(c) Define and explain the function of *any two* of the following terms with respect to PERL (i) an associative array (ii) variable $1 (iii) split operator.

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

Group – E

8. (a) (i) An algorithm like Chou-Fasman for secondary structure prediction (CFSSP) calculates helical and strand propensities such as presented below.

Amino acid	Helical(α) propensity	Strand (β) propensity
Glu	1.59	0.52
Val	0.90	1.87
Gly	0.43	0.58
Pro	0.34	0.31

Interpret the above data in terms of the secondary structural elements (SSEs) that these residues adopt. What are the two extra parameters that the GOR method adopted to improve the accuracy of the prediction?
(ii) Three consensus secondary structure prediction programs PSIPRED PHD and JPred have prediction accuracies above 75%. Cite the primary computational and physico-chemical reasons for this improvement. Interpret the accuracy score mathematical relationship, $Q_3 = (C_h+C_s+C_c)/3$. How is the helical correlation coefficient, C_h, calculated?

(b) What does the acronym PHYRE2 represent? "Fold recognition (threading) is based on similarities between nonhomogeneous folds" — explain the principles behind fold recognition by drawing the stepwise implementation flow chart for PHYRE2. What is the role of multiple sequence alignment in this procedure?

(c) What are the principles underlying ab initio structure prediction? Why hasn't it acquired wider use?

[(2 + 1) + (1 × 3)] + (1 + 2 + 1) + 2 = 12

9. (a) Explain the differing assumptions and results inherent between rigid and flexible molecular docking. Why is computational complexity increased in the case of flexible docking? How can bridging hydrogen bonds complicate docking calculations? What would be the methodology to carry out an ideal docking in which both ligand and receptor are able to explore their conformational flexibility? Explain your answers.

(b) What does the equation below represent in the subject of computer aided drug design in general and molecular docking in particular? Define the terms
$\Delta G_{bind} = \Delta G_{solvent} + \Delta G_{conf} + \Delta G_{int} + \Delta G_{rot} + \Delta G_{1/r} + \Delta G_{vib}$

(c) Enlist the three areas where bioinformatics procedures participate in drug discovery and development. Briefly and pointwise elaborate on any one of the three areas.

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

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) Local alignments are more used when_____
(a) there are totally similar and equal length sequences
(b) dissimilar sequences are suspected to contain regions of similarity
(c) similar sequence motif with larger sequence context
(d) all are evolutionarily related.

(ii) 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) These matrices are logarithmic identity values
(d) The year it was developed was 1992.

(iii) Which of the following is untrue regarding the gap penalty used in dynamic programming?
(a) Gap penalty is added for each gap that has been introduced
(b) The gap score defines a penalty given to alignment when we have insertion or deletion
(c) Gap penalty is subtracted for each gap that has been introduced
(d) Gap open and gap extension have been introduced when they are continuous.

(iv) Which of the following is valid for regular expression motif in finding consensus N-glycosylation sites in proteins?
(a) Asn followed by anything but Pro (b) Asn followed by anything but Ser
(c) Asn followed by anything but Thr (d) None of the above.

(v) The HEX docking correlation generates up to 25000 candidate ligands using
(a) a Fast Fourier transform
(b) a SPF shape density representation to N = 20th order polynomial
(c) a Gaussian interface for homology modelling
(d) a Taylor series expansion for molecular dynamics calculation.

(vi) In a PDB entry 6E3D for Mycobacterium tuberculosis DppA it is documented that a PubMed ID is not available. This implies that which of the following choices are valid?
(a) The structure does not have a suitable atomic resolution
(b) The structure has been deposited very recently to Protein Data Bank
(c) The structure obtained is based on NMR data
(d) The protein has >20,000 M.W.

(vii) A neural network is specified by which of the following characteristics?
(a) Topology of its connections (b) Weights of its nodes
(c) Decision formulae of its nodes (d) None of the above.

(viii) A coiled coil is a secondary structural element with two or more
(a) interacting alpha-helices (b) beta-sheets
(c) knots (d) all of the above.

(ix) Which of the following represents a method to test calculated phylogenies?
(a) Jackknifing (b) Position dependent inventory
(c) Support vector machines (d) Dotplot diagonalization.

(x) Target selection in drug discovery and development using principles of bioinformatics involve
(a) differential genomics and proteomics
(b) knowledge of prokaryotic and viral genomes
(c) metabolic pathways specific to microorganisms
(d) lead compound database.

Group – B

2. (a) Using examples of biological databases and portals, explain the differences between flat file, RDBMS and OODBMS formats. Use one of your examples to explain the need for a transition from one format to another in database construction.

(b) How would you relate between structures and function analysis in Biocomputing tool development? Give intrinsic connections between the two as examples.

(c) What role does bioinformatics play in computational studies of protein-ligand interactions?

(4 + 3) + 3 + 2 = 12

3. (a) What are the three major characteristics of relational databases? What problems associated with RDBMS led to the development of object oriented databases?

(b) Briefly describe the varieties of database queries. What is the relevance of each?

(c) How is reorganization of data in a biological database achieved?

(d) Use two examples each from NCBI, Uniprot, PDB and EMBL to support your answers in (a), (b), and (c) above.

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

Group – C

4. (a) "Multiple sequence alignment result reveals more biological information than many pair wise alignments can" — analyze the following statement citing three justifications.

(b) Mention one global alignment based multiple sequence alignment based method.

(c) State the situation where this procedure is suitable.

(d) Mention the features of this method.

6 + 1 + 2 + 3 = 12

5. (a) Define sequence alignment. Describe them with suitable examples.

(b) Mention with suitable reasons why alignment of sequences are necessary.

(c) What are the different types of gaps that can be assigned to a sequence analysis?

(d) "Often the scoring matrix in dynamic programming is known as the substitution matrix" — justify the statement.

(e) When would you choose to use BLOSUM62 and BLOSUM45 in BLAST and why?

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

Group – D

6. (a) What are the unique characteristics of a scripting language (e.g. PYTHON or PERL)? What are the two specific complementary aspects of machine learning used as a computational approach? Give two examples of specific advanced machine learning algorithms that have been used for specialized bioinformatics applications. Cite the specific applications also in your answer.

(b) For what types of data are the following markup languages specialized (i) VRML (i) CNL (iii) BSML (iv) LOGML?

(c) A regular expression is a consensus sequence pattern obtained from a multiple sequence alignment to derive motifs and domains. What is the regular expression for a protein phosphorylation motif? Interpret the following regular expression for a different motif written as E-X(2)-[FHM]-X(4)-{P}-L.

(1 + 2 + 2) + (1 × 4) + (1 + 2) = 12

7. (a) What sort of computational modules does the web portal www.bioperl.org typically contain? Write a simple PERL program to translate a nucleotide sequence into an amino acid sequence according to the standard genetic code. What are the three types of data structures that appear in this program? How is PERL's strength at character string handling illustrated through this example?

(b) A pair of sequences of hexokinase expressed in *Staphylococcus aureus* and *Oryza sativa* are stored in two files in v.pep and p.pep. Using proper syntax, write a PERL program that calculates percentage of identity between the two sequences. What pattern recognition resources of PERL are utilized in this program?