Register Number:……..

KUMARAGURU college of technology
character is life

B.TECH DEGREE EXAMINATIONS: NOV/DEC 2024

(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.	Enlist the difference between Hamming and Levenshtein Distance with use case.	CO1	[K2]
2.	List the types of gap penalty	CO1	[K2]
3.	A research team collected data on the height (in cm) of 10 plants over a period of 3 weeks and recorded the following measurements: 15, 18, 14, 20, 22, 15, 17, 19, 21, 16. A) Calculate the variance and standard deviation of the heights. B) Analyze the spread of the data. How would the variability in the data influence decisions related to plant growth conditions?	CO2	[K3]
4.	With an example discuss the concept behind Type-I and Type -II error.	CO3	[K3]
5.	What is the key difference between correlation and regression?	CO2	[K3]
6.	For the given scenario - Justify which statistical test would be applied - A researcher is investigating whether there is a significant difference in the median expression levels of a specific gene between two independent groups of patients: one group receiving a standard treatment and another receiving a new experimental treatment. The data collected does not meet the assumptions of normality required for parametric tests.	CO3	[K3]
7.	Enlist five packages for data visualization in R.	CO4	[K2]
8.	How do you overcome class imbalance in a dataset?	CO4	[K3]

9. Given a nucleotide sequence of RNA, utilize a computational tool to predict its secondary structure based on thermodynamic parameters. CO6 [K2]

10. How would you address missing dataset using R-scripts? CO5 [K3]

Answer any FIVE Questions:-
PART B (5 x 16 = 80 Marks)
(Answer not more than 400 words)

11. a) Describe how you would set up and execute an Illumina sequencing experiment to analyze the genomic DNA from a specific organism. 16 CO1 [K3]

12. a) A researcher is studying the relationship between the amount of fertilizer applied and the crop yield of a specific crop variety. The researcher collects data from 10 different farms, as given below: Calculate the Pearson correlation coefficient between the amount of fertilizer applied and the crop yield. Interpret the result in the context of the study. 8 CO2 [K3]

Fertilizer kg/ha	50	60	80	40	70	90	100	110	120	130
Crop yield (ton/ha)	2.1	2.5	3	1.8	2.8	3.5	3.7	4	4.5	4.8

b) A biologist is investigating whether there is a significant difference in the growth rates of two different strains of bacteria when exposed to a specific antibiotic. The growth rate data (in mm) for each strain is collected from independent samples, but the data does not follow a normal distribution. The growth rates for the two strains are as follows: Using a suitable non-parametric test, determine if there is a significant difference in the growth rates of the two strains at a 0.05 significance level. 8 CO2 [K3]

Strain A (mm)	5	6	4	8	5
Strain B (mm)	3	7	6	4	5

13. a) A medical researcher is studying the effects of three different drug treatments on the blood pressure levels of patients suffering from hypertension. The blood pressure readings (in mmHg) after treatment for each group are recorded as follows: Using a One-Way ANOVA, determine if there is a significant difference 16 CO3 [K3]

in the mean blood pressure levels among the different drug treatments at a significance level of alpha = 0.05 .

Drug A	140	142	138	141
Drug B	135	137	136	134
Drug C	130	132	129	131

14. a) A cell biologist is trying to count cells from fluorescence images. Using R-script help him in finding the total cell population. 16 CO4 [K4]

15. a) Using the dplyr package in R, perform the following tasks based on a given dataset named plant_data, which contains the following columns: plant_id, species, height_cm, leaf_area_cm^2, and growth_stage. 4+4 +4+ 4 CO5 [K3]

A) Filter the dataset to include only plants of the species ``Fern'' that are taller than 30 cm. Display the resulting dataframe.

B) Summarize the average leaf area for each growth stage. Present the output in a tidy format.

C) Create a new column called growth_category that categorizes plants into ``Tall'' (height > 50 cm), ``Medium'' (30<height < 50 cm), and ``Short'' (height < 30 cm). Display the first 10 rows of the updated dataframe.

D) Calculate the total number of plants for each species and display the results in a tidy format.

16. a) You are provided with a fuzzy fluorescent microscopy image that requires enhancement for better visualization of cellular structures. 4+4 +8 CO6 [K4]

A) Analyze the impact of Gaussian blur and sharpening filters on the clarity of the image.

B) Discuss how the choice of filter parameters (radius and sigma) affects the final output.

C) Write an R script that demonstrates the process of applying these filters, and explain the steps involved in your code.

****