ADVANCED BIOINFORMATICS
(BIOT 5201)

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 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

(ii) 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.

(iii) Which of the following is a wrong statement regarding Gene Prediction Using Markov Models and Hidden Markov Models?
(a) A zero-order Markov model assumes each base occurs independently with a given probability
(b) In a Markov model the conditional probability of a particular sequence position depends on k alternate positions
(c) A Markov model describes the probability of the distribution of nucleotides in a DNA sequence
(d) Markov models and HMMs can be very helpful in providing finer statistical description of a gene

(iv) Which of the following is incorrect regarding the advantages of Molecular data for phylogenetics study?
(a) More clear cut and robust phylogenetic trees can be constructed with the molecular data
(b) Sampling bias is involved
(c) They are easier to obtain as compared to fossil records
(d) They are more numerous than fossil records

(v) 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.

(vi) Which of the following types of algorithms are used in threading based protein tertiary structure prediction?
(a) Pairwise energy based (b) Profile based
(c) Fitting sequence to a structure database (d) All of the above

(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 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.

(ix) 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’’

(x) 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

Fill in the blanks with the correct word

(xi) _____ is a data retrieval based bioinformatics tool.

(xii) Clustal ω does _____ alignment procedures.

(xiii) Prokaryotic promoter preduiction tool is _____.

(xiv) Global alignment algorithm follows _____ algorithm.

(xv) In the equation MR= [(n2-1)/(n2+2) X (MW)/d for molar refractivity, n represents _____.

Group - B

2. (a) Mention the application of sequence alignment. *[(CO3)(Analyse/HOCQ)]*

(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. *[(CO3)(Remember/LOCQ)]*

(c) Mention the name of the statistical indicator in the result of the above mentioned programme Mention with an example how it is related to raw alignment score. *[(CO3)(Apply/IOCQ)]*

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

3. (a) What is the basis of ab initio based gene prediction? What are the two important conceptual steps in ab-initio gene prediction? *[(CO4)(Understand-LOCQ)]*

(b) Itemize the involvement of different gene signals in this procedure. *[(CO4)(Understand-analyze/IOCQ)]*

(c) "Presence of just a start codon is sufficient to initiate the beginning of the frame of translation". Using an example from bacterial gene prediction, evaluate this statement on a scientific-technical basis. *[(CO4)(Evaluate-HOCQ)]*

(d) 'In order to evaluate the accuracy of a prediction program (for genes, proteins), a performance evaluation of the said program is a requirement. What are the two parameters that are essential for this performance evaluation? Define the parameters mathematically briefly explaining their significance. How does one quantitatively summarize the two parameters into a single "summarizing parameter"? Explain your answer. *[(CO4)(Understand-apply -IOCQ)]*

3 + 2 + 3 + 4 = 12

Group - C

4. (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. *[(CO3)(Analyse/HOCQ)]*

(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. *[(CO2)(Apply/IOCQ)]*

6 + 6 = 12

5. (a) Define Markov chain, zero order and first order Markov model cite relationships among zero order model and first order model. *[(CO4)(Remember/LOCQ)]*

(b) Briefly describe role of the following factors in context of Hidden Markov Model. - Observed and non-observed factors. *[(CO4)(Remember/LOCQ)]*

(c) Graphically represent the three states in a Hidden Markov Model. *[(CO4)(Analyse/IOCQ)]*

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

Group - D

6. (a) What are the important cellular functions and biomedical applications of transmembrane (TM) proteins? Use a schematic of the positive-inside rule to explain why and how special algorithms are needed to solve the secondary structure of TM proteins. Cite two specific factors that can improve the accuracy of algorithms for predicting the secondary structure of TM proteins. *[(CO5)(Analyse/HOCQ)]*

(b) Comparatively tabulate the methodological differences in building a tertiary structure model of a protein using homology modelling vs. threading. What are the two specific quantitative requirements for successful fold recognition by threading? *[(CO5)(Remember/LOCQ)]*

(c) What two principles form the basis of ab-initio protein structure prediction? Outline and briefly explain the steps of the ab-initio protein structure prediction algorithm ROSETTA. What are two current methodological limitations of such ab-initio algorithms? *[(CO5)(Apply/IOCQ)]*

(2 + 2 + 1) + (2 × 2) + 3 = 12

7. (a) Explain in a tabular fashion the essential methodological differences between homology modeling and threading? What are the requirements for successful fold recognition? *[(CO3)(Analyse/HOCQ)]*

(b) Much of bioinformatics is centered on forecasting and prediction. Enumerate the specific steps involved in the following two methods for threading-based scoring functions (i) empirical pattern of residue neighbors (ii) energy function ranking *[(CO4)(Remember/LOCQ)]*

(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. *[(CO2)(Apply/IOCQ)]*

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

Group - E

8. (a) The Hantsch multiple regression analysis based equation relating the molar concentration of a potential drug to physico-chemical parameters and steric/size effects is given by
Log 1/C = -aπ2+ bπ+ρσ +cEs+d|S +e.
Define the terms in this equation. *[(CO6)(Analyse/HOCQ)]*

(b) Aspirin, acetaminophen (paracetamol), ibuprofen are non steroidal anti inflammatory drugs (NSAIDs). What are their specific therapeutic uses? Name the enzymes (two) that these drugs are targeted against. What are the functions of these two enzymes? How can you selectively target one of these enzymes so as to reduce the possibility of adverse side effects of any one of these drugs? What is the mechanism for the adverse side effect of aspirin? *[(CO6)(Remember/LOCQ)]*

(c) Docking and subsequent scoring of ligand-target complexes are important steps between a total library and testing a small number of compounds for further validation. Answer these two questions relevant to the above statement: (i) Name a docking experiment that aims to understand the specificity of a ligand/potential drug(ii) Draw a simple representative "inverted funnel" to show how a large library of potential lead compounds can be reduced to an acceptable number for validation testing. *[(CO6)(Apply/IOCQ)]*

2 + (1 + 2 + 2 + 1 + 1) + (1 + 2) = 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:
Log (1/C) =k1log P -k2(log P)2 + k3σ + k4. *[(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	39.6	36.5	23.9