Advances in Applied Mathematics 应用数学进展, 2022, 11(10), 6951-6958
Published Online October 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.1110737

Hans汉斯

基于t-SNE与UMAP降维的细胞分类及差异化基因筛选研究

李元夫

南京信息工程大学数学与统计学院，江苏 南京

收稿日期：2022年9月11日；录用日期：2022年10月2日；发布日期：2022年10月11日

摘　要

单细胞RNA测序技术已经广泛地应用于细胞异质性等关键生物学问题的研究中，与此同时该技术的发展也为基因数据分析提出了很大的挑战。本文基于t-SNE和UMAP两种非线性降维方法，对单细胞RNA数据进行降维、聚类并与线性主成分降维聚类结果进行对比，得出结论：UMAP方法针对单细胞RNA数据降维聚类的效果更为理想。最后以UMAP非线性降维聚类的结果为例筛选出不同细胞类别中的显著差异化基因。

关键词

单细胞RNA测序，t-SNE，UMAP，显著差异化基因

Cell Classification and Differential Gene Screening Based on t-SNE and UMAP Dimension Reduction

Yuanfu Li

School of Mathematics and Statistics, Nanjing University of Information Science & Technology, Nanjing Jiangsu

Received: Sep. 11th, 2022; accepted: Oct. 2nd, 2022; published: Oct. 11th, 2022

Abstract

Single-cell RNA sequencing technology has been widely used in key biological problems such as cell heterogeneity, and at the same time, the development of this technology also poses great challenges in gene data analysis. In this paper, based on two nonlinear dimensionality reduction methods, t-SNE and UMAP, the dimensionality reduction and clustering of single-cell RNA data were carried out and compared with the results of linear principal component dimensionality reduction

文章引用: 李元夫. 基于 t-SNE 与 UMAP 降维的细胞分类及差异化基因筛选研究[J]. 应用数学进展, 2022, 11(10): 6951-6958. DOI: 10.12677/aam.2022.1110737

clustering. The conclusion was drawn that the UMAP method was more ideal for the dimensionality reduction clustering of single-cell RNA data. Finally, the results of UMAP nonlinear dimensionality reduction clustering were taken as an example to screen out the significantly differentiated genes in different cell categories.

Keywords

Single-Cell RNA Sequencing, t-SNE, UMAP, Significantly Differentiated Genes

1. 引言

近些年来，单细胞 RNA 测序技术在众多领域都有着重大进展，已然成为了研究细胞动力学的一种强有力的工具。单细胞 RNA 测序技术不但能够帮助基因学者更好地探索单细胞水平的基因表达谱，了解在不同组织微环境下细胞表达及功能差异，也为发现罕有细胞、细胞谱系分化轨迹提供了前所未有的机会。目前，单细胞 RNA 测序已成为研究细胞异质性的关键生物学问题的首选。该类技术的进步为单细胞 RNA 数据分析提供大量数据的同时，也在基因数据分析方向上提出了挑战，其中包括如何根据测序信息进行准确的细胞分类，并找到细胞差异化表达的基因等一系列问题。

在单细胞 RNA 测序数据的细胞聚类问题上，Kiselev，Kirschner，Schaub [1]等人提出了一种具有一致性的单细胞聚类算法 SC3，该方法先对基因数据应用主成分分析法将基因维度降到总基因的 4%到 7%，然后对所有主成分进行 k-means 聚类，最后对所有结果进行一致性聚类，经对比该聚类算法具有良好的稳定性；而 Guo，Hui，Steven [2]等人提出了一种新的单细胞 RNA 序列分析流程 SINCERA，在该分析过程中运用了无监督的分层聚类，并运用 t 检验和 Wilcoxon 秩和检验筛选出差异性表达基因；Yang，Liu，Lu [3]等人提出了一种迭代聚类算法 SAIC，该聚类算法需要提前选取聚类的数目和阈值，然后再进行聚类，并通过方差分析选出显著差异性表达基因。

伴随着机器学习技术的飞速发展，包括主成分分析在内的一些机器学习方法广泛地应用于基因数据分析领域。以常用的线性主成分分析方法为例，在对细胞基因数据分析时，同类细胞会被同一主成分解释，而不同主成分间的线性无关则代表着细胞间的部分差异。但基因表达数据具有强稀疏的特殊性，也即基因的维度远远大于细胞的维度，这与线性主成分样本数据服从高斯分布的前提不符，使得上述方法在单细胞 RNA 测序数据降维聚类的问题上受到限制。因此需要选择非线性降维方法来对单细胞 RNA 测序数据进行降维聚类分析。

本文应用 t-SNE 和 UMAP 两种非线性降维聚类方法，对单细胞 RNA 测序数据进行降维聚类分析，并将聚类结果与线性主成分分析方法降维聚类的结果进行对比，通过比较聚类图中细胞类内、类间的距离属性得出 UMAP 方法降维聚类效果更为理想，进而在其聚类结果的基础上筛选出显著差异化表达基因。

2. 基于 t-SNE 与 UMAP 的降维方法

2.1. t-SNE 降维

2.1.1. 随机邻近嵌入(Stochastic Neighbor Embedding, SNE)

在介绍 t 分布－随机邻近嵌入(t-distributed Stochastic Neighbor Embedding, t-SNE)方法之前，首先介绍

随机邻近嵌入(Stochastic Neighbor Embedding, SNE)方法。SNE 首先将数据点间的高维欧式距离转换成了代表着相似度的条件概率[4]。将数据点 x_i 到数据点 x_j 的相似度用条件概率 $p_{j|i}$ 来表示，也就是说 x_i 将会以概率 $p_{j|i}$ 挑选 x_j 为其临近点。那么在以 x_i 为中心点的 Gauss 分布下，可以定义 x_i 与 x_j 的相似度为：

$$p_{j|i}=\frac{\exp\left(-\left\|x_i-x_j\right\|^2/2\sigma_i^2\right)}{\sum_{k\neq i}\exp\left(-\left\|x_i-x_k\right\|^2/2\sigma_i^2\right)}, \tag{1}$$

其中 σ_i 是以 x_i 为中心点的 Gauss 分布的方差，$\|\cdot\|$ 代表欧式范数。如果数据点距离越近，那么 $p_{j|i}$ 越大；如果数据点距离越远，那么 $p_{j|i}$ 越趋近于无穷小。

对于和高维空间中的点 x_i 和 x_j 相对应的在低维空间中的点 y_i 和 y_j，同样可得点 y_j 到点 y_i 的相似度 $q_{j|i}$，在 Gauss 分布下，方差为常数值 $\frac{1}{\sqrt{2}}$ 时，

$$q_{j|i}=\frac{\exp\left(-\left\|y_i-y_j\right\|^2\right)}{\sum_{k\neq i}\exp\left(-\left\|y_i-y_k\right\|^2\right)}. \tag{2}$$

显然，如果点 y_j 到点 y_i 可以完全模拟高维数据点 x_j 到数据点 x_i 的相似度，那么相应的条件概率 $q_{j|i}$ 和 $p_{j|i}$ 应该完全相同。若考虑高维空间下所有点与 x_i 的相似度，则可构成一个条件概率分布 $P_i=\left(p_{k|i}\right)_{k\neq i}$，同理在低维空间存在一个条件概率分布 Q_i 且应该与 P_i 一致。由此可选用 Kullback-Leibler 距离(KL 距离)来衡量两个分布的相似性，构造(3)式代价函数 C，运用梯度下降的方法来使 C 达到最小来确定低维空间，具体形式如下：

$$C=\sum_i KL\left(P_i\|Q_i\right)=\sum_i\sum_j p_{j|i}\log\frac{p_{j|i}}{q_{j|i}}, \tag{3}$$

其梯度为：

$$\frac{\partial C}{\partial y_i}=2\sum_j\left(p_{j|i}-q_{j|i}+p_{i|j}-q_{i|j}\right)\left(y_i-y_j\right). \tag{4}$$

2.1.2. t-分布随机邻近嵌入(t-distributed Stochastic Neighbor Embedding, t-SNE)

虽然 SNE 也可以对数据进行降维，但是 SNE 注重的是局部结构而非全局结构。由于 t 分布是重尾分布，对异常点不是十分敏感，所以在对称 SNE 的基础上将低维空间中的分布选取为 t 分布。这样一来，在降维过程中，t 分布的处理会使同一簇中的数据更加紧密，不同簇之间的数据更加稀疏，效果显著提高。自由度为 1 的 t 分布重新定义的 $q_{j|i}$ 形式如下：

$$q_{j|i}=\frac{\left(1+\left\|y_i-y_j\right\|^2\right)^{-1}}{\sum_{k\neq i}\left(1+\left\|y_i-y_k\right\|^2\right)^{-1}}, \tag{5}$$

相应的梯度形式为：

$$\frac{\partial C}{\partial y_i}=4\sum_j\left(p_{j|i}-q_{j|i}\right)\left(y_i-y_j\right)\left(1+\left\|y_i-y_j\right\|^2\right)^{-1}. \tag{6}$$

2.2. UMAP 降维

均匀流行逼近和投影(Uniform Manifold Approximation and Projection, UMAP)是一种基于黎曼几何和

代数拓朴的理论框架构建的流行学习算法。它依据高维空间映射到低维空间相似度的定性结论，将高维数据的拓扑结构进行低维映射以达到降维结果[5]。

在 UMAP 降维过程中，设高维数据点 $X=\{x_1,\cdots,x_n\}$，低维数据点 $Y=\{y_1,\cdots,y_n\}$。令 $d:X\times X\to R^+$ 为其度量空间，给定一个超参数 k，可以得到 x_i 在 d 下的近邻集合 $\{x_{i1},\cdots,x_{ik}\}$。那么高维模糊拓扑结构可使用指数概率分布表示如下：

$$p_{j|i}=\mathrm{e}^{-\frac{d(x_i,x_j)-\rho_i}{\sigma_i}}, \tag{7}$$

其中，ρ_i 是点 x_i 到首个最近邻数据点的距离。σ_i 是 x_i 最近邻数据点的直径。需要注意的是，这并不是一个对称函数，所以应该将该函数对称化：

$$p_{j|i}\triangleq p_{j|i}+p_{i|j}-p_{j|i}p_{i|j}. \tag{8}$$

在高维分布中建立模糊拓扑后，需要在低维分布中构建概率分布：

$$q_{j|i}=\left(1+a\left(y_i-y_j\right)^{2b}\right)^{-1}, \tag{9}$$

其中超参数 $a=1.93,\ b=0.79$ [6]。

UMAP 希望相关数据点在投影空间中尽可能地靠近，而不相关数据点尽可能地远离。因此，引入如下函数：

$$\begin{cases}\text{Attractive}=p_{j|i}(X)\log\left(\dfrac{p_{j|i}(X)}{q_{j|i}(Y)}\right)\\ \text{Repulsive}=\left(1-p_{j|i}(X)\right)\log\left(\dfrac{1-p_{j|i}(X)}{1-q_{j|i}(Y)}\right)\end{cases}. \tag{10}$$

在上式中，$p_{j|i}$ 是高维分布中数据点的权重，$q_{j|i}$ 是低维分布中数据点的权重。算法首先对数据集中相似的数据点施加 Attractive (引力)，并对非相似的数据点施加 Repulsive (斥力)，接着通过模拟退火优化算法逐渐减小 Attractive 和 Repulsive，最终达到收敛。

3. 实例分析

在进行正式的降维聚类工作之前，我们对单细胞 RNA 测序数据[7]进行前期的预处理工作，在描述性分析工作中发现该基因数据具有较强的稀疏性，又根据坏死细胞和破损细胞的基因表现特征对随细胞进行筛选和质量控制，最后对数据进行标准化处理，得到基因种类 8315 个、细胞个数 750 个的单细胞测序数据。接下来对处理过后的基因数据进行降维聚类分析。

3.1. 降维聚类结果

为比较 t-SNE 与 UMAP 两种非线性降维聚类方法在基因测序数据应用的优越性，我们先对预处理后的单细胞 RNA 测序数据进行线性主成分降维聚类，聚类结果见图 1。

而后我们对预处理后的单细胞 RNA 数据依 UMAP 及 t-SNE 两种非线性方法进行降维聚类[8]，细胞聚类结果见图 2。

观察上文中两细胞聚类散点图我们可以发现三种降维聚类方法都将细胞分为 0~4 共五个类别，但从聚类效果来看，图 1 中线性主成分降维聚类结果类内点位距离较大、类间点位距离较小、类边缘界限模糊不清，且不同细胞类别之间有不同程度的位置重合。图 2 中左图为应用 UMAP 方法后的降维聚类的结

果，右图为应用 t-SNE 方法后降维聚类的结果，通过对比可以明显的观察到 UMAP 方法的聚类结果中不同细胞类别类间距离更大，类内距离更小，且细胞点位分布更密集；而相较于 UMAP 聚类结果，t-SNE 方法的聚类结果中类间细胞点位距离较小、类内细胞点位距离较大且细胞点位分布较为分散。综上所述，我们认为非线性 UMAP 降维聚类的效果最为理想。

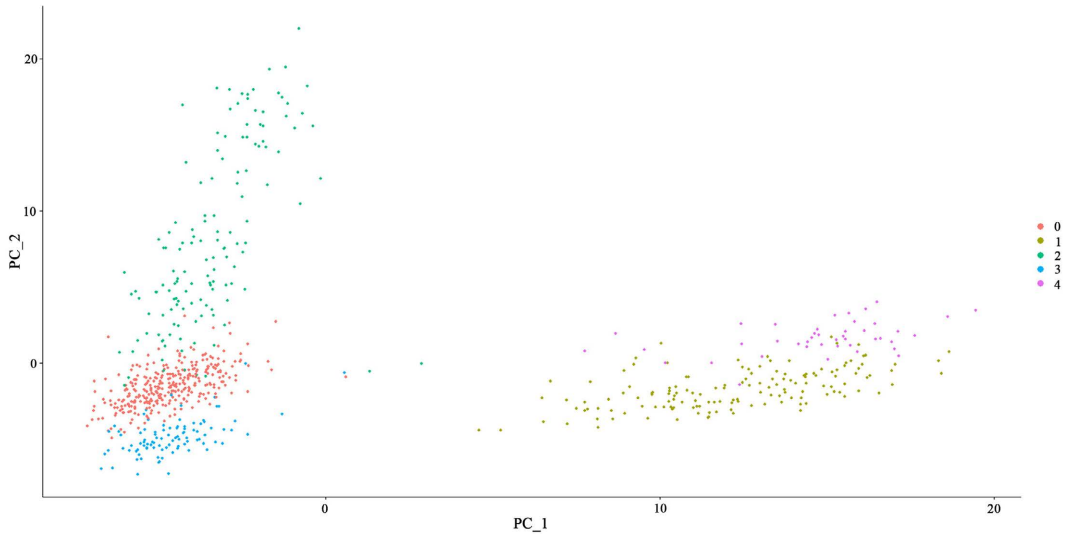

Figure 1. Cell scatter plot based on linear principal component dimensionality reduction (PCA) clustering
图 1. 基于线性主成分降维(PCA)聚类的细胞散点图

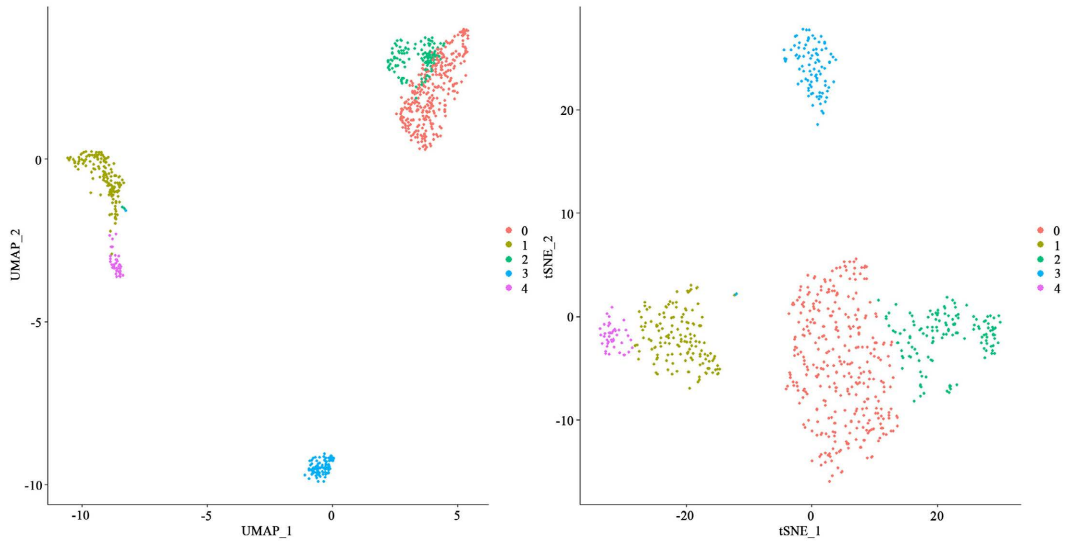

Figure 2. Cell scatter plot based on UMAP and t-SNE nonlinear dimensionality reduction clustering
图 2. 基于 UMAP 和 t-SNE 非线性降维聚类的细胞散点图

3.2. 显著差异性基因筛选结果及分析

细胞分类的不同往往可能源于其显著表达基因的差异化，UMAP 非线性降维方法的聚类结果能有效帮助探索识别具有显著差异性的基因。我们计算基因在不同细胞类别表达的细胞比例及其差值、基因在不同类别之间表达量差异的倍数，并依照假设：

H0：基因在两组别之间没有显著性差异；

H1：基因在两组数据之间有显著性差异。

进行 Wilcoxon 秩和检验，在此结果上进一步筛选出各类别的显著差异化基因。

以类别 0 的显著差异化基因为例生成部分差异基因列表(见表 1)，其中基因在类别 0 的细胞内与其他类别细胞内的含量差值(Difference)为降序排列，也即基因表达的差异化依表顺序依次递减。

表 1 中各基因的 avg_log2FC 值均为正值，说明这些基因在类别 0 的细胞群体内平均表达的折叠度高于其他类别的细胞，且由 Wilcoxon 秩和检验得出的 p_val 值远小于 $\alpha = 0.05$ 的临界值，也即拒绝基因在两组别之间没有显著差异的原假设。基因在 0 类别细胞内与其他类别细胞内的含量差值(Difference)均为正值，这也进一步佐证表中基因为 0 类别细胞区别于其他类别细胞的显著差异性基因。

Table 1. Some significantly different genes with tag category 0
表 1. 部分标签类别为 0 的显著差异性基因

gene	p_val	avg_log2FC	pct.1	pct.2	Difference
CD3D	2.42E−53	1.435291	0.884	0.252	0.632
IL7R	1.41E−42	1.576791	0.687	0.185	0.502
LDHB	6.67E−78	1.941558	0.942	0.442	0.5
IL32	7.69E−30	0.974124	0.827	0.34	0.487
CD3E	2.14E−39	1.140736	0.784	0.299	0.485

注：gene：基因名称；FC (fold change)：组间基因表达量的差异倍数；avg_logFC：基因在组间的平均表达量取 log2，正值表明该基因在当前组中表达更高；pct.1：在当前类别细胞中检测到该基因表达的细胞比例；pct.2：在其它类别细胞中检测到该基因表达的最大细胞比例；p_val：Wilcoxon 秩和检验所得 p 值；Difference：pct.1 − pct.2 [9]。

因此，我们分别取每组表达最高且与其他类别细胞之间含量的差值最大的前五个基因作为显著差异性基因的结果，它们分别是 CD3D、IL7R、LDHB、IL32、CD3E (类别 0)；S100A8、LGALS2、FCN1、S100A9、LST1 (类别 1)；NKG7、CST7、GZMA、CCL5、CTSW (类别 2)；CD79A、CD79B、MS4A1、TCL1A、LINC00926 (类别 3)；FCGR3A、IFITM3、CD68、CFD、CFP (类别 4)。

3.3. 基于基因分布热图、分布点图的差异性基因验证

为了以更加直观的方式展示上述各基因的显著差异性，我们绘制了部分含有不同显著差异性基因的细胞分布热图和部分显著差异性基因的分布点图。

我们选择每组显著差异性基因的前两个共计 10 个基因绘制分布热图，如图 3 所示，其中图例中颜色深浅为表示基因分布的密度标识：数值越大、颜色越深，说明差异性基因在对应分类区域中密度越大。从图中我们可以清晰地观察到，每个基因在细胞聚类图中的分布都较为集中，说明该基因能够较为准确的代表所属聚类区域细胞的基因特征属性。

而以 IL7R、S100A8、NKG7、CD79A、FCGR3A 这五个显著差异基因为例，观察其基因分布点图，如图 4 所示，坐标轴的横坐标为显著差异性基因名称，纵坐标为细胞类别，其中点的大小代表含有该基因的细胞比例，颜色深浅代表平均表达水平，通过比较图中分布点的颜色和大小，我们可以清楚地判断上述基因在各自不同的细胞类别内有着较高的表达水平，且表达程度远远大于其他组别。这也进一步说明筛选出的差异化基因在细胞分类的过程中起到了显著作用。

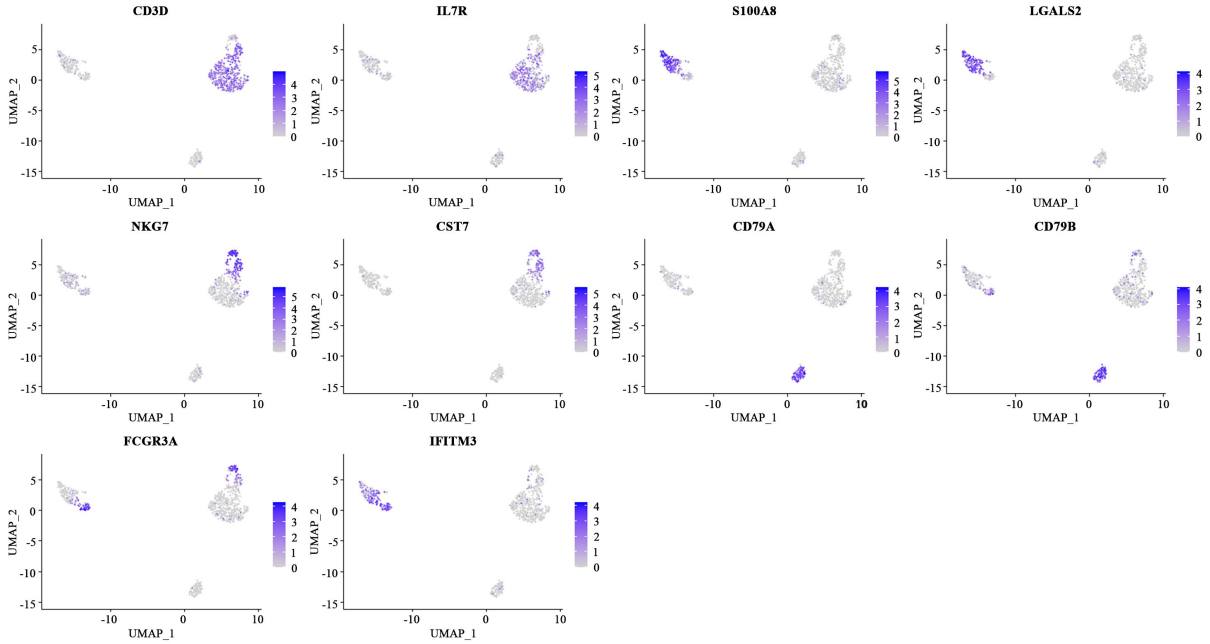

Figure 3. Heat map of significantly different gene distribution in different groups
图 3. 不同组别下显著差异性基因分布热图

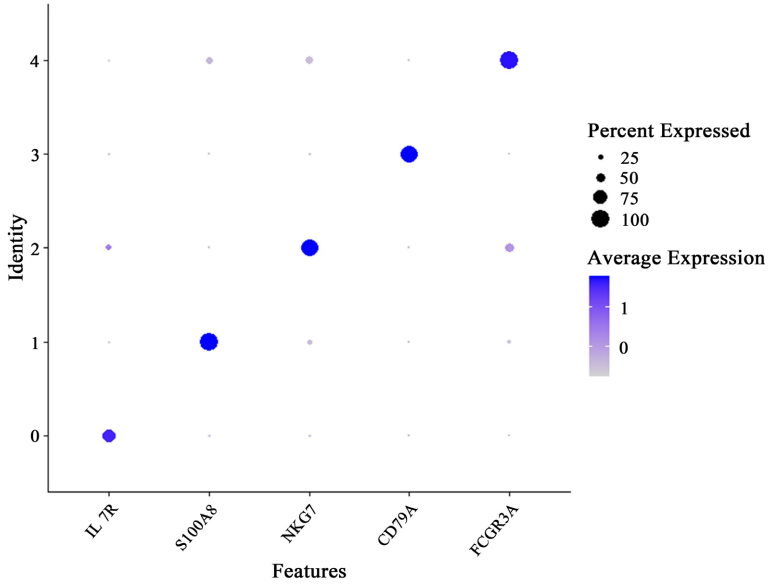

Figure 4. Dot plot of the distribution of significantly different genes in different classes
图 4. 显著差异性基因在不同类别中的分布点图

4. 模型评价

通过对上文中单细胞 RNA 测序数据降维聚类的结果进行对比和对不同降维聚类模型特点进行归纳总结，我们得出以下结论：

线性主成分降维聚类方法受线性模型的限制，在特征选择的过程中没有进行基因筛选，降维后的基因数据对细胞分类的解释能力相对减弱，且细胞分类的区分效果在基因投影之后有所削减，反而可能使

部分细胞混杂在一起难以准确区分[10]。

不同于线性主成分降维聚类的方法，t-SNE 和 UMAP 可以直接将高维空间的结构特征投影到低维空间中，通俗地讲，就是用平面或立体空间内点的疏密远近表现其在原本多维度状态下的疏密远近，这样能够更大程度上保留细胞原有的基因的差异化特征。而 UMAP 又是一种非线性流形学习算法，相比于线性主成分降维方法只适用于高斯分布样本的限制，UMAP 在非高斯大容量样本上有着良好的性能。在数据降维过程中，UMAP 可以保留样本全局结构，最大程度上维持样本数据完整性，并且不需要人工去选择核函数，规避了人工选择核函数对其降维性能的影响，因此，UMAP 比 t-SNE 和线性主成分降维方法有更强的工程适用性。

基金项目

国家自然科学基金面上项目：超高维复杂数据统计降维研究(11771215)，2018.1~2021.12。

参考文献

[1] Kiselev, V.Y., Kirschner, K., Schaub, M.T., *et al*. (2017) SC3: Consensus Clustering of Single-Cell RNA-Seq Data. *Nature Methods*, **14**, 483-486. https://doi.org/10.1038/nmeth.4236

[2] Guo, M., Wang, H., Potter, S.S., *et al*. (2015) SINCERA: A Pipeline for Single-Cell RNA-Seq Profiling Analysis. *PLOS Computational Biology*, **11**, e1004575. https://doi.org/10.1371/journal.pcbi.1004575

[3] Yang, L., Liu, J., Lu, Q., *et al*. (2017) SAIC: An Iterative Clustering Approach for Analysis of Single Cell RNA-Seq Data. *BMC Genomics*, **18**, 689-697. https://doi.org/10.1186/s12864-017-4019-5

[4] Van der Maaten, L. and Hinton, G. (2008) Visualizing Data Using t-SNE. *Journal of Machine Learning Research*, **9**, 2679-2605.

[5] Wang, Y., Huang, H, Rudin, C, *et al*. (2021) Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMAP, and PaCMAP for Data Visualization. *Journal of Machine Learning Research*, **22**, 1-73. https://doi.org/10.48550/arXiv.2012.04456

[6] 顾君垚, 丁强, 夏宇栋, 江爱朋, 丁晓雯. 基于 UMAP-AdamDD 的冷水机组故障诊断方法[J]. 低温与超导, 2022, 50(1): 81-87.

[7] http://mas.ruc.edu.cn/syxwlm/MASkx/5da681cd2206452ebebc141ff5121548.htm

[8] Kiselev, V.Y., Andrews, T.S. and Hemberg, M. (2019) Challenges in Unsupervised Clustering of Single-Cell RNA-Seq Data. *Nature Reviews Genetics*, **20**, 273-282. https://doi.org/10.1038/s41576-018-0088-9

[9] Suvà, M.L. and Tirosh, I. (2019) Single-Cell RNA Sequencing in Cancer: Lessons Learned and Emerging Challenges. *Molecular Cell*, **75**, 7-12. https://doi.org/10.1016/j.molcel.2019.05.003

[10] 吴德亮. 基于降维与聚类的单细胞 RNA 测序数据分析[D]: [硕士学位论文]. 哈尔滨: 哈尔滨工业大学, 2018.