END SEM
Practical BioInformatics
BIO221

Instructions: **Total Marks: 100**

1. Fill the front page with the correct details **Time: 2 Hrs**
2. Only **one option** in MCQs is correct**.**
3. MCQs must be submitted on the Answer Sheet; **responses marked on the Question Paper will NOT be considered**.

MCQs/Close end (1 Mark Each) (10 Marks)

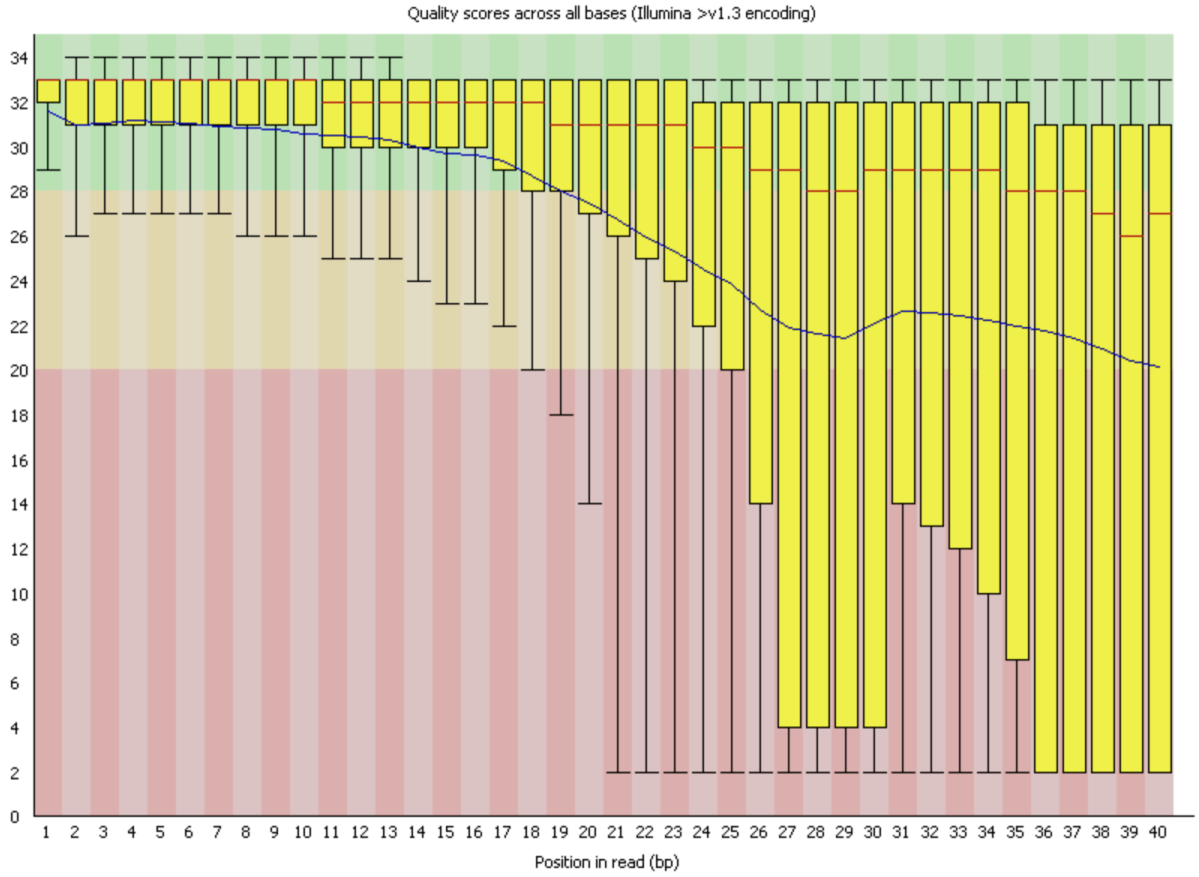
1. Match the sequence of biological data repositories matching the following information (i) genetic blueprint of life (ii) messages read from the genetic blueprint (iii) three-dimensional workhorses of the cell that perform function.
    a. PDB, GenBank, GEO
    b. GEO, GenBank, PDB
    c. PDB, GEO, GenBank
    d. GenBank, GEO, PDB
2. What is the GC content of the DNA sequence "ATGCATGCATGCATGC"?
    a. 25%
    b. 50%
    c. 75%
    d. 100%
3. A Type I error occurs when:
    a. The null hypothesis is not rejected when it is false
    b. The alternative hypothesis is accepted when it is true
    c. The sample size is too small
    d. The null hypothesis is rejected when it is true
4. What is **Phred Score** used for?
    a. Assessing enrichment of biological processes in differentially expressed genes
    b. Assessing the quality of DNA synthesis during PCR amplification
    c. Assessing the quality of sequenced reads
    d. Calculating statistical significance of gene expression levels.
5. What is the primary role of mRNA in the central dogma?
    a. Carries amino acids to the ribosome
    b. Provides a template for DNA replication
    c. Carries genetic information from DNA to the ribosome
    d. Removes introns from pre-mRNA molecules
6. BLOSUM, the default matrix in BLAST was constructed using distantly related sequences. (**True**/False)

7. PAM matrix, as constructed by Dayhoff and Dayhoff used families of closely related sequences (**True**/False)
8. A mutation that is least likely to result in a change in the amino acid sequence of a protein is called a **Silent/Synonymous** mutation.
9. Confidence about the sequence decreases with higher sequencing depth. True/**False**
10. Every data point lies close to the center of mass of the distribution in high-dimensional datasets (True/**False)**

Brief question: (10*3 = 30)

1. **Explain the pattern seen in the attached FastQC report image (attached page) and explain its occurrence. What kind of sequencing approach is used to avoid such a pattern?**

ANS:



For each position a BoxWhisker type plot is drawn. The elements of the plot are as follows:

- The central red line is the median value
- The yellow box represents the inter-quartile range (25-75%)
- The upper and lower whiskers represent the 10% and 90% points
- The blue line represents the mean quality

The y-axis on the graph shows the quality scores. The higher the score the better the base call. The background of the graph divides the y axis into very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red). The quality of calls on most platforms will degrade as the run progresses, so it is common to see base calls falling into the orange area towards the end of a read.

Occurrence :

The most common reason for this pattern to occur is a general degradation of quality over the duration of long runs. In general sequencing chemistry degrades with increasing read length and for long runs you may find that the general quality of the run falls to a level where a warning or error is triggered.

The illumina sequencing approach can be used to avoid this pattern, as illumina generates highly accurate reads with high sequencing quality score (phred score).

2. **Explain how dimensionality reduction is achieved in Principal Components Analysis?**

Ans: Principal Component Analysis (PCA) is a method used to reduce the dimensionality of a dataset while retaining its essential information. Here's how PCA achieves dimensionality reduction:

1. Standardization:
    - PCA starts by standardizing the data. This involves centering the data (subtracting the mean) and scaling it (dividing by the standard deviation). Standardization ensures that all variables have equal importance in the analysis.
2. Covariance Matrix:
    - After standardization, PCA calculates the covariance matrix of the standardized data. The covariance matrix shows how each variable in the dataset changes concerning every other variable.
3. Eigenvalue Decomposition:
    - PCA performs eigenvalue decomposition on the covariance matrix. This decomposition breaks down the covariance matrix into eigenvectors and eigenvalues.
    - Eigenvectors represent the directions of maximum variance in the data and correspond to principal components.
    - Eigenvalues indicate the magnitude of variance explained by each eigenvector (principal component).
4. Selection of Principal Components:

- PCA ranks the eigenvectors (principal components) based on their corresponding eigenvalues in descending order. The eigenvector with the highest eigenvalue captures the most variance, followed by the next highest, and so on.
- By selecting the top N eigenvectors (where N is determined based on the desired amount of variance retention), PCA reduces the dimensionality of the data while retaining important patterns and structures.

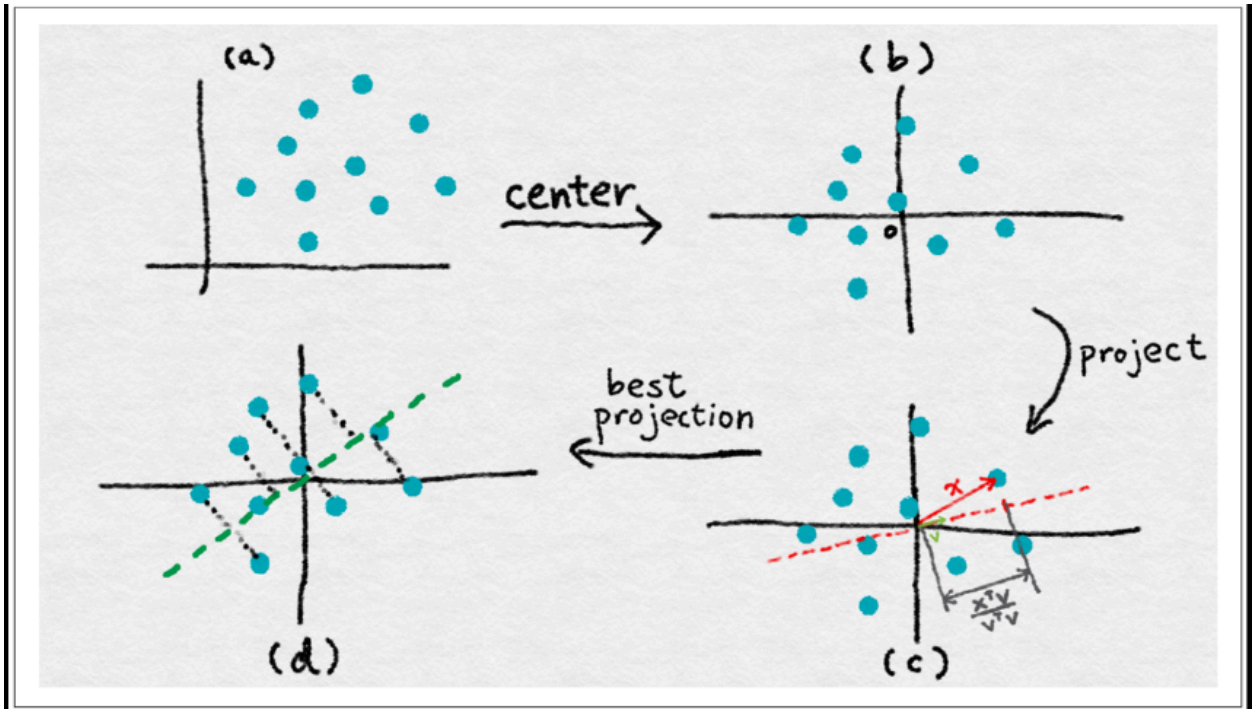
5. Projection:
    - Finally, PCA projects the original data onto the selected principal components. This projection transforms the data from its original high-dimensional space to a lower-dimensional space defined by the chosen principal components.
    - The reduced-dimensional data preserves much of the original data's variability, making it easier to analyze or use in machine learning models without sacrificing critical information.

In essence, PCA reduces dimensionality by identifying principal components that capture the most significant variance in the data, allowing for a more manageable representation of the dataset while retaining important information.

3. Eigenvectors:
    a. Eigenvectors are vectors that represent the directions of maximum variance in a dataset. In PCA, these vectors are derived from the covariance matrix of the standardized data.
    b. Each eigenvector corresponds to a principal component and indicates the direction in which the data varies the most.
4. Principal Components:
    a. Principal components are new variables created as linear combinations of the original variables in the dataset.
    b. The first principal component (PC1) is aligned with the eigenvector that has the highest eigenvalue, representing the most significant source of variability in the data.
    c. Subsequent principal components (PC2, PC3, and so on) capture progressively less variance and are aligned with the next highest eigenvectors.

Each eigenvector corresponds to a principal component, and they represent the directions of maximum variance in the data. The eigenvectors are arranged with respect to their corresponding eigenvalues, with the first eigenvector (PC1) capturing the most variance in the data, followed by PC2, PC3, and so on, each capturing progressively

less variance. So, in PCA, when we refer to eigenvectors or principal components same.



5. Give at least 3 salient differences between RNA-Seq and Microarray analysis.

RNA-Seq	**Microarray**
RNA sequencing is an accurate and high-throughput method.	Microarray is a robust, reliable, high throughput method.
Expensive	Low-cost
More data generation. Hence, the process is complex	Prob based, Knowledge of sequences is required
High sensitive	Limited sensitivity
Gives relative expression levels	Does not give absolute quantification of gene

	expression
Bias is low compared to microarray.	This is a biased method since it depends on hybridization.
Repositories: GEO, ENA, SRA	Repositories: GEO, ArrayExpress

Detailed questions:(4*15=60)

1. **What is Dayhoff's model, and how does it contribute to understanding protein evolution?**

Answer - Dayhoff's model is a statistical framework used to estimate the rates of amino acid substitutions in protein sequences over evolutionary time.

$$R_{ij} = \frac{M_{ij}}{f_i}.$$

Interpretation of Relatedness Score 'R'

- 1: substitution occurs as often as can be expected by chance
- > 1: Alignment of two residues occurs more often than expected by chance (e.g., a conservative substitution of serine for threonine)
- <1: Alignment is not favored

It assumes that amino acid substitutions occur at a constant rate and provides a basis for understanding the evolutionary relationships between proteins. By analyzing the patterns of amino acid substitutions, researchers can infer the evolutionary history of proteins, identify conserved regions, and predict protein function based on sequence conservation. Dayhoff's model is a valuable tool in molecular evolution research, aiding in the analysis of protein evolution and the prediction of protein structure and function.

Illustrate the central dogma of molecular biology using a diagram, with each phase depicting a specific biological process. Annotate each phase with the corresponding database utilized for information retrieval/analysis.

2. Complete the below table:

Type	Query	NO. of database searches	Database
BLASTP	Protein	1	Protein
BLASTN			
BLASTX			
TBLASTN			
TBLASTX	DNA		DNA

ANS:

Type	**Query**	**NO. of database searches**	**Database**
BLASTP	Protein	1	Protein
BLASTN	**DNA**	**1**	**DNA**
BLASTX	**DNA**	**6**	**Protein**
TBLASTN	**Protein**	**6**	**DNA**
TBLASTX	DNA	**36**	DNA

3. **Write the steps for assessing biological differences in gene expression between two classes (e.g. healthy vs cancer). Provide Python code snippets for each of these steps.**

Ans:

1. **Data Collection:** RNA-seq raw data collected from database: (NCBI, GEO)

2. **Data Preparation:** Raw gene expression data usually need to be appropriately cleaned to remove noise, and unwanted biological effects. This also involves normalization, a tailored mathematical approach to enable valid comparison between groups.

3. **Differential gene expression:** Differentially expressed genes (DEGs) are genes exhibiting significant changes in expression between experimental groups. DEGs can be identified with distinct statistical methods that assess both magnitude and statistical significance of the difference between groups (i.e., '*fold-change*' and '*p-value*', respectively)

4. **Statistical Analysis:** Conduct statistical tests to identify significant differences in gene expression between the two classes.

4. **Visualization:** Visualize the results to gain insights and interpret the findings.

snippet:

1. **Import Required Libraries:** The code starts by importing necessary libraries such as NumPy, pandas, Matplotlib, and ttest_ind function from scipy.stats.
2. **Load Gene Expression Data:** It loads gene expression data from a tab-separated file into a pandas DataFrame. The index column is specified as 0, indicating that the first column contains row labels.
3. **Extract Control and Treatment Groups:** It identifies control and treatment groups by filtering columns whose names start with 'CTRL' and 'SWTX143', respectively.
   Calculate Mean Expression: It calculates the mean expression for each gene across control and treatment groups.
4. **Calculate Log2 Fold Change:** It computes the log2 fold change by taking the log2 ratio of treatment mean expression to control mean expression.
   Calculate p-values using t-test: It performs a t-test for each gene to calculate the p-value, assuming unequal variances between control and treatment groups.
5. **Add Log2 Fold Change and p-values to DataFrame:** It adds the computed log2 fold change and p-values to the DataFrame for further analysis.
6. **Define Thresholds for Regulation and Significance:** It defines thresholds for fold change and p-value to identify differentially expressed genes.
   Identify Differentially Expressed Genes: It identifies genes that are significantly upregulated, downregulated, or show no significant change based on the defined thresholds.
7. **Create Volcano Plot:** It generates a volcano plot to visualize the log2 fold change versus the negative logarithm of p-values. Genes are color-coded based on their regulation and significance levels.
8. **Annotate a Random Gene:** For demonstration, we've annotated a random gene of choice in the dataset. But you can probably try to annotate significantly upregulated and downregulated genes.

Code:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from scipy.stats import ttest_ind
```

```
# Load gene expression data
gene_expression_df = pd.read_csv('GeneExpData.txt', sep='\t', index_col=0)

# Extract control and treatment groups
control_groups = gene_expression_df.columns.str.startswith('CTRL') #
Assuming control column names startswith 'CTRL'
treatment_groups = gene_expression_df.columns.str.startswith('TRT') #
Assuming treatment column names startswith 'TRT'

# Calculate mean expression
control_mean = gene_expression_df.loc[:, control_groups].mean(axis=1)
treatment_mean = gene_expression_df.loc[:, treatment_groups].mean(axis=1)

# Calculate log2 fold change
log2_fold_change = np.log2(treatment_mean / control_mean)

# Calculate p-values using t-test
p_values = [ttest_ind(gene_expression_df.loc[gene, control_groups].values,
                      gene_expression_df.loc[gene,
treatment_groups].values,
                      equal_var=False)[1] for gene in
gene_expression_df.index]

# Add log2 fold change and p-values to the dataframe
gene_expression_df['log2_fold_change'] = log2_fold_change
gene_expression_df['p-value'] = p_values

# Define thresholds for regulation and significance
fold_change_threshold, p_value_threshold = 1, 0.05

# Identify differentially expressed genes
upregulated = gene_expression_df[(gene_expression_df['log2_fold_change'] >
fold_change_threshold) & (gene_expression_df['p-value'] <
p_value_threshold)]
downregulated = gene_expression_df[(gene_expression_df['log2_fold_change']
< -fold_change_threshold) & (gene_expression_df['p-value'] <
p_value_threshold)]
no_change = gene_expression_df[abs(gene_expression_df['log2_fold_change'])
<= fold_change_threshold]
```

4. **Enumerate the steps in Position Specific Iterated BLAST (PSI-BLAST) (12 marks). Give at least one special use case for PSI Blast and explain why is it suitable (3 marks)**

ANS: Five Steps of Position Specific Iterated BLAST

1. A normal BLASTP search uses a scoring matrix (such as BLOSUM62)
2. A multiple sequence alignment from an initial BLASTP-like search using composition-based statistics.
3. Creation of PSSM, a specialized, individualized search matrix (also called a profile) based on that multiple alignment
4. Query to search the database again with PSSM rather than original query and calculation of statistical significance as described in E-value
5. The search process is continued iteratively, typically about five times. At each step a new profile is used as the query

Need of PSI-BLAST:
In summary, PSI-BLAST is a powerful tool for detecting remote homologs and uncovering evolutionary relationships among proteins. By iteratively refining the search based on position-specific information, it can reveal biologically meaningful similarities that may be missed by standard BLAST searches.
Example scenario:

1. PSI-BLAST can be used to identify members of a protein family across different species. For example, if you have a known protein sequence associated with a specific function or domain, PSI-BLAST can help identify homologous proteins with similar functions in other organisms.