DOI: 10.3724/SP.J.1096.2012.10527

基于人工神经网络的生物组织质谱成像分类与识别方法

熊行创*[1,2] 方 向[2] 欧阳证[3] 江 游[2] 黄泽建[2] 张玉奎[1]
[1](北京理工大学生命科学与技术学院,北京 100081) [2](中国计量科学研究院,北京 100013)
[3](美国普渡大学韦尔登生物医学工程学院,西拉法叶 47907)

摘 要 生物组织质谱成像技术不仅能够展示组织的生物分子信息,而且能直观地显示分子空间分布,是当今生物质谱的研究热点。如何对生物组织质谱成像的数据进行基于生物分子的有效分类与识别是该领域关注的重要问题,特别对于病变组织与其邻近非病变组织的区分与识别和生物组织功能区域的划分与鉴定具有重要的意义。本研究开发出一种新的分类与识别方法。其流程是,首先进行质谱成像数据预处理,应用无监督的自组织特征映射网络区分组织样品区与非组织区域,提取组织区域的质谱数据,应用有监督的学习向量量化网络对已知类别数据进行学习训练,建立模型;应用模型对未知样品进行识别。应用本方法对 6 个膀胱癌患者的膀胱癌变组织与邻近非癌变组织的质谱成像数据进行分类与识别,结果显示,癌变组织判错率低于 23.38%,而非癌变组织判错率低于 9.08%,表现出较高的准确度;对 3 片邻近的小鼠大脑切片质谱成像数据进行白质与灰质区域划分,将中间的 1 片用于训练,两边的 2 片用于验证,结果显示,自组织特征映射网络的分类结果与学习向量量化网络的预测结果不一致率低于 4%。本方法基于生物分子的质谱成像组织区域分类与识别,具有较高准确度和操作简便等优点,在临床医学研究领域有大规模的应用潜能。

关键词 质谱成像; 分类与识别; 自组织特征映射网络; 学习向量量化网络

1 引 言

生物质谱成像(Mass spectrometry imaging,MSI) 是近几年快速发展起来的生物分子成像技术,不仅能展示组织的生物分子信息,而且可直观显示分子空间分布[1~3]。MSI 广泛应用于从生物细胞到生物组织的蛋白、多肽和脂质分子的成像研究,包括药物及其代谢物在组织内的分布研究、生物医学诊断、分子病理研究、以及三维生物分子质谱成像研究等[4~10]。

如何对生物组织质谱成像的测试数据进行基于生物分子的有效分类与识别,是 MSI 研究领域关注的重要问题,也是利用质谱成像提供分子生物空间信息的关键。特别对于病变组织与其邻近非病变组织的区分与识别,癌变与非癌变的判定,以及癌变的早、中、晚期识别具有重要意义。同时对生物组织功能区域的划分与鉴定、功能区域边界的划分与认定等,同样具有重要意义。

选择分类与识别依据的变量类型直接关系到分类与识别模型的成败。可以作为判别的变量类型有:疾病标志物的有无、单一分子的含量差异和多分子的复合差异。利用疾病标志物的质谱成像分类与识别很简单直观,但由于疾病标志物难以寻找和发现已知的标志物种类太少,依赖于此的质谱成像分类应用研究过于狭窄。另外,应用单一分子含量差异能够得到对比显著的质谱成像图,但其结果通常不可靠。因为样品分析环节中存在诸多可以导致单一分子含量产生显著变化的因素,容易掩盖样品间的本质差异,而且样品个体间的本身差异太大,以至于判定阈值难以适用。多分子的复合差异,相对单一分子差异,能够显著增强其可靠性。本研究采用多分子的复合差异作为分类与识别的依据变量。

基于多分子复合差异分类与识别的系统方法包括:提取多分子复合差异的特征信息,并应用已知样品的特征信息进行模型训练获得判别规则,再将判别规则应用于其它未知样品的有效区分与识别。这类方法的研究才刚刚起步。当前,美国普渡大学 Cooks 教授研究组对人膀胱癌组织与邻近非癌变组织进行解吸附电喷雾离子化(Desorption electrospray ionization, DESI) 质谱成像分析,应用多元统计偏最小

2011-05-24 收稿;2011-06-27 接受
本文系科技支撑计划(Nos. 2009BAK58B03,2009BAK59B03) 资助
* E-mail: xingchuangxiong@gmail.com

二乘判别分析(Partial least-square discriminate analysis,PLS-DA)方法进行训练和判别,取得了很好的结果[11]。然而,在PLS-DA方法判断的过程中需要人为选定参与训练和判别主成分的数量(这将直接影响最终的判定结果),而且整个过程相对复杂,对应用该方法的人员提出了较高的专业背景要求。

本研究建立了从质谱原始数据处理到基于人工神经网络的生物组织质谱成像分类与识别方法。充分利用自组织特征映射网络(Self-organizing feature map,SOFM)无监督、自组织自学习网络特点来区分样品区与非样品区。SOFM相对其它的自组织网络(竞争层网络)既可以学习训练数据输入向量的分布特征,也可以学习训练输入向量的拓扑结构,具有聚类速度快、结果精确等特点[12,13]。在获取了样品区域后,再应用学习向量量化网络(Learning vector quantization,LVQ)进一步对样品区的生物组织进行有监督的学习训练,建立模型,应用模型对其它未知样品进行类别识别。LVQ有一个创建原型的优势,其结果易于解释,在模式识别和优化领域有着广泛的应用[12,14]。应用6个膀胱癌患者的膀胱癌变组织与邻近非癌变组织的质谱成像数据和3片临近的小鼠大脑切片质谱成像数据测试本方法的效果。从测试数据看,本方法有效、简便、实用,具有大规模应用的潜能。

2 实验部分

生物组织质谱成像的分类与识别方法的流程示于图1。本方法的总体策略是,为了获得判别规则模型的高稳定性,剔除与组织样品本质特征无关的信息,包括剔除弱小质谱峰信号,以减少参与判别的无意义变量;应用无监督SOFM方法将非样品区排除在外,即减少无关的变量和无关的样品采样点的干扰,增加可靠性。方法的具体步骤如下:首先将原始质谱数据重构成质谱图像数据。质谱图像数据格式可以是科学图像Analyze 7.5格式数据或标准通用的imzXML格式。本研究采用本课题组开发的imgGenerate软件(http://msimaging.net)实现此操作。对质谱成像数据进行基线扣除,信号剔除(剔除出现机率小的质谱峰信号)等预处理操作,以排除化学噪声的干扰和增强样品区域与非样品区的对比度。质谱成像的各离子图像的浏览可以用Biomap软件(http://www.maldi-msi.org)或MSI-View软件(http://msimaging.net)查看。本研究采用MSI-View软件实现此步骤的操作。

对预处理后的质谱成像数据进行无监督SOFM分类。根据每个样品点的分子组成和含量的差异与相似度分类,分为样品区和非样品区两类。质谱峰是体现每个样品点的分子组成和含量的关键数据。但是采用质谱相对峰强还是绝对峰强,分类效果差别较大。在对比样品区与非样品区时,以质谱绝对峰强为变量产生的差异明显大于相对峰强产生的差异,数据测试也验证了以质谱绝对峰强作为输入变量的分类效果要明显优于相对峰强。因此,在应用无监督SOFM进行样品区与非样品区分类时,采用质谱绝对峰强作为输入变量。本研究应用Matlab(The Mathworks, Natick, MA, USA)的人工神经网络函数开发的SOFM分类软件实现样品区与非样品区的区分。在具体应用SOFM分类时,可以直接选用1×2神经元结构将整个分析区域分为组织样品区与非组织样品区两类,也可以选用2×2或2×3的神经元结构,最后根据网络拓扑结构较容易的区分样品区与非样品区。

对样品区内的质谱成像数据进行有监督的LVQ训练,建立识别模型。由于建立后的识别模型要用于在相同实验条件下获得的其它测试数据的判别,因此要求用于建立模型和用于判别的输入变量稳定可靠,否则建立的模型没有意义。由于实验随机误差等因素,组织中生物分子的质谱绝对峰强容易产生较大的波动(即便是同一样品的重复实验),因此不可以作为LVQ输入变量。而组织中生物分子构成不同,生物分子的质谱相对峰强也不同,实验随机误差能够影响分子的质谱绝对峰强,但是难以改变质谱相对峰强的本质特征,因此以质谱相对峰强可以作为LVQ输入变量。采用1对或多对已知样品的数据进行训练,建立模型。建立好的识别模型由在相同实验条件获得的已知样品进行验证,再对未知样品进行预测。本研究应用Matlab的人工神经网络函数开发出LVQ训练、识别模型软件工具来实现对组织区域类型的判断。

3 结果与讨论

3.1 人膀胱癌变组织的质谱成像分类与识别

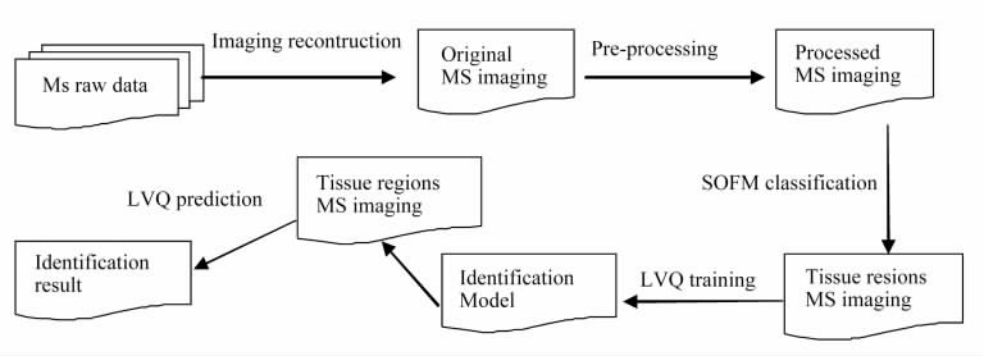

图 1　生物组织质谱成像分类与识别方法流程示意图

Fig. 1　Workflow diagram of classification and identification for biological imaging mass spectrometry data in this method

由美国普渡大学 Cooks 教授研究组提供的 6 个膀胱癌患者的癌变组织与邻近非癌变组织质谱成像数据用于测试此方法的效果。这 6 对癌变组织及非癌变组织均由提供该组织的临床医生进行过严格确认。质谱原始数据均由 DESI 离子源(负离子模式)结合 LTQ 线性离子阱质谱仪(Thermo Fisher Scientific San Jose. CA, USA)进行质谱成像分析获得[11]。

对原始质谱数据(质量范围 m/z 150 ~ 1000,单位质量分辨)进行图像重构和数据预处理,剔除化学噪声(剔除质谱峰出现概率小于 1% 的信号),获取 46 个有明显意义的质谱峰。应用以质谱绝对峰强为变量的 SOFM 方法(2 × 2 神经元结构)对所有样品进行组织区域和非组织区域分类。为了验证在应用 SOFM 方法区分组织区域与非组织区域时,以质谱绝对峰强为变量的策略优于以相对峰强的策略,采用这两种策略对 12 片组织数据分别进行 SOFM 分类,将两种策略获得的结果与相应的组织染色图进行对照。数据显示,相对于以质谱相对峰强为变量的策略,以绝对峰强为变量的策略的 SOFM 方法更容易将组织区域与非组织区域分开。其中一例的对比结果示于图 2。

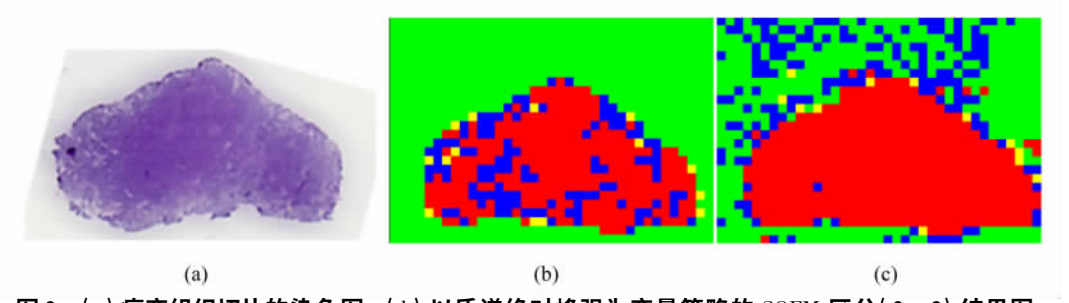

图 2　(a)癌变组织切片的染色图,(b)以质谱绝对峰强为变量策略的 SOFM 区分(2 × 2)结果图,(c)以质谱相对峰强为变量策略的 SOFM 区分(2 × 2)结果图。组织区域(红色、蓝色和黄色)与非组织区域(绿色)在图(b)中区分明显,而在图(c)中却相对难以区分

Fig. 2　(a) H & E stained tissue section of the tumor tissue, (b) the classification result map using self-organizing feature map (SOFM) (2 × 2 neuronal structure) with absolute peak intensity as input vector, (c) the classification result map using SOFM (2 × 2 neuronal structure) with relative peak intensity as input vector. In Figure (b), the tissue regions (in red, blue and yellow) can be clearly classified from the non-tissue regions (in green). That is difficult to be obtained in Figure (c)

对样品区内的数据信息进行分析,癌变组织与邻近非癌变组织的典型平均质谱图示于图 3。在癌变组织中,m/z 537.7 和 563.5 等脂肪酸与 m/z 788.7(glycerophosphoserines 18: 0/18: 1)和 m/z 885.7(glycerophosphoinositols 18: 0/20: 4)等脂质分子的含量相对比值明显高于正常组织中的相对比值。

随机选取 1 对样品数据应用 LVQ 网络进行训练,建立模型,并用这 6 对已知类型样品数据对模型进行检验。6 对组织的总离子强度图像和应用 LVQ 判断结果示于图 4。从图 4 可见,每对组织的脂质分子绝对峰强都不相同,因此绝对峰强不能作为判断模型的输入参数;癌变组织脂质分子总含量明显高于非癌变区域,与文献[15]一致;判错率低,错判的主要分布在分子信号较弱的区域。

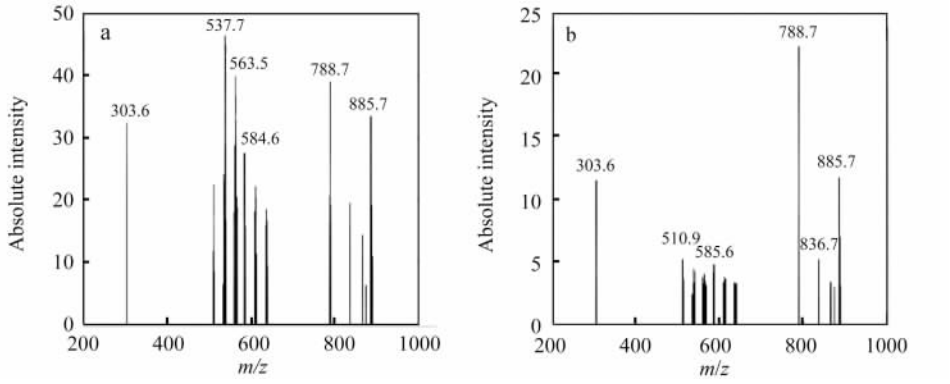

图 3 人膀胱癌变组织与邻近非癌变组织的平均质谱图:(a) 癌变组织,(b) 邻近非癌变组织

Fig. 3 Mean mass spectra plots of human cancerous and adjacent normal bladder tissue sample: (a) cancerous tissue samples, (b) adjacent normal tissue samples

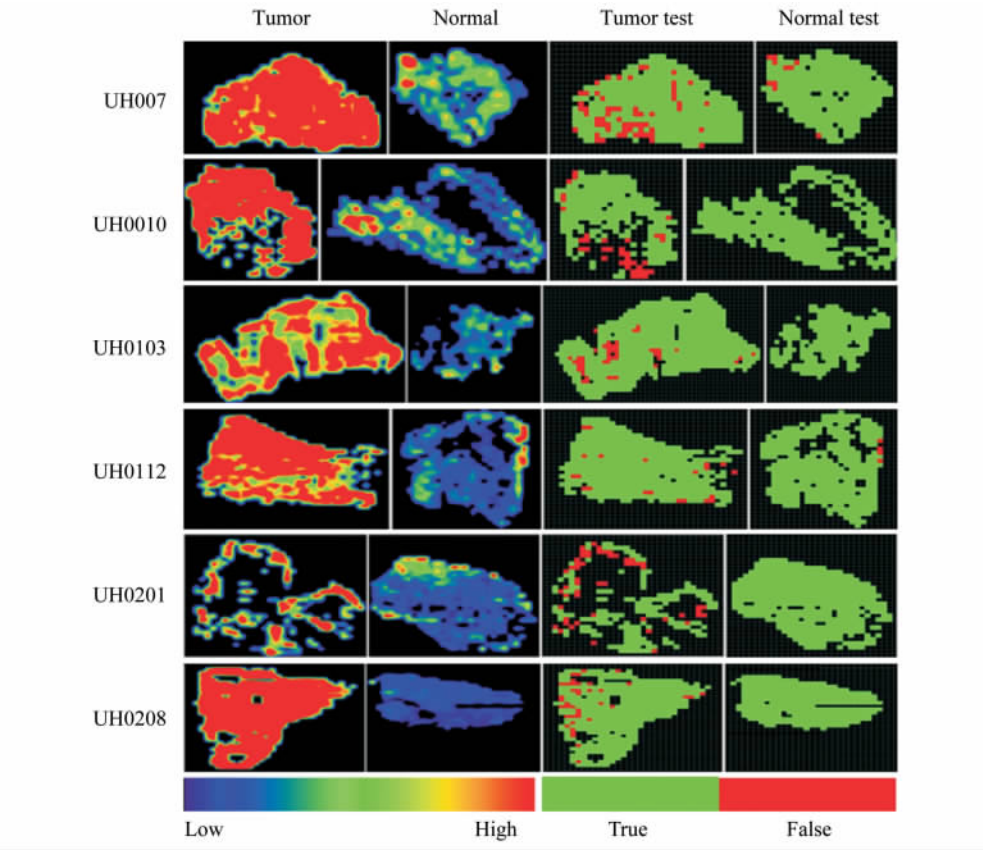

图 4 6 对人膀胱癌变组织与邻近非癌变组织的总离子强度图像和应用 LVQ 判断结果图。在 LVQ 判断结果图中,绿色表示判对的区域,红色表示判错的区域

Fig. 4 Total ion image maps of six pairs of human cancerous and adjacent normal bladder tissue samples, and the corresponding result maps identified by learning vector quantization (LVQ) . In the result maps by LVQ, the right identified regions are indicated in green and the regions of mis-judgment are with red

统计判错率,数据示于表 1。由 1 对样品数据训练建立的模型识别 6 对已知的样品类别,癌变判错率最大值 23.38%,均值 12.32%;非癌变判错率最大值 3.85%,均值 0.82%。由 2 对样品数据训练建立的模型识别 6 对已知的样品类别,癌变判错率最大值 10.73%,均值 5.11%;非癌变判错率最大值 9.08%,均值 2.08%。模型的识别正确率较高,癌变判错率高于非癌变判错率。而且 2 对样品数据训练的模型相对于 1 对样品数据训练的模型,识别正确率较高,服从训练样品数据信息越丰富,模型识别准确性越高的

规律。

3.2　小鼠大脑组织切片的质谱成像分类与识别

3 张空间邻近的小鼠大脑组织切片的质谱成像数据由美国普渡大学 Cooks 教授研究组提供。这 3 张切片的编号是 N260，N273 和 N294，其中 N273 在中间，距离 N260 和 N294 分别是 0.26 和 0.28 mm，具有相似的生物分子空间分布。这 3 张组织切片数据均由 DESI 离子源(负离子模式)结合 LTQ 线性离子阱质谱仪进行质谱成像分析获得[9]。

表 1　应用 LVQ 方法识别 6 对人膀胱癌变组织与邻近非癌变组织判错统计表

Table 1　Error rate table of identification of six pairs of human cancerous and adjacent normal bladder tissue samples using learning vector quantization

样品 Samples	1 对样品数据训练模型 One pair of data set used for training		2 对样品数据训练模型 Two pairs of data set used for training	
	癌变判错率 Error rate of identification for cancerous regions (%)	非癌变判错率 Error rate of identification for normal regions (%)	癌变判错率 Error rate of identification for cancerous regions (%)	非癌变判错率 Error rate of identification for normal regions (%)
UH0007	15.58	3.85	1.04	9.08
UH0010	14.18	0	10.73	0.64
UH0103	6.98	0	6.84	0.38
UH0112	4.84	1.07	2.07	2.40
UH0201	23.38	0	9.09	0
UH0208	8.95	0	0.89	0
平均值 Mean	12.32	0.82	5.11	2.08

对原始质谱数据(质量范围 m/z 150 ~ 1000，单位质量分辨)进行图像重构和数据预处理。应用以质谱绝对峰强为变量的 SOFM 方法对组织切片数据进行样品区和非样品区分类，获得每个切片的样品区域。

3 张切片数据的总离子强度图示于图 6 左列。从总离子强度质谱图像中可以大体看出 2 个不同的区域。进一步对样品区内的数据信息进行无监督模式的 SOFM 分类(以每个样品点的质谱相对峰强为输入变量)，获取白质和灰质区域。这两个区域的典型平均质谱图示于图 5。在白质区域 m/z 888.8 (sulfatide 24: 1) 具有较高的相对峰强，而在灰质区域 m/z 834.4 (phosphatidylserine 18: 0/22: 6) 具有较高的相对峰强。

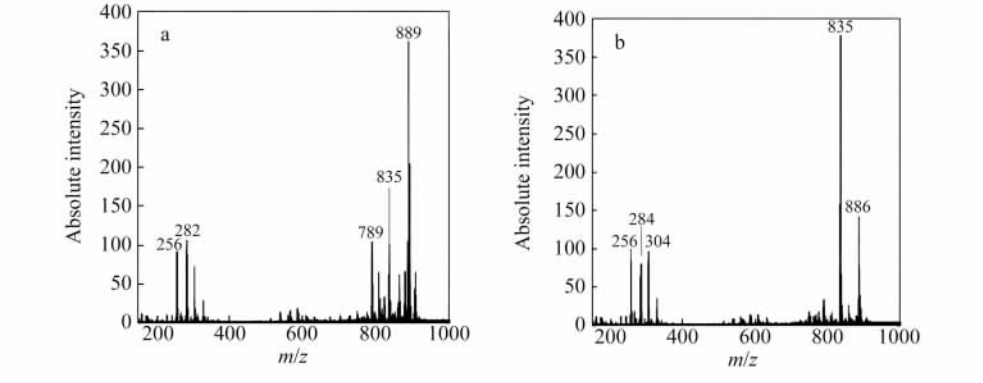

图 5　小鼠大脑切片组织的平均质谱图。(a) 白质区域质谱，(b) 灰质区域质谱

Fig. 5　Mean mass spectral plots of a mouse brain section: the mass spectral of white matter in mouse brain section, (b) mass spectral of gray matter in mouse brain section

将由 SOFM 对中间切片 N273 分类的结果数据用来作为 LVQ 网络的训练参数建立模型，再应用该模型来预测这 3 张切片上的白质与灰质区域。然后将预测结果与 SOFM 分类的结果对比。N260，N273 和 N294 这 3 张切片的不一致率分别是 1.75%，2.71% 和 4.00%。SOFM 分类结果示于图 6 中间列，LVQ 预测结果与 SOFM 分类结果对比图示于图 6 右列。从图 6 可见，LVQ 预测与 SOFM 分类结果一致性较高，不一致的结果多发生在白质与灰质相连接的区域(此区域的分子组成相对含量没有像白质与灰质区域那样特征明显)。

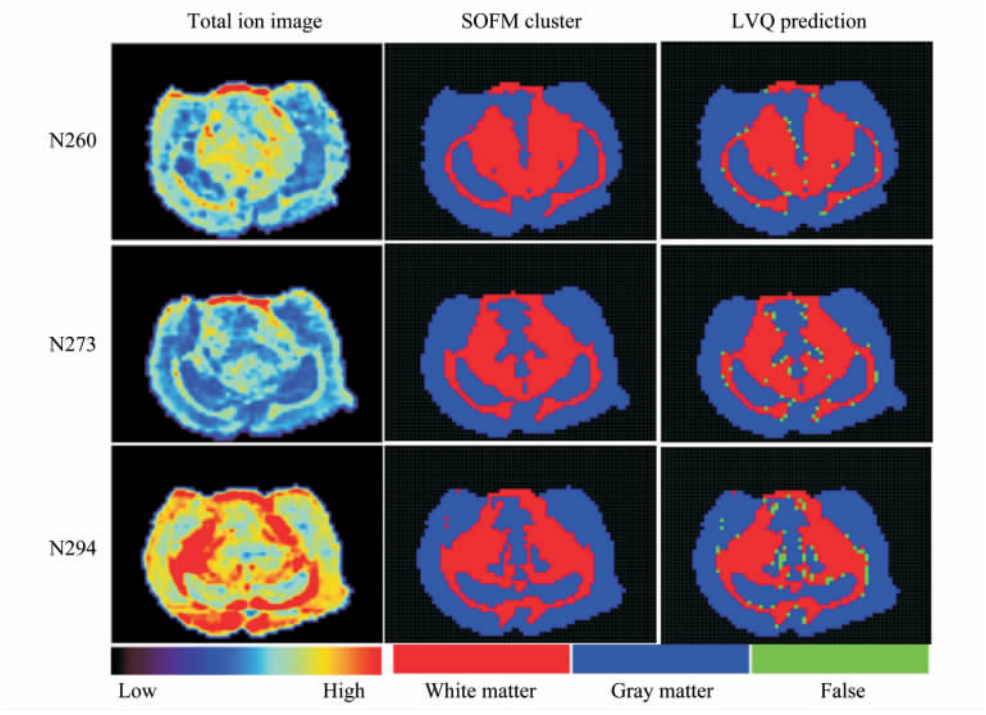

图 6　3 张邻近的小鼠大脑组织切片的总离子图像、应用 SOFM 分类图和 LVQ 判断结果图。SOFM 分类图和 LVQ 判断结果图中，红色表示白质区域，蓝色表示灰质区域，绿色表示 SOFM 分类结果与 LVQ 判断结果不一致的区域

Fig. 6　Total ion image maps of three adjacent slices of mouse brain tissue, the corresponding result maps classified by self-organizing feature map (SOFM) and the result maps identified by learning vector quantization (LVQ). In the result maps, the regions of white matter are in red, the regions of gray matter are in blue, and the regions in green indicate the inconsistent result by classification with SOFM and identification with LVQ

4　结　论

组合应用人工神经网络技术实现生物组织质谱成像的分类与识别方法，基于多分子复合差异，通过剔除弱小质谱峰信号和应用无监督 SOFM 方法把非样品区排除在外等措施，减少无关变量和无关样品采样点的干扰，增加可靠性，获得高准确度。本方法不需要判别阈值设置、关键变量选择等人为操作，因而操作简便。同时，本方法可以精确到具体单个采样点的识别，具有高精度的优点。在经过大量样品验证后，拥有作为常规工具应用于基于生物分子成像临床医学研究和生命科学研究的潜能。

致 谢　感谢美国普渡大学 Cooks 教授研究组提供的质谱成像原始数据。

References

1　Pacholski M L, Winograd N. *Chemical Reviews*, **1999**, 99(10): 2977 ~ 3006
2　McDonnell L A, Heeren R M A. *Mass Spectrometry Reviews*, **2007**, 26(4): 606 ~ 643
3　Caprioli R M. *Proteomics*, **2008**, 8(18): 3679 ~ 3680
4　Sinha T K, Khatib-Shahidi S, Yankeelov T E, Mapara K, Ehtesham M, Cornett D S, Dawant B M, Caprioli R M, Gore J C. *Nature Methods*, **2008**, 5(1): 57 ~ 59
5　Chaurand P, Rahman M A, Hunt T, Mobley J A, Gu G, Latham J C, Caprioli R M, Kasper S. *Molecular & Cellular Proteomics*, **2008**, 7(2): 411 ~ 423
6　Altelaar A F M, Luxembourg S L, McDonnell L A, Piersma S R, Heeren R M A. *Nature Protocols*, **2007**, 2(5): 1185 ~ 1196

7 LIU Hui, CHEN Guo-Qiang, WANG Yan-Ying, LI Zhi-Li. *Chinese J. Anal. Chem.*, **2011**, 39(1): 87 ~ 90
刘 辉, 陈国强, 王艳英, 李智立. 分析化学, **2011**, 39(1): 87 ~ 90
8 Nemes P, Barton A A, Vertes A. *Anal. Chem.*, **2009**, 81(16): 6668 ~ 6675
9 Eberlin L S, Ifa D R, Wu C, Cooks R G. *Angewandte Chemie-International Edition*, **2010**, 49(5): 873 ~ 876
10 Eberlin L S, Dill A L, Costa A B, Ifa D R, Cheng L, Masterson T, Koch M, Ratliff T L, Cooks R G. *Anal. Chem.*, **2010**, 82(9): 3430 ~ 3434
11 Dill A L, Eberlin L S, Costa A B, Zheng C, Ifa D R, Cheng L A, Masterson T A, Koch M O, Vitek O, Cooks R G. *Chemistry-a European Journal*, **2011**, 17(10): 2897 ~ 2902
12 Fritzke B. *Neural Networks*, **1994**, 7(9): 1441 ~ 1460
13 Kohonen T, Kaski S, Lagus K, Salojarvi J, Honkela J, Paatero V, Saarela A. *IEEE Transactions on Neural Networks*, **2000**, 11(3): 574 ~ 585
14 Ahalt S C, Krishnamurthy A K, Chen P K, Melton D E. *Neural Networks*, **1990**, 3(3): 277 ~ 290
15 Dill A L, Ifa D R, Manicke N E, Costa A B, Ramos-Vara J A, Knapp D W, Cooks R G. *Anal. Chem.*, **2009**, 81(21): 8758 ~ 8764

Artificial Neural Networks Method of Classification and Identification for Mass Spectrometry Imaging Data of Biological Tissue

XIONG Xing-Chuang*[1,2], FANG Xiang[2], OU Yang-Zheng[3], JIANG You[2], HUANG Ze-Jian[2], ZHANG Yu-Kui[1]
[1](*School of Life Science, Beijing Institute of Technology, Beijing* 100081)
[2](*National Institute of Metrology, Beijing* 100013)
[3](*Weldon School of Biomedical Engineering, Purdue University, West Lafayette* 47907, *USA*)

Abstract Mass spectrometry imaging (MSI), the combination of molecular mass analysis and spatial information, provides visualization of molecules on complex biological surfaces, thus is currently getting a significant amount of attention in the mass spectrometric community. One important problem in this researching field is how to develop an effective method of classification and identification for MSI data, especial for identifying the cancerous tissue from adjacent normal tissue and classifying the different functional regions in a complex biological tissue. For this purpose, we developed a new method, containing image reconstruction from raw mass spectral data, MSI data pre-processing, classification of tissue regions from background regions by self-organizing feature map and identification of special interesting regions from the whole tissue regions by learning vector quantization. The MSI data of six pairs (12 tissue samples) of human cancerous and adjacent normal bladder tissue samples were used to test the effect of this method. The result showed an error rate of less than 23.38% for identification of cancerous regions and an error rate of less than 9.08% for identification of the adjacent normal regions. The method was also tested to classify white matter and gray matter regions of three adjacent slices of mouse brain tissue. The slice in the middle was used to train and to establish an identification model; the other two slices were used to test the model. The inconsistent rate of the identification results by using self-organizing feature map is less than 4% comparing with the results using learning vector quantization. This indicated that the method could be performed simply and efficiently, to extend the capability of MSI, and underline its potential to be a regular tool applied to study on clinical application.

Keywords Mass spectrometry imaging; Classification and identification; Self-organizing feature map; Learning vector quantization

(Received 24 May 2011; accepted 27 June 2011)