known. In addition, the sequences of the original DNA that these proteins are derived from is also known. Why is a better alignment expected using the protein sequences? Explain your answer. [(CO2)(Argue/HOCQ)]

(b) In performing alignment of protein sequences, the parameters considered are match/mismatches as also the similarities between amino acids. How are these parameters represented in practice? Protein alignments also make use of various substitution matrices. Justify the conditions under which one type of substitution matrix is preferred over the other. [(CO2)(Evaluate/HOCQ)]

(c) What is the basic problem solving mechanism that is incorporated in the BLAST algorithm? Explain the steps of Iterated BLAST. Itemize the advantage(s) of an iterated BLAST search compared to a simple BLAST search. [(CO2)(Explain/LOCQ)]

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

5. (a) A program has been written for generation of random DNA sequences. It is expected that there is a 25% sequence identity between pairs of random sequences in gapped alignments. However when the alignment related calculation is done, more than 25% sequence identity is obtained. Explain the difference in the identity value.

[(CO2)(Analysis/HOCQ)]

(b) Out of the two protein sequences of catalase one sequence is of *Vibrio cholerae* and another of *Drosophilla melanogastor*. Mention the approaches to be followed that reveal similarity among themselves. [(CO4)(Understand/LOCQ)]

(c) Mention the role of gap in sequence alignment. [(CO2)(Understand/LOCQ)]

(d) The default values of opening gap penalty is much higher and extention gap penalty – state why this type of discrimination is found. [(CO2)(Analyze/IOCQ)]

(e) In another situation if you are given one sequence what technique will you be using to find out its relationship with others? [(CO2)(Consider/HOCQ)]

3 + 2 + 2 + 2 + 3 = 12

Group - D

6. (a) Assume a situation where nucleotide sequence is stored in a file as x.pl you are given an information that in order to start the transcription you need to find out whether the start codon is located there or not. Write a perl script for this problem.

[(CO4)(Consider/HOCQ)]

(b) Write a program where a nucleotide sequence stretch is stored in the array variable, find the length of the sequence find out its complimentary sequence.

[(CO4)(Remember/IOCQ)]

6 + 6 = 12

7. (a) Explain the function of the following operators: split and join in PERL using a suitable example. [(CO4)(Explain/LOCQ)]

(b) Write a perl script to open a file which contains a single nucleotide sequence. Read the sequences from the file and append a new sequence to the file. [(CO4)(Compose/HOCQ)]

(c) Write PERL programs to obtain DNA sequence from a given mRNA sequence. [(CO4)(Examine /IOCQ)]

3 + 6 + 3 = 12

Group - E

8. (a) (i) Explain the basis of 3-state secondary structure prediction for proteins. How is the accuracy factor Q_3 for secondary structure prediction mathematically expressed? *Your answer should be in the term of correlation coefficients*.

(ii) What are the advantages of including multiple sequence alignment as part of a secondary structure prediction algorithm? How does use of a machine learning method like support vector machine (SVM) improve the accuracy of a secondary structure prediction algorithm? [(CO5)(Understand-analyze/IOCQ)]

(b) "In a structure-template based 3D-homology modelling of a protein, potential energy calculations are applied to parts of the model (e.g loop modelling) to improve the quality of the structural model" Explain the reasons why additional structural modelling is necessary. How is this energy minimization procedure carried out on the final model? Rationalize why *only limited energy minimization is* recommended in this step and how is it actually carried out. [(CO5)(Reasoning-analyze/HOCQ)]

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

9. (a) "A computational approach towards studies of protein-ligand interactions can be done through quantitative analysis of relevant binary physico-chemical interactions and bioinformatics based procedures." Explain this statement using the simple example of a protein P binding to a ligand L forming a binary adduct P-L. *Your answer should be based on quantitative logic and analysis*. [(CO6)(Understand-apply/IOCQ)]

(b) Write out the steps of a generalized homology modelling based method that is used for the tertiary structure prediction of a protein. Choose a specific bioinformatics tool that is used for this purpose : What was/were the primary algorithm(s) used? How has this algorithm been used for the purpose of annotating a gene? [(CO5)(Understand-Analyse/IOCQ)]

6 + 6 = 12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	17.71	50	32.29

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