Register Number:……..

KUMARAGURU
college of technology
character is life

B.TECH DEGREE EXAMINATIONS: APRIL / MAY 2023

(Regulation 2018)

Sixth Semester

BIOTECHNOLOGY

U18BTI6204: Biological Data Analysis

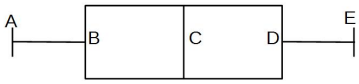
COURSE OUTCOMES

CO1: Understand and apply the biological annotation for macromolecules; apply and interpret the structural analysis of macromolecules using high throughput experiment.
CO2: Apply and interpret the biological data through fundamental statistical analysis.
CO3: Apply and interpret biological data related with hypothesis testing
CO4: Explore and infer biological data using visualization.
CO5: Understand and apply R-programming for biological data analysis
CO6: Provide optimal solution and statistics to biological problems

Time: Three Hours **Maximum Marks: 100**

Answer all the Questions:-
PART A (10 x 2 = 20 Marks)
(Answer not more than 40 words)

1.	List any two tools to analyse protein-protein interactions	CO1	[K_2]
2.	"Nanopore vs Illumina sequencing" - Which one would you prefer and why?	CO1	[K_3]
3.	Investigator-A takes a random sample of 100 men age 18-24 in a community. Investigator-B takes a random sample of 1,000 such men. Which investigator will tend to get a bigger standard deviation (SD) for the heights of the men in his sample?	CO2	[K_3]
4.	What is Type-I and Type-II errors in statistical hypothesis testing?	CO2	[K_2]
5.	Identify the points in the below box plot and name the points A to E.	CO3	[K_2]



6.	Deduce the motive behind the two-sample hypothesis-testing problem	CO3	[K_2]
7.	Difference between correlation and regression	CO3	[K_2]
8.	Write the importance of Mann-Whitney test?	CO4	[K_2]
9.	When do you apply high-pass filter for an image?	CO5	[K_3]
10.	Write a R-script to plot a histogram for sample size n= 10000	CO6	[K_4]

Answer any FIVE Questions:-
PART B (5 x 4 = 20 Marks)
(Answer not more than 80 words)

11.	How does miRNA-mRNA binding regulate protein synthesis?	CO1	[K_3]
12.	Enlist the limitations of a central tendency.	CO2	[K_2]
13.	30 patients with hypertriglyceridemia were randomized between two different therapeutic regimens: sample A, comprising 15 patients with lipid-lowering diet only, and sample B, comprising 15 patients with lipid-lowering diet plus oral gemfibrozil. The following is the investigator's hypothesis: Could oral gemfibrozil increase triglyceride-lowering properties of a lipid-lowering diet?	CO3	[K_3]
14.	You are working with a inverted fluorescent microscope for imaging a live cell. However, the output of the cell is too fuzzy due to some technical issues with the objective lens. Use a suitable R-package and reduce the noise in the image. Explain the various steps in obtaining a better image.	CO5	[K_4]
15.	What happens if you explore facet on a continuous variable? Consider the graph is plotted on a mpg data set, x=displ, y = hwy and the facet-grid is on the cty.	CO4	[K_3]
16.	Using suitable R-package to perform the following task A) Given a DNA sequence - calculate the "GC" percentage B) Convert a DNA sequence to RNA C) Split the entire DNA sequence of k-mer size =3 D) Given a list of protein sequence in FASTA extract the protein name only	CO6	[K_2]

Answer any FIVE Questions:-
PART C (5 x 12 = 60 Marks)
(Answer not more than 300 words)

17.	a)	A proteomic scientist aims to predict the protein's secondary structure solely based on its sequential information. However, the only available information is that there is a limited degree of homology between the protein sequences. Perform the task by assisting the scientist.	CO1	[K_4]
	b)	You were asked to develop a computational tool for identifying miRNA:mRNA interactions. Assess the necessary features to incorporate in the tool to ensure accuracy and efficiency.	CO1	[K_4]

18. a) A clinical psychologist has run a between-subjects experiment comparing two treatments for depression (cognitive-behavioral therapy (CBT) and client-centered therapy (CCT) against a control condition. Subjects were randomly assigned to the experimental condition. After 12 weeks, the subject's depression scores were measured using the CESD depression scale. The data are summarized as follows: CO3 [K3]

	N	Mean	SD
Control	40	21.4	4.5
CBT	40	16.9	5.5
CCT	40	19.1	5.8

Employ one-way ANOVA with level of significance (α) = 0.01 for the test

b) Write a R-script to test the level of significance for the above data set using 1-way ANOVA. CO6 [K4]

19. a) Perform correlation analysis for the following data and provide suitable inference on the correlation co-efficient calculated. CO2 [K3]

Car	1	2	3	4	5	6	7	8	9	10
MPG	23	18	32	25	20	30	27	19	22	28
Weight (in lbs)	2800	3500	2200	2900	3800	2400	2700	4000	3100	2300

b) Write a R-script to identify the correlation and with suitable package generate a correlation plot for the same data set above. CO6 [K4]

20. a) Demonstrate the relationship model between predictor and response variables. CO4 [K4]
The predictor vector stores the height of the person and the response stores the weight of the persons.
Height of the persons = 152, 175, 139, 187, 129, 137, 180, 162, 151, 130
Weight of the persons = 62, 80, 55, 90, 48, 56, 75, 73, 63, 49
Substantiate you solution for the following using R-script
a) Print the summary of the relationship.
b) Determine the weight of the person whose height is 170cm.
c) Visualize the regression graphically with appropriate aesthetics.

21. a) The following data shows the age at diagnosis of type II diabetes in young adults. Is the age at diagnosis different for males and females? CO3 [K_4]

Males: 19,22,16,29,24

Females: 20,11,17,12

22. a) Given a gene dataset containing the following variable - Gene, SampleID , Expression CO5 [K_2]

A) Perform data cleaning and exploratory analysis

B) Remove any rows with missing data

C) Calculate the average expression level for each gene across the sample

D) Identify top 10 genes with highest expression levels

E) Group the data by sample ID and

F) Calculate the total expression level of each sample.

****