(1)

Q.P. Code: 33983

S.Y.B.Sc. Biotechnology Semester IV (Revised) Examination

Biotechnology- Biostatistics and Bioinformatics

Q 1 Do as directed (Any fifteen) **15**

1. **d. printer** is not an input device.
   a. mouse b. keyboard c. stylus d. printer
2. **c. algorithm**is a finite, precise, unambiguous sequence of instructionscapable of being carried out by a machine in a finite time.
   a. logarithm b. flowchart c. algorithm d. recipe
3. A memory cell which does not lose the bit stored in it when no power is supplied to the cell is known as a **a. non-volatile cell** .
   a. non-volatile cell b. volatile cell c. battery cell d. permanent cell
4. **d. Flash drive** is an example of EEPROM.
   a. CD-ROM b. Floppy disk c. Magnetic tape d. Flash drive
5. *AND*, *OR* and *NOT* are examples of **b. Boolean** operators.
   a. computer b. Boolean c. Euclidean d. Newtonian
6. **c. ftp** is not an internet protocol.
   a. http b. https c. ftp d. www
7. **d. SGD.** is a specialized database for baker's yeast.
   a. Pubmed b. Taxonomy c. ProSite d. SGD.
8. **Algorithm used to compute local alignment.**
   Smith- Waterman
   What is Phylogenetic Analysis?
9. Phylogenetic analysis is also known as molecular taxonomy. It uses the representation of evolutionary information in the form of phylogenetic trees.
10. **Define Algorithm.**
    A set of steps that define computational steps at abstract level
    What is a sub-sequence?
11. If there are two sequences A & B and if A is entirely identical to any portion of B then A is said to be a sub-sequence of B.
12. **Define Noise.**
    Random dots on a dot plot is Noise.
    What is E-value?
13. E-value is a parameter that describes the no-of hits one can expect to see by chance when searching a database of a particular size. It is also probability of Z-score.
14. **What is a Probe?**
    Query sequence is called as a probe.
15. -1
16. X on Y
17. 4

(3)

18. Type II error- failing to reject the null hypothesis when it is actually false.

19. True

20. Null hypothesis is an assumption which states that there is no significant difference between the means. It is denoted by H_0

Q. 2 A With suitable examples explain - primary and secondary databases **08**

1. Primary Database [Definition with explanation and example **[4 Marks]**

2. Secondary Database [Definition with explanation and example **[4 Marks]**

Q. 2 B Explain the need for classifying proteins based on motifs and patterns. **07**

Several databases attempt to use structural similarities of proteins for their classification. Proteins are classified to reflect both structural and evolutionary relatedness. Many levels exist in the hierarchy, but the principal levels are family, superfamily and fold.

The demarcation of boundaries between these levels is to some degree subjective. The evolutionary classification is generally conservative: where any doubt about relatedness exists. The new divisions at the family and superfamily levels. Thus, some researchers may prefer to focus on the higher levels of the classification tree, where proteins with structural similarity are clustered.

The different major levels in the hierarchy are:

Family: Clear Evolutionarily Relationship Proteins clustered together into families are clearly evolutionarily related. Generally, this means that pair-wise residue identities between the proteins are 30% and greater. However, in some cases, similar functions and structures provide definitive evidence of common descent in the absence of high sequence identity; for example, many globins form a family though some members have sequence identities of only 15%.

3

Superfamily: Probable Common Evolutionary Origin Proteins that have low sequence identities, but whose structural and functional features suggest that a common evolutionary origin is probable, are placed together in super families. For example: actin, the ATPase domain of the heat shock protein, and hexakinase together form a superfamily.

Fold: Major Structural Similarity Proteins are defined as having a common fold if they have the same major secondary structures in the same arrangement and with the same topological connections. Different proteins with the same fold often have peripheral elements of secondary structure and tum regions that differ in size and conformation. In some cases, these differing peripheral regions may comprise half the structure. Proteins placed together in the same fold category may not have a common evolutionary origin: the structural similarities could arise just from the physics and chemistry of proteins favouring certain packing arrangements and chain topologies.

Relevant explanation as per the above information.....3M

Examples of Database:

CATH, SCOP with explanation 4M

OR

Q. 2 C Elaborate on the databases classifying proteins based on structure. **08**

1. CATH with explanation [4 Marks]

   The CATH database is a hierarchical classification of protein domain structures, which form clusters at four major structural levels, which include Class (C), Architecture (A), Topology (T) and Homologous superfamily (H). Class, derived from secondary structure content, is assigned for more than 90% of protein structures automatically. Architecture, which describes the gross orientation of secondary structures, independent of connectivity, is

currently assigned manually. The topology level clusters structures according to their topological connections and numbers of secondary structures. The homologous super families cluster proteins with highly similar structures and functions. The assignments of structures to topology families and homologous super families are made by sequence and structure comparisons.

2. SCOP with explanation [4 Marks]

The SCOP (Structural Classification of Proteins) database created by manual inspection and abetted by a battery of automated methods, aims to provide a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known. As such, it provides a broad survey of all known protein folds, detailed information about the close relatives of any particular protein, and a framework for future research and classification.

Q. 2 D Comment on RasMol as a protein visualization software. **07**

Major points:[1M each]

- Molecular graphics visualization tool
- 3D Structure Analysis
- Reads PDB and renders
- Any other relevant point.

Q. 3 A Explain BLAST tool for sequence alignment. **08**

Introduction to BLAST- 2 M
BLAST algorithm three steps with suitable diagram(representation)-5 M
E-values and significance of BLAST- 1 M

Q. 3 B What do you understand by Dot plots? How does it differ from dynamic programming? **07**

Introduction to dot plots- 1M;
Mention with example by taking two sequences- 3M
Introduction to dynamic programming- 1M
Mention how Dynamic programming uses backtracking and testing different paths using various parameters such as gap penalties- 3M

5

OR

Q. 3 C **PSI-BLAST is a popular program for exploring protein family relationships. Discuss the areas of applications of this program.** **08**

Introduction to PSI-BLAST: 1M
Steps used by PSI-BLAST: 5M
Applications: 2M

Q. 3 D How can you compute global alignment using the Needleman-Wunsch algorithm? **07**

Introduction To global Alignment: 1M
Take any example of two nucleotide sequences or protein sequences and scoring , backtracking and alignment can be shown- 6M

Q. 4 A Find Coefficient of correlation for the following data:

Table 4 Marks

X	Y	X^2	Y^2	XY
1	1	1	1	1
2	2	4	4	4
3	3	9	9	9
4	4	16	16	16
5	5	25	25	25
$\Sigma X=15$	$\Sigma Y=15$	$\Sigma X^2 = 55$	$\Sigma Y^2 = 55$	$\Sigma XY = 55$

08

Formula **(1 Mark)**

$$r = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)\left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}}$$

Calculation and Answer **(3 Marks)**

r = 55 – 15 X 15/5

square root of (55– 15 X 15/5) (55– 15 X 15/5)

r = 55-45

square root of (55- 45) (55-45)

r = 10/ square root of (10) (10)

r = 10/ square root of 100

r = 10/10

r = 1

Q. 4 B Explain Chi-square with a suitable example.

Concept – 2 marks

The chi square distribution is a **non-parametric** theoretical or mathematical distribution which has wide applicability in statistical work. The Greek letter χ is used to define this distribution. It is always **positive.** The chi square test is based on the difference between the **observed** and the **expected** values for each category. **Degree of freedom** is V= N- 1 or in case of contingency table (R-1) (C-1) **07**

Uses – 3 marks

Test goodness of fit test

Test homogeneity of the variables

Test independence of attributes. ,

The chi square statistic is defined as $\chi 2 = \Sigma\underline{(Oi - Ei)^2}/ Ei$

Where Oi is the observed number of cases in category i, and Ei is the expected number of cases in category i.

Example -2 marks

OR

Q. 4 C Calculate regression coefficients b_{xy} and b_{yx} for the following data and calculate x when y =15.

x	Mean = 10	Standard deviation is 8	Coefficient of correlation is +1
y	Mean = 20	Standard deviation is 10	

b_{xy} = r X Std. Deviation of x series/ Std. Deviation of y series **08**

= 1 X 8/10

= 0.8

b_{yx} = r X Std. Deviation of y series/ Std. Deviation of x series

= 1 X 10/8

= 1.25

Given y = 15 x = ? so the line is x on y

x- Mean of x $= b_{xy}$ (y- Mean of y)

x- 10 = 0.8 (15-20)

x- 10 = 0.8 (-5)

x = - 4 + 10

x = 6

Q. 4 D What is t-test? Give its types. **07**
-test - explanation 2 marks formula 1 mark.
two types (with example) – single mean – 1 mark
Two means - unpaired t test and paired t test – 3 marks.

Q. 5 Write Short notes on **any three** of the following **15**

a. Operating system.
Definition……..1M
Role……..3M

b. World Wide Web.
Definition……..1M
Explanation…..3M

c. **Gapped BLAST.**
Introduction to gapped BLAST- 2M
Need to use gapped BLAST-3M

d. **Multiple Sequence Alignment**
MSA Introduction: 1M
Goal and methods used in MSA- 4M

e. **Types of correlation.**

Positive correlation – Perfect positive correlation - +1 and Strong to moderately positive correlation **(2 Marks)**
Negative correlation – Perfect negative correlation - -1 and Strong to moderately positive correlation **(2 Marks)**
No correlation- independent variables value 0 **(1 Mark)**