(b) For a homology model based prediction algorithm like SWISS-MODEL, the predicted model has to be refined by *limited* energy minimization. What is the role of this step? Explain how this model refinement can be done utilizing GROMOS.

(c) Suppose that you are trying to evaluate by threading whether a sequence of length M is likely to have a folding pattern of a protein of known structure of length N>M. How many different alignments of the sequences are possible? How do *ab-initio* protein structure prediction algorithms work and what are their methodological limitations? Explain with examples.

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

Group - E

8. (a) Draw a schematic to represent ligand based drug design.What are its essential differences from structure based drug design?

(b) Explain the different parameters in the QSAR equation $\log (1/C) = k_1 II - k_2 II + k_3\sigma + k_4$

(c) Explain why the partition coefficient log P remains the benchmark molecular descriptor for drug design. Explain its significance. Name 2 other molecular descriptors that you would use to ascertain the properties of a drug.

(d) What is the most widely used technique for deriving a QSAR equation? Give a simple illustration of this technique.

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

9. (a) Define the 'classic' approach to combinatorial synthesis. Use a diagrammatic representation to illustrate the "split-mix" approach to combinatorial synthesis.

(b) Cite two methods of improving Lipinski rule of 5 to estimate the "drug-like"properties of a molecule.

(c) Give a simple 'inverted funnel' table of how a library of leads is reduced to an acceptable size.

(d) What are the three traditional categories of CASP based protein structure prediction?

3 + 3 + 3 + 3 = 12

M.TECH/BT/2ND SEM/BIOT 5201/2018

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 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) A machine learning algorithm is a
(a) computational approach to detect patterns
(b) computational approach that uses progressive optimization
(c) computational approach that optimizes internal parameters of an algorithm
(d) all of the above.

(ii) The prediction of transmembrane segments in membrane proteins depends upon
(a) use of hydrophobicity
(b) use of neural networks
(c) use of evolutionary information
(d) all of the above.

(iii) For selection as a starting template in homology modelling, a database protein should have _____% sequence identity with the query sequence
(a) 20 (b) 30 (c) 40 (d) 50.

(iv) Which of the following are classical isosteres
(a) CH_3, NH_2, OH, F (b) CH_3, C_2H_5, C_3H_7
(c) S-S (d) none of the above.

(v) In conformational energy calculations, weaker interactions involve which of the following energy terms ?
(a) bond stretching (b) bond angle bend
(c) torsional angle change (d) Van der Waals forces.

(vi) In a neural network, a weight is regarded as a
(a) constant (b) pattern identifier
(c) an input (d) a learning process.

(vii) Homologation in drug design refers to the
(a) effect of carbon chain length on drug potency
(b) effect of carbon chain length on toxicity
(c) effect of aromatic groups on drug potency
(d) effect of aromatic groups on toxicity.

(viii) *Unequal* evolutionary rates between sequences are corrected using a conversion step .Which of the following formulae is used for this purpose?
(a) $d_{AB}' = d_{AB} - ¼ \times (r_A + r_B)$ (b) $d_{AB}' = d_{AB} - ½ \times (r_A + r_B)$
(c) $d_{AB}' = d_{AB} - 1/3 \times (r_A + r_B)$ (d) $d_{AB}' = (d_{AB}/3) - 1/2 \times (r_A + r_B)$.

(ix) A false negative is
(a) false match that is incorrectly identified as a true match (by an algorithm)
(b) false match that is correctly ignored
(c) true match that fails to be recognized by an algorithm
(d) none of the above.

(x) Leucine zipper domains are
(a) a special type of β-sheets (b) a special type of coiled coil
(c) a special type of lipid bilayer (d) a β barrel membrane protein.

Group - B

2. (a) Mention what you mean by content sensor and signal sensor in respect to gene prediction programme.

(b) *Ab-initio* based approaches in eukaryotic system of gene prediction programs rely on the several features. Describe those.

(c) Mention the characteristics of eukaryotic gene content sensor citing one example of such software used for this purpose.

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

3. (a) Mention the application of sequence alignment.

(b) Name the program in sequence analysis which is based on finding high-scoring ungapped segments among related sequences based on the query sequence. Briefly describe steps of the said procedure which the said programme follows.

(c) Mention the name of the statistical indicator in the result of the above mentioned programme. Mention how it is related to raw alignment score.

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

Group - C

4. (a) Why is parsimony a good assumption for phylogenetic analysis? On what principle is the parsimony method based? Cite ONE main advantage and disadvantage of a maximum parsimony method.

(b) The following set of 4 sequences are provided: Seq W: ACAGGAT ; Seq X: ACACGCT; Seq Y: GTAAGGT; Seq Z: GCACGAC; Using the cladistic method of Maximum parsimony find the best 3 phylogenetic trees. What are the *informative* sites that are required for maximum parsimony of these data sets? Find out the weight matrix for the Fitch parsimony method and draw the corresponding generalized unrooted tree based on this method.

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

5 (a) Name the phylogenetic tree that shows evolutionary divergence. Explain with reasons why this is the case.

(b) What are the number of unrooted trees that can be generated from six taxa? Cite the usable advantages of rooted trees over unrooted trees.

(c) Which one of gene or species phylogeny is necessary for effective phylogenetic tree construction?

(d) What is the necessity for formal statistical tests in calculated/plotted phylogenies? Explain two of the adopted methods for statistical evaluation.

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

Group - D

6. (a) What characteristics of RNA secondary structure make them important as drug targets? Depict a RNA structure with its SSEs. Explain the role of supersecondary structures in RNA hairpin-bulge interactions.

(b) Explain the methods in DALI and VAST for structure comparison. What is the role of the Z-score to evaluate structure alignment?

(c) In comparing the hierarchical protein classification databases SCOP and CATH, what conclusions can you draw about the *quantitative criteria* to classify them?

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

7. (a) Using a table, compare the steps used in protein tertiary structure prediction by the methodologies of homology modelling and threading. What are the *requirements* for a successful fold recognition by threading?