Data Science Lab 3: Project 4

Mosaic Data Integration

Due: 17th March 2025, 11:59pm

Problem Introduction

Advancements in single-cell multi-omics technologies are improving our understanding of cellular heterogeneity by providing multiple perspectives on biological systems. Single-cell RNA sequencing (scRNA-seq) is a high-throughput method that measures gene expression at the single-cell level, it generates a cell $\times$ gene matrix, where each cell is a data point and each gene is a feature. In contrast, CITE-seq (Cellular Indexing of Transcriptomes and Epitopes by sequencing) is a multi-omics approach that also profiles surface proteins from the cell. In CITE-seq experiments, we get a cell $\times$ protein matrix, again the cells are data points and each protein is a feature.

Mosaic data integration is used to merge multi-modal single-cell data (e.g., scRNA-seq and protein data in CITE-seq), creating a unified representation across different data types. Integrating RNA and protein data enhances our understanding of the biological sample. Some of the genes and proteins will have correspondence as they are related. Remember the central dogma of biology - from gene, RNA is produced which gives rise to protein.

The goal of this assignment is to develop a computational method that learns joint representation (dimension reduction) for gene expression and protein datasets and performs clustering on the inferred reduced dimensions. To learn more details on the problem, read the following paper.

- Chen S. et al. "Integration of spatial and single-cell data across modalities with weakly linked features" Nature Biotechnology Volume 42, 1096–1106 (2024).

Dataset Description

The dataset presented here comprises Peripheral Blood Mononuclear Cells (PBMCs) acquired using the CITE-seq technique. The following files are provided

- scRNA-seq dataset(10000*2000): .csv file with 10000 cells and 2000 highly variable genes
- Protein dataset (10000*400): .csv file with 10000 cells and 400 protein markers

- protein_gene_conversion.csv: .csv file with information on mapping of protein markers and gene names.

The dataset includes cells from 8 different cell types: which means, when you perform clustering, you need to look for 8 clusters.

Task Description

The summary of the tasks is as follows.

- First, you need to develop a computational method that learns 10-dimensional latent space representation for both the scRNA-seq dataset as well as the Protein dataset, resulting in output with 20000*10 dimensions. When you are performing joint dimension reduction of the two datasets, the cells from the same cell type (irrespective of whether it RNA or protein is profiled) should be close in the latent representation.

- Second, you need to perform clustering on the latent space representation using algorithm of your choice (leiden, louvain, kmeans, etc). Please note that the number of clusters in the dataset is 8. In case you are using the Louvain or Leiden algorithm, you should use a resolution that results in 8 clusters. The clustering output should have the following format.

Listing 1: Format of output csv file

```
Id,Expected
AAACAAGTATCTCCCA-1_rna,1
AAACAATCTACTAGCA-1_rna,2

AAACACCAATAACTGC-1_protein,3
AAACAGAGCGACTCCT-1_protein,2
AAACAGCTTTCAGAAG-1_protein,1
```

- Third, you need to plot the latent space representation using UMAP (Uniform Manifold Approximation and Projection). UMAP is primarily a dimensionality reduction technique that is commonly used to visualize high-dimensional data by reducing it to 2D or 3D. The UMAP algorithm are available

  - in Scanpy package, Python `https://scanpy.readthedocs.io/en/stable/api/generated/scanpy.pl.umap.html`
  - in Seurat package, R `https://satijalab.org/seurat/reference/runumap`

  - Helpful tutorials : `https://scanpy.readthedocs.io/en/stable/tutorials.html`, `https://satijalab.org/seurat/`

- Lastly, you will be able to calculate Adjusted Rand Index (ARI) as a measure of your clustering accuracy by submitting to Kaggle.

Deliverables

The deliverables for the assignment are the following

1. Clustering predictions in the form of the csv above. These results will be evaluated on the kaggle leaderboard.
2. A UMAP plot demonstrating the clustering.
3. A short report describing the steps taken to solve the assignment. Describe in brief the algorithms you have used and the performance of your algorithms on the dataset. The write-up should mention the names and roll numbers of the team members.

Evaluation

We will be evaluating your submitted clustering predictions based on the ARI evaluation metric. These results will be evaluated on the Kaggle leaderboard. The link to the Kaggle competition is `https://www.kaggle.com/competitions/assignment-data-science-lab`

15 points are for the leaderboard and 5 points for the report.

Submission Deadline

March 17^{th} 11:59 PM.