中国激光

结合多次DBSCAN和层次聚类算法的膜蛋白单分子定位超分辨图像分割

杨建宇[1]，胡芬[1]，邢福临[1]，董浩[1]，侯梦迪[1]，李任植[1]，潘雷霆[1,2,3,4*]，许京军[1,3]

[1]南开大学弱光非线性光子学教育部重点实验室，物理科学学院，泰达应用物理研究院，天津 300071；

[2]南开大学细胞应答交叉科学中心，药物化学生物学国家重点实验室，生命科学学院，天津 300071；

[3]南开大学深圳研究院，广东 深圳 518083；

[4]山西大学极端光学协同创新中心，太原 山西 030006

摘要 膜蛋白在细胞膜上的时空分布形式决定了其活性状态及功能，在调控细胞生命活动过程中起着重要作用。单分子定位超分辨成像(SMLM)技术为在纳米尺度解析膜蛋白的空间分布提供了可能，但分辨率的极大提升对图像准确聚类分割提出了更高要求。基于密度的空间聚类算法(DBSCAN)是常用的聚类方法之一，但其对于膜蛋白分布不均匀的SMLM超分辨图像的分割效果往往不太理想。本文提出了一种结合多次DBSCAN和层次聚类的混合聚类算法，该算法以DBSCAN方法为分割基础，通过进一步的面积阈值分析和层次聚类，在保持超分辨点簇图像精确聚类识别的前提下，仍能保留每个点簇内的多次定位信号。将该算法应用于模拟数据集和实验数据分割得到的轮廓系数等性能普遍优于传统DBSCAN算法。这种混合聚类方法为膜蛋白SMLM超分辨图像的聚类分割提供了新思路和新方法，有助于更精准地分析膜蛋白在纳米尺度上的空间分布信息。

关键词 生物光学；单分子定位超分辨成像；超分辨图像分割；膜蛋白；基于密度的空间聚类算法；层次聚类算法

中图分类号 O436 **文献标志码** A **DOI**：10.3788/CJL221242

1 引 言

细胞膜上富含多种蛋白质，这些蛋白质参与许多关键的细胞活动过程，如信号传导、物质运输等。特定膜蛋白的表达量、空间分布等特性与其功能息息相关。膜蛋白的尺寸以及它们之间的距离都在纳米尺度，传统的光学显微镜受限于衍射极限无法在单分子水平上对膜蛋白的空间分布特性展开深入研究。电子显微镜的空间分辨率虽然很高，但只能给出形貌信息，不具有特异性，无法识别特定膜蛋白的空间分布信息。因此，亟需可对膜蛋白的空间分布模式进行解析且兼具特异性和高成像分辨率的技术[1]。为了突破衍射极限，诞生了两类荧光超分辨成像技术，其中一类是基于照明光场改造的受激发射损耗显微成像术(STED)[2-3]和结构光照明显微术(SIM)[4-5]，另一类是基于单分子定位成像(SMLM)[6-8]的随机光学重建显微术(STORM)[9]和光激活定位显微术(PALM)[10]。

SMLM由于具有最高的分辨率，在膜蛋白精准空间分布研究方面发挥着不可替代的作用。Rossboth等[11]利用PALM证明了T细胞受体被激活后会发生寡聚化，从随机分布形式转变为小于衍射极限的团簇状分布，以新的视角揭示了T细胞快速识别抗原的原理。Pritchard等[12]通过STORM发现小鼠血管平滑肌细胞中的兰尼碱受体同样呈纳米团簇状分布，杜氏肌营养不良小鼠细胞受体簇的平均面积比野生型小鼠细胞大，导致患病小鼠细胞的大电导钾离子通道活性增强，发生钙火花的次数增多，进而导致了血管功能障碍。Yan等[13]利用STORM发现葡萄糖转运受体GLUT1在细胞中呈团簇状分布，团簇的平均直径约为250 nm，被环糊精破坏后团簇直径减小至约130 nm。此外，SMLM通过荧光闪烁定位得到的点簇还包含更深层次的信息，比如每个点簇的定位次数与蛋白质的多聚度之间存在关联[14-15]。SMLM在膜蛋白成像和组织特性分析领域具有无可比拟的优越性。

针对SMLM图像数据深层信息的提取，尤其是对图像点簇结构的聚类分析，已经出现了多种方法，如K

收稿日期：2022-09-14；**修回日期**：2022-09-28；**录用日期**：2022-10-08；**网络首发日期**：2022-10-20

基金项目：广东省基础与应用基础研究重大项目(2020B0301030009)、国家重点研发计划(2022YFC3400600)、国家自然科学基金(11874231，32227802，12174208，31870843)、中国博士后科学基金(2020M680032)、天津市自然科学基金(20JCYB-JC01010)、中央高校基本科研业务费(2122021337，2122021405)

通信作者：*plt@nankai.edu.cn

函数(Ripley's K function)分析法[16]、关联函数分析法[17]、贝叶斯方法[18]、密度聚类方法[19-20]、泰森多边形法[21-22]等。其中1996年提出的基于密度的空间聚类算法(DBSCAN)[19]被认为是密度聚类技术的开创性方法,之后根据不同的使用场景衍生出了多种基于DBSCAN的聚类方法[23]。DBSCAN在激光清洗[24]、激光雷达[25-27]等方向均有应用。近年来,DBSCAN方法在SMLM图像数据提取中发挥了重要作用,如:Siddig等[28]利用DBSCAN方法解析了小鼠脑内突触前活动区促代谢型谷氨酸受体4的纳米级空间分布;Meng等[29]利用DBSCAN算法提取了微管相关蛋白tau(MAPT)的纳米级尺寸,揭示了过度磷酸化MAPT蛋白的自聚集效应。

传统DBSCAN方法可以快速地从有噪声点的图像数据中提取出聚类信息,而且分析区域的选取非常自由灵活;然而,在处理点簇密度不均匀的SMLM膜蛋白图像数据时,该方法难以进行较好的识别与分割。为此,本文提出了一种结合多次DBSCAN和层次聚类的混合聚类方法,该方法可以有效提高点簇SMLM图像数据的提取精度。

2 基本原理

2.1 SMLM图像数据的提取

SMLM图像数据的获取流程如图1所示。在较低激光强度照射条件下找到感兴趣的样本视野[如图1(a)所示],将激光功率密度切换到约2 kW/cm^2,使该视野范围内的荧光分子发生随机"闪烁",实现两个相邻点之间交替稀疏发光。在一段时间内连续采集上万张该视野下的闪烁荧光图像,得到一系列单分子定位图像的原始数据,用于后续分析处理,如图1(b)所示。对每一帧图像中的点扩散函数进行高斯拟合,以确定其质心坐标,再将所有定位得到的坐标点整合到一起,得到点云图像,如图1(c)所示。将点云图像进行渲染,得到最终的SMLM图像,如图1(d)所示。

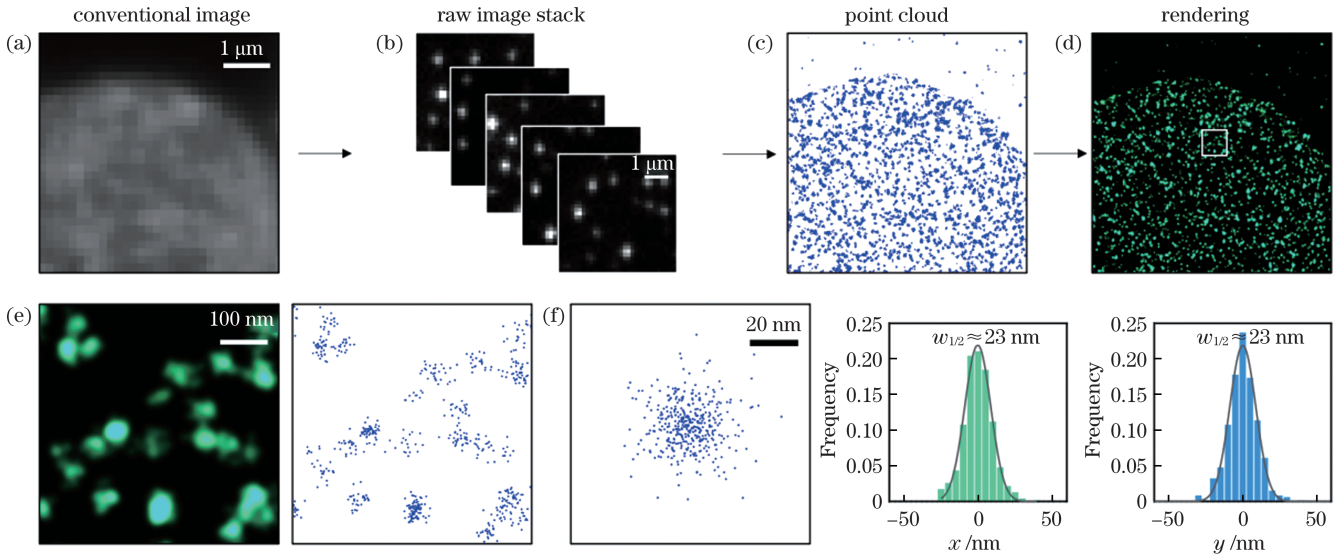

图1 SMLM成像流程及图像特征。(a)普通荧光成像;(b)SMLM数据采集;(c)点云图像重建;(d)点云图像渲染;(e)渲染图像的局域放大图;(f)SMLM分辨率

Fig. 1 SMLM imaging process and SMLM image characteristics. (a) Conventional fluorescence image; (b) SMLM data acquisition; (c) point cloud image reconstruction; (d) point cloud image rendering; (e) enlarged view of selected region in rendered image; (f) SMLM spatial resolution

从图像上看,渲染后的点构成了SMLM图像的基本组成单元,如图1(e)所示。为了确定SMLM图像的分辨率,将20个较为孤立的点簇质心重合,得到呈高斯分布的闪烁点簇,其x、y方向的半峰全宽$w_{1/2}$均约为20 nm,此即SMLM成像的横向分辨率,如图1(f)所示。SMLM成像的膜蛋白可能呈现为聚集的簇状分布或随机分布,而较为分散的随机分布形式容易导致定位后得到的点云数据分布不均匀,传统方法不能较好地对其进行聚类分析。因此,亟须开发新的方法,以便对随机分布膜蛋白的SMLM图像进行精确识别与分割。

2.2 DBSCAN原理

DBSCAN算法需要输入邻域半径ε以及该邻域半径内包含的最少点数量M两个参数,以获得SMLM数据中的簇[19]。该算法首先选取区域中的任意一点作为起始点,若该点同时满足DBSCAN所需的两个参数,则该点被归类为核心点。找到核心点后,算法继续以相同的标准对该核心点ε邻域内的其他点依次进行判断,若某点在ε邻域内依然满足最少点数量M,则其继续被标记为核心点并加入簇中,否则判定为边界点并加入簇中,直到没有符合条件的点为止。如果数据中的某个点既不满足核心点的要求,也不满足边界点的要求,则该点被归类为噪声点。DBSCAN算法通过不

断地重复上述过程，遍历数据中所有的点完成对SMLM数据点簇的识别。该算法的优势在于：一是可以基于固定的参数直接给出指定区域内簇的数量（细胞膜上膜蛋白的密度往往是未知的，同时分析区域大小的选择也需要综合实际情况，因此簇的总量也是未知的）；二是该算法计算时的延展扩张特性可以使其识别出具有特殊形状的簇，只要选取适当的参数即可对SMLM图像实现理想分割。但该算法也存在明显的缺点：当膜蛋白分布较为随机时，容易出现点密度分布不均匀的情况，单一的参数设置可能并不适用于区域内所有的簇，参数选择不当既可能导致一些点被错误地归为噪声点，也可能将两个距离较近的簇合并成一个簇[30]。

2.3 层次聚类原理

层次聚类算法首先将每个点当作单独的聚类，然后按照一定的规则（如欧氏距离）进行计算，找到每个点的最近邻点，随后将两个类连接起来合并得到新的类；不断重复这一过程，直到所有点全部归到一个类为止。这一操作使得点与点之间的关系产生了层次。层次聚类可以更好地发现球形簇，且层次聚类不会产生噪声点[31]；但层次聚类在SMLM数据应用中的缺点是聚类数量需要人为给定，然而在大多数情况下，人们很难对区域内的膜蛋白数量作出准确判断。

2.4 多次DBSCAN和层次聚类结合的方法

对于膜蛋白SMLM数据点簇密度分布不均匀的问题，可以通过改变邻域半径参数ε的方法解决。加入可变邻域半径后的DBSCAN算法得到的聚类数量的准确度虽然相比传统DBSCAN有了较大提升，但同时也由于为了识别更加相邻的点簇而产生了较多的噪声点。因此，在传统DBSCAN的基础上，本文引入层次聚类对点簇进行二次分类。以模拟数据集D31[32]为例，本文方法的具体步骤如下：

1）输入点云数据，如图2(a)所示。

2）输入较大参数（ε_1、M_1）执行第一次DBSCAN算法。由于参数较大，该过程会将图像中过于离散的点当作噪声点去除，同时将一部分距离较近的点簇合并到一起，如图2(b)所示。选择较大参数的标准是该参数分割后的图像中既有被单独分割的点簇，也有未被分割而导致合并的点簇。若参数选择得过小[如图3(b)所示]，就会使后续面积阈值的选取出现困难，进而可能导致错误的聚类。

3）计算每个初步识别的簇的面积（A），并用其除以平均面积（$\langle A\rangle$）进行归一化。选取适当的阈值参数进行阈值划分[如图2(c)所示]，并提取出面积较大的簇[如图2(d)所示]。在图2(c)所示阈值图y轴上单个簇的面积值与被错误合并簇的面积值之间选取阈值。

4）输入较小或相等的DBSCAN参数执行第二次DBSCAN计算，此次的M_2参数与M_1相同，邻域参数ε_2选取小于或等于ε_1。对每一个初步分割的簇依次从ε_2至ε_1执行循环。不同的ε参数会导致被分割的点簇数量不同。若ε_2选择得太小，可能导致大量甚至所有数据点被归类为噪声；若ε_2选择得过大，则分割结果与第一次分割结果相同。因此，本文选择从ε_2至ε_1过程中可分割的点簇数的最大值作为层次聚类的聚类参数，并执行层次聚类，如图2(e)所示。

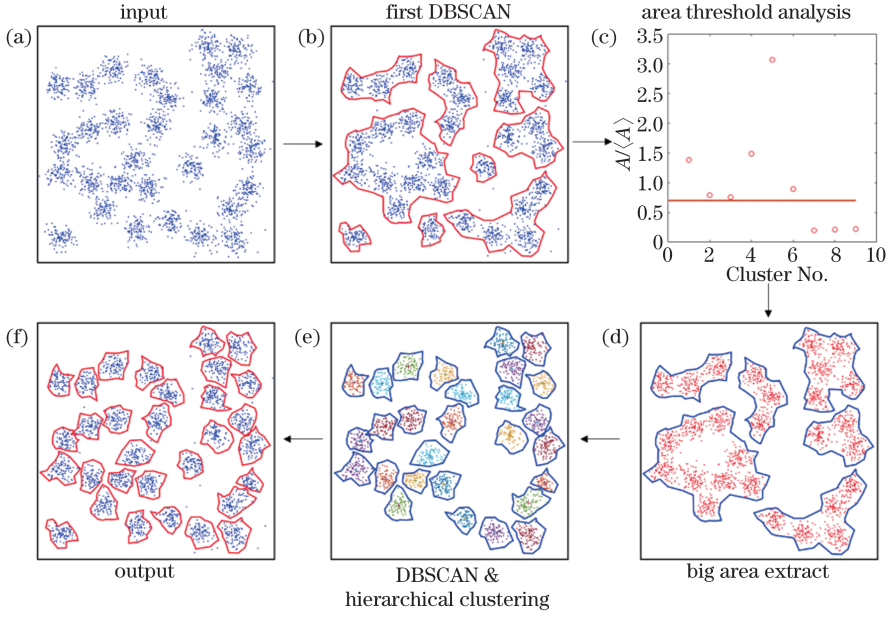

图2 结合多次DBSCAN和层次聚类进行图像分割的流程图。(a)数据输入；(b)第一次DBSCAN分割；(c)簇面积分析和阈值设定；(d)超阈值面积簇的提取；(e)二次DBSCAN分割和层次聚类；(f)聚类结果输出

Fig. 2 Flowchart of image segmentation by combination of multi-step DBSCAN and hierarchical clustering algorithm. (a) Input of original data; (b) the first DBSCAN segmentation; (c) cluster area analysis and threshold setting; (d) super-threshold area clusters extraction; (e) secondary DBSCAN segmentation and hierarchical clustering; (f) output of segmentation result

5) 将两次聚类的结果进行合并，得到最终的聚类结果，如图 2(f)所示。

本文算法相比于传统 DBSCAN 算法有两个优势：一是可以识别分割密度不均匀的点簇，二是可以缓解传统 DBSCAN 算法在参数选择上的敏感性。对于传统 DBSCAN 算法，当最少点数量 M 一定时，调整 ε 参数分别为 5、6、7、8，聚类效果如图 3(c)～(f)所示。当 ε=7 时，聚类效果最好；当 ε=5 时，传统 DBSCAN 算法将大量点归为噪声点且丢失了很多点簇的识别；当 ε=6 时，传统 DBSCAN 算法可以将点簇正确识别，但将很多点归为噪声点；当 ε=8 时，传统 DBSCAN 算法因将两个簇合并为一个而导致了错误分割。改进后的算法只需考虑聚类后的个数而不用考虑噪声点是否过多的问题，因此参数的选择更加宽泛。

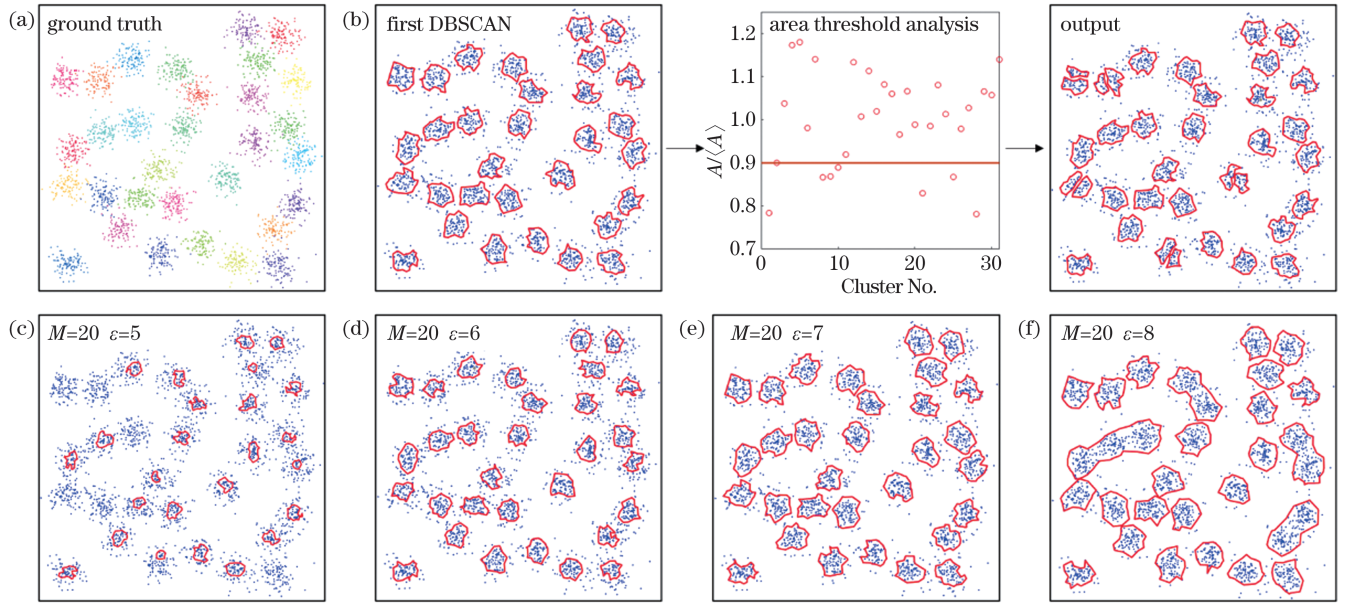

图 3 参数选择对聚类性能的影响。(a)D31 数据集的真实值；(b)参数选择过小对改进聚类算法的影响；(c)～(f)不同参数对传统 DBSCAN 算法性能的影响

Fig. 3 Influence of parameter selection on clustering performance. (a) Ground truth of D31 dataset; (b) influence of selecting too small parameter on improved clustering algorithm; (c)-(f) influence of different parameters on performance of traditional DBSCAN algorithm

3 分析与讨论

为了验证改进聚类算法在点簇识别和聚类中的有效性，以及相较于传统 DBSCAN 算法的优缺点，首先使用文献中的模拟数据集 D31[32] 和 S2[33] 进行测试。D31 数据集为随机分布的高斯簇，每个簇有 100 个点，如图 4(a)所示。对于 D31 数据集，两种算法均可以正确识别相同数量的点簇，如图 4(b)、(c)所示，但传统 DBSCAN 算法为了将点簇分辨开，不得不将大量点簇周围的点定义为噪声。相较于数据集给出的真实值，改进聚类算法对 D31 数据集识别的纯净度为 95.64%（传统 DBSCAN 算法为 86.52%），调整兰德系数为 0.9186（传统 DBSCAN 算法为 0.6463）。此外，本文还使用轮廓系数和噪声率对数据进行分析，结果显示：改进聚类算法的轮廓系数相较于传统 DBSCAN 算法有所提升，噪声率较传统 DBSCAN 算法亦有所下降，如表 1 所示。S2 数据集由 15 个高斯簇组成，共 5000 个点，如图 4(d)所示，该数据集的点簇中心比 D31 更聚集，点簇外围比 D31 更分散。对于 S2 数据集，改进聚类算法相比传统 DBSCAN 算法分割得更精准，如图 4(e)、(f)所示。改进聚类算法对 S2 数据集识别的纯净度为 95.52%（传统 DBSCAN 算法为 77.38%），调整兰德系数为 0.9128（传统 DBSCAN 算法为 0.6777），轮廓系数和噪声率也分别有一定上升和下降，如表 1 所示。改进聚类算法由于引入多次循环对每个错误合并的簇进行重新识别分割，可能会导致计算量和计算时间增加。

为了测试改进后的聚类算法在实验中的性能，选择了三种不同分布模式下的膜蛋白图像进行实验。对于膜蛋白相对均匀且离散分布的 SMLM 图像[如图 5(a)所示]，K 函数峰值对应的 x 轴坐标值为 27 nm，如图 5(b)所示。传统 DBSCAN 算法和改进聚类算法均能有效且准确地识别膜蛋白点簇的位置，改进聚类算法的轮廓系数略有提升，噪声率略有下降，如图 5(a)所示。对于膜蛋白随机分布的 SMLM 图像[如图 5(c)所示]，部分点簇之间的距离恰好位于 SMLM 分辨率附近，点云图像存在一定程度的非均匀性，K 函数峰值对应的 x 轴坐标值为 34 nm，如图 5(d)所示。若 ε 参数选取得过大，传统 DBSCAN 算法无法分辨距离较近的

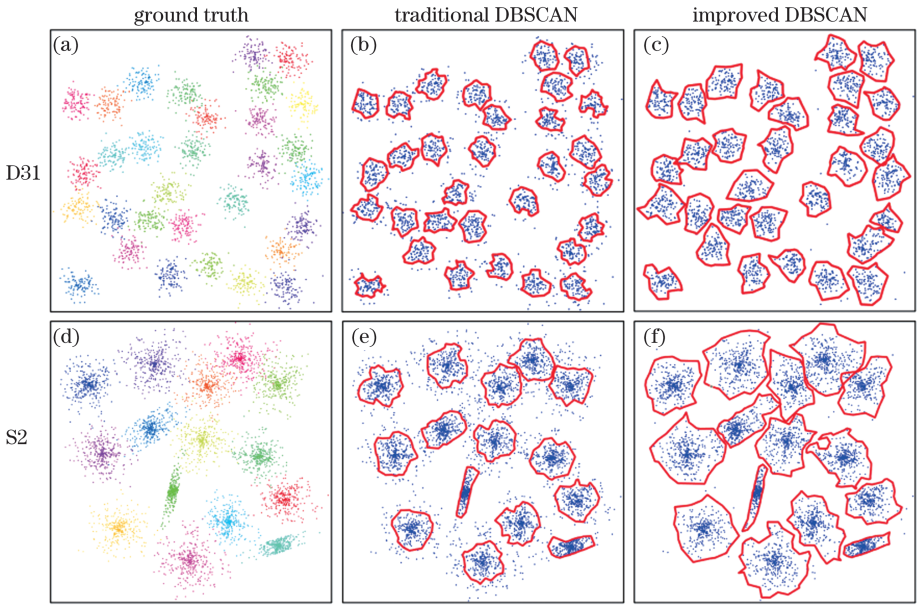

图 4 模拟数据集的聚类分割。(a)D31 数据集的真实值;(b)传统 DBSCAN 算法对 D31 数据集的分割;(c)改进聚类方法对 D31 数据集的分割;(d)S2 数据集的真实值;(e)传统 DBSCAN 算法对 S2 数据集的分割;(f)改进聚类算法对 S2 数据集的分割

Fig. 4 Clustering segmentation of simulated datasets. (a) Ground truth of D31 dataset; (b) segmentation of D31 dataset by traditional DBSCAN method; (c) segmentation of D31 dataset by improved clustering method; (d) ground truth of S2 dataset; (e) segmentation of S2 dataset by traditional DBSCAN algorithm; (f) segmentation of S2 dataset by improved clustering algorithm

表 1 传统 DBSCAN 算法和改进聚类算法的性能对比

Table 1 Performance comparison between traditional DBSCAN and improved clustering algorithm

Sample	Number of parameter		Identified cluster		Purity /%		Adjusted Rand index		Silhouette coefficient		Noise ratio /%		Running time /s	
	Trad.	Imp.	Trad.	Imp.	Trad.	Imp.	Trad.	Imp.	Trad.	Imp.	Trad.	Imp.	Trad.	Imp.
D31	2	4	31	31	86.52	95.64	0.6463	0.9186	0.4400	0.5606	12.96	0.48	0.0433	0.1358
S2	2	4	15	15	77.38	95.52	0.6777	0.9128	0.5091	0.6069	10.60	0.46	0.0842	1.2282
Fig. 4 (a)	2	4	14	14					0.7080	0.7306	3.93	2.08	0.0032	0.0385
Fig. 4(b)	2	4	19	23					0.3544	0.5725	16.42	0.08	0.0047	0.0583
Fig. 4(c)	2	4	11	11					0.4215	0.5351	14.98	3.67	0.0070	0.0911

点簇;若 ε 参数选取得过小,则会漏掉部分游离的膜蛋白点簇,如图 5(c)中箭头所指。改进后聚类方法成功识别了闪烁次数较少的膜蛋白点簇,分离了相对位置接近分辨率的点簇,获得了更加准确的膜蛋白空间分布信息。对于膜蛋白分布更加不均匀的 SMLM 图像,如图 5(e)所示,其点簇更为弥散,K 函数峰值对应的 x 轴坐标值为 50 nm,如图 5(f)所示。传统 DBSCAN 算法为了将点簇识别出来而错误地将较多的定位点归为噪声点,影响了后续的数据分析,而改进聚类算法可以在识别准确的前提下尽可能保留每个簇内的定位信息,如图 5(e)所示。

判断非均匀分布膜蛋白图像聚类分割效果的优劣主要取决于两点:一,是否识别到了密度较低部分的点簇;二,是否对高密度部分点簇实现了有效分割。改进聚类算法对于更复杂的膜蛋白分布有更优异的分割表现,其具备三个优点:1) 可以将距离接近分辨率的膜蛋白簇分开;2) 孤立且闪烁较少的膜蛋白簇可以被识别并保留下来;3) 避免过度去除所谓噪声点,从而更好地保留点簇定位次数信息。

改进聚类算法的输入参数较多,导致参数的选择相对困难,但参数的选择和调整并非完全没有规律。对于 SMLM 数据,成像的分辨率约为 20 nm,如图 1(f)所示,因此在执行第一次 DBSCAN 时,可以先将 ε_1 设置在 20~25 nm 附近,而参数 M 则可以根据实际图像中孤立单分子点的最少闪烁次数决定。执行完第一次 DBSCAN 操作后,一般会产生点簇合并的情况,此时可以将 ε_2 设置在 10 nm 附近,该邻域半径与点簇的半峰全宽相近,不宜设置得过小。此外,分析图 5 还可以发现 K 函数峰值对应的 x 轴坐标信息可能可以作为改进聚类算法的参数,达到简化参数的目的。

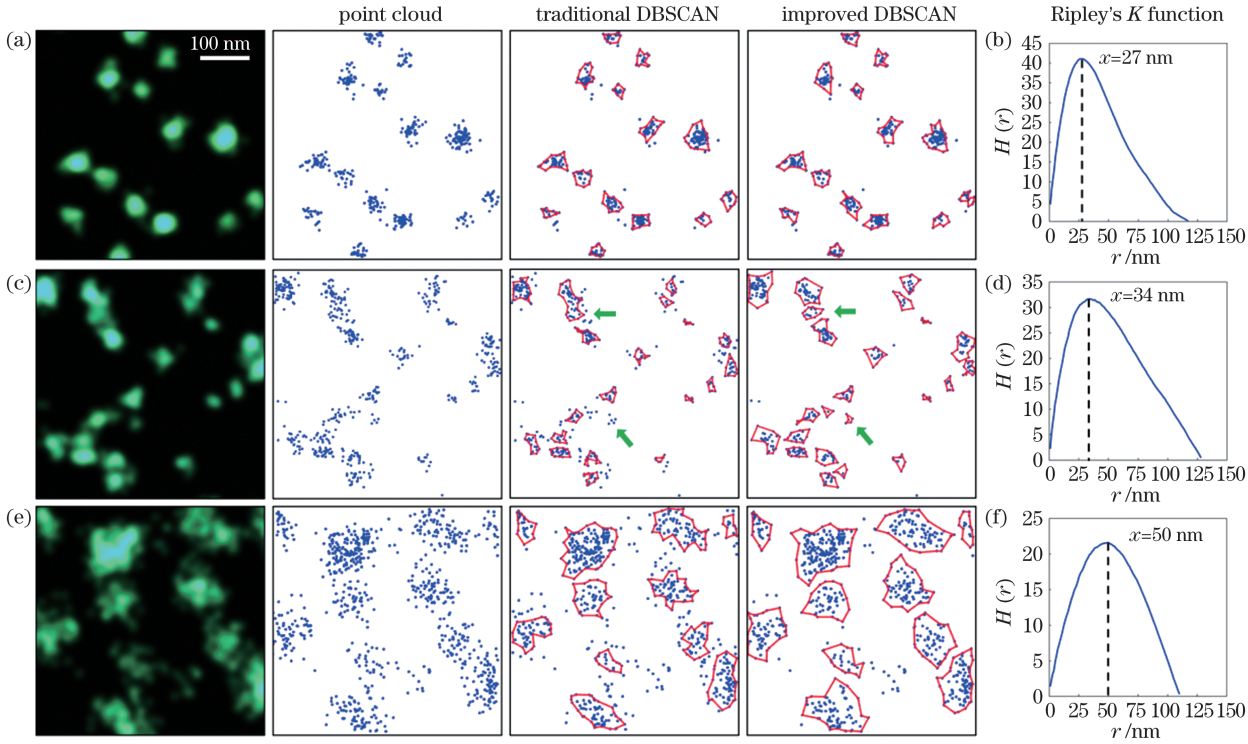

图 5 膜蛋白 SMLM 超分辨图像的聚类分割。(a)均匀膜蛋白数据的分割效果;(b)均匀膜蛋白数据的 K 函数;(c)随机分布膜蛋白数据的分割效果;(d)随机分布膜蛋白数据的 K 函数;(e)非均匀分布膜蛋白数据的分割效果;(f)非均匀分布膜蛋白数据的 K 函数

Fig. 5 Clustering segmentation for SMLM image of membrane proteins. (a) Segmentation effect of uniform distributed membrane protein SMLM data; (b) Ripley's K function of uniform distributed membrane protein SMLM data; (c) segmentation effect of random distributed membrane protein SMLM data; (d) Ripley's K function of random distributed membrane protein SMLM data; (e) segmentation effect of non-uniform distributed membrane protein SMLM data; (f) Ripley's K function of non-uniform distributed membrane protein SMLM data

4 结 论

膜蛋白在细胞表面分布的多样性和复杂性导致 SMLM 成像得到的点云密度不均匀。本文根据膜蛋白 SMLM 图像数据点簇的特征,在传统 DBSCAN 算法的基础上引入了面积阈值划分和二次聚类划分。结果显示,相较于传统 DBSCAN 算法,改进聚类算法显著提高了对已知点簇数据集和膜蛋白 SMLM 实验图像的聚类分割能力,并且在保持一定降噪能力的同时很大程度地还原了每个点簇的定位信息。但该算法仍然存在两个缺点:一是参数设置较多,未来可以考虑加入其他自适应算法来确定参数,在不降低聚类分割精度的前提下减小对参数的依赖;二是该算法引入的多次聚类循环导致计算量偏大,计算时间过长。对于其他需要即时演算获得结果的应用场景来说,对膜蛋白的识别还是应以聚类精准为前提。综上,本文算法在保留一定降噪能力的同时较好地实现了对点簇密度不均匀膜蛋白的 SMLM 图像的精确分割,为研究膜蛋白在细胞膜上纳米尺度的空间分布特性提供了新方法。

参 考 文 献

[1] Stone M B, Shelby S A, Veatch S L. Super-resolution microscopy: shedding light on the cellular plasma membrane[J]. Chemical Reviews, 2017, 117(11): 7457-7477.

[2] Hell S W, Wichmann J. Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy[J]. Optics Letters, 1994, 19(11): 780-782.

[3] Klar T A, Jakobs S, Dyba M, et al. Fluorescence microscopy with diffraction resolution barrier broken by stimulated emission[J]. Proceedings of the National Academy of Sciences of the United States of America, 2000, 97(15): 8206-8210.

[4] Gustafsson M G L. Surpassing the lateral resolution limit by a factor of two using structured illumination microscopy[J]. Journal of Microscopy, 2000, 198(2): 82-87.

[5] Gustafsson M G L. Nonlinear structured-illumination microscopy: wide-field fluorescence imaging with theoretically unlimited resolution[J]. Proceedings of the National Academy of Sciences of the United States of America, 2005, 102(37): 13081-13086.

[6] Sigal Y M, Zhou R B, Zhuang X W. Visualizing and discovering cellular structures with super-resolution microscopy[J]. Science, 2018, 361(6405): 880-887.

[7] 杨建宇, 潘雷霆, 胡芬, 等. 随机光学重构显微术及其应用研究进展[J]. 红外与激光工程, 2017, 46(11): 1103008.
Yang J Y, Pan L T, Hu F, et al. Stochastic optical reconstruction microscopy and its application[J]. Infrared and Laser Engineering, 2017, 46(11): 1103008.

[8] 杨建宇, 董浩, 邢福临, 等. 单分子定位超分辨成像技术进展及应用[J]. 激光与光电子学进展, 2021, 58(12): 1200001.
Yang J Y, Dong H, Xing F L, et al. Single-molecule localization super-resolution microscopy and its applications[J]. Laser & Optoelectronics Progress, 2021, 58(12): 1200001.

[9] Rust M J, Bates M, Zhuang X W. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM)[J]. Nature Methods, 2006, 3(10): 793-796.

[10] Betzig E, Patterson G H, Sougrat R, et al. Imaging intracellular

fluorescent proteins at nanometer resolution[J]. Science, 2006, 313(5793): 1642-1645.
[11] Rossboth B, Arnold A M, Ta H S, et al. TCRs are randomly distributed on the plasma membrane of resting antigen-experienced T cells[J]. Nature Immunology, 2018, 19(8): 821-827.
[12] Pritchard H A T, Pires P W, Yamasaki E, et al. Nanoscale remodeling of ryanodine receptor cluster size underlies cerebral microvascular dysfunction in Duchenne muscular dystrophy[J]. Proceedings of the National Academy of Sciences of the United States of America, 2018, 115(41): E9745-E9752.
[13] Yan Q Y, Lu Y T, Zhou L L, et al. Mechanistic insights into GLUT1 activation and clustering revealed by super-resolution imaging[J]. Proceedings of the National Academy of Sciences of the United States of America, 2018, 115(27): 7033-7038.
[14] Khater I M, Nabi I R, Hamarneh G. A review of super-resolution single-molecule localization microscopy cluster analysis and quantification methods[J]. Patterns, 2020, 1(3): 100038.
[15] Wu Y L, Tschanz A, Krupnik L, et al. Quantitative data analysis in single-molecule localization microscopy[J]. Trends in Cell Biology, 2020, 30(11): 837-851.
[16] Ripley B D. Modelling spatial patterns[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1977, 39(2): 172-192.
[17] Sengupta P, Jovanovic-Talisman T, Skoko D, et al. Probing protein heterogeneity in the plasma membrane using PALM and pair correlation analysis[J]. Nature Methods, 2011, 8(11): 969-975.
[18] Rubin-Delanchy P, Burn G L, Griffié J, et al. Bayesian cluster identification in single-molecule localization microscopy data[J]. Nature Methods, 2015, 12(11): 1072-1076.
[19] Ester M, Kriegel H P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]// Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, USA. Menlo Park: AAAI Press, 1996: 226-231.
[20] Mazouchi A, Milstein J N. Fast optimized cluster algorithm for localizations (FOCAL): a spatial cluster analysis for super-resolved microscopy[J]. Bioinformatics, 2015, 32(5): 747-754.
[21] Levet F, Hosy E, Kechkar A, et al. SR-Tesseler: a method to segment and quantify localization-based super-resolution microscopy data[J]. Nature Methods, 2015, 12(11): 1065-1071.
[22] Andronov L, Orlov I, Lutz Y, et al. ClusterViSu, a method for clustering of protein complexes by Voronoi tessellation in super-resolution microscopy[J]. Scientific Reports, 2016, 6: 24084.
[23] Bushra A A, Yi G M. Comparative analysis review of pioneering DBSCAN and successive density-based clustering algorithms[J]. IEEE Access, 2021, 9: 87918-87935.
[24] 孙兰香, 王文举, 齐立峰, 等. 基于激光诱导击穿光谱技术在线监测碳纤维复合材料激光清洗效果[J]. 中国激光, 2020, 47(11): 1111003.
Sun L X, Wang W J, Qi L F, et al. Online monitoring of laser cleaning effect of carbon fiber composite materials based on laser-induced breakdown spectroscopy technology[J]. Chinese Journal of Lasers, 2020, 47(11): 1111003.
[25] 邵靖滔, 杜常清, 邹斌. 基于点云簇组合特征的激光雷达地面分割方法[J]. 激光与光电子学进展, 2021, 58(4): 0428001.
Shao J T, Du C Q, Zou B. Lidar ground segmentation method based on point cloud cluster combination feature[J]. Laser & Optoelectronics Progress, 2021, 58(4): 0428001.
[26] 张长勇, 陈治华, 韩梁. 基于改进DBSCAN的激光雷达障碍物检测[J]. 激光与光电子学进展, 2021, 58(24): 2428005.
Zhang C Y, Chen Z H, Han L. Obstacle detection of lidar based on improved DBSCAN algorithm[J]. Laser & Optoelectronics Progress, 2021, 58(24): 2428005.
[27] 王祝, 王智, 张旭, 等. 基于二维激光雷达的自适应阈值聚类分割方法[J]. 中国激光, 2021, 48(16): 1610005.
Wang Z, Wang Z, Zhang X, et al. Adaptive threshold clustering segmentation method based on two-dimensional lidar[J]. Chinese Journal of Lasers, 2021, 48(16): 1610005.
[28] Siddig S, Aufmkolk S, Doose S, et al. Super-resolution imaging reveals the nanoscale organization of metabotropic glutamate receptors at presynaptic active zones[J]. Science Advances, 2020, 6(16): eaay7193.
[29] Meng J X, Zhang Y, Saman D, et al. Hyperphosphorylated tau self-assembles into amorphous aggregates eliciting TLR4-dependent responses[J]. Nature Communications, 2022, 13: 2692.
[30] Schubert E, Sander J, Ester M, et al. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN[J]. ACM Transactions on Database Systems, 2017, 42(3): 19.
[31] Murtagh F, Contreras P. Algorithms for hierarchical clustering: an overview[J]. WIREs Data Mining and Knowledge Discovery, 2012, 2(1): 86-97.
[32] Veenman C J, Reinders M J T, Backer E. A maximum variance cluster algorithm[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(9): 1273-1280.
[33] Fränti P, Virmajoki O. Iterative shrinking method for clustering problems[J]. Pattern Recognition, 2006, 39(5): 761-775.

Clustering Segmentation for Single-Molecule Localization Super-Resolution Image of Membrane Protein by Combining Multi-Step DBSCAN and Hierarchical Clustering Algorithm

Yang Jianyu[1], Hu Fen[1], Xing Fulin[1], Dong Hao[1], Hou Mengdi[1], Imshik Lee[1], Pan Leiting[1,2,3,4*], Xu Jingjun[1,3]

[1]*Key Laboratory of Weak-Light Nonlinear Photonics, Ministry of Education, School of Physics, TEDA Institute of Applied Physics, Nankai University, Tianjin* 300071, *China*;
[2]*Frontiers Science Center for Cell Responses, State Key Laboratory of Medicinal Chemical Biology, College of Life Sciences, Nankai University, Tianjin* 300071, *China*;
[3]*Shenzhen Research Institute of Nankai University, Shenzhen* 518083, *Guangdong, China*;
[4]*Collaborative Innovation Center of Extreme Optics Shanxi University, Taiyuan* 030006, *Shanxi, China*

Abstract

Objective There are a variety of functional proteins localized on the cell membrane that participate in many crucial cellular

processes, such as signal transduction and transmembrane transport. The spatiotemporal distribution of specific membrane proteins largely determines their activity states and functions. It is known that the sizes of membrane proteins and the distances between them are both on a nanometer scale. Owing to diffraction limits, traditional optical microscopy cannot provide the spatial distribution of membrane proteins at the single-molecule level. Therefore, imaging techniques with strong specificity and high resolution are urgently required to reveal the precise spatial distribution of membrane proteins. Nowadays, single-molecule localization microscopy (SMLM) offers new opportunities to resolve the detailed distribution information of membrane proteins at the nanoscale, while the great improvement in spatial resolution also presents higher demands for accurate clustering segmentation of images. Density-based spatial clustering of applications with noise clustering (DBSCAN) is one of the most commonly used clustering methods; however, it shows relatively poor performance in clustering segmentation in SMLM images of membrane proteins with heterogeneous density. To address this issue, we propose a novel clustering method using a combination of a multi-step DBSCAN and a hierarchical clustering algorithm. This improved clustering method is based on the traditional DBSCAN method, which combines area threshold analysis and hierarchical clustering.

Methods In the present work, we improved the traditional DBSCAN method by introducing a variable neighborhood radius and hierarchical clustering to perform precise image clustering segmentation in the original image (Fig. 2). First, we inputted a relatively large parameter (ε_1, M_1) to perform the DBSCAN calculation. Owing to this relatively large parameter, the excessively discrete points in the original image were removed as noise points. Meanwhile, some of the close-point clusters merged together. Subsequently, the area of each preliminarily identified cluster was calculated and divided by the average area for normalization. Based on the acquired normalized values, we selected an appropriate threshold parameter for extracting clusters with a relatively large area. Subsequently, secondary DBSCAN was performed by the input of a smaller or equal parameter (ε_2, M_1; $\varepsilon_2 \leqslant \varepsilon_1$). For each point cluster extracted in the second step, the calculation was looped from ε_2 to ε_1. The parameter showing the maximum number of divisible point clusters in the output during the looped process from ε_2 to ε_1 was selected as the clustering parameter for the next hierarchical clustering. Finally, we combined the above two DBSCAN results to obtain the final clustering segmentation result.

Results and Discussions We tested this improved clustering method on both simulated and experimental SMLM data. For the simulation datasets, we chose the D31 and S2 datasets from previous studies as our test objects (Fig. 4). The purity of the improved method on the D31 dataset was 95.64% (86.52% for the traditional DBSCAN method), and the adjusted Rand index was 0.9186 (0.6463 for the traditional DBSCAN method). In addition, the silhouette coefficient and noise ratio were used to analyze the two datasets. Compared with the traditional DBSCAN method, the silhouette coefficient of the improved method significantly increased, and the noise ratio decreased (Table 1). For the S2 dataset, the improved method also exhibited a more accurate segmentation effect than the traditional DBSCAN method. The identification purity of the improved method for the S2 dataset was 95.52% (77.38% for the traditional DBSCAN method), and the adjusted Rand index was 0.9128 (0.6777 for the traditional DBSCAN method). The silhouette coefficient and noise ratio increased and decreased, respectively (Table 1). For the experimental SMLM data, we tested the clustering segmentation effect of the improved method on the uniform, random, and non-uniform SMLM images of membrane proteins (Fig. 5). Similarly, the improved clustering method has a higher accuracy and silhouette coefficient, and a lower noise ratio (Table 1). However, it is regrettable that the time consumption of the improved clustering method is higher than that of the traditional DBSCAN method for both the simulated and experimental datasets (Table 1).

Conclusions Based on the characteristics of the point clusters in SMLM images of membrane proteins, we proposed a novel clustering method that combines area threshold segmentation and multi-step clustering segmentation based on the traditional DBSCAN algorithm. When we applied this method for the image segmentation of simulated datasets as well as experimental SMLM data of membrane proteins, the obtained parameters, including purity, adjusted Rand index, silhouette coefficients, and noise ratio, were generally improved compared with those of the traditional DBSCAN method. On the premise of accurate clustering recognition of super-resolution images and a certain noise reduction ability, the localization information of each cluster can be preserved as much as possible. Our method exhibites a good clustering segmentation ability, especially for SMLM images of membrane proteins with heterogeneous densities. This improved clustering method provides novel insights into the segmentation of membrane protein SMLM images, which is expected to facilitate research into the nanoscale spatial distribution of various membrane proteins.

Keywords bio-optics; single-molecule localization microscopy (SMLM); super-resolution image segmentation; membrane protein; density-based spatial clustering of applications with noise clustering (DBSCAN); hierarchical clustering algorithm