Institute of Biochemistry and Biophysics, Department of Bioinformatics

Final Project

Discovering Biological Signatures: A Comprehensive Analysis of Single-Cell Gene Expression Data

Background:

In this project, you will perform an unsupervised analysis of single-cell RNA sequencing (scRNA-Seq) data to gain biological insights from an immune system dataset that includes gene expression information across various cell types. The analysis will involve preprocessing raw data, applying dimensionality reduction techniques, performing clustering to group similar cells, and conducting differential expression analysis to explore the structure and dynamics of the single-cell transcriptome.

1. Preprocessing:

You will be provided with raw count data from 7551 individual cells. Before conducting any downstream analysis or machine learning tasks, several preprocessing steps are necessary to ensure the data is clean and suitable for further analysis.

Tasks:

1. Quality control and filtering:
    - For each cell, calculate the mitochondrial gene percentage (MT percentage).
    - Create violin plots for MT percentage, UMI (Unique Molecular Identifier) count, and the number of genes detected.
    - Filter out low-quality cells based on your chosen quality control metrics (e.g., high MT percentage, low UMI count). Justify the selection of your metrics and discuss why you have chosen them.
    - Re-create the violin plots after filtering to confirm that the data quality has improved.
2. Normalization:

    - Normalize the expression values using a method of your choice (e.g., log transformation, scaling, or normalization to library size). discuss your chosen method and its appropriateness for scRNA-Seq data.
    - Filter out low-quality genes based on your criteria, such as gene expression levels and variance across cells. Explain why you selected these filtering thresholds.
    - Standardize the data by scaling each gene’s expression to have zero mean and unit variance. Discuss why this step is essential for the analysis and its impact on downstream processes.

---

2. Cell Population Clustering

In this step, you will cluster the cells into distinct **population** groups (e.g., B cells, NK cells, T cells) based on their gene expression profiles. Multiple clustering algorithms will be applied and evaluated to determine the most effective method for grouping cells.

Tasks:

1. Model Training:
    - Select 4 different clustering algorithms (e.g., K-means, Hierarchical Clustering, DBSCAN, or Spectral Clustering, etc.).
    - Train each model on the processed gene expression data to predict the cell populations. You may need to perform feature selection or dimensionality reduction (e.g., PCA) to improve the clustering results.
2. Model Evaluation:
    - Evaluate each model using the following performance metrics:
        - Silhouette Score
        - Dunn Index
        - Calinski-Harabasz Index
        - Adjusted Rand Index (ARI)
        - Normalized Mutual Information (NMI)
    - Compare the clustering models based on these performance metrics. Discuss which model provides the best clustering results and why it is most appropriate for this dataset.
3. Dimensionality Reduction for Visualization:
    - Apply non-linear dimensionality reduction techniques such as t-SNE (t-distributed Stochastic Neighbor Embedding) or UMAP (Uniform Manifold Approximation and Projection).

- Visualize the lower-dimensional representation of the cells and examine how they are distributed. Discuss any patterns observed, particularly in relation to cell type or other biological features.

---

3. Differential Expression Analysis

In this step, you will perform differential expression analysis to identify genes that are significantly different between the clusters identified in the previous step. For each cluster, you will determine a set of genes that are enriched or depleted compared to other cell populations.

Tasks:

- For each cluster, report at least five differentially expressed genes.
- Search relevant biological databases (e.g., Gene Ontology, PubMed) to determine if these genes have been previously reported as biologically specific to the corresponding cell population. Discuss whether these findings align with existing literature and their biological significance.

Keynotes:

You can complete the project in groups of 1 or 2 members. Each group must announce their members' names in the Skype group.

You are required to create a comprehensive Jupyter Notebook report that includes your code, results, and answers to the provided questions. The report should also contain brief explanations of your code and any relevant analysis. Send your answers as a zip file via email to Mahdi.Anvari7@ut.ac.ir by **February 14, 2025**. Make sure to include your name and student number in the subject line of the email.

A presentation of your project is mandatory and will take place between **February 15-17, 2025**.

Good luck