DOI:10. 20033/j. 1003-7241. (2025)12-0120-06

基于自监督的单细胞 ATAC-seq 数据智能聚类算法

宋宇航，李　猛，刘姿邑
（西安工程大学计算机科学学院，陕西 西安 710600）

摘　要：单细胞水平的染色质可及性测序技术（scATAC-seq）在研究表观遗传景观中的细胞异质性方面显示出巨大的潜力。针对 scATAC-seq 数据高稀疏性和高维度的特点，提出了一种基于自监督的单细胞 ATAC-seq 数据智能聚类算法。首先，通过两个并行的编码器对原始数据进行特征选择和特征增强以获得低维且有效的嵌入表示。然后，利用聚类模块生成的伪标签和分类网络的分类结果构建分类损失，实现自监督学习。最后，使用谱聚类算法对嵌入表示进行聚类。对比实验结果表明，该算法在 4 个数据集和 3 个评价指标上优于大部分对比算法。参数敏感性实验和收敛性实验进一步验证了所提算法的鲁棒性和快速收敛的能力。
关键词：聚类；自监督；scATAC-seq；深度神经网络
中图分类号：TP311. 13　　**文献标志码**：A　　**文章编号**：1003-7241(2025)12-0120-06

Data intelligent clustering algorithm for single-cell ATAC-seq data based on self-supervised

SONG Yuhang, LI Meng, LIU Ziyi
(School of Computer Science, Xi′an Polytechnic University, Xi′an 710600, China)

Abstract: Single-cell assay for transposase-accessible chromatin sequencing (scATAC-seq) shows great promise for studying cellular heterogeneity in epigenetic landscapes. Addressing the challenges of high sparsity and dimensionality in scATAC-seq data, this paper introduces a self-supervised intelligent clustering algorithm for single-cell ATAC-seq data. Firstly, two parallel encoders are employed to perform feature selection and enhancement on the raw data, obtaining a low-dimensional and effective embedding representation. Subsequent- ly, pseudo-labels generated by the clustering module and classification results from the classification network are utilized to construct a classification loss, achieving self-supervision. Finally, spectral clustering is applied to cluster the embedding representation. Comparative experimental results indicate that the proposed algorithm outperforms the comparison algorithms across four datasets and three evaluation metrics. Sensitivity experiments and convergence experiments further validate the robustness and fast convergence capability of the proposed algorithm.
Keywords: clustering; self-supervised; scATAC-seq; deep neural network

0　引言

染色质可及性的早期研究技术是染色质可及性测序技术（assay for transposase-acce-ssible chromatin with high-throughput sequencing, ATAC-seq）[1]，该方法通过改良的 Tn5 转座酶，捕捉开放染色质区域并在这些区域的两端连接测序接头，成为表观遗传学和基因调控研究中的重要工具。

近年来，单细胞转座酶可及染色质测序技术（scATAC-seq）的发展进一步推动了研究的深度和广度。相比传统 ATAC-seq，scATAC-seq 通过单细胞分辨率解析染色质可及性，不仅揭示了神经系统、肿瘤组织以及细胞发育等复杂生物系统的细胞多样性，还提供了探索细胞谱系发育过程中染色质状态动态变化的可能性[2]。研究人员借助这一技术，能够更精确地鉴定细胞亚型及其基因调控特性，为解读生命过程的复杂机制提供了新的视角。

然而，尽管 scATAC-seq 技术前景广阔，其在应用过程中仍面临许多技术挑战。例如，现有的无监督分析方法在细胞聚类效果上存在一定局限性，难以准确反映细胞的真实状态。此外，scATAC-seq 数据通常包含大量噪音，同时具有高维稀疏的特性，这为数据降维与特征提取带来了额外的困难[3]。如何在保持生物学信息完整性的同时，提高数据处理的效率和精确性，是目前研究中的重要课题。

聚类在计算机视觉、模式识别和生物信息学等领域至关重要[4]。由于目前针对 scATAC-seq 数据的方法通常需要在数十万维的细胞矩阵上执行线性降维，例如 SVD，因此将分析扩展到数百万个细胞具有非常大的挑战性。此外，使用 scATAC-seq 数据集对复杂组织中的细胞类型或状态进行无监督鉴定准确度较低。为了更好地对 scATAC-seq 数据进行聚类分析，我们需要根据 scATAC-seq 数据特点开发新的聚类方法。

为了解决上述问题，本文提出了基于自监督的单细胞 ATAC-seq 数据智能聚类算法（self-supervised intelligent

* 基金项目：陕西省自然科学基础研究计划（2024JC-YBMS-473）；陕西省教育厅重点项目（22JS019）
收稿日期：2024-01-11

clustering algorithm,SICA),主要贡献和创新点包含以下三方面。

1) 设计了一种新颖的特征降维方法:使用两个并行的编码器得到高维数据的候选特征和特征系数,并以此为依据获得有效的低维嵌入;

2) 引入了自监督的学习方法:在低维嵌入后引入一个分类网络,利用聚类模块生成的伪标签来促进分类网络的学习;

3) 通过在不同细胞数量和细胞类型的四个基准 scATAC-seq 数据集上测试 SICA,证明了 SICA 能够以自监督的方式揭示细胞亚群之间的差异并识别细胞类型。

1 相关工作

在过去几年中,已经提出了多种计算方法来解决 scATAC-seq 分析中的挑战。

1.1 基于加权聚类的方法

scABC[5]是一种用于 scATAC-seq 数据的无监督聚类的统计方法。它依赖于基因组区域内读取计数的模式来聚集细胞。为了解决稀疏性问题,scABC 首先通过不同 reads 数量估计细胞权重,然后应用加权的 K-medoids 聚类[6]将细胞划分为不同的组。为了提高分类精度,scABC 方法为每个聚类计算地标,代表每个集群的原型细胞,并选择每个集群中最具代表性的峰值作为特征。然后,scABC 根据 Spearman 相关性将细胞分配给最接近的地标。该方法在处理细胞异质性,尤其是细胞群体不均衡时,表现良好。然而,scABC 对高测序深度的地标样本有较高依赖,对于具有较多缺失值的 scATAC-seq 数据,可能无法准确计算 Spearman 秩。

1.2 基于概率模型的方法

cisTopic[7]是一个基于主题模型的无监督贝叶斯框架,用于将区域划分为调控主题,并根据其调控主题的贡献对细胞进行聚类。它们认为对细胞和调控区域进行共同优化的聚类可以改进对细胞状态的识别。cisTopic 采用潜在狄利克雷分布(LDA)[8]和坍缩吉布斯采样器[9]作为概率模型,通过同时优化 cell-topic 概率和 region-topic 概率来识别不同细胞中富集的顺式调控主题。这种概率模型的引入使得对细胞调控主题的推断更加灵活,为细胞类型的鉴定提供了更多的信息。

1.3 基于特征变换和降维的方法

Cusanovich 等人提出了一种分析流程[10],该分析流程迭代地执行频率逆文档频率变换(TF- IDF)和奇异值分解(SVD),以获得 scATAC-seq 数据的低维表示。尽管他们使用组合索引策略避免了细胞的物理分离,但是没有为数据处理及其下游分析建立一个明确的生物信息学分析流程。

Scasat[11]引入了另一种基于 Jaccard 相似性度量和多维缩放(MDS)的生物信息学分析流程来降低 scATAC-seq 数据的维度。它可以在几个简单的步骤中处理测序数据,而下游分析明确地考虑了染色质可及性的二元性质。

SnapATAC[12]以无偏的方式识别细胞类型,而不需要群体水平的峰值注释,从而在复杂组织中识别稀有细胞类型时具有卓越的灵敏度。另外它通过 Nyström 方法[13,14],显著降低了 CPU 和内存的使用,能够分析百万以上的大规模数据集。它主要通过评估细胞之间的染色质开放性的差异来区分细胞类型,每个细胞的染色质开放性通过二进制向量进行表示。所有细胞之间的染色质开放性差异最终通过 Jaccard 相似度矩阵的形式进行体现,然后应用主成分分析(PCA)进行降维。

1.4 基于深度学习的方法

最近深度生成模型已经成为表征学习和数据生成的强大框架[15]。

SCALE 是一种将变分自编码器(VAE)框架和高斯混合模型相结合的方法[16]。它利用 VAE 学习 scATAC-seq 数据的潜在特征,然后使用 K-means 对潜在特征进行聚类。

scDEC[17]通过以无监督的方式同时学习细胞的深度嵌入和聚类,它由一对生成对抗网络(GAN)组成。这种对称且配对的 GAN 架构最近已成功应用于图像样式转换[18]和密度估计[19]。

他们将这种 GAN 结构应用于无监督聚类的新任务,并将其应用于单细胞基因组数据的分析。scDEC 不需要借助外部方法来对潜在特征进行聚类,而是直接对细胞聚类过程进行建模。因此它可以同时学习潜在特征和细胞聚类。

2 基于自监督的 scATAC-seq 数据智能聚类算法

当涉及处理 scATAC-seq 数据时,目前大多数计算方法中,对细胞的低维表示求解和聚类的过程是分开进行的。然而,这种分离的方式可能导致前期的求解结果并不完全有利于后续的聚类过程。本文提出了一种名为 SICA 的算法模型,它引入了自监督学习的概念。在网络训练过程中,SICA 利用已有的结果来约束学习过程,从而提高了聚类的性能。

2.1 网络架构

本文提出了一种新的网络架构,同时学习输入数据的候选特征和特征系数,以突出候选特征中更有意义的特征。此外,利用训练过程中的结果进行聚类,生成伪标签给分类模块作为实际值来对训练过程进行自我监督,从而实现对整个网络性能的提升。SICA 算法的网络架构如图 1 所示。

SICA 的网络架构由卷积编码器模块和自监督模块组成。卷积编码器模块用于得到数据矩阵的候选特征 $\boldsymbol{Z}^{'} \in \boldsymbol{R}^{D\times N}$ 和特征系数 $\alpha \in R^{N}$。然后,为了强调 $\boldsymbol{Z}^{'}$ 的特征向量,特征系数按元素乘以候选特征中的每一个向量,得到一个低维的嵌入表示 $\boldsymbol{Z}$。$\boldsymbol{Z}$ 可以定义为

$$\boldsymbol{Z} = [\alpha_1 z_1^{'}, \alpha_2 z_2^{'}, \cdots, a_N z_N^{'}] \tag{1}$$

其中 $z_i^{'}$ 和 α_i 分别是候选特征和特征系数中的元素。

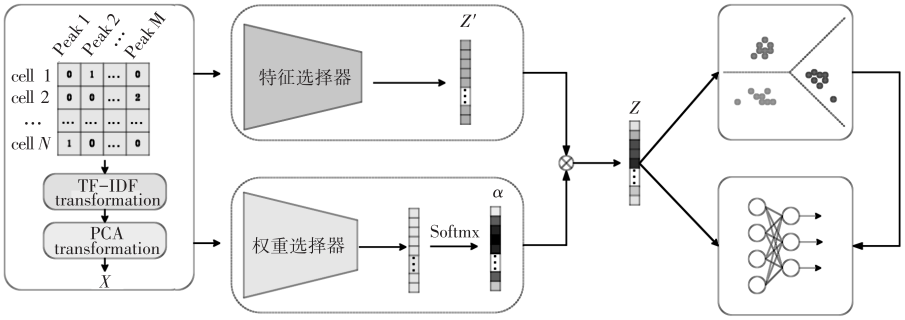

图 1　SICA 的网络架构图

自监督模块包括聚类模块和分类模块，聚类模块利用 $\boldsymbol{Z}$ 进行聚类，得到的聚类结果作为伪标签提供给分类模块。分类模块对数据进行分类，同时利用伪标签和分类得到的结果构建分类损失，实现自监督学习。

2.2　预处理

在将 scATAC-seq 数据输入 SICA 模型之前，进行了统一的预处理。首先，筛选出在超过 3% 细胞中至少有一次读取计数的峰（Peak），以排除在细胞中频次较低的峰，减少噪音。接着，对原始的 scATAC-seq 计数矩阵进行 TF-IDF 转换。TF-IDF 变换在信息检索和文本挖掘领域有广泛的应用[20,21]。在这一过程中，我们首先计算了每个细胞的总读数，并将其与原始读数矩阵进行比值运算，以获得“项频率”。接着，计算了所有细胞中每个开放区域的“文档反转频率”，并对其进行了对数转换，再与“术语频率”相乘。TF-IDF 转换的目的是提升在细胞中出现较少的峰的重要性，使其得到更高的权重。最后，为了进一步减少数据维度，我们应用 PCA 将数据的维数降至 20。

2.3　并行编码器

近年来，针对 scATAC-seq 数据，已经提出了多种聚类算法。此类数据通常包含数万到数十万个特征，其中许多特征可能不相关，且数据高度稀疏。因此，识别能够有效表达高维数据的特征至关重要。然而，在保持数据全局性质的同时，找到合适的特征是一个挑战。许多深度学习聚类方法采用自编码器进行特征选择，以挖掘潜在的聚类特征，但有时这些潜在特征的数量可能超过原始数据中的特征数量。此外，由于无监督学习无法利用样本标签，标签信息难以有效地参与特征提取过程。

为了提高聚类性能，我们加入了权重选择器，使潜在空间的特征向量更容易区分。权重选择器生成一个特征系数，在特征选择器得到的候选特征中突出更有意义的特征。

2.4　自监督学习

自监督学习因其能够避免标注大规模数据集的成本而受到欢迎。它能够采用自定义的伪标签作为监督，并将学习到的表示用于若干下游任务。

SICA 通过编码器模块得到了一个低维的嵌入表示 $\boldsymbol{Z}$。接下来，在 $\boldsymbol{Z}$ 上应用谱聚类可以得到数据样本的聚类结果，该结果可以用来作为数据集的伪标签。尽管它并不是在所有样本上都是正确的，但是它依然包含有用的信息。

由于编码器模块采用权重选择器来突出更好的特征，因此 $\boldsymbol{Z}$ 包含足够的信息来预测数据样本点的标签。所以分类模块通过聚类模块生成的伪标签作为分类的预期结果，就可以用来监督特征提取及整个网络的学习。

定义自监督的损失为

$$L_S = \frac{1}{N}\sum_{i=1}^{N}\left(\ln\left(1+e^{-y_i^T q_i}\right)+\lambda \left\| y_i-\mu_{c(y_i)}\right\|_2^2\right) \tag{2}$$

其中，y_i 为分类网络的输出，q_i 为对应聚类模块生成的伪标签。$c(y_i)$ 为该数据点所在聚类簇的索引，$\mu_{c(y_i)}$ 为对应聚类簇索引对应的聚类中心。自监督损失由两部分构成，一类是分类结果和伪标签的误差，用交叉熵表示，另一部分是聚类模块中样本点距离所属聚类簇中心的误差。

2.5　网络模型训练

SICA 算法的训练过程以下 2 个阶段：

1）训练阶段。随机初始化 SICA 网络架构中的训练参数。将数据矩阵 $\boldsymbol{X}$ 作为一个批量，使用 ADAM 优化器对整个网络进行训练。训练阶段的损失函数为

$$L = \frac{1}{N}\sum_{i=1}^{N}\left(\ln\left(1+e^{-y_i^T q_i}\right)+\lambda \left\| y_i-\mu_{c(y_i)}\right\|_2^2\right) \tag{3}$$

其中，λ 大于零。

2）聚类阶段。待步骤 1 的损失函数达到收敛后，得到低维嵌入 $\boldsymbol{Z}$，将其作为相似度矩阵，应用谱聚类算法得到聚类结果。

SICA 算法的算法步骤如下。

算法 1　SICA 算法

输入　数据矩阵 $\boldsymbol{X}$，超参数 λ，训练阶段迭代次数 T，学习率 β。

输出　聚类结果。

步骤 1　随机初始化卷积编码器的参数，随机初始化 $\boldsymbol{Z}$，令 $t=0$；

步骤 2　repeat

步骤 3　优化式（3），更新 SICA 的网络参数；

步骤 4　$t=t+1$；

步骤 5　until $t=T$；

步骤 6　将数据矩阵 $\boldsymbol{X}$ 输入训练好的 SICA 中得到低维嵌入 $\boldsymbol{Z}$；

步骤 7 对 Z 应用谱聚类算法得到聚类结果。

3 实验

3.1 数据集及评价指标

实验在 InSilico[22]、Forebrain[23]、Splenocyte[24]和 All blood[25]共 4 个数据集上进行。数据集的详细信息如表 1 所示。

表 1 数据集信息

数据集	细胞数	特征数	聚类个数
InSilico	1 377	68 069	6
Forebrain	2 088	140 102	8
Splenocyte	3 166	77 453	12
All blood	2 034	455 057	13

本文采用了 3 个常用的聚类算法评价指标，分别是 NMI(normalized mutual information)和 ARI(adjusted rand index)，Hom(homogeneity)。NMI 和 Hom 的取值范围均为[0,1]，ARI 取值范围为[1,-1]。数值越高代表聚类效果越好。

3.2 参数设置

对于卷积层，在两个维度上都将步长设置为 2，并使用 Sigmoid 作为激活函数。对于两个并行的编码器，本文使用统一的卷积核大小和通道数。使用学习率为 0.000 1 的 Adam 优化器训练网络参数。SICA 在各个数据集上的卷积核的大小和通道数如表 2 所示。

表 2 参数信息

数据集	编码器	
	卷积核大小	通道数
InSilico	3×3,3×3,3×3	10,20,30
Forebrain	3×3,3×3,3×3	10,20,30
Splenocyte	3×3,3×3,3×3	10,20,30
All blood	3×3,3×3,3×3	10,20,30

对聚类个数、batch-size 这样的训练参数，我们通过使用轮廓系数来选择最佳模型。然后对模型进行微调。

以 InSilico 数据集为例，事先并不知道它包含多少个类别。因此我们预先选择了一组可能的聚类个数，范围从 2 到 10。然后对于每个聚类个数 k，保持其他参数不变，将 k 值输入到 SICA 模型中进行训练。最后从 SICA 得到低维的嵌入表示，通过谱聚类计算表示相应的轮廓系数值。实验结果如图 2 所示。

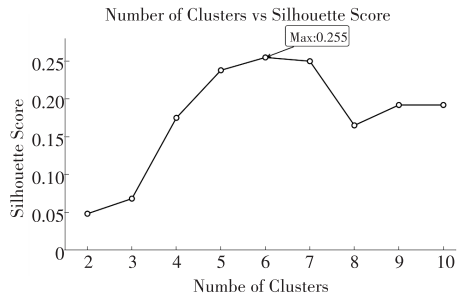

图 2 k 值的轮廓系数实验结果

如图所示，当聚类个数达到 6 时，轮廓系数的值达到最大。因此我们选择 6 作为最终的聚类数，类似的，我们可以对其他的训练参数进行类似的操作，然后进行微调。

3.3 对比实验

为了验证算法的聚类效果，将 SICA 与 6 种基准方法(scABC、SCALE、cisTopic、Cusanovich2018、Scasat、SnapATAC 和 scDEC)进行对比。评估方法性能的标准是 1)是否可以在低维空间中清晰地分离不同的细胞亚群；2)是否可以通过聚类准确地推断出真实的细胞类型标签。为了解决第一个问题，我们应用每种方法进行降维或提取潜在特征。对于所有的数据集，将维度设置为 20。

为了解决第二个问题，我们通过对比实验来验证。实验结果如表 3 到表 5 所示，将最优值用加粗标记，次优值用下划线标记。

表 3 不同数据集的 NMI 值

算法	数据集			
	InSilico	Forebrain	Splenocyte	All blood
scABC	0.797	0.610	0.670	0.461
SCALE	0.786	0.706	0.682	0.425
cisTopic	0.781	0.646	0.634	**0.532**
Cusanovich2018	0.638	0.616	0.730	0.514
Scasat	0.758	**0.730**	0.652	0.419
SnapATAC	0.717	0.646	0.658	0.467
scDEC	0.854	0.638	0.829	0.520
SICA	**0.879**	0.653	**0.860**	0.508

表 4 不同数据集的 ARI 值

算法	数据集			
	InSilico	Forebrain	Splenocyte	All blood
scABC	0.857	0.491	0.640	0.227
SCALE	0.857	0.664	0.461	0.215
cisTopic	0.738	0.556	0.377	0.299
Cusanovich2018	0.506	0.479	0.437	0.281
Scasat	0.638	**0.670**	0.371	0.215
SnapATAC	0.573	0.461	0.407	0.233
scDEC	0.907	0.543	0.871	0.309
SICA	**0.908**	0.547	**0.889**	**0.322**

表 5 不同数据集的 Hom 值

算法	数据集			
	InSilico	Forebrain	Splenocyte	All blood
scABC	0.839	0.568	0.724	0.443
SCALE	0.798	0.730	0.766	0.467
cisTopic	0.846	0.682	0.736	**0.574**
Cusanovich2018	0.739	0.640	0.844	0.562
Scasat	0.848	**0.748**	0.754	0.538
SnapATAC	0.801	0.646	0.766	0.508
scDEC	0.866	0.624	0.835	0.532
SICA	**0.883**	0.649	**0.854**	0.521

由表 3 到表 5 可以发现，在大部分的比较中，SICA 算法获得了最佳的聚类性能，这表明了该算法较其他算法的优越性。具体来说：1)基于深度学习的聚类算法的聚类性能明显优于传统的聚类算法，这表明非线性映射有助于提取适合于 scATAC-seq 数据的聚类表示。例如，在 InSilico

数据集上,本文提出的 SICA 算法比传统的 scATAC-seq 数据聚类算法 scABC 的 NMI、ARI 和 Hom 值分别提高了 8.2%、5.01%和 4.4%。2)SICA 算法优于其他基于深度学习的 scATAC-seq 数据聚类算法。这是由于 SICA 算法可以同时学习输入数据的候选特征和特征系数,提升深度模型提取特征的质量。并且 SICA 算法通过聚类模块生成的伪标签来促进分类模块的训练,提升了分类网络的性能。

3.4 参数敏感性实验

为了探究参数 λ 对模型性能的影响,在实验中调整参数 λ 的范围为 0.001 到 1 000。使用 In-Silico 数据集进行敏感性实验,实验结果如图三所示。

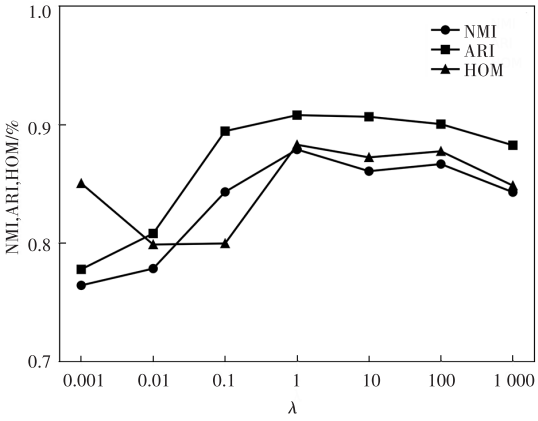

图 3 λ 的敏感性实验结果

实验结果显示,SICA 算法在 λ 为 1 到 1 000 的范围内能较好的进行聚类。具体来说,λ 是聚类损失的超参数,它通常用于衡量样本间的相似性。较大的 λ 可以使模型更加关注簇内的一致性,即同一簇内的样本更加相似,有助于提高聚类效果。并且聚类损失用于将样本划分到不同的簇,λ 较大时,模型将更加关注于样本的正确划分,降低噪声对于聚类效果的干扰。

3.5 收敛性实验

图 4 给出了 SICA 算法在 InSilico 数据集上的收敛曲线。可以发现,随着迭代数的增加,损失值不断下降,并在经过了 800 次迭代之后,损失值趋于稳定。因此,该算法具有较好的收敛性。

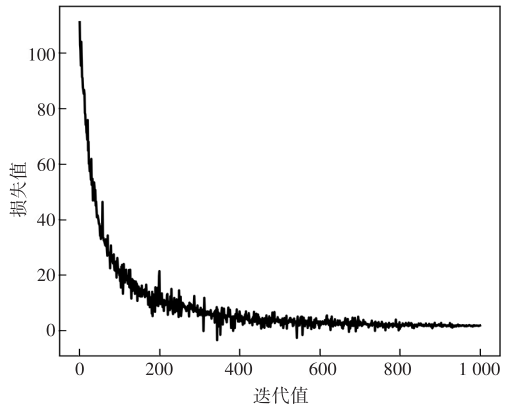

图 4 在 InSilico 数据集上的收敛曲线

4 结束语

本文提出了一种基于自监督学习的单细胞 ATAC-seq 数据智能聚类算法(SICA)。该算法通过同时学习输入数据的候选特征和特征系数,有效地促使并行编码器生成更加有意义的嵌入表示,从而显著提升聚类的准确性和效果。通过聚类模块生成的伪标签能够进一步指导分类模块的学习,这种自监督的策略帮助模型在缺乏标注数据的情况下,优化其训练过程。实验结果表明,SICA 算法在聚类性能方面远超现有的其他算法,显示出其在处理复杂的 scATAC-seq 数据时的优势。

在未来的工作中,计划引入字典学习方法。通过构建一个完备的字典,原始数据可以通过字典中的少量原子进行稀疏表示,这一技术有助于提取数据中的重要特征。字典学习不仅能帮助模型更好地识别数据中的结构性信息,还能进一步提升聚类结果的精度。

参考文献

[1] 郭晶鑫. 单细胞技术在 Lepr^{+}骨髓基质细胞分析中的应用及单细胞 ATAC-seq 分析软件的开发[D]. 杭州:浙江大学, 2022.

[2] SATPATHY A T, GRANJA J M, YOST K E, et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cellexhaustion[J]. Nature biotechnology, 2019, 37(8):925-936.

[3] CHEN H, LAREAU C, ANDREANI T, et al. Assess ment of computational methods for the analysis of single-cell ATAC-seq data[J]. Genome biology, 2019, 20(1):1-25.

[4] 吴浩, 罗少辉, 李颖昕,等. 基于聚类挖掘的科技数据价值动态监测方法[J]. 自动化技术与应用, 2024, 43(2):81-84, 106.

[5] ZAMANIGHOMI M, LIN Z, DALEY T, et al. Unsupervised clustering and epigenetic classification of single cells [J]. Nature communications, 2018, 9(1):2410.

[6] STUDER M. WeightedCluster library manual [J]. Pract. Guide Creat. Typol. Trajectories Soc. Sci, 2013,2013(24):2296-1658.

[7] BRAVO GONZÁLEZ-BLAS C, MINNOYE L, PAPASO KRATI D, et al. cisTopic:cis-regulatory topic modeling on single-cell ATAC-seq data[J]. Nature methods, 2019, 16(5):397-400.

[8] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation [J]. Journal of machine Learning research, 2003, 3(1):993-1022.

[9] GRIFFITHS T L, STEYVERS M. Finding scientific topics[J]. Proceedings of the National academy of Sciences, 2004, 101 (Sup. 1): 5228-5235.

[10] CUSANOVICH D A, DAZA R, ADEY A, et al. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing[J]. Science, 2015, 348(6237):910-914.

[11] BAKER S M, ROGERSON C, HAYES A, et al. Classifying cells with Scasat, a single- cell ATAC-seq analysis tool[J]. Nucleic acids research, 2019, 47(2):10.

[12] FANG R, PREISSL S, LI Y, et al. Comprehen sive analysis of single cell ATAC-seq data with SnapATAC[J]. Nature communications, 2021, 12(1):1337.

[13] KUMAR S, MOHRI M, TALWALKAR A. Sampling methods for the Nyström method[J]. The Journal of Machine Learning Research, 2012, 13(1):981-1006.

(下转第 188 页)