期末作业

SegGCN: 基于语义分割预处理的图卷积空间转录组聚类算法

502022370071 庄镇华[1]

1. 南京大学
E-mail: zhuangzh@lamda.nju.edu.cn

摘要 空间转录组学是基因组技术的集合，这些技术能够对具有空间定位信息的组织进行转录组学分析。分析空间转录组学数据在计算上具有挑战性，因为从各种空间转录组学技术收集的数据通常很嘈杂，并且在组织位置之间显示出实质性的空间相关性。本文基于 SpaGCN 方法实现了一种整合基因表达、空间位置和组织学的图卷积算法 SegGCN，创新性地使用 U-net 语义分割网络进行预处理步骤，将组织学图像分割生成的伪标签作为空间位置的第三维度，通过图卷积，SegGCN 算法从其相邻点聚集每个点的基因表达，能够更好地识别具有相关表达和组织学的空间域。我们使用 SegGCN 算法分析了 DLPFC 数据集，结果表明，与原始 SpaGCN 算法相比，SegGCN 算法具有更优的性能和更好的鲁棒性。

关键词 空间转录组, 聚类分析, 生物信息

1 引言

空间转录组测序技术的最新进展使基因表达谱能够与组织中的空间信息结合。了解组织中不同细胞的相对位置对于理解疾病病理学至关重要，因为空间信息有助于理解细胞的基因表达如何受到其周围环境的影响。空间转录组可以结合显微成像和测序技术在获得基因表达数据的同时最大程度的保留样本的空间位置信息，使得揭示异质组织的复杂转录结构成为可能，增强了我们对疾病中细胞机制的理解。

在空间转录组测序研究中，一个重要的步骤是识别空间域，即在基因表达和组织学上空间相关的区域。传统的聚类方法如 K-means 和 Louvain 方法 [1] 仅将基因表达数据作为输入，缺乏对空间信息和组织学信息的建模。最近，一些用于解释基因表达的空间依赖性的方法已经被提出，例如，stLearn [2] 在聚类之前使用从组织学图像中提取的特征以及相邻取样点的空间表达来归一化基因表达数据；BayesSpace [3] 采用贝叶斯方法进行聚类，对模型施加先验，给予物理上接近的点更高的权

期末作业: 庄镇华. 期末作业. 生物信息学.
Zhenhua ZHUANG. Final Assignment. Bioinformatics.

重。虽然这些方法可以将取样点或细胞聚集成不同的组，但这些方法大都只使用了空间信息和基因表达数据，或者简单直观地使用了组织学信息，而没有深入探索组织学信息的使用方法以及组织学信息能够提高聚类性能的本质。

因而本文基于 SpaGCN [4] 方法实现了一种整合基因表达、空间位置和组织学的图卷积算法 SegGCN，创新性地使用 U-net [5] 语义分割网络进行预处理步骤，将组织学图像分割生成的伪标签作为空间位置的第三维度，通过图卷积，SegGCN 算法从其相邻点聚集每个点的基因表达，能够更好地识别具有相关表达和组织学的空间域。

2 方法

2.1 U-net 语义分割算法

在图像分割任务特别是医学图像分割中，U-net 无疑是最成功的方法之一，该方法采用的编码器-解码器结构和跳跃连接是一种非常经典的设计方法。当前许多新的卷积网络设计方式仍延续了 U-Net 的核心思想，加入了新的模块或者融入其他设计理念。U-net 使用全卷积神经网络 (FCN)。全卷积神经网络 (FCN) 与卷积神经网络 (CNN) 的不同点在与 FCN 将 CNN 最后的全连接层替换为卷积层，因此 FCN 中可以输入任意尺寸的图片，输出任意尺寸的图片，即为一个端到端网络。图2左边网络为收缩路径：使用卷积和最大池化操作。右边为扩张路径：与左侧的特征图相结合，然后逐层上采样。最后再经过两次的卷积得到特征图，再用 1x1 的卷积做分类。

U-net 网络支持少量的训练数据，医学方向所能够训练的数据相对较少而需要进行检测的数据又较多，因此可以使用数据增强操作。U-net 网络在不同的生物医学图像分割中有很好的表现，并且数据增强使得仅仅需要很少的带标注数据。U-net 也是较早的使用多尺度特征进行语义分割任务的算法之一，其 U 形结构也启发了后面很多算法。

2.2 SegGCN 空间转录组聚类算法

算法整体流程如算法1和图1所示。具体步骤如下文所述，本文对已有算法 SpaGCN 的预处理模块进行修改，即创新点在于第二部分，将基于语义分割模型 U-net 网络预处理组织学图像数据，将组织学图像的信息整合进聚类过程以获得优异的效果。

算法 1 SegGCN: 基于语义分割预处理的图卷积空间转录组算法

1: 数据预处理：将每个采样点的基因表达归一化。

2: 将原始数据转化为图结构数据：无向加权图的两点距离由点 u 和 v 在组织切片中的物理位置，以及这两个点的组织学信息决定。即使用 U-net 语义分割网络进行预处理步骤，将组织学图像分割生成的伪标签作为空间位置的第三维度。最终的两点距离和边权重定义为：

$$d(u,v)=\sqrt{(x_u-x_v)^2+(y_u-y_v)^2+(z_u^*-z_v^*)^2},\ w(u,v)=\exp\left(-\frac{d(u,v)^2}{2l^2}\right)$$

3: 图卷积层：使用主成分分析降低维数。接着图卷积网络根据 G 中指定的边权重聚合基因表达信息。

4: 通过聚类进行空间转录组识别：基于上述图卷积层的输出，迭代地采用无监督聚类算法来将采样点聚类到不同的空间域。

输出: 聚类结果

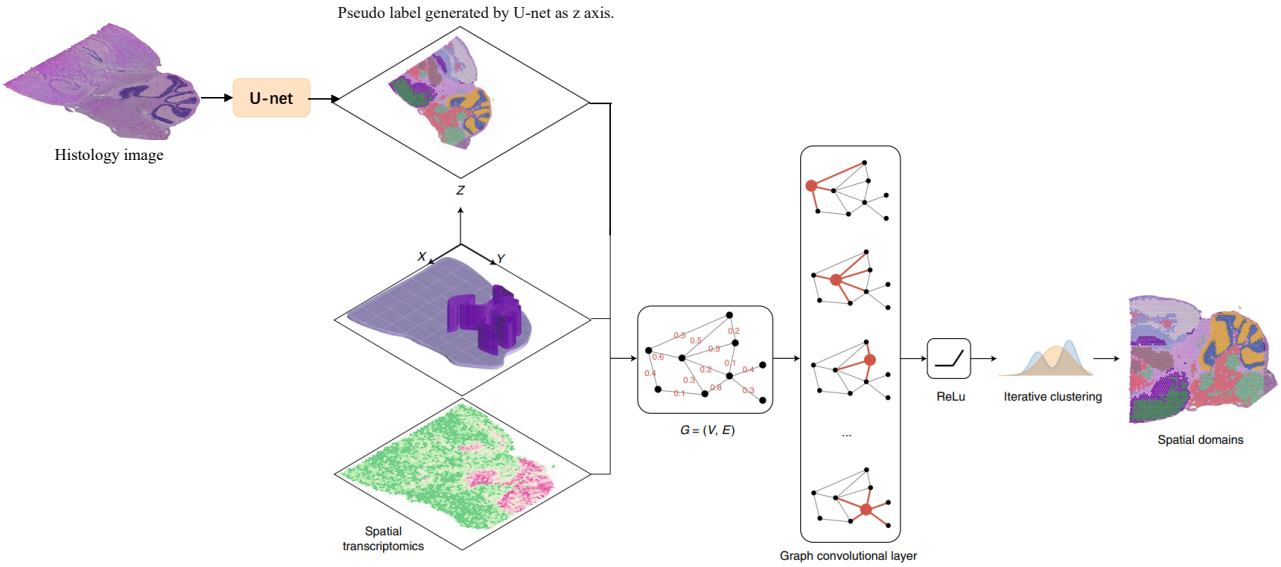

图 1　SegGCN 的工作流程 SegGCN 使用图卷积网络 (GCN) 整合基因表达、空间位置和组织学信息，然后使用无监督迭代聚类将取样点聚合到不同的空间域。GCN 基于无向加权图，其中每两个点之间的边缘权重由两点之间的欧几里德距离确定，欧几里德距离由空间坐标（x，y）和组织学图像经过 U-net 分割网络生成的第三维坐标伪标签 z 定义。

数据预处理　SegGCN 将空间基因表达和组织学图像数据作为输入。空间基因表达数据以及每个采样点的二维坐标存储在具有 N 个采样点和 D 个基因 UMI 计数的 N × D 矩阵中。将每个采样点的基因表达归一化，即将每个基因的 UMI 计数除以给定点中所有基因的总 UMI 计数，乘以 10000，然后取自然对数。

将原始数据转化为图结构数据　SegGCN 将基因表达和组织学图像数据预处理后转换为加权无向图 G(V, E)。在这个图中，$v \in V$ 表示一个点，V 中的每两个顶点通过一条具有指定权的边相连。

关于两点之间距离的计算。图中任意两个顶点 u 和 v 之间的距离反映了两个对应点的相似性。该距离由两个因素决定: 点 u 和 v 在组织切片中的物理位置，以及这两个点的组织学信息。

如图2，关于组织学信息的利用，SegGCN 方法创新性地使用 U-net [5] 语义分割网络进行预处理步骤，将组织学图像分割生成的伪标签作为空间位置的第三维度，通过图卷积，SegGCN 算法从其相邻点聚集每个点的基因表达，能够更好地识别具有相关表达和组织学的空间域。

最后，每两个点 u 和 v 之间的欧几里得距离计算, 点 u 和点 v 之间的边权重分别被定义为:

$$d(u,v) = \sqrt{(x_u - x_v)^2 + (y_u - y_v)^2 + (z_u^* - z_v^*)^2},\ w(u,v) = \exp\left(-\frac{d(u,v)^2}{2l^2}\right)$$

图卷积层　SegGCN 使用主成分分析（PCA）降低预处理的基因表达矩阵的维数。前 50 个主成分被用作输入，接着图卷积网络根据 G 中指定的边权重聚合基因表达信息。图卷积层可以被写为

$$f(X, A) = \delta(AXB)$$

其中 X 是从 PCA 特征矩阵，B 是表示卷积层参数矩阵，$\delta(\cdot)$ 是非线性激活函数。

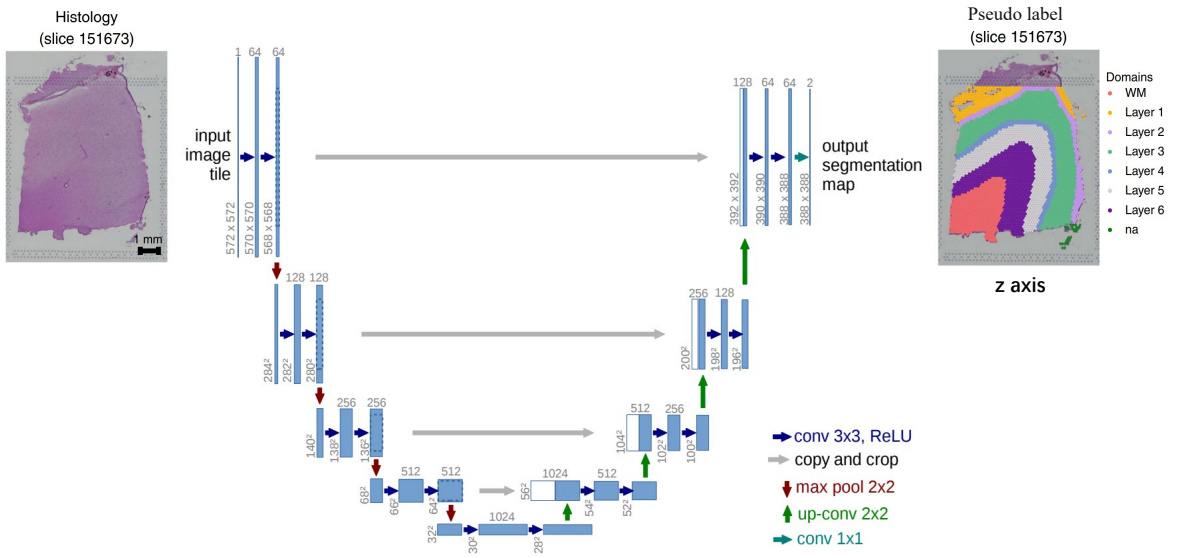

图 2 组织学图像经过 U-net 分割网络生成第三维坐标伪标签 z 用于计算 GCN 无向加权图的权重。

通过聚类进行空间转录组识别 基于上述图卷积层的输出，SegGCN 迭代地采用无监督聚类算法来将采样点聚类到不同的空间域。该步骤得到的每个簇都被认为是一个特定空间域，其包含在基因表达和组织学上相似的采样点。简便起见，无监督聚类算法采用与算法 [4] 相同的方法。

3 实验测试

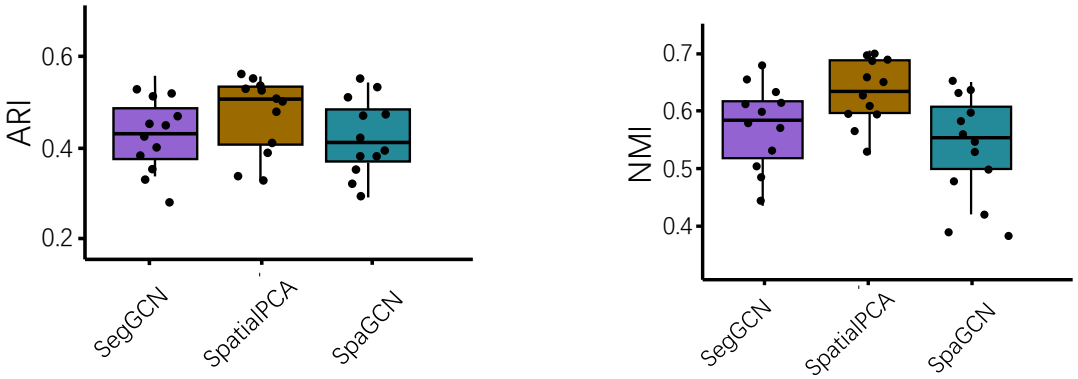

图 3 SegGCN、SpatialPCA 和 SpaGCN 算法在 DLPFC 数据集上的实验指标

3.1 数据集与实验指标

数据集包括 Visium 平台上对 3 个个体收集的 12 份人类背外侧前额叶皮质 (dorsolateral prefrontal cortex, DLPFC) 样本，每个样本包含大约 4,000 个采样点，这些采样点被人工注释为六个 DLPFC 层或白质。

遵循 Maynard 等人 [6] 的方法，我们使用调整的 Rand 指数 (ARI) 和归一化互信息指数 (NMI) 来度量聚类标签和人工注释之间的相似性，进而得到实验测试指标。

3.2 实验结果

实验结果如图3所示。可以看到，相对于原始方法 SpaGCN，SegGCN 算法在 ARI 和 NMI 指标上都有一定的提升，即均值变高，方差变小，这说明使用 U-net 语义分割网络进行预处理步骤，将组织学图像分割生成的伪标签作为空间位置的第三维度，可以使聚类获得更优的性能和更好的鲁棒性。

然而，相比于目前最优的 SpatialPCA 算法，SegGCN 算法仍有一定的差距，但是 SegGCN 对组织学图像进行语义分割获取伪标签的做法较为新颖，对后续的工作有一定的启发意义。

4 结束语

总结 空间转录组测序技术的最新进展使基因表达谱能够与组织中的空间信息结合。了解组织中不同细胞的相对位置对于理解疾病病理学至关重要，本文基于 SpaGCN 方法实现了一种整合基因表达、空间位置和组织学的图卷积算法 SegGCN，创新性地使用 U-net 语义分割网络进行预处理步骤，将组织学图像分割生成的伪标签作为空间位置的第三维度，通过图卷积，SegGCN 算法从其相邻点聚集每个点的基因表达，能够更好地识别具有相关表达和组织学的空间域。我们使用 SegGCN 算法分析了 DLPFC 数据集，结果表明，与原始 SpaGCN 算法相比，SegGCN 算法具有更优的性能和更好的鲁棒性。

展望 然而，相比于 SOTA 算法 SpatialPCA，SegGCN 算法仍有一定的差距。因而，后续的研究方向可从以下内容展开：即可以探究带约束条件的聚类方法，或者使用 t-SNE、UMAP 等聚类方法对基因表达进行预聚类，并将获得的信息整合到正式聚类过程中以获得优异的效果。

参考文献

1 V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, p. P10008, 2008.

2 D. Pham, X. Tan, J. Xu, L. F. Grice, P. Y. Lam, A. Raghubar, J. Vukovic, M. J. Ruitenberg, and Q. Nguyen, "stlearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues," *BioRxiv*, pp. 2020–05, 2020.

3 E. Zhao, M. R. Stone, X. Ren, J. Guenthoer, K. S. Smythe, T. Pulliam, S. R. Williams, C. R. Uytingco, S. E. Taylor, P. Nghiem *et al.*, "Spatial transcriptomics at subspot resolution with bayesspace," *Nature Biotechnology*, vol. 39, no. 11, pp. 1375–1384, 2021.

4 J. Hu, X. Li, K. Coleman, A. Schroeder, N. Ma, D. J. Irwin, E. B. Lee, R. T. Shinohara, and M. Li, "Spagcn: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network," *Nature methods*, vol. 18, no. 11, pp. 1342–1351, 2021.

5 O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241.

6 K. E. Maynard, L. Collado-Torres, L. M. Weber, C. Uytingco, and A. E. Jaffe, "Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex," *Cold Spring Harbor Laboratory*, no. 3, 2020.

7 L. Shang and X. Zhou, "Spatially aware dimension reduction for spatial transcriptomics," *Nature Communications*, vol. 13, no. 1, p. 7203, 2022.

8 K. Dong and S. Zhang, "Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder," *Nature communications*, vol. 13, no. 1, p. 1739, 2022.