基于对比学习增强自编码器的单细胞转录组聚类方法

1 引言

单细胞转录组测序技术能够在单个细胞水平上解析基因表达模式，在发育生物学[1]、自身免疫[2]及癌症研究[3]等领域展现出巨大的应用价值。然而，由于单细胞转录组数据普遍具有高度稀疏性、噪声较大以及维度高等特点，对其进行直接分析面临较大挑战。聚类分析能够在无监督条件下根据基因表达特征对细胞进行分组，从而揭示细胞间潜在的异质性与功能差异，因此已成为单细胞转录组数据分析中最为核心的方法之一。通过聚类分析，研究人员可以识别具有相似表达模式的细胞群体，进而探究细胞在组织结构中的组成关系，以及其在生理过程或疾病发生发展中的动态变化[4]。

传统的单细胞转录组聚类方法通常采用主成分分析[5]（Principal Component Analysis, PCA）对高维表达数据进行降维，并结合聚类算法（如 K-均值算法、Leiden 算法等）对细胞进行分组。然而，PCA 本质上是一种线性降维方法，假设数据分布于线性子空间中，在当前单细胞数据规模和复杂性不断提升的背景下[6]，其对复杂非线性细胞结构的刻画能力受到明显限制。此外，PCA 对异常值和噪声较为敏感，在处理具有高噪声、高稀疏性特征的单细胞转录组数据时稳定性较差，容易将技术噪声误识为主要成分，从而影响特征表示质量。

随着机器学习方法的发展，越来越多的研究将深度学习技术引入单细胞转录组聚类分析，以应对更加复杂的数据分布。其中，自编码器作为一种有效的表示学习模型，能够通过重构输入数据的方式学习其潜在低维表示，在细胞特征提取方面得到了广泛应用[7-9]。然而，传统自编码器的训练目标主要聚焦于最小化重构误差，缺乏对潜在空间中样本判别性结构的显式约束，导致其学习到的特征在聚类等下游任务中的表现仍存在一定局限。

为提升自编码器学习到的潜在表示在聚类任务中的判别能力，本文在传统自编码器框架基础上引入对比学习机制，通过联合优化重构损失与对比损失，引导模型在保留表达信息的同时，学习更加紧凑且具有良好可分性的聚类友好特征表示。具体而言，对同一细胞的表达数据施加不同的数据增强操作以构造正样本对，并通过编码器映射至潜在空间；对比

损失鼓励同一细胞不同视图的表示保持相似，同时拉开不同细胞之间的距离。模型训练过程中联合优化重构损失与对比损失，从而在保留原始表达信息的同时，增强潜在特征的判别性。实验结果表明，引入对比学习的自编码器在聚类任务中能够获得更加紧凑且分离良好的细胞簇。

2 方法

2.1 基线模型——自编码器

2.2 对比学习

3 实验

3.1 实验设置

3.2 结果

3.3 可视化分析

4 总结

参考文献

[1] Semrau S, Goldmann J E, Soumillon M, Mikkelsen T S, Jaenisch R, Van Oudenaarden A. Dynamics of lineage commitment revealed by single-cell transcriptomics of differentiating embryonic stem cells[J]. Nature communications, 2017, 8(1): 1096.

[2] Gaublomme J T, Yosef N, Lee Y, Gertner R S, Yang L V, Wu C, Pandolfi P P, Mak T, Satija R, Shalek A K, et al. Single-cell genomics unveils critical regulators of Th17 cell pathogenicity [J]. Cell, 2015, 163(6): 1400-1412.

[3] Patel A P, Tirosh I, Trombetta J J, Shalek A K, Gillespie S M, Wakimoto H, Cahill D P, Nahed B V, Curry W T, Martuza R L, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma[J]. Science, 2014, 344(6190): 1396-1401.

[4] Jiang S, Wang C, Sun Q, Zhang Z. A robust multi-scale clustering framework for single-cell RNA-seq data analysis[J]. Scientific Reports, 2025, 15(1): 18543.

[5] Pearson K. LIII. On lines and planes of closest fit to systems of points in space[J]. The London, Edinburgh, and Dublin philosophical magazine and journal of science, 1901, 2(11): 559-572.

[6] Lopez R, Regier J, Cole M B, Jordan M I, Yosef N. Deep generative modeling for single-cell transcriptomics[J]. Nature methods, 2018, 15(12): 1053-1058.

[7] Tian T, Wan J, Song Q, Wei Z. Clustering single-cell RNA-seq data with a model-based deep learning approach[J]. Nature Machine Intelligence, 2019, 1(4): 191-198.

[8] Li X, Wang K, Lyu Y, Pan H, Zhang J, Stambolian D, Susztak K, Reilly M P, Hu G, Li M. Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis[J]. Nature communications, 2020, 11(1): 2338.

[9] Tian T, Zhang J, Lin X, Wei Z, Hakonarson H. Model-based deep embedding for constrained clustering analysis of single cell RNA-seq data[J]. Nature communications, 2021, 12(1): 1873.