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) An advantage of HMMs over profiles is
(a) probability modelling in HMMs has more predictive power
(b) profile modelling is not able to differentiate between insertion and deletion states.
(c) probability modelling is 3X faster than profile modelling
(d) HMM allows variations in the query sequence.

(ii) GeneBank and SWISSPROT are example of
(a) primary database (b) secondary database
(c) composite database (d) none of these.

(iii) Which of the following is a markup language?
(a) Java (b) Python (c) PERL (d) HTML.

(iv) BLOSUM matrices are used for
(a) multiple sequence alignment (b) pairwise sequence alignment
(c) Phylogenetic Analysis (d) all the above.

(v) Which of the following categories is included in critical assessment of structure prediction (CASP)?
(a) template based modelling (b) fold recognition
(c) new fold determination (d) all of the above.

(vi) Which of the following options is involved in local alignment?
(a) finding local regions with the highest level of similarity between two sequences
(b) two sequences that have similarity over the whole length
(c) two sequences where alignment is carried over the whole length
(d) two closely related sequences.

(vii) Which of the following steps is involved in threading?
(a) a method to score models
(b) a method to score alignments
(c) a method to score sequences on the basis of homology
(d) none of the above.

(viii) Which of the following choices is included in a docking calculation?
(a) 1 ligand-1 protein (b) many ligands-1 protein
(c) 1 ligand-many protein (d) all of the above.

(ix) A PSSM is defined as a table that contains
(a) probability information of amino acids or nucleotides
(b) information about ungapped multiple sequence alignments
(c) a matrix of raw frequencies of each residue
(d) all of the above.

(x) Which of the following energies is calculated by a molecular mechanics program?
(a) heats of formation (b) steric energy
(c) strain energy (d) all of the above.

Group – B

2. (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 2 examples each from NCBI, Uniprot, PDB and EMBL to explain your answers in (a), (b), and (c) above.

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

3. (a) Mention the major characteristics of GenBank format.

(b) How is database quality control ensured? Use examples to highlight your answer.

(c) Define machine learning. What are complementary aspects of a machine learning algorithm?

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

Group – C

4. (a) What is a typical architecture of a hidden Markov Model (HMM) that represents a multiple sequence alignment? The representation should be properly labelled.

   (b) What are scoring matrices in pairwise alignment? Why are scoring matrices more complicated for amino acids than for nucleotides? What are the differences between PAM and BLOSUM matrices? What is the statistical significance of pairwise sequence alignment? Define P-value.

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

5. (a) Draw a scheme of a typical progressive alignment procedure such as Clustal-ω for MSA. What are some of its drawbacks? How do newer generation algorithms address such drawbacks in Clustal?

   (b) Define the parameters that are used for gene prediction accuracy. Use a line diagram to represent them. What is the definition of a correlation coefficient?

   (c) Draw a sequence diagram to show the differences between global and local sequence alignment. Write out the mathematical equations that represent sequence similarity and sequence identity.

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

Group – D

6. (a) What is scoping?

   (b) Mention how 'my variable' is used in scoping in subroutine in PERL.

   (c) Write a program in PERL using a subroutine and my variable to include an element and delete an element from two different variables.

   2 + 2 + (4 + 4) = 12

7. (a) In the FASTA database similarity search tool what is the nature of "hashing" strategy employed? Define any applicable terms and stepwise enumerate the steps involved in the FASTA alignment algorithm.

   (b) Define a regular expression. Outline the rules to describe a motif sequence pattern as a regular expression. Based on the above rules, what is the interpretation of a motif written as E-X(3)-[VYM]-X(5)-{N}-L?

   (c) How was conventional determination of open reading frames accomplished?

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

Group – E

8. (a) What are the computational approaches to protein 3D structure modelling and prediction? Briefly distinguish between the three approaches. How are they categorized into knowledge based and ab-initio methods?

   (b) Why is side chain refinement an important part of protein structural modelling?

   (c) In evaluating a predicted model of an unknown protein, what are the physicochemical features that are normally considered? How are stereochemical properties statistically compared in evaluating such a predicted model?

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

9. (a) Enumerate pointwise how bioinformatics principles have been applied to knowledge based drug design.

   (b) Define molecular docking.

   (c) What are the core parameters that are computed by a docking calculation?

   (d) Tabulate the different types of docking calculations and the information provided by each.

   (e) How does bioinformatics assist in the target selection step of drug discovery?

   3 + 2 + 1 + 3 + 3 = 12