Practical Bioinformatics

Mid Sem (TOTAL OF 50 POINTS)

February 27, 2021

Instructions:

- Please turn on the camera.
- Do your end sem questions individually.
- Make one PDF format file for writing your analysis/results.
- ***Do not* zip your submissions.**
- State your assumptions (if any) in your question.
- If the solution requires you to use paper, paste a good quality image of the solution in the document that you are submitting.
- **All questions are compulsory**
- **Only one option is correct for MCQs**
- **Correct MCQ will be awarded +4 and the Incorrect choice carries -1 mark**
- **Answer the subjective questions in bullet points. Keep your responses crisp and to the point.**

---

1. Gene Ontologies are …
    a. undirected cyclic graphs of relationships between genes
    b. undirected acyclic graphs of relationships between genes
    c. directed acyclic graphs used to organize information about genes into hierarchical relationships
    d. directed cyclic graphs used to organize information about genes into hierarchical relationships

    [4]

2. Suppose we have a sample of five values of hemoglobin A1c (HgbA1c) obtained from a single diabetic patient. HbA1c is a serum measure often used to monitor compliance among diabetic patients. The values are 8.5%, 9.3%, 7.9%, 9.2%, and 10.3%.
    a. What is the standard deviation for this sample?
    b. What is the standard error for this sample? [2 + 2]

3. Log transformation of microarray data
    a. is primarily done for data-visualization
    b. is only used for single channel microarray designs
    c. converts expression data into Z-scores
    d. is primarily done to transform non-normally distributed data to normal distribution

    [4]

4. Which of the following is FALSE for a Box and Whisker plot used for EDA?
    a. It shows the quartiles of data with the median represented as a thick line
    b. It shows the density of data points falling in the distribution
    c. It shows the outliers on both sides of the distribution
    d. It shows the inter-quartile range of the distribution, hence indicating the spread

[4]

5. What is the central-limit theorem? Why is it important in statistics? [3]

6. In a dataset with >2 classes that follow a normal distribution, ___ is used as a test of significance of difference of means

    a. Wilcoxon's rank-sum test
    b. Kruskal Wallis test
    c. Student's t-test
    d. Analysis of Variance (ANOVA)

[4]

7. Give at least 1 technical reason for each:
    a. The read count for the same gene may vary between two samples although it is not differentially expressed.
    b. The read counts of two genes may vary within the same sample although they are expressed at the same level.
    c. The same biological sample used using the same microarray platform may give a different expression at different times.

[1+1+1=3]

8. Which among the two sequences might possibly represent the transcript of a gene
    a) 5'AAGCGTGATTGCAC3'
    b) 5'GUGCAAUCACGCUU3'

Write the sequence of the cDNA from above for the preparation of the microarray probe.

[2+2 = 4]

9. Enlist and briefly explain the steps involved in conducting Exploratory Data Analysis upon microarray data available from the Gene Expression Omnibus.

[10]

10. Assume that you have a pre-processed microarray data on which EDA has already been conducted. Enlist the use case scenarios and assumptions for the different types of statistical tests of significance used to extract differentially expressed genes from this data.

[10]