Project-4-Group-3

Sunny
Roll No. 231080094

Jayant Vikash Jha
Roll No. 220479

Anoop Patel
Roll No. 231080017

Laxmi Agarwal
Roll No. 231080053

1 Abstract

Advancements in single-cell multi-omics technologies have revolutionized our understanding of cellular heterogeneity. This project focuses on the integration of single-cell RNA sequencing (scRNA-seq) and CITE-seq protein data using a computational approach for joint representation learning and clustering. The primary objective is to develop a dimensionality reduction technique that maps gene expression and protein data into a 10-dimensional latent space while ensuring that cells of the same type remain closely clustered. Subsequently, clustering algorithms such as Leiden, Louvain, or K-Means are applied to classify the cells into eight predefined clusters. The clustering performance is evaluated using the Adjusted Rand Index (ARI), and results are visualized using UMAP for interpretability. The project contributes to the growing field of mosaic data integration, offering insights into multi-modal single-cell analysis.

2 Dataset

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

- **scRNA-seq dataset**: 10000 cells and 2000 highly variable genes
- **Protein dataset**: 10000 cells and 400 protein markers
- **protein_gene_conversion**: information on mapping of protein markers and gene names.

2.1 Methodology

We utilized RNA and protein datasets for multimodal integration using the MaxFuse method. The RNA dataset contained whole transcriptomic expression profiles, while the protein dataset comprised targeted proteomic measurements. Initial preprocessing steps included:

- **Feature Selection**: Identifying highly variable features across modalities.
- **Linked Feature Identification**: Establishing a correspondence between protein markers and their coding genes.
- **MaxFuse Integration Pipeline**: MaxFuse was employed for cross-modal data fusion through an iterative matching and embedding approach. The integration proceeded in three stages:

  Stage 1: Initial Cross-Modal Matching
  Graph Construction: A fuzzy nearest-neighbor graph was built for each modality based on all available features.

  Fuzzy Smoothing: The signal-to-noise ratio of linked features was improved by averaging within local neighborhoods in the graph.

  Linear Assignment: Cross-modal distances between smoothed linked features were computed, and initial cell matching was performed using a linear assignment algorithm.

Stage 2: Iterative Refinement
Joint Embedding Learning: A canonical correlation-based embedding was generated using matched cell pairs.

Embedding Smoothing: The joint embedding was further refined using graph-based smoothing.

Matching Update: A new round of linear assignment was applied to improve matching accuracy. This process iterated until convergence.

Stage 3: Final Output Generation
Pivot Selection: High-confidence matched cell pairs were retained as pivots.

Propagation: Unmatched cells were assigned to their nearest pivot-based match.

Final Joint Embedding: A joint low-dimensional representation of all cells across both modalities was constructed.

Evaluation and Validation

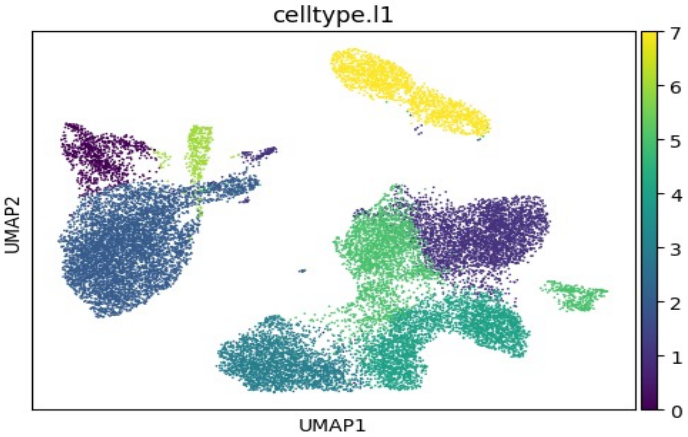
To assess integration quality, we computed:

Cell-Type Matching Accuracy: Fraction of correctly matched cell types across modalities.

UMAP Visualization: Qualitative assessment of modality mixing and biological signal preservation.

3 Conclusion

The MaxFuse method successfully integrated RNA and protein datasets, preserving cell population structure while enhancing cross-modal alignment. The approach effectively handled weakly linked features and provided a biologically meaningful joint representation of the data.



The UMAP plot clearly shows distinct clusters, meaning the model has done a good job of grouping similar cell types together. With a score of 0.81827, the clustering performance is strong, though there are some overlaps, suggesting that certain cell types might share similarities.

A few smaller, isolated clusters could represent rare cell types or noise in the data. Overall, this confirms that unsupervised learning can effectively analyze and classify biological data, especially in single-cell RNA sequencing (scRNA-seq). While the results are promising, fine-tuning the model or trying other methods like t-SNE could further improve accuracy. To make the findings more reliable, comparing the clusters with actual biological labels would be a useful next step.

4 Reference

Chen, S., Zhu, B., Huang, S., Hickey, J. W., Lin, K. Z., Snyder, M., Greenleaf, W. J., Nolan, G. P., Zhang, N. R., Ma, Z. (2024). Integration of spatial and single-cell data across modalities with weakly linked features. **Nature Biotechnology, 42, 1096–1106**.