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) Which of the following characteristics are necessary to evaluate the usefulness of a biological database?
(a) curation (b) annotation
(c) quality control (d) all of the above.

(ii) Which one of the following characteristics describes a relational database format?
(a) has a text file separated by a delimiter
(b) uses a set of tables to organize data
(c) stores data as objects
(d) describes complex hierarchical relationships.

(iii) Homology modelling predicts 3D protein structures based on
(a) structure homology with known 3D structures of proteins
(b) sequence homology with known 3D structures of proteins
(c) sequence homology with known native folds of proteins
(d) sequence homology with template structures only.

(iv) Which ONE of the following properties of an artificial neural network gives it its unique computational flexibility?
(a) has *hidden* layers
(b) unlimited assembly and connection of neurons
(c) weighted strengths of the connections feasible
(d) the weights may be regarded as variables.

(v) In the following equation $\text{Log K} = 2.83\,\sigma + 0.65 \log \text{P} - 4.8\,I_1 + 7.5$, σ represents
(a) a binding constant
(b) a partition coefficient
(c) electron withdrawing/donating strength of a chemical substituent
(d) parameter for steric clash.

(vi) An open reading frame (ORF) begins with
(a) an ATG initiation codon (b) ribosomal binding site
(c) an expressed sequence tag (d) a TATA box.

(vii) Local alignments are more prevalent 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 exist
(d) all the sequences are evolutionarily related

(viii) Which of the following represents a homology modelling for proteins bioinformatics tool?
(a) ANOLEA (b) GMQE
(c) SWISS MODEL (d) COMBINE Analysis.

(ix) Which of the following are applications of DNA microarrays?
(a) specialized diagnosis of disease (b) pathogen resistance
(c) target selection for drug design (d) all of the above.

(x) An alignment of two sequences is performed using
(a) dot matrix analysis (b) dynamic programming algorithm
(c) word or K-tuple method (d) all of these.

Group – B

2. (a) What are the defining characteristics of databases? Use two examples each from different databases to highlight the concepts of database annotation and database quality control.

(b) How can false positive annotations cause database pollution? What are the other ways by which database errors can occur? Give an example of such an error.

(c) If you were doing a standard similarity searching of your query sequence with sequences in a particular database what other fundamental functionality based drawback(s) must you deal with? Highlight your answer with one example from the world of proteins.

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

3. (a) How have bioinformatics based methods/techniques assisted in the development of tools that find use in forensics and criminology? Your answer should include specific reference to technologies and statistical methods developed and their applications.

(b) Name four specific categories of function analysis applications of bioinformatics. Based on your knowledge explain which one of these applications has expanded the range of bioinformatics applications maximally. Explain your answer. Using one example within function analysis applications, show the intrinsic binary connection between structure analysis and function analysis.

(c) Using the example of the *Arabidopsis* proteome explain how bioinformatics methods have helped in categorizing the proteome's varied functions specific to plants.

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

Group – C

4. (a) With the help of a diagram establish the relationship among the following items: safe zone, twilight zone and midnight zone with reference to protein sequence alignment.

(b) Define a scoring matrix. Why is a scoring matrix necessary in a sequence alignment program?

(c) Explain with reasons the operational definition of the terms PAM80 and PAM 40.

4 + (2 + 2) + 4 = 12

5. (a) For multiple sequence alignment assessment the progressive alignment method is suitable for comparing sequences of similar lengths - write with suitable reasons to justify the comment.

(b) Explain one method which is used for multiple sequences alignment process with the help of a schematic representation of the procedure.

(c) Mention the several limitations in this approach.

3 + (3 + 3) + 3 = 12

Group – D

6. (a) Itemize the applications of BioPython in computational biology

(b) What are the distinguishing characteristics of PERL?

(c) In gene prediction determination of an open reading frame (ORF) is a key goal. Assume one sequence is provided. Using PERL program, how will you test whether initiation codon is present or absent in the sequence? Write the steps of the program.

3 + 3 + 6 = 12

7. (a) PERL consists of different types of variable – mention the types and state them with suitable example.

(b) Write a program where two user defined sequences are joined.

(3 + 6) + 3 = 12

Group – E

8. (a) Explain stepwise the operation of a neural network that is used for the secondary structure prediction of proteins. Use the example of one SSE prediction tool and a diagram to explain your answer if necessary.

(b) Briefly explain the role of the third hidden (jury) layer in such a secondary structure prediction algorithm.

(c) What characteristics of a bidirectional recurrent neural network (BRNN) improves the accuracy of a SSE prediction algorithm over the usual unidirectional network? Give an example of a BRNN based protein SSE prediction bioinformatics tool.

(d) Define a rotamer and briefly explain how side chain refinement is carried out.

4 + 2 + 3 + 3 = 12

9. (a) What is the basis of *ab initio* structure prediction of proteins? Explain the all atom protein modelling method. Why were the early generation *ab initio* algorithms inaccurate in predicting protein structure?

(b) What are the common error sources in protein secondary structure prediction? Cite the range of prediction accuracy for such methods.
Itemize two ways of reaching robust (>80%) prediction accuracies in secondary structure prediction algorithms.

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

Department & Section	Submission Link
BT	https://classroom.google.com/c/MjM3MzczODM5NzA0/a/Mj1MDk3NjMyMTIz/details