SDAEC 算法在单细胞测序数据批次校正中的应用*

王文杰[1] 李 康[1] 谢宏宇[2△]

【提 要】 **目的** 提出深度堆叠降噪自编码嵌套聚类(stacked denoising auto encoder embedded cluster,SDAEC)算法并用于单细胞 mRNA 测序(single cell mRNA sequence,scRNA-seq)数据的批次效应移除,对其移除批次效应性能进行评估。**方法** 基于单细胞数据具有高维度、高稀疏性及高度非线性误差特点,通过将单细胞 Louvain 聚类算法嵌入堆叠降噪自动编码器(stacked denoising auto encoder,SDAE)算法中,形成 SDAEC 算法,用于单细胞测序数据的批次效应移除。结合实际卵巢癌组织 scRNA-seq 数据,利用分布邻域嵌入(t-distributed stochastic neighbor embedding,tSNE)、k 最近邻批次效应检测(k-nearest-neighbor batch-effect test,kBET)、调整兰德系数(adjusted rand index,ARI)、标准化互信息(normalized mutual information,NMI)、平均轮廓宽度(average silhouette width,ASW)评价其移除批次效应性能。**结果** 利用 SDAEC 方法对 scRNA-seq 数据批次效应移除性能高于 Combat、相互最近邻(mutual nearest neighbors,MNN)、分布匹配残差网络(maximum mean discrepancy distribution-matching residual networks,MMD-ResNet)和基于零膨胀负二项的方差提取法(zero-inflated negative binomial-based wanted variation extraction,ZINB-WaVE)。**结论** SDAEC 算法能够移除 scRNA-seq 数据的批次效应,提高 scRNA-seq 数据下游分析的有效性,具有实际应用价值。

【关键词】 深度堆叠降噪自编码嵌套聚类 单细胞测序 批次效应 卵巢癌
【中图分类号】 R195.1 【文献标识码】 A **DOI** 10.11783/j.issn.1002-3674.2024.04.005

SDAEC Method and its Application in Batch Effect Removal for Single Cell mRNA Sequence

Wang Wenjie, Li Kang, Xie Hongyu(*Department of Medical Statistics, Harbin Medical University, Harbin* 150081)

【**Abstract**】 **Objective** To propose a deep stacked denoising auto encoder embedded cluster(SDAEC) algorithm and apply it to single cell mRNA sequence(scRNA-seq) data to remove the batch effect, and further to evaluate the performance of its batch effect removal. **Methods** Based on the characteristics of high dimension, high sparsity and high non-linear error of single-cell data, the algorithm of single cell Louvain clustering was embedded into stacked denoising auto encoder(SDAE) algorithm, and formed a SDAEC algorithm, which was used to batch effect removal for scRNA-seq data. SDAEC algorithm was utilized to scRNA-seq data of actual ovarian cancer tissue for batch effect removal, t-distributed stochastic neighbor embedding(tSNE), k-nearest-neighbor batch-effect test(kBET), adjusted rand index(ARI), normalized mutual information(NMI) and average silhouette width(ASW) were used to evaluate the performance of removing batch effect. **Results** The performance of SDAEC was better than Combat, mutual nearest neighbors(MNN), maximum mean discrepancy distribution-matching residual networks(MMD-ResNet) and zero-inflated negative binomial-based wanted variation extraction(ZINB-WaVE) in removing batch effect of scRNA-seq. **Conclusion** SDAEC algorithm can remove the batch effect of scRNA-seq data and improve the validity of downstream analysis of scRNA-seq data.

【**Key words**】 Stacked denoising auto encoder embedded cluster; Single cell mRNA sequence; Batch effects; Ovarian cancer

传统的基于 mRNA 测序技术只能得到肿瘤组织内所有细胞的平均表达数据,忽略了组织内部异质性,难以从肿瘤微环境上进行更深层次的研究。单细胞 mRNA 测序(single cell mRNA sequence,scRNA-seq)技术可在单个细胞水平上检测肿瘤组织中所有细胞的基因表达量,是目前解决肿瘤异质性问题的新手段[1-2]。然而在实际肿瘤单细胞研究中,新鲜的样本采集往往是在多个时间点上分批次完成的,在样品采集、预处理以及检测等过程中不可避免地产生批次效应。这种批次效应属于系统误差,其以高度非线性的方式与生物学变异、检测的随机误差相互耦合在一起,对不同细胞和功能亚群的聚类和鉴别、细胞分化的轨迹分析和生物标志物的筛选都会产生极大的影响[3-4]。scRNA-seq 数据主要特点是高维度、高稀疏性及高度的非线性误差,难以通过简单的算法移除其中的“批次效应”[5]。深度学习算法通过设置多层非线性拟合模块,将特征学习过程嵌入到模型中,让模型能自主学习 scRNA-seq 数据的批次效应特征。在移除批次效应的同时,保留 scRNA-seq 数据中蕴含的具有生物学意义的变异[6]。本文基于深度堆叠降噪自编码嵌套聚类算法(stacked denoising auto encoder embedded cluster,SDAEC),对 scRNA-seq 数据进行批次

* 基金项目:国家自然科学基金(82003551);浙江省自然科学基金(LTGY24H160008)
1. 哈尔滨医科大学卫生统计学教研室(150081)
2. 浙江大学医学院附属妇产科医院临床研究中心
△通信作者:谢宏宇,E-mail:xiehongyu@ zju.edu.cn

效应移除，提高批次校正性能。

原理与方法

1. SDAEC 算法

降噪自动编码器（denoising auto encoder, DAE）[7]，在自动编码器（auto encoder, AE）[8]训练数据加上一层噪声，避免自动编码器在样本量小的情况下模型过拟合问题，通过干扰原始数据中的噪声，学习到更真实、稳定的数据特征，扩大了数据集间的差异，使模型具有更高的泛化能力[9]。考虑到 scRNA-seq 数据的多批次性、高维度性和高稀疏性，模型需要更强的非线性拟合能力和构造力，对于 scRNA-seq 数据分析需将多个 DAE 结构串联堆叠起来，构成堆叠降噪自动编码器（stacked denoising auto encoder, SDAE）[10]。本研究拟通过将单细胞 Louvain 聚类算法[11]嵌入到上述的 SDAE 中，形成 SDAEC 算法，在自学习过程中逐步迭代地优化聚类和特征表示，可同时解决特征学习和聚类问题，用来处理 scRNA-seq 数据的批次问题。该算法通过将聚类分析嵌入到深度模型的整个训练过程中，使模型准确提取到反映细胞真实生物学效应的特征，解耦混杂进来的批次效应，在迭代聚类的过程中逐步完成批次校正。算法流程如图 1 所示。

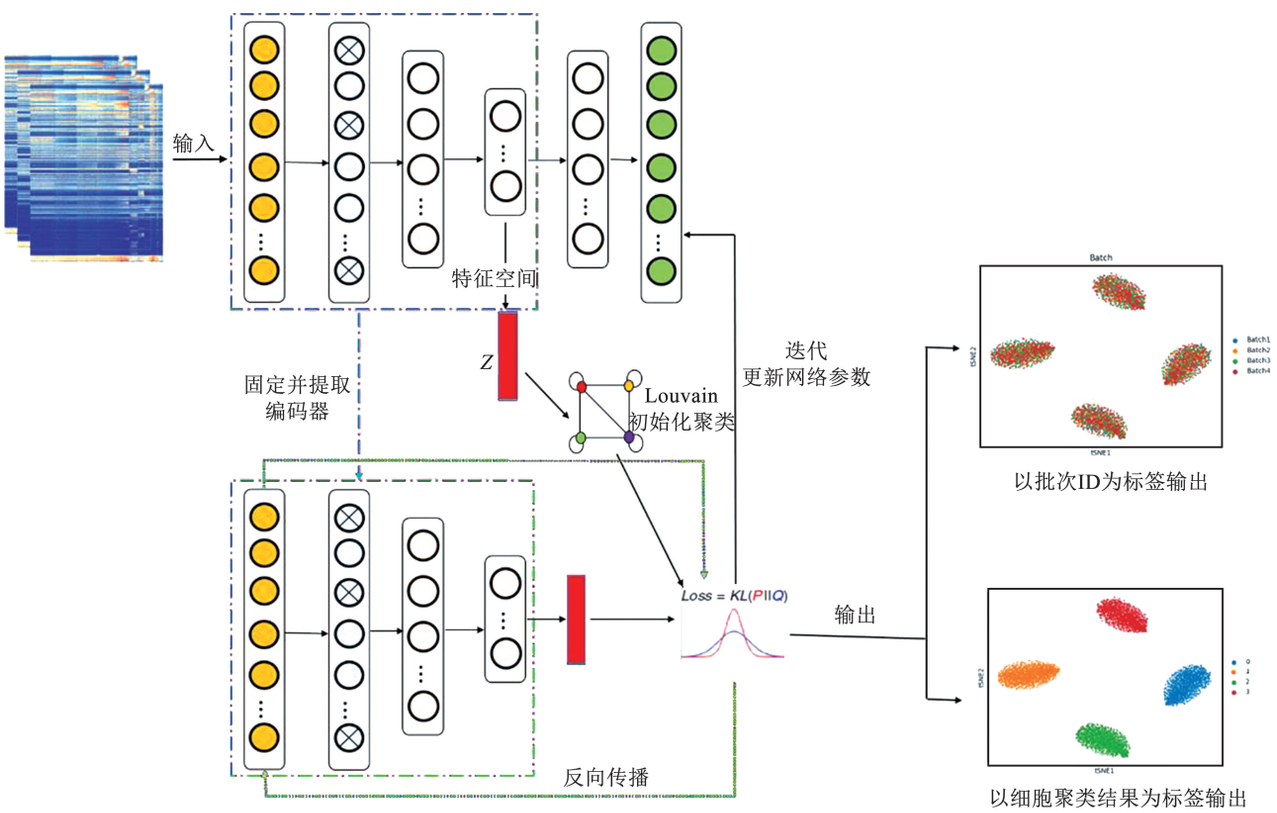

图 1　深度堆叠降噪自编码嵌套聚类算法的示意图

SDAEC 算法的具体流程如下：

第一步：通过 SDAE 进行参数初始化。含有批次效应的 scRNA-seq 数据输入到 SDAE 网络中，首先为每一个输入添加一层噪声，然后输入编码器过程。通过编码器的逐层训练，层层初始化，每一层都作为训练的编码器来重构上一层的输出。解码层则通过反向传播和梯度下降算法，对网络结构的参数进行调整和优化。形成初始化后的堆叠降噪自编码网络结构。编码器的最后一层隐藏层作为原始数据低维特征空间的初始映射，获得特征空间 Z，再利用 Louvain 算法[11]在特征空间 Z 中进行聚类，根据增量模块度 ΔQ 的结果，得到初步聚类的类别数，记为 K，初始对应的各类质心记为 $\{\mu_j : j=1,..,K\}$，并作为本研究方法的初始聚类。

第二步：迭代聚类。通过迭代方式，在以下两步间交替计算直到收敛，从而提高聚类效果。①首先为每一个细胞计算嵌入点（embedded points）和质心间的软聚类分配，软聚类可解释为将细胞 v 分配给类别 j 的概率。使用 t 分布作为内核去测量细胞 v 的嵌入点 z_v 与类别 j 的相似性，如下：

$$q_{vj}=\frac{(1+\| z_v-u_j \|^2/\alpha)^{-1}}{\sum_{j'}(1+\| z_v-u_{j'} \|^2/\alpha)^{-1}} \quad (1)$$

其中 α 为 t 分布的自由度。因为在无监督学习中无法交叉验证 α，所以不做估计，统一设置为 1。②通过一个辅助目标分布 P，从高置信度的细胞软聚类分配中不断学习，从而优化聚类。辅助分布 P 定义为：

$$p_{vj}=\frac{q_{vj}^2/\sum_{v=1}^{n}q_{vj}}{\sum_{j=1}^{K}(q_{vj}^2/\sum_{v=1}^{n}q_{vj})} \quad (2)$$

将目标损失函数 L 定义为细胞 v 的软分配 q_v 和

辅助分布 p_v 之间的 Kullback-Leibler(KL)离散损失：

$$L=KL(P||Q)=\sum_{v=1}^{n}\sum_{j=1}^{K}p_{vj}\log\frac{p_{vj}}{q_{vj}} \quad (3)$$

其中，深度堆叠降噪自编码网络通过迭代，最小化 L 来实现整个网络结构的参数微调。辅助分布 P 可以通过给高置信度的细胞给予更多权重而提高聚类纯度。由于目标分布 P 由 Q 定义，则最小化 L 的过程就是一个自训练的过程，同时，P_{vj} 为细胞 v 属于类别 j 的概率，这一概率可用于衡量每一个细胞类别分配的概率。

第三步：优化 KL 离散损失。使用随机梯度下降联合优化聚类质心 $\{\mu_j:j=1,\cdots,K\}$ 和 SDAE 网络的参数，每一细胞 v 的嵌入点 z_v 与每一聚类质心 μ_j 在特征空间所对应的目标离散损失函数 L 的梯度分别为：

$$\frac{\partial L}{\partial z_v}=\frac{\alpha+1}{\alpha}\sum_{j=1}^{K}\left(1+\frac{z_v-\mu_j^2}{\alpha}\right)^{-1}\times(p_{vj}-q_{vj})(z_v-\mu_j) \quad (4)$$

$$\frac{\partial L}{\partial \mu_v}=\frac{-(\alpha+1)}{\alpha}\sum_{v=1}^{n}\left(1+\frac{z_v-\mu_j^2}{\alpha}\right)^{-1}\times(p_{vj}-q_{vj})(z_v-\mu_j) \quad (5)$$

然后将梯度传递给 SDAE 神经网络，并用于标准反向传播算法中，以计算 SDAE 网络的参数。在每次迭代过程中，当损失不再减小或达到 epoch 数的阈值时，则更新辅助分布 P，并用新的辅助分布来优化聚类的质心和自编码器的参数。当前后两步聚类过程中的增量模块度 $\Delta Q=0$，细胞所属类别发生变化的比例小于设定的条件时，迭代过程停止，从而得到最终的网络结构和细胞聚类结果。

SDAEC 算法的网络结构嵌入了单细胞 Louvain 聚类分析，用于还原细胞生物学变异，在此过程中可完成批次校正。由于深度网络结构参数调整的目标损失函数是根据迭代过程中的细胞聚类辅助目标分布和 KL 离散损失函数来计算的，深度网络提取的特征可以看成是 scRNA-seq 数据中反映细胞生物学变异最本质的特征，此时的特征表示已经解耦了批次效应。原始的 scRNA-seq 数据经过最终的深度网络结构的层层映射，在特征空间 Z 中校正了批次效应。

2. 单细胞批次效应校正评价体系

本研究各批次校正方法的评价体系包括：①批次校正和聚类结果的可视化评价：t 分布邻域嵌入(t-distributed stochastic neighbor embedding，tSNE)算法[12]；②批次效应校正的定量评价：k 最近邻批次效应检测(k-nearest-neighbor batch-effect test，kBET)[13]，kBET 值越大，说明各批次校正效果越好；③细胞聚类的定量评价：调整兰德系数(adjusted rand index，ARI)[14]，取值范围为[-1，1]，其值越大代表聚类结果越准确，批次移除效果越好；标准化互信息(normalized mutual information，NMI)用于评价两个聚类结果的相似性[15]，其取值为[0，1]，其值越大代表聚类标签和真实标签越接近，聚类结果越准确；平均轮廓宽度(average silhouette width，ASW)是衡量真实标签与原始数据之间的一致性评价指标[16]，取值范围为[0，1]，其值越大，表明细胞 v 与所属集群越匹配。这里的集群可以是以细胞类型标记的集群，此时的细胞标签 ASW 记为 cASW，cASW 越大，代表不同种类的细胞分离效果越好，细胞聚类效果越好。当然集群也可以是以批次标签标记，此时的批次标签 ASW 记为 bASW，其值越大，代表不同批次的细胞分离效果越好，即批次校正结果越差。

本研究的 SDAEC 算法过程及各单细胞批次效应评价指标的计算均使用 python 软件实现。

实例分析

从 GEO 数据库下载卵巢癌组织样本的 scRNA-seq 数据(GSE147082)，该数据包含了 6 例浆液性卵巢癌患者的卵巢癌/输卵管组织样本的测序结果[17]。所有样本均使用 Illumina NextSeq 500 单细胞 mRNA 测序平台检测。该数据根据各类细胞的基因表达模式及生物学实验准确验证了各细胞的所属类别标签(10 类，如图 2)。本研究首先通过 tSNE 细胞散点图评价 SDAEC 算法的批次移除能力和效果，利用 kBET、bASW 指标定量评价批次混合能力。将 SDAEC 算法处理后的细胞聚类结果与真实标签进行对比，评价 SDAEC 算法对于细胞分类生物学变异的还原程度，通过 ARI、NMI 和 cASW 指标定性评价其处理后的聚类结果，聚类结果越准确说明 SDAEC 算法的批次移除越成功。同时，通过 SDAEC 算法批次处理后的数据，进一步探索各类细胞间的差异表达基因，检测到的差异表达基因越多，说明有更多的生物学信号被保留。SDAEC 的处理结果与 Combat[18]、相互最近邻(Mutual nearest neighbors，MNN)[19]、分布匹配残差网络(maximum mean discrepancy distribution-matching residual networks，MMD-ResNet)[20]和基于零膨胀负二项的方差提取法(zero-inflated negative binomial-based wanted variation extraction，ZINB-WaVE)[21]等其他单细胞批次处理方法进行综合比较。

各批次校正方法处理后的 tSNE 细胞聚类散点图如图 2 所示。左侧图以批次信息为标签，中间图以 Louvain 细胞聚类结果为标签，右侧图以各细胞真实标签作图。根据左侧图判断各方法的批次处理效果，从未经批次校正的原始数据结果可以看出，该数据确实存在较为明显的批次效应，各批次的细胞大部分单独聚成一类或多类，使得相同类型的细胞由于批次效应的存在无法有效聚集在一起。上皮性癌细胞(右侧图棕色部分)分散在各患者上，形成 6 堆，无法聚成一类。同样的情况也见于 Combat 算法校正后的结果，

各批次的癌细胞虽然相比未校正的结果，在视觉上能更加相互靠近，但依然聚成多类。

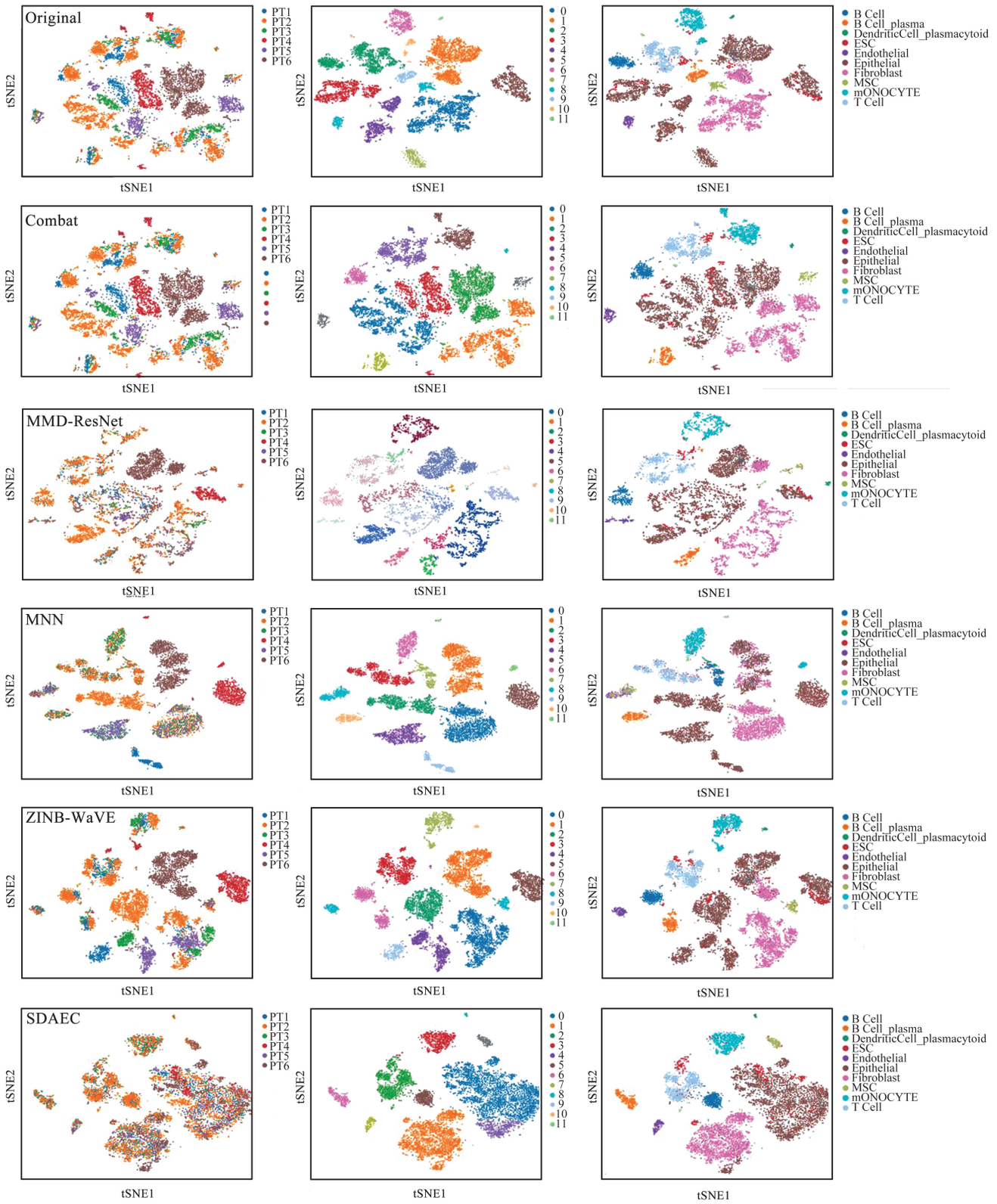

图 2　卵巢癌组织样本 scRNA-seq 数据（GSE147082）批次校正前后的 tSNE 散点图

MMD-ResNet 算法对多个批次的数据一一对齐校正，虽然各批次间相同类型的细胞能够混合起来，但多次校正过程可能破坏了细胞固有的生物学变异，导致最终聚类出 14 类结果，与实际真实细胞分类结果相距甚大。MNN 算法处理后，批次效应得到一定程度的校正，各批次的成纤维细胞（右侧图粉红色部分）、单核细胞（右侧图浅蓝色部分）、T 细胞（右侧图灰色部分）在批次校正后能充分混合在一起。但上皮性癌细胞仍然在不同批次间分散开来，说明此时的癌细胞仍残留一定程度的批次效应。ZINB-WaVE 算法校正

后，批次效应同样得到一定的校正，各批次间的成纤维细胞、单核细胞在批次校正后也能互相靠近，在聚类分析时各形成一类。相比 MNN 算法，这些细胞在不同批次间混合得并不够充分，仍存在明显的批次间隙。而 SDAEC 算法处理后的数据最终聚成了 10 类细胞，与真实的细胞分类情况一致。无论是癌细胞、成纤维细胞、单核细胞、B 细胞、T 细胞等细胞类别上，各批次间的细胞在聚类过程中均能相互均匀混合在一起。说明此时的批次效应已经得到有效处理，深度网络结构提取到了反应细胞类别间固有生物学变异的特征，并最终获得更加准确的细胞聚类结果。

各批次校正评价指标如表 1 所示，指标直观反映了各算法的批次处理能力。Combat、MNN、ZINB-WaVE 算法的 kBET 值均不高，提示这些算法处理后的细胞依然有较明显的批次间隙。SDAEC 算法在 kBET 及 bASW 指标上远远优于其他算法的结果，kBET 值达到了 0.966，表明此时各批次间相同类型的细胞混合均匀，批次效应得到有效校正。同时，在细胞聚类准确性的评价指标上，Combat、MMD-ResNet 和 ZINB-WaVE 算法处理后的聚类结果，由于仍残留不同程度的批次效应，导致部分细胞归属类别仍不清晰，与真实标签的分类结果存在较大差异，其 ARI 和 NMI 也都处于较低水平。SDAEC 算法的 ARI 和 NMI 值分别为 0.896 和 0.912，表明 SDAEC 算法处理后数据，在处理掉批次的同时，有效还原了细胞类别间固有的生物学变异，聚类结果更贴近真实标签的分类情况。

表 1 卵巢癌组织样本 scRNA-seq 数据（GSE147082）批次校正前后的评价指标结果

方法	kBET	bASW	cASW	ARI	NMI
Original	0.246	0.204	0.213	0.403	0.586
Combat	0.452	0.181	0.357	0.634	0.721
MMD-ResNet	0.872	0.066	0.300	0.577	0.694
MNN	0.783	0.102	0.384	0.683	0.767
ZINB-WaVE	0.655	0.136	0.337	0.618	0.713
SDAEC	0.966	0.021	0.415	0.896	0.912

为进一步评价各批次处理方法对细胞固有生物学信息的保留情况，对各批次校正方法处理后的数据研究，筛选了上皮性癌细胞与其他类型细胞之间的差异表达基因。各算法处理后所得到的差异表达基因数量如图 3 所示。由于批次效应被有效地处理，SDAEC 算法处理后所得到的差异表达基因的数量也远远高于其他算法，达到了 537 个基因。未经校正的原始数据只能筛选到 106 个差异表达基因。Combat、MMD-ResNet、MNN 和 ZINB-WaVE 算法筛选到的差异表达基因个数分别为 209、186、348、272，SDAEC 算法得到的差异表达基因数量是 MNN 算法的 1.5 倍多。因此，SDAEC 算法在准确校正批次效应的同时，暴露出了更多反映细胞生物学变异的特征，扩大了各类细胞间表达模式的差异，保留了更多有用的生物学信号，从而能够获得更加贴近真实标签的聚类结果。

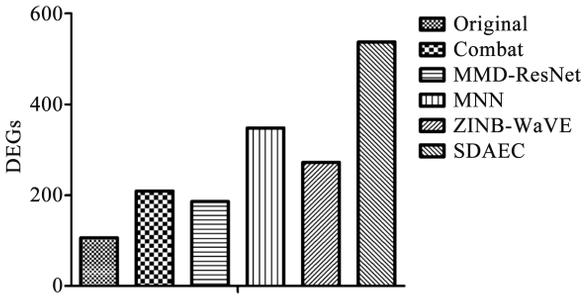

图 3 卵巢癌组织样本 scRNA-seq 数据批次处理前后的肿瘤细胞差异表达基因数量

讨 论

基于深度自编码神经网络和细胞聚类的 Louvain 算法，提出了旨在移除 scRNA-seq 批次效应的 SDAEC 算法。该算法的网络结构参数根据细胞聚类的损失函数进行调整，保证了网络结构能提取到反映细胞生物学变异的真实特征，在细胞迭代聚类过程中，逐步移除批次效应。实际卵巢癌组织样本 scRNA-seq 数据的研究显示，相比于其他算法，使用 SDAEC 算法进行批次处理后，各批次间相同细胞能更加均匀混合，无明显批次间隙，其细胞聚类结果也更加符合真实情况。同时揭示了更多生物学信息，获得了更多的差异表达基因。

同时，通过对比实际单细胞测序数据的细胞真实标签，批次标签，以及各批次处理方法处理后得到的细胞聚类标签，可以看到，其他算法对批次效应均存在不同程度的错误处理：①未能解除不同批次间同种细胞类型的批次效应，导致同一种细胞由于残留的批次效应而无法聚集在一起；②将细胞生物学效应误认为是批次效应，强行整合，导致结果将不同类型的细胞聚集在一起。上述两种错误的批次处理方式可能与各方法自身的假设和特征提取能力有关。Combat 方法主要用来处理“bulk”RNA 测序数据，其应用于 scRNA-seq 数据的前提假设是：各批次中细胞类型的构成和比例需完全一致，且不同批次间基因平均表达的差异都归因于检测技术差异[22]。在实际 scRNA-seq 研究中，很难满足上述假设。当 scRNA-seq 数据存在多个批次效应时，利用 MNN 方法容易导致过度校正问题[23]。ZINB-WaVE 估计的结果更加稳健。但由于其是利用降维后的数据进行校正，无法利用个体基因进行下游的差异基因表达分析或假时间轨迹分析[21]。MMD-ResNet 也是一种能够基于深度学习的批次校正方法，但仅适用于样本间批次效应较小的校正[20]。SDAEC 算法在进行 scRNA-seq 数据批次效应移除中

也存在一定的局限性:①SDAEC 算法是基于深度学习框架结合批次移除和细胞聚类问题给出的新模型,在模型构建过程中,需要确定网络结构的超参数(如隐藏层数量,隐藏层节点数等),目前主要通过多次实验的经验方法来确定。但这些超参数选择与实际处理的数据密切相关,在综合考虑计算时间、批次处理效果等因素下找到的参数并不一定是该数据本身的"最优"参数。如何通过数据本身的特点,自适应地获得模型的"最优"参数,需要进一步研究。②深度网络作为一个"黑匣"过程,我们只看到 SDAEC 算法处理后的结果移除了批次效应,但并不知道数据经过的具体运算来解耦批次效应和细胞生物学效应之间复杂的非线性关联,因此在算法的应用解释上相对不如其他算法明确。

参 考 文 献

[1] Suv ML, Tirosh I. Single-cell RNA sequencing in cancer: lessons learned and emerging challenges[J]. Molecular cell, 2019, 75(1): 7-12.

[2] Baslan T, Hicks J. Unravelling biology and shifting paradigms in cancer with single-cell sequencing[J]. Nature Reviews Cancer, 2017, 17(9): 557-569.

[3] Leek JT, Scharpf RB, Bravo HC, et al. Tackling the widespread and critical impact of batch effects in high-throughput data[J]. Nature Reviews Genetics, 2010, 11(10): 733-739.

[4] Liotta LA, Petricoin EF, Veenstra TD, et al. High-resolution serum proteomic patterns for ovarian cancer detection[J]. Endocrine-Related Cancer, 2004, 11(4): 585-587.

[5] Chi W, Deng M. Sparsity-penalized stacked denoising autoencoders for imputing single-cell RNA-Seq data[J]. Genes, 2020, 11(5): 532.

[6] Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(8): 1798-828.

[7] Lu X, Tsao Y, Matsuda S, et al. Speech enhancement based on deep denoising autoencoder[C]. ISCA, 2013:436-440.

[8] Wang Y, Yao H, Zhao S. Auto-encoder based dimensionality reduction[J]. Neurocomputing, 2016, 184:232-242.

[9] Majumdar A. Blind denoising autoencoder[J]. IEEE transactions on neural networks and learning systems, 2018, 30(1): 312-317.

[10] Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks[J]. Science(New York, NY), 2006, 313(5786): 504-507.

[11] Blongel VD, Guillaume JL, Lambiotte R, et al. Fast unfolding of communities in large networks[J]. Journal of statistical mechanics: theory and experiment, 2008, 2008(10): P10008.

[12] Linderman GC, Rachh M, Hoskins JG, et al. Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data[J]. Nature methods, 2019, 16(3): 243-245.

[13] Bttner M, Miao Z, Wolf FA, et al. A test metric for assessing single-cell RNA-seq batch correction[J]. Nature methods, 2019, 16(1): 43-49.

[14] Santos JM, Embrechts M. On the use of the adjusted rand index as a metric for evaluating supervised classification[J]. DBLP, 2009, 175-184.

[15] Strehl A, Ghosh J. Cluster ensembles-a knowledge reuse framework for combining multiple partitions[J]. Journal of machine learning research, 2002, 3(Dec): 583-617.

[16] Batool F, Hennig C. Clustering with the average silhouette width[J]. Computational Statistics & Data Analysis, 2021, 158:107190.

[17] Olalekan S, Xie B, Back R, et al. Characterizing the tumor microenvironment of metastatic ovarian cancer by single-cell transcriptomics[J]. Cell Rep, 2021, 35(8): 109165.

[18] Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods[J]. Biostatistics, 2007, 8(1): 118-127.

[19] Haghverdi L, Lun AT, Morgan MD, et al. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors[J]. Nature biotechnology, 2018, 36(5): 421-427.

[20] Shaham U, Stanton KP, Zhao J, et al. Removal of batch effects using distribution-matching residual networks[J]. Bioinformatics(Oxford, England), 2017, 33(16): 2539-2546.

[21] Risso D, Perraudeau F, Gribkova S, et al. A general and flexible method for signal extraction from single-cell RNA-seq data[J]. Nature communications, 2018, 9(1): 1-17.

[22] Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods[J]. Biostatistics(Oxford, England), 2007, 8(1): 118-127.

[23] Haghverdi L, Lun ATL. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors[J]. Nature Biotechnology, 2018, 36(5): 421-427.

(责任编辑:张　悦)