Course: Introduction to Bioinformatics
University: Sharif University of Technology
Semester: Fall 2022

Project Phase ONE

Project Goal:

One of the main causes of diseases is the improper expression of certain genes, such as the overexpression or underexpression of critical genes. Using **microarray** technology, we can measure gene expression levels in various tissues. By analyzing microarray data obtained from healthy and diseased individuals, we aim to identify potential genetic contributors to the disease.

Data Used:

The project utilizes microarray data related to Acute Myeloid Leukemia (AML), a type of blood cancer. Data Link: GEO Data - GSE48558

Project Requirements:

1. Install the **R programming language** and the associated IDE, **RStudio**.
2. Watch the required tutorial videos (Sessions 9–12) for working with the data and performing analyses.

Deliverables:

The following should be submitted:

1. Written code.
2. A comprehensive report including the methodology and answers to the questions below:
    - Use **microarray** data to explain the method, outputs, and data formats.
    - Identify **phenotype labels**: Healthy samples labeled as "normal" and diseased samples labeled as "AML patient."
    - Normalize and preprocess the data. Check for data quality and report the steps and necessary changes.
    - Perform dimensionality reduction using three methods (**PCA**, **MDS**, **t-SNE**) and compare the results. Choose the best method and justify your choice.
    - Analyze group similarities based on the **source name** field and visualize the relationships using appropriate plots.