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 one of the following represents an example of database pollution?
(a) Sequence conservation correctly predicts functional conservation
(b) Non divergence of sequences and functional conservation
(c) Progressive integration of databases
(d) Misinterpretation of search algorithms

(ii) Prediction errors for secondary structure prediction in proteins arise predominantly from WHICH ONE of the following choices?
(a) missed β strands (b) missed α helices
(c) short lengths of alpha helices and strands (d) a+c

(iii) A Z score of 0 implies
(a) observed similarity is no better than the average of random permutations of the sequence
(b) observed similarity has not arisen by chance
(c) is expressed in terms of mean and variance
(d) none of the above

(iv) If you were to quantify olfactory perception (sense of smell) by mapping that chemical perception space using multidimensional methods, which of the following algorithms would you be typically using?
(a) a systems language based algorithm (b) a Markov chain
(c) a Hidden Markov model (d) a self-organizing map (SOM)

(v) A neural network for structure prediction of proteins is a machine learning process exhibits which one of the following features?
(a) input is a nucleotide sequence
(b) output is the probability of a residue to adopt a particular structure
(c) the training process involves computation with unoptimized hidden layers
(d) requires a structure of variables and nodes that are not connected

(vi) 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) Partially similar, different length and conserved region containing sequence

(vii) In sequence alignment by BLAST, each word from query sequence is typically _____ residues for protein sequences and _____ residues for DNA sequences.
(a) ten, eleven (b) three, three (c) three, eleven (d) three, ten

(viii) Which of the following is untrue regarding BLAST and FASTA?
(a) FASTA is faster than BLAST
(b) FASTA is the most accurate
(c) BLAST has limited choices of databases
(d) FASTA is more sensitive for DNA-DNA comparison

(ix) SWISSPROT is related to
(a) Portable data (b) Swiss bank data
(c) Sequence data bank (d) Sequence data

(x) While running the BLAST the E value shows 3 it means
(a) Three proteins have been found in database that are similar to the query sequence
(b) Three proteins have been found in database that are similar to the query sequence
(c) There will be three matches in the database that are found by chance
(d) The match in the amino acid sequence is perfect except for the amino acids in the 3 positions.

Group- B

2 (a) How can structure and function analysis as bioinformatics experiment categories lead to applications in specific areas of biotechnology and biomedical sciences? [CO6-(Explain-IOCQ)]

(b) Analyze the following statement "If errors do enter databases in either data or annotation, they tend to propagate into other databases and are difficult to remove" in the context of **quality control** in biological databases. Use two examples to highlight your answer. [CO1 (Analyze-IOCQ)]

(c) Explain the *bioinformatics* based reasons for the creation of secondary databases. [CO1 (Remember-understand-LOCQ)]

3 + 6 + 3 = 12

2. (a) "DNA sequence determines protein sequence". Explain this statement in the context of one type of biological data analysis as part of organized bioinformatics activity. What are the three other primary processes that make up organized bioinformatics activity? Give examples of biological activity that lead to such data analysis. [(CO1) Analyze/ IOCQ)]

(b) (i) Based on the fact that virtually all work in the biomedical field depends on databases, it is obvious that the quality of data directly impinges on the

quality of research. What are the two general approaches to improving database quality? Use two examples of *quality checks in biological databases* that are provided for within the databases themselves to explain your answer which should be specific in its details. [(CO1)-(understand/IOCQ)]

(ii) The PDB ID 1FU2 representsa crystal structure of a variant of the T3R3 human insulin-Zinc complex. It has the following wwPDB validation metrics with their (values): Clashscore (131), Ramachandran outliers (1.1%), Sidechain outliers (29.3%). Qualitatively comment on how these metrics help in improving the quality of the entry and the database itself.
[(CO1) (Quantify/HOCQ)]

(c) Briefly define the concept of database interoperability with an example of a biological database's contents. [(CO1)-(Remember/LOCQ)]

4 + (3 + 3) + 2 = 12

Group - C

4. (a) An sequence alignment tool starts with building up of words , checking in databases, substitution matrix play a major role here. Name the tool and discuss the procedure in detail, state the application of this sequence homology search tool. What is the significance of E value in the result and explain how it is different from that of the score. [(CO2) (explain/IOCQ)]

(b) Analyse a situation where a nucleotide sequence is given - how will you distinguish ORFs and state why ORFs are important in sequence annotation?
[(CO2) (Analyse/IOCQ)]

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

5. (a) The following two DNA sequences ACGTCCTTCATT and GTCTCATG have been provided. The assignment is to align the two based on a scoring scheme that has been provided. Find the optimal alignment citing clearly all the steps involved. Assume the following parameters:
- match = +1,
- a mismatch = 0,
- gap opening costs =−10. [(CO2) (Calculate/IOCQ)]

(b) Discuss the applications of BLAST in sequence homology searches. How is the E value determined? What is its significance in a BLAST result? How is it different from the score? [(CO2) (Analyze/IOCQ)]

6 + 6 = 12

Group - D

6. Write PERL programs to do the followings using lexical loop: **(3 × 4 = 12)**
(i) Obtain DNA sequence from a given mRNA sequence.
(ii) To obtain reverse complementary strand of any given DNA sequence.
(iii) To count the no. of elements of an array.
(iv) To cut the last letter of all the elements of an array. [(CO4) (Create/HOCQ)]

7. (a) Write PERL programs where you can "read a file" and "print the contents of the file" in the *reverse order.* [(CO4)(Understand/LOCQ)]

   (b) Write a program where a sequence is obtained from the user; include a check as to whether an open reading frame is present or absent. [(CO4)(Analyze/IOCQ)]

6 + 6 = 12

Group - E

8. (a) What is the necessity for classification of proteins according to their 3D structures in structural databases? Use a labelled flowchart or an example of a duly categorized protein to represent the construction of the SCOP structural database at different levels. Based on this representation answer the following (i) what is the unusual characteristic of the SCOP database? and (ii) what are the unique features of a SCOP defined superfamily?

[(CO1)-remember-understand-LOCQ]

   (b) "A neural network is defined as a machine learning process that necessitates a structure of multiple layers of interconnected variables or nodes". Use a schematic representation of a secondary structure prediction (SSP) algorithm for a protein sequence to illustrate the above definition making sure you assign the correct meaning to variables, nodes, input/output, hidden layers and training in the context of a protein sequence and artificial neural networks.

[(CO5) (Analyze/IOCQ)]

   (c) Estimation of the prediction accuracy of any protein secondary structure prediction algorithm is generally achieved through a cross validation score Q_3. This commonly used quality index is sometimes defined as Q = true positives +true negatives/total number of residues. (i) Interpret this equation in the context of a 3 state secondary structure prediction for a protein (ii) Cite three sources of errors that reduces the accuracy of prediction based on *assignment of correct/incorrect secondary structural elements*. (iii) Cite three sources of errors *based on the scoring system itself*. [(CO5) (Remember-understand-LOCQ)]

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

9. (a) In a stepwise homology modelled structural template based protein tertiary structure prediction algorithm, backbone model building is one of the steps. Answer the following questions with respect to this protein tertiary structure prediction procedure (i) What does backbone model building achieve? (ii) Why is choice of one template structure preferred over multiple ones? How would the r value be impacted by choice of multiple templates? What manipulations of multiple templates would be necessary for a best fit model of the predicted protein to emerge? [(CO5) (Understand-analyse/IOCQ)]

   (b) What does CASP stand for and what were the reasons why such a mechanism had to be established for accurate protein modelling? What are the two sub-structural elements of a protein that contribute maximally to the accuracy of a protein model and is connected to the applications of a protein model? Name the

three traditional CASP categories for protein structure prediction and the nature of target each is suited for. [(CO5 and CO6) (Remember-Understand/LOCQ)]

(c) A distance dependent dielectric function that is sigmoidal is sometimes used for sophisticated protein conformational energy calculations. An example of such a function is $\varepsilon_{eff}(r) = \varepsilon_r - \varepsilon_r - 1/2[(rS)^2 + 2rS + 2]e^{-rS}$. Define the terms in the equation. In what part of a protein conformational energy calculation does such a function come in and for what type of proteins and solvents is its use warranted?

[(CO6)(Analyse/IOCQ)]

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

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	41.66%	30%	30%

Course Outcome (CO):

After completing the course, the students will be able to:

1. Gain and analyze knowledge about genes and proteins obtained through primary, secondary and specialized databases (e.g. NCBI, PDB).
2. Learn and apply principles and methodologies of pairwise and multiple sequence alignment towards biological problems (e.g. Smith Waterman, Needleman and Wunsch, CLUSTAL algorithm).
3. Learn and apply principles of gene prediction algorithms with respect to prokaryotic gene systems (e.g. Hidden Markov Model based gene annotation).
4. Learn and apply PERL for bioinformatics data interpretation (e.g. sequence analysis, protein to DNA translation).
5. Learn and apply principles and algorithms for secondary and tertiary structure prediction of globular and fibrous proteins (e.g. homology modeling, fold recognition methodologies).
6. Use introductory applications of bioinformatics procedures and protein structure prediction techniques to molecular modeling, molecular docking and virtual screening using representative examples.

*LOCQ: Lower Order Cognitive Question; IOCQ: Intermediate Order Cognitive Question; HOCQ: Higher Order Cognitive Question

Department & Section	Submission Link
BT	https://classroom.google.com/c/NDQ1NDI3MzgxMzg3/a/NDY0Mjg0MDM0NTA0/details