M.TECH/BT/2ND SEM/BIOT 5201/2022

ADVANCED BIOINFORMATICS
(BIOT 5201)

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

Figures out of the right margin indicate full marks.

Candidates are required to answer Group A and <u>any 5 (five)</u> from Group B to E, taking <u>at least one</u> 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) One of the application subfields of bioinformatics is function analysis. Which of the following is an example of bioinformatics function analysis?
(a) Gene and promoter prediction (b) Motif discovery
(c) Metabolic pathway modelling (d) Sequence alignment.

(ii) Which of the following is untrue about the Unweighted Pair Group Method Using Arithmetic Average?
(a) The simplest clustering method is UPGMA, which builds a tree by a sequential clustering method
(b) Given a distance matrix, it starts by grouping two taxa with the largest pairwise distance in the distance matrix
(c) The distances between this new composite taxon and all remaining taxa are calculated to create a reduced matrix
(d) The grouping process is repeated and another newly reduced matrix is created.

(iii) Which one of the following combinations is a proper choice for protein structure prediction methods?
(a) ab initio and homology based methods
(b) Equilibrium and kinetic methods
(c) Monte Carlo and molecular dynamics methods
(d) None of the above.

(iv) A GO (gene ontology) annotation of a protein provides which of the following categories of information?
(a) Biological process (b) Cellular component
(c) Molecular function (d) All of the above.

(v) The Eisenberg "3D-profiles" method for testing the suitability of a predicted protein structure model is based upon a statistical analysis of known protein structures. In this method, the score for a residue I in an environment j is calculated using which one of the following choices?
(a) Score=Σ ln (N-6) X N_g (b) Score = ln (P(i:j)/P_i)
(c) Score= ln (S(i:j)/S_j (d) None of the above.

(vi) A Hidden Markov model (HMM) has better predictive power than profiles because
(a) its probability modelling is worse
(b) it is able to differentiate between insertion and deletion states
(c) a single gap penalty score determines insertion or deletion
(d) lower numbered PAM matrices are appropriate for comparing distantly related sequences.

(vii) Which of the following cannot be considered as a "rule-based" approach for protein secondary structure prediction?
(a) Chou Fasman method of secondary structure prediction
(b) Neural networks based method of secondary structure prediction
(c) Monte Carlo-simulated annealing lattice model
(d) Information theory based GOR model.

(viii) Which of the following properties influence the ability of a potential drug to cross the blood brain barrier?
(a) Molar refractivity (MR) (b) Buried surface area
(c) Dielectric constant of p-dioxane (d) None of the above.

(ix) Which of the following is multiple sequence alignment tool?
(a) CHIME (b) RASMOL (c) CLUSTALW (d) PDB.

(x) The Tanimoto coefficient T is frequently used as an index to quantify structural similarity between potential therapeutic drug structures; T is defined by which of the following expressions?
(a) T= N_σ - N_β (b) T = N_{11} /n-N_{00} (c) T = c logP- 524 (d) T= S-ln P.

Group- B

2. (a) Define the following terminologies homologous, paralogous and write the relationship among them in relation to sequence alignment.
[(CO1)(Understand-IOCQ)]

(b) Illustrate the three zones of protein sequence alignment with the help of a suitable graphical representation. [(CO3)(Apply-IOCQ)]

(c) For quantitative assessment between sequences the scoring matrices play a major role. — Justify the statement with suitable example.
[(CO3)(Evaluate-HOCQ)]

(d) Make a comparative analysis between the following scoring matrices: PAM and BLOSUM. [(CO3)(Evaluate-HOCQ)]

3 + 3 + 3 + 3 = 12

3. (a) Use a properly labelled schematic representation to describe the features of Clustal. [(CO3)(Understand/LOCQ)]

   (b) One of the important features of Clustal is its flexibility in using multiple substitution matrices. Explain stepwise how this is achieved in practice. [(CO3)(Understand/LOCQ)]

   (c) Cite another distinguishing feature of Clustal and discuss briefly how this feature is reflected in the operation of Clustal. [(CO3)(Understand/LOCQ)]

6 + 3 + 3 = 12

Group - C

4. (a) For phylogenetic analysis DNA sequence choose synonymous and nonsynonymous substitution–explain how this approach reveal negative and positive selection events. [(CO4)(Evaluate-HOCQ)]

   (b) "Choice of a molecular marker is one of the first of a key series of steps in molecular phylogenetic tree construction." Discuss the different types of molecular markers that are in existence and their specific purposes with respect to sequence alignment. [(CO3)(Explain-IOCQ)]

   (c) Define homoplasy. [(CO3)(Knowledge-IOCQ)]

4 + 6 + 2 = 12

5. (a) Find out the best three trees following the Maximum Parsimony method based on the following set of sequences: SeqW: ACAGGAT/ Seq X: ACACGCT/ SeqY: GTAAGGT/ SeqZ: GCACGAC. Choose the informative sites that are needed for maximum parsimony for the set of these data. [(CO3)(Design–HOCQ)]

   (b) Draw the weight matrices for both Fitch and Transversion Parsimony method; draw the generalized unrooted trees following both these methods based on the above data set. [(CO3)(Design–HOCQ)]

   (c) Tabulate the advantages and disadvantages of this phylogenetic method. [(CO3)(Design–HOCQ)]

(2 + 1) + 5 + (2 + 2) = 12

Group - D

6. (a) Comment briefly on why structure prediction of membrane proteins is of such practical importance. What are the structural reasons responsible for rendering algorithms for globular proteins unusable for transmembrane proteins? Briefly outline (using a diagram) the special assumptions/rules that are necessary for correct prediction of the structure of membrane proteins with transmembrane α-helices. Name one bioinformatic software tool that is used for predicting the location of transmembrane helices. What two bioinformatics based steps can improve the prediction accuracy of algorithms that predict the secondary structure of transmembrane proteins. [(CO4)(Understand-Analyze/IOCQ)]

   (b) What are the utilities of alignment/comparison of protein structures? Name the two methods that are typically applied for comparing protein structures. Explain the GDT method for comparing the similarity of protein structures.

What are the conditions under which this method is applicable?
[(CO4)(Remember-Understand/LOCQ)]

(c) Why has so much attention been focused on RNA secondary structure prediction? Draw a schematic diagram of a hypothetical RNA molecule containing the four types of RNA loops. What are the higher level structures present in RNA? Itemize the steps in the comparative approach for RNA secondary structure prediction. How can this algorithm be further sub-divided into two types based on sequence alignment of RNA sequences? Use examples wherever necessary. [(CO4)(Analyze/IOCQ)]

4 + 3 + 5 = 12

7. (a) Name three protein classification databases (including their full names) that are derived databases from PDB. Use a table to represent the biological hierarchy of any one of these three protein classification databases.
[(CO2)(Remember-Understand/LOCQ)]

(b) What are the unique structural characteristics of beta-barrel membrane proteins that make the corresponding prediction techniques difficult? Outline the steps of one algorithm that is typically used for the predicting transmembrane beta-barrel regions. What are the reasons why this particular algorithm is favoured over other similar type algorithms? [(CO3&CO4)(Understand-Analyze/LOCQ)]

(c) What are the accuracy improvements achieved in secondary structure prediction of proteins through the combined use of multiple sequence alignment (MSA) and neural networks (NNs)? What are the technical reasons behind this enhanced accuracy? Name two public domain webservers for secondary structure prediction of proteins that use such consensus methods.
[(CO1)(Analyze/IOCQ)]

4 + 4 + 4 = 12

Group - E

8. (a) How are drug targets classified? What are the two major classes of drug targets? Explain the biological/physiological reasons why these two classes constitute the maximum known drug targets? [(CO6)(Understand-Analyze/IOCQ)]

(b) Many proteins that have been isolated from pathogens have corresponding human homologues. You have developed a method for comparing the parameters for specificity determination in the binding sites of two homologous proteins. How would you use this method for selecting relevant drug targets?
[(CO5)(Analyze/HOCQ)]

(c) Explain how principles of *molecular modelling* have been applied for the development of ANY ONE analgesic drug. [(CO6)(Apply-Analyse/IOCQ)]

5 + 4 + 3 = 12

9. (a) Illustrate molecular docking between a ligand and a receptor using a diagram. What are the primary measurables of docking? Use a table to represent three types of docking calculations. [(CO6)(Remember-Understand/LOCQ)]

(b) QSAR equations were first used to rationalize biological activity by relating such activity to a molecule's electronic characteristics and hydrophobicity. Use the above statement to define and analyze the significance of the parameters of the following QSAR equation:
$\text{Log}\,(1/C) = k_1 \log P - k_2(\log P)^2 + k_3\sigma + k_4$. [(CO6)(Analyze/LOCQ)]

(c) (i) Draw a flowchart for the Combinatorial chemistry process starting with library design. If the molecular weight of a building block (BB) in a combinatorial chemistry process is 150, how many BBs are necessary to produce final library members of 300-750 molecular weight? (ii) Itemize the underlying scientific-technical reasons for the greater drug development incentive to produce small organic compounds (MW <700) by techniques of molecular diversity than bio-oligomers. [(CO6)(Analyse/IOCQ)]

3 + 4 + (3 + 2) = 12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	31.25	41.67	27.08

Course Outcome (CO):

After the completion of the course students will be able to

- Use acquired knowledge about different bioinformatics experiment categories (e.g. sequence, structure analysis) and their applications in new biology) (e.g. genomics, proteomics)
- Learn organization and characteristics of primary and specialized databases and portals, introduction to new applications of databases/portals towards study of metabolic pathways and systems biology
- Learn and apply sequence alignment methodologies (including comparison of applicable heuristic and dynamic algorithms) for pairwise and multiple sequence alignment and molecular phylogenetics
- Learn and apply bioinformatics based software tools (and the algorithms underlying them) for annotation and structure prediction of prokaryotic and eukaryotic genes, RNA secondary structure prediction and secondary structure prediction of globular, fibrous and membrane proteins (e.g. use of artificial neural network and Hidden Markov model based algorithms for these purposes)
- Principles and applications of homology, fold recognition, and ab initio based algorithms for tertiary structure prediction of proteins, application of protein tertiary structure prediction towards problems of protein folding and design.
- Learn and apply the principles of molecular modelling and energy minimization for small molecule -protein and protein-protein binding; learn the principles and methodologies of computer aided drug design.

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