Contrastive Learning in Single-cell Multiomics Clustering

Bingjun Li
bingjun.li@uconn.edu
University of Connecticut
Storrs, Connecticut, USA

Sheida Nabavi
sheida.nabavi@uconn.edu
University of Connecticut
Storrs, Connecticut, USA

1 Introduction

Recent advancements in single-cell multiomics sequencing technology present new opportunities for researchers. However, the integrative analysis of the multiomics data poses new challenges, especially in cell clustering, a crucial step for any downstream analysis [5]. A key challenge is the alignment of multimodal omic features during fusion. A commonly adopted solution is adversarial training by implementing a discriminator of different omic features [1]. However, discriminators have several drawbacks affecting real-world performance [8]. In this study, we propose to use contrastive learning for better omic alignment by forcing different clusters of latent features to be separable and compact in the same space. We also aim to incorporate prior knowledge of interactions across genomics entities, specifically the gene regulatory network (GRN) for better clustering. Prior studies have shown GRN's important role in cell type classification [3, 4]. To our best knowledge, no end-to-end clustering method that incorporates GRN exists [1].

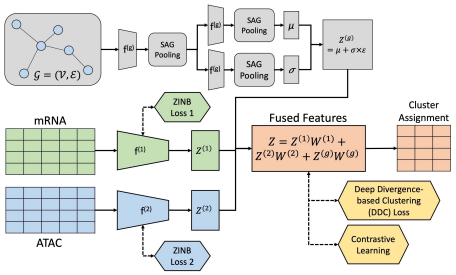

Figure 1: The overall structure of the model is plotted.

We proposed scGEMOC, a novel end-to-end single-cell multiomics clustering method that integrates prior knowledge and contrastive learning. Our contributions are: 1) utilizing the prior knowledge of the GRN, as a pseudo omic; 2) using contrastive learning for better omic alignment in the feature space; 3) developing a scalable model structure that can accommodate additional omics.

2 Method & Results

As shown in Figure 1, there are three key components in scGEMOC: GRN pseudo omic, contrastive learning module, and ZINB loss function module. ZINB loss function regulate the individual feature extractors for each omic data to better model the sparse distribution of omic data. We use shallow fully-connected networks as feature extractor on regular omic data and a variational graph encoder on the GRN pseudo omic. Contrastive learning module regulates the feature alignment in the latent feature space. The fused feature is obtained by weighed addition with learnable omic weights. Deep divergence-based clustering (DDC) loss ensures orthogonal cluster assignments and compact clusters in space.

Contrastive learning aims to make distinct clusters in space for easy clustering. It assigns positive pair and negative pair to each pair of omic features. It aligns different omic features and makes them separable by forcing the latent features of positive pairs close to each other and latent features of negative pairs far away from each other [8]. In scGEMOC, the omic features of cells from the same cluster are assigned positive pairs, and omic features of cells from different clusters are assigned negative pairs.

Table 1: Clustering Results of scGEMOC and Baseline Models

Model	10xPBMC			CellLine			BMMC		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
scGEMOC	**0.552**	**0.620**	**0.416**	**0.893**	0.719	**0.771**	**0.574**	**0.635**	**0.466**
MoClust	0.501	0.613	0.339	0.721	0.634	0.556	0.557	0.603	0.420
GLUE+KMeans	0.426	0.510	0.373	0.571	0.462	0.397	0.336	0.319	0.275
sigDGCNb	0.504	0.457	0.363	0.848	**0.760**	0.764	0.469	0.542	0.412

We conducted experiments on three public single-cell multiomics mRNA and ATAC datasets: 10xPBMC, CellLine, and BMMC dataset [2, 6, 7]. We tested the proposed model against three state-of-the-art baseline models: MoClust, GLUE, and sigDGCNb [1, 9, 10], as shown in Table 1. scGEMOC outperforms all baselines on all three datasets in clustering accuracy (ACC), normalized mutual information (NMI), and adjusted rand index (ARI), except for NMI on the CellLine dataset. To test key components' contribution, we conducted an ablation study with different combinations of them. We found the performance deteriorates as more components are turned off, and ZINB loss has the biggest performance impact.

3 Conclusion

We proposed scGEMOC, a scalable single-cell multiomics clustering method that leverages GRN information and contrastive learning. scGEMOC consistently outperforms existing models on various datasets, showing the benefits of contrastive learning for omic alignment and the potential of incorporating GRN data.

References

[1] Zhi-Jie Cao and Ge Gao. 2022. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat. Biotechnol.* (2022).
[2] 10X Genomics. [n. d.]. *PBMC from a healthy donor, single cell multiome ATAC gene expression demonstration data by Cell Ranger ARC 2.0.0.*
[3] Bingjun Li and Sheida Nabavi. 2023. A Multimodal Graph Neural Network Framework for Cancer Molecular Subtype Classification. *arXiv:2302.12838* (2023).
[4] Bingjun Li, Tianyu Wang, and Sheida Nabavi. 2021. Cancer molecular subtype classification by graph convolutional networks on multi-omics data. In *BCB*. 1–9.
[5] Chen *et al.* 2019. Assessment of computational methods for the analysis of single-cell ATAC-seq data. *Genome Biol.* 20, 1 (Dec. 2019), 241.
[6] Chen *et al.* 2019. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat. Biotechnol.* 37, 12 (2019), 1452–1457.
[7] Luecken *et al.* 2021. A sandbox for prediction and integration of dna, rna, and proteins in single cells.
[8] Trosten *et al.* 2021. Reconsidering Representation Alignment for Multi-view Clustering. In *IEEE/CVF CVPR*. IEEE, Nashville, TN, USA, 1255–1265.
[9] Yuan *et al.* 2023. Clustering single-cell multi-omics data with MoClust. *Bioinform.* 39, 1 (2023), btac736.
[10] Tianyu Wang, Bingjun Li, and Sheida Nabavi. 2021. Single-cell RNA sequencing data clustering using graph convolutional networks. In *IEEE BIBM*. 2163–2170.

BCB '23, September 3–6, 2023, Houston, TX, USA

ACM ISBN 979-8-4007-0126-9/23/09.
https://doi.org/10.1145/3584371.3613010