🧬 Bioinformatics Assignment I

From GEO Data to Protein Function and Sequence Relationships

Organizational remarks

Please solve the following task either in plain Python, or as Jupyter notebook. Answer the questions as comments, or attach the answers as separate file (e.g. PDF). Send the results to pawel.tumanski@pw.edu.pl before Wednesday, October 29th, 23:59. Should you have any questions, feel free to drop me an e-mail.

Objective

In this assignment, you will integrate multiple bioinformatics techniques to trace the path from gene expression data to protein function analysis. You will:

- Retrieve gene expression data from NCBI GEO
- Extract and translate DNA sequences
- Identify protein function using BLAST
- Explore sequence similarity using alignment algorithms and substitution matrices
- Apply dimensionality reduction (PCA/t-SNE/UMAP) to visualize relationships between proteins

Suggested GEO Datasets

You may choose one of the following datasets (or another GEO dataset of interest):

GEO Accession	Organism	Study Focus
GSE70970	Homo sapiens	Breast cancer vs. normal tissue
GSE10245	Homo sapiens	Lung adenocarcinoma vs. normal
GSE73072	Homo sapiens	Host response to influenza infection
GSE40279	Homo sapiens	Aging-related expression changes
GSE81547	Arabidopsis thaliana	Salt-stress response in plants

Tasks

1. Data Retrieval

Select a GEO dataset and identify one or two differentially expressed genes. Briefly describe the dataset statistically (size, number of features, their mean, variance etc.), its biological context, and your chosen gene(s).

Deliverable: Short paragraph describing dataset (accession, organism, condition) and chosen gene(s).

2. Sequence Extraction

Download the nucleotide sequence(s) of your selected gene(s) from NCBI in FASTA format.
Deliverable: FASTA file(s) of gene sequence(s).

3. Translation to Protein

Translate DNA → amino acid sequence(s).
Deliverable: Protein FASTA file(s).

4. Functional Annotation via BLAST

Use BLASTp to identify homologous proteins and infer function. Record top 10 hits with identity, E-value, and description.
Deliverable: Table of BLAST results + short summary of inferred protein function.

5. Sequence Alignment and Substitution Matrices

Select 3–5 BLAST hits and align them with your protein sequence
Deliverable: Alignment outputs and a brief comparison (≤150 words). Which algorithm did you choose and which substitution matrix? Why?

6. Dimensionality Reduction of Alignment Scores

Take first 100 hits from BLAST. Compute pairwise alignment scores using the method and scoring function of choice. Then each protein can be described by the vector of length 100 of score similarities to all other proteins (including the analyzed protein). Normalize the vectors, apply the dimensionality reduction (UMAP/PCA/t-SNE) and visualize the data as 2D plot. Identify clusters and calculate the assess the cluster correctness using Silhouette score.
Deliverable: 2D plot (PCA/t-SNE/UMAP) with clusters plotted in different colors. What is the motivation for dimension reduction method choice and what is the optimal number of clusters? Why?

Assessment Criteria

Component	Weight
GEO dataset & gene selection	10%
Sequence retrieval & translation	10%
BLAST analysis & interpretation	25%
Pairwise alignments & matrix comparison	25%
Dimensionality reduction & visualization	20%
Report clarity and presentation	10%