9. (a) What is the basis of the hybrid method for *structure comparison* of proteins? Name the parameters that are employed to increase alignment accuracy in this hybrid method. Why is dynamic programming used in such a hybrid method?

   (b) Using a simple 2D potential energy diagram, write down the steps of a conventional Monte Carlo algorithm. What is the role of the phenomenological parameter, T, in the calculation?

   (c) Draw an accurate flow chart to represent the critical technologies in the Combichem (combinatorial chemistry) process. Your chart must include the knowledge areas involved from library design to hit confirmation/interpretation.

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

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

   (ii) Which of the following are classical isosteres
   (a) $-CH_2-$, -CONHR, -benzene, pyridine (b) CH_3, NH_2, OH, F, Cl
   (c) carbonyl group and its analogs (d) all of the above.

   (iii) 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.

   (iv) The tight affinity criteria for an inhibitor drug arises from which of the following logical choices?
   (a) Tight binding is necessary for efficacy at lower concentrations
   (b) Tight binding is necessary for efficacy at higher concentrations
   (c) Tight binding ensures specificity of the inhibitor drug
   (d) Tight binding causes more possible side effects from the drug.

(v) 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.

(vi) Pharmacophore generation / identification involves
(a) identification of common structures of many pharmacologically active compounds
(b) identification of chromophores
(c) identification of fluorophores
(d) identification of bioactive compounds of dissimilar therapeutic activity.

(vii) Which of the following corrects for unequal evolutionary rates between sequences by using a conversion step? This conversion requires the calculations of "r values" and "transformed r values" using the following _____formula.
(a) $d_{AB}'=d_{AB}-1/4\ X\ (r_A+r_B)$ (b) $d_{AB}'=d_{AB}-1/2\ X\ (r_A+r_B)$
(c) $d_{AB}'=d_{AB}-1/3\ X\ (r_A+r_B)$ (d) $d_{AB}'=d_{AB}/3-1/2\ X\ (r_A+r_B)$.

(viii) The UNIQUE mathematical step in a Metropolis Monte Carlo method based *simulated annealing* calculation involves which of the following?
(a) generation of a random set of values of x to provide a starting conformation
(b) calculation of the energy of the new conformation at x" if energy is increased/T raised
(c) calculation of the energy of the new conformation at x" if energy is decreased/T lowered
(d) perturbation of the variable from X to X".

(ix) Which of the following choices is a wrong statement regarding the conventional determination of open reading frames?
(a) without the use of specialized programs, prokaryotic gene identification can rely on manual determination of ORFs and major signals related to prokaryotic genes
(b) prokaryotic DNA is first subject to conceptual translation in all six possible frames, two frames forward and four frames reverse
(c) a stop codon occurs in about every twenty codons by chance in a noncoding region
(d) prokaryotic DNA is first subject to conceptual translation in all six possible frames, three frames forward and three frames reverse .

(c) What are the steps in a general homology modeling procedure for protein tertiary structure prediction? Explain the process of limited energy minimization that is an essential part of a homology modeling procedure.

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

7. (a) Homology, comparative modelling and fold recognition methods use energy functions and are based on template based modelling (TBM). What are the three main factors that have made TBM techniques successful? Use one example to illustrate each of the factors you have cited as reasons for TBM's success.

(b) Give a detailed, properly labelled Phyre 2 pipeline development flowchart showing the four stages. What functions are performed in the fold library scanning stage?

(c) How are the structures of transmembrane helices and signal sequences found by Hidden Markov models (HMMs)? What is the key assumption in such methods? Name a bioinformatics tool that is based on an HMM that performs the above function.

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

Group - E

8. An example of a QSAR equation to relate a possible lead molecule's biological activity to its electronic characteristic and hydrophobicity is given by $\log(1/C) = k_1 \log P - k_2 (\log P)^2 + k_3\sigma + k_4$.
(i) Define and explain the different terms in the above equation.
(ii) Extensive efforts have been made to develop computer programs that estimate values for various molecular descriptors based on chemical structure and reactivity. How have such programs been useful?
(iii) One of the fundamental assumptions of molecular docking is using a scoring function to approximate the binding free energy for ligand binding to the receptor molecule. Write out the full mathematical expression for such a docking scoring function that includes the various energy contributions to binding. Plot the "piecewise linear potential" for docking.
(iv) How do metabolic pathways specific to microorganisms help in identification of targets against metabolic disease? Use your knowledge of pathway databases (e.g. KEGG, STRING) to answer this question.

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

(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) Mention what do 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.

(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) Draw a detailed labelled architecture of a neural network (NN) that is utilized for eukaryotic gene prediction. Name a webserver based tool that does eukaryotic gene prediction using NNs.

(b) Itemize the distinguishing characteristics of the following two gene prediction based bioinformatic tools: GLIMMER and GENEMARK. Which is preferred over the other and under what circumstances?

(c) P, Q, R, S are four taxa for phylogenetic tree construction based on clustering methods. The respective distances are PQ=0.40, PR=0.35, PS=0.60, QR=0.45, QS=0.70 and RS=0.55. Construct a phylogenetic tree based on this data following any one clustering based method. Show stepwise how the final phylogenetic tree is developed.

(d) Name one bioinformatics phylogenetics software platform that is based on the clustering method you adopted in part (a). Itemize the advantages and disadvantages of this method.

6 + 6 = 12

(c) Itemize the steps required to develop a functional Hidden Markov Model(HMM) that best represents a sequence alignment. (*hint:* explain the steps of the 'training' process). How is this process similar to building a PSSM?

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

Group - C

4. (a) Define a phylogram and a cladogram using branch lengths as markers.

(b) How can biological data be used for constructing a molecular phylogenetics tree?

(c) What is the role of statistical models for constructing DNA phylogenies? Explain the distinction between the Jukes Cantor and the Kimura substitution models.

(d) Use an appropriate substitution model to calculate the evolutionary distance d_{AB}, between two sequences A and B that differ by 30%. An additional conditionality is that 20% of the sequence changes are a result of transitions and 10% of sequence changes are a result of transversions. Explain your calculation defining all parameters and cite what factor(s) made you choose this particular substitution model?

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

5. (a) P, Q, R, S are four taxa for phylogenetic tree construction based on clustering methods. The respective distances are PQ=0.40, PR=0.35, PS=0.60, QR=0.45, QS=0.70 and RS=0.55. Construct a phylogenetic tree based on this data following any one clustering based method. Show stepwise how the final phylogenetic tree is developed.

(b) Name one bioinformatics phylogenetics software platform that is based on the clustering method you adopted in part (a). Itemize the advantages and disadvantages of this method.

6 + 6 = 12

Group - D

6. (a) Explain in a tabular fashion the essential methodological differences between homology modeling and threading? What are the requirements for successful fold recognition?

(b) Much of bioinformatics is centered on forecasting and prediction. Enumerate the specific steps involved in the following two methods for threading-basedscoring functions (i) empirical pattern of residue neighbors (ii) energy function ranking.