DOI: 10.19756/j.issn.0253-3820.201083

基于随机森林算法的原位质谱快速鉴别肺癌的方法研究

欧阳永中*[1] 曾玉庭[2] 郭伟清[1] 邓金连[1] 魏益平[3]

[1](佛山科学技术学院环境与化学工程学院，佛山 528000)

[2](佛山科学技术学院食品科学与工程学院，佛山 528000) [3](南昌大学附属第二医院胸心外科，南昌 330006)

摘 要 随机森林(Random forest，RF) 算法是一种基于决策树的机器学习算法，具有良好的分类与变量筛选性能，因而在生物医学高维数据分析中应用广泛。本研究开发了一种基于 RF 算法的原位质谱快速鉴别肺癌的模型和方法，通过构建液体辅助表面解吸常压化学电离质谱技术平台(DAPCI-MS) ，结合 RF 算法，在常温常压条件下，直接实现对未处理人体肺鳞癌组织切片的准确鉴别与区分，并获取肺癌区别于正常组织的生物特征标记物。研究表明，当决策树数目 n_{tree} = 100 时，对人体肺鳞癌组织与邻近正常组织的区分准确率可达到 100%。与其它分类方法相比，本模型具有稳健性高、分类效果好、泛化能力强等特点，为实现复杂基质的人体肺癌组织与相邻正常组织的区分提供了一种快速、准确和可靠的分类模型。

关键词 随机森林算法; 表面解吸常压化学电离质谱技术; 肺癌组织切片; 特征生物标记物

1 引 言

肺癌是目前我国癌症发病率和死亡率最高的肿瘤[1]，其中，非小细胞肺癌(Non-small cell lung cancer，NSCLC) 约占所有肺癌的 80%，75%的患者发现患病时已处于中晚期，导致 5 年肺癌生存率非常低。目前，外科切除手术仍是大多数早期非小细胞肺癌的最佳治疗方法[2]。但是，在恶性肿瘤临床手术过程中，实现最佳手术效果的最大障碍之一是难以在短时间内确定区分肿瘤组织和邻近正常组织的清晰边界，不完全切除恶性肿瘤会导致局部术后复发症等系列问题[3]。因此，建立一种快速、准确区分肿瘤组织与邻近正常组织区域的诊断方法，对于辅助临床手术中恶性肿瘤的诊断和彻底根治性切除具有重要的应用价值。

目前，临床诊疗主要借助 CT 筛查[4]、胸片(CXR) [5]、正电子发射断层成像(PET) [6]、核磁共振成像(MRI) [7] 等医学影像技术进行肺癌筛查与癌症边界的鉴定。尽管基于冰冻切片的组织病理学实验目前仍是临床区分癌症与正常组织的金标准，但易受到样品处理过程复杂、干扰严重、处理时间长(>1 h) 等因素的制约[8,9]。自 2004 年以来，以电喷雾解吸电离(Desorption electrospray ionization，DESI) [10] 为代表的原位电离源质谱技术(Ambient mass spectrometry，AIMS) 相继出现[11~20]，由于其在无需样品预处理和大气压操作条件下，可直接实现复杂生物组织样本分析，拓展了现代质谱技术在生物医学[11~14]、临床诊断[15~20] 等领域的应用范围。近年来，空气辅助电喷雾解吸电离(Air flow-assisted desorption electrospray ionization，AFA-DESI) [21]、组织喷雾电离质谱(Tissue spray ionization-mass spectrometry，TSI-MS) [22]、表面解吸常压化学电离(Surface desorption atmospheric pressure chemical ionization，DAPCI) [23,24] 等原位电离质谱技术，结合偏最小二乘-线性判别分析(PLS-LDA) 或主成分分析方法(PCA) ，在人体肺癌组织切片的快速分析和鉴别方面的应用研究取得了进展。由于电离质谱信号的稳定性不佳、质谱数据复杂等因素，癌症与癌旁组织的区分准确度有待提升。

随机森林(Random forest，RF) 算法[25] 是一种基于决策树的集成学习(Ensemble learning) 算法，主要用于处理分类和回归问题[26,27]。RF 算法可处理海量数据和高维问题，提供变量重要性度量和相似度矩阵等有用信息[28]，具有训练速度快、分类效果好、不易过拟合、对包含奇异值和噪声的数据预测结果稳健性较好等特点[29,30]，并且能够借助多维标度分析技术(Multidimensional scaling，MDS) 将样本的

2020-02-18 收稿; 2020-05-06 接受
本文系国家自然科学基金项目(No. 21405013) 资助
* E-mail: ouyang7492@163.com

相似度矩阵可视化[31]。本研究旨在构建一种基于 RF 算法的原位质谱快速鉴别肺癌的分类模型。通过改进表面解吸常压化学电离质谱技术(DAPCI-MS)[32],结合 RF 算法,在常温常压操作环境下,直接实现对未处理人体肺鳞癌组织切片的准确区分与生物特征标记物的提取。

2　实验部分

2.1　仪器与试剂

自制的液体辅助 DAPCI 离子源(图 1)与商业购置的 LTQ 线性离子阱质谱仪(美国 Thermo Fisher 公司)耦合,DAPCI 离子源的内部构造与仪器参数详见文献[25];CM1950 冰冻切片机(德国徕卡公司)。

甲醇(色谱纯,美国天地有限公司);去离子水利用纯水仪(美国 Thermo Fisher 公司)制备。样本组织源于南昌大学第二附属医院,–80℃超低温储存。

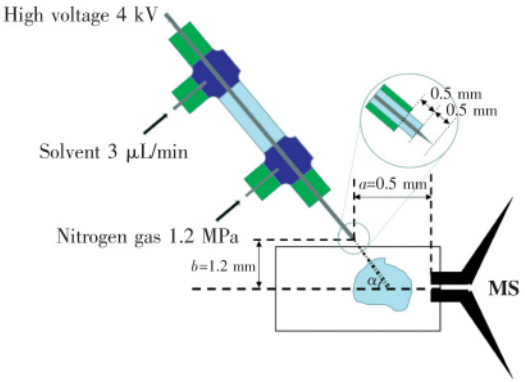


图 1　液体辅助表面解吸常压化学电离源(DAPCI)示意图

Fig. 1　Schematic diagram of liquid-assisted surface desorption atmospheric pressure chemical ionization (DAPCI) source

2.2　实验方法

2.2.1　样本　本研究已得到南昌大学第二附属医院的院内审查委员会医学伦理委员会的批准,得到了患者签署的知情同意书,并且所有临床研究均根据赫尔辛基宣言的原则进行。本研究共招募 15 位男性和 5 位女性患者,其中,16 例中分化患者,4 例低分化患者(见电子版文后支持信息表 S1)。患者的标准是经病理诊断的非小细胞肺鳞癌,并且没有伴随的恶性肿瘤,无其它肺部疾病和术前化学疗法或放疗史。每个患者都有两个匹配的肺鳞癌组织和相邻的正常肺组织样本对。

2.2.2　样本制备　将切片机设置在–20℃的条件下,预先运行 2 h 以上,进行肺癌组织样品切片的制作。实验前需将样本由超低温冰箱内取出,解冻至 4℃后,在真空干燥器内干燥约 15 min 后进行实验。将样品组织切成 10 μm 的厚度,处理好的切片固定在玻璃载玻片上,用于质谱分析。利用 CM1950 冰冻切片机将肺癌组织样本切成厚度为 10 和 6 μm 的薄片,分别置于玻璃载玻片上,直接用于质谱分析和标准染色法对比实验(组织病理学中通过苏木精/伊红(H&E)染色分析,并记录光学影像图,用于区分恶性肿瘤和正常组织)。在载玻片上做好标记,其中,癌症标记为 CA,正常组织标记为 CAB,并编号。

2.2.3　质谱参数　正离子模式,扫描范围 m/z 50~1000,离子源电压为 4 kV,离子传输管温度为 250℃。以甲醇-水(55∶45,V/V)混合溶液为离子源萃取剂,以 3 μL/min 流速通过石英毛细传输管,并使用鞘气流(N_2)以 1.2 MPa 压力雾化生成微滴。高压放电针尖距待测样品表面 1.2 mm,与待测样品表面夹角为 50°,放电针针尖与质谱进样口之间的距离为 5.5 mm。

2.2.4　RF 算法　RF 算法是 Breiman[25] 于 2001 年提出的一种基于决策树(Classification and regression tree, CART)的组合分类器。通过自助法(Bootstrap)[33] 重采样技术,由原始样本集中重复随机抽取同原始数据样本集个数相同的多个样本构成样本子集,利用每个样本子集构建决策树,然后融合多棵决策树的预测结果。在自助采样过程中,每棵决策树建立时只使用了初始训练集 63.2%的样本,剩余的 36.8%的样本可作为验证集对泛化性能进行“包外估计”,此数据称为袋外数据(Out-of-bag),可用于取代测试集进行误差估计。因此,RF 无需再进行交叉验证或者单独的测试集获取测试集误差,可用袋外数据误差(Out-of-bag classification error,OOB error)取代。还可用 OOB error 作为评价指标优化参数。RF 算法的随机性主要体现在数据采样和特征选择的随机性,通过优化决策树数量(n_{tree})和分裂变量数目(m_{try})进行模型优选。

2.2.5　数据处理　所有实验均使用由 XCalibur 2.0 软件(Thermo Fisher Scientific,San Jose,CA,USA)控制的 LTQ 线性离子阱质谱仪进行。待实验完成后,将全扫描质谱数据导入至 Excel 文件,利用 Matlab

(7.8.0,Mathworks,Inc.,Natick,MA)中的 Tree Bagger 函数进行 RF 分析,建立的模型将样本的分类结果以相似度矩阵的形式输出,并将 RF 算法得到的相似度矩阵通过多尺度分析(MDS)进行可视化。

3 结果与讨论

3.1 DAPCI-MS 一级扫描质谱分析

前期研究表明,在人体正常组织中,脂类(尤其是磷脂类化合物)的成分和含量发生显著变化是肿瘤性病变的一个重要信号[24,34]。图 2 为正离子模式下 DAPCI-MS 直接分析未处理人体肺鳞癌组织和相邻正常组织(A6 患者)的全扫描质谱图。图 2A、2B 和 2C 分别为空白背景、癌症组织和邻近正常组织在质谱扫描范围内(m/z 50~1000)的质谱分析结果。图 2B 和 2C 均为扣掉背景信号后的扫描平均结果。

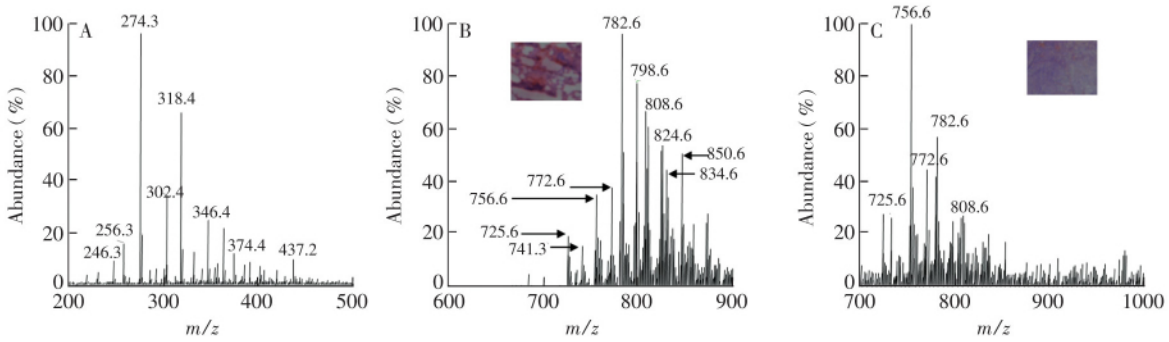


图 2 正离子模式下 DAPCI-MS 人体肺鳞癌组织全扫描质谱图:(A)空白背景载玻片;(B)恶性肿瘤组织区域;(C)与肿瘤相邻的正常组织区域。内插小图标分别为放大 200 倍的癌症和正常组织冰冻切片的 H&E 染色图

Fig.2 Full scan mass spectra of human squamous cell carcinoma lung tissue and adjacent normal tissue samples using DAPCI-MS in positive ion mode: (A) blank glass; (B) tumor tissue regions; (C) normal tissue regions adjacent to tumor tissue regions. The insets display the corresponding H&E-stained sections and the amplified figures (×200) showing different histopathological classes of lung tissues

由图 2B 和 2C 可知,DAPCI-MS 检测的离子信号主要集中在质量扫描范围(m/z 700~900)内,癌症和正常组织的磷脂酰胆碱(PC)和鞘磷脂(SM)类化合物的相对丰度或强度存在显著差异(其余 9 个组织样品中也观察到类似情况)。肺癌组织中磷脂酰胆碱化合物离子峰丰度,如 m/z 798.6([PC(34:1)+K]$^+$)、m/z 782.6([PC(34:1)+Na]$^+$)、m/z 808.6([PC(36:2)+Na]$^+$)、m/z 824.6([PC(36:2)+K]$^+$)、m/z 834.6([PC(38:3)+Na]$^+$)和 m/z 850.6 [PC(38:3)+K]$^+$等明显比邻近正常组织高,而邻近正常组织中磷脂酰胆碱类化合物(PC)(m/z 756.6 [PC(32:0)+Na]$^+$、m/z 772.6 [PC(32:0)+K]$^+$和 m/z 184.1($C_5H_{15}NO_4P^+$))和鞘磷脂化合物(m/z 725.6 [SM(16:0)+Na]$^+$)的离子相对丰度比癌症组织中高。正常组织中大量磷脂酰胆碱的产生和部分鞘磷脂的减少可能与肿瘤性病变有关,但是否为癌症的生物标记物还需进一步验证。这些化合物的结构已通过碰撞诱导裂解实验(CID)确认(电子版文后支持信息图 S1),与文献[24]报道结果一致。这些磷脂类化合物可视为区分癌症与正常组织的潜在生物标记物,表明 DAPCI-MS 是一种可直接检测肺癌组织异质性基质复杂化合物的有效手段。

3.2 RF 算法区分肺癌与邻近正常组织

为处理肺癌组织在 DAPCI-MS 直接分析中产生的大量高维质谱数据,采用 RF 算法实现癌症边界的界定和特征生物标记物的提取。采取每隔 0.2 mm 等间距采集数据的方法,分别对每个患者的癌症组织和正常组织区域采集 40 个质谱数据点,共 20 名肺癌确诊患者,共计采集 1600 个质谱数据点。将被标记的原始质谱数据集随机分成训练集和检验集,50%作为训练集用于训练分类模型,50%作为检验集评估分类模型性能。RF 算法最重要的两个参数是决策树的棵数 n_{tree} 和分裂属性集中属性个数 m_{try}。m_{try} 采用了 Breiman[25] 建议的默认值,而 RF 算法中决策树的构建是模型建立的核心,决策树的数量直接

影响随机森林分类算法的运算速度和分类效果。

由图 3A 可知，n_{tree} 在 100～500 之间，OOB 趋于稳定，故 $n_{tree}=100$ 时，分类误差低于 0.005，为最优模型。通过 MDS 对得到的相似度矩阵进行降维，得到肿瘤与临近正常组织的可视化结果（图 3B）。由图 3B 可见，肺癌组织与相邻正常组织能够完全分开，区分准确率达到 100%。

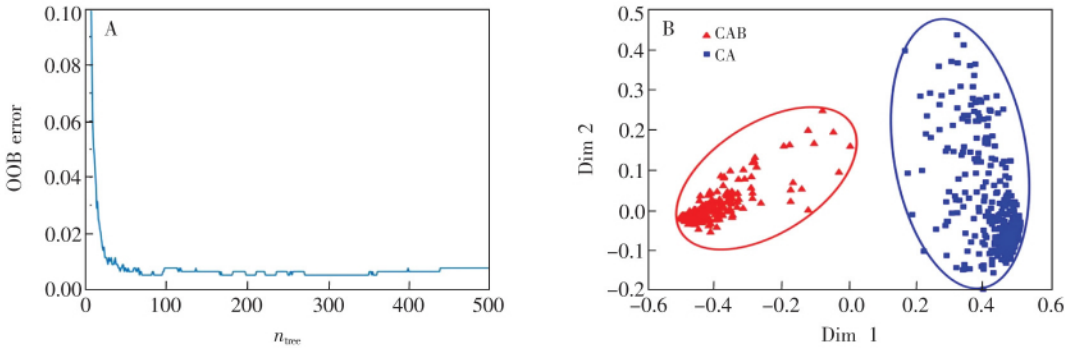


图 3　（A）随机森林决策树数目对袋外数据误差的影响；（B）肺癌与正常组织的随机森林多维标度二维分类图

Fig.3　(A) Influence of number of decision trees on out-of-bag classification error (OOB) in random forest (RF); (B) 2D graph of proximity of RF for distinguishing lung cancer from normal tissue using multidimensional scaling (MDS). CA: cancer tissue; CAB: normal tissue

为进一步验证所建模型的准确性，利用外部验证集进行验证。表 1 给出了 RF 算法模型混淆矩阵的分类结果，分类准确率达到 100%。预测类别和真实类别完全一致，这表明对于不同患者的癌症组织混合样本，RF 算法模型能够有效鉴别肺癌与相邻正常组织，且优于 PLS-LDA 预测的结果（错误率为 2.16%）[24]。此结果表明RF算法具有很好的分类效果，同时也展示了 DAPCI 对复杂基质样本良好的电离解吸和对复杂基质的耐受能力。RF 算法所有预测结果都经过病理组织冰冻切片的 H&E 染色图对比分析，预测与实际完全一致。

表 1　随机森林的混淆矩阵分类结果

Table 1　Results of classification of confusion matrix of RF

		预测的类别 Predicted class	
		CA	CAB
实际的类别 True class	CA	400	0
	CAB	0	400

注：每行表示实际的类别，每列表示随机森林判定的类别

The row indicates real classification; the column indicates predicted classification

3.3　RF 算法的特征变量重要性评估

RF 算法可通过调整变量顺序得到的 OOB 预测错误率衡量特征变量的重要性。如图 4 所示，采用节点不纯度的平均减少值作为度量变量重要性的指标，该值越高，表示该变量对分类的影响越大。由图 4 可知，根据横坐标节点不纯度 Gini 的平均减少值，对区分癌症与正常组织影响较大的前 15 个重要变量依次排序为：(m/z 782.6) [PC(34∶1)+Na^+]、(m/z 810.6) [PC(36∶1)+Na]$^+$、(m/z 808.6) [PC(36∶2)+Na]$^+$、(m/z 806.6) [PC(36∶3)+Na]$^+$、(m/z 834.6) [PC(38∶3)+Na]$^+$、(m/z 725.6) [SM(16∶0)+Na]$^+$、(m/z 798.6) [PC(34∶1)+K]$^+$、(m/z 756.6) [PC(32∶0)+Na]$^+$、(m/z 772.6) [PC(32∶0)+K]$^+$、(m/z 824.6) [PC(36∶2)+K]$^+$、(m/z 809.6) [SM(22∶0)+Na]$^+$、(m/z 826.6) [PC(36∶1)+K]$^+$、(m/z 780.6) [PC(34∶2)+Na]$^+$、(m/z 850.6) [PC(38∶3)+K]$^+$、(m/z 784.6) [PC(36∶3)+H]$^+$。上述磷脂酰胆碱（PC）和鞘磷脂（SM）类化合物对分类的贡献为 99.95%。因此，特

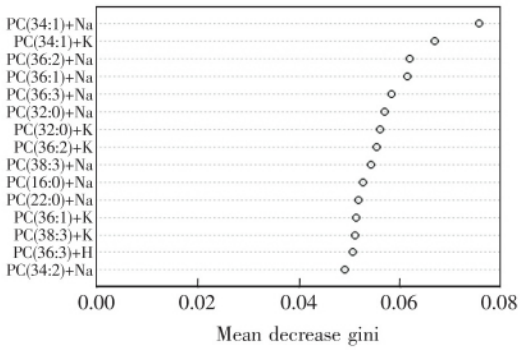


图 4　肺癌组织区分正常组织的变量重要性排序

Fig.4　Variables importance ranking for distinguishing lung cancer from normal tissues

征化合物可视为区分癌症与正常组织的潜在生物标记物，并进行了串联质谱实验验证(电子版文后支持信息图 S1)。

除部分特征离子化合物可直接由对比癌症与健康组织的一级质谱图获得外，如 *m/z* 782.6、808.6、834.6、725.6、798.6、756.6、772.6、824.6 和 850.6 等，其它特征生物标记物(如 *m/z* 810.6、806.6、809.6、826.6、780.6 和 784.6)都是通过 RF 获得。RF 具有良好的分类效果和变量筛选性能，具备处理大量高维数据的能力，尤其是通过分类效果评估解释变量的重要性，这对从复杂基质样本质谱分析中产生的海量大数据中获取特征生物标记物非常重要。此外，RF 能准确获取特征生物标记化合物对癌症的早期筛查或临床手术辅助界定癌症区域边界，对提高癌症治愈率具有重要的临床应用价值。

4 结 论

基于 RF 算法构建了一种原位质谱快速鉴别肺癌分类模型，在常温常压条件下，成功实现了对未处理人体肺鳞癌组织切片和相邻正常组织的准确诊断与区分。结果表明，RF 算法具有分类效率好、准确度高和不易过拟合等优点，解决了直接质谱分析肺癌组织中产生的大量高维质谱数据的难题，尤其是通过分类效果评估解释变量的重要性，克服了基于质谱峰强度对比的传统方法难以由高维大数据中获得特征化合物的弊端，以及 PCA、PLS-LSA 等其它机器学习算法在处理高维大数据中的局限性，为准确挖掘特征生物标记物提供了一种可靠和高效的方法，也为肺癌等基质复杂的恶性肿瘤组织的深入研究提供了一种新的分析手段。

References

1 ZHENG Rong-Shou, SUN Ke-Xin, ZHANG Si-Wei, ZENG Hong-Mei, ZOU Xiao-Nong, CHEN Ru, GU Xiu-Ying, WEI Wen-Qiang, HE Jie. *Chinese Clinical Oncology*, **2019**, 41(1): 19-28
郑荣寿, 孙可欣, 张思维, 曾红梅, 邹小农, 陈 茹, 顾秀瑛, 魏文强, 赫 捷. 中华肿瘤杂志, **2019**, 41(1): 19-28
2 Molina J R, Yang P, Cassivi S D, Schild S E, Adjei A A. *Mayo Clin. Proc.*, **2008**, 83(5): 584-594
3 Reck M, Heigener D F, Mok T, Soria J C, Rabe K F. *Lancet*, **2013**, 382(9893): 709-719
4 Field J K, Hansell D M, Duffy S W, Baldwin D R. *Lancet Oncol.*, **2013**, 14: e591-e600
5 Beek E J R, Mirsadraee S, Murchison J T. *World J. Radiol.*, **2015**, 7: 189-193
6 Toloza E M, Harpole L, McCrory D C. *Chest*, **2003**, 1: S137-S146
7 Edelman R R, Hatabu H, Tadamura E, Li W, Prasad P V. *Nat. Med.*, **1996**, 2: 1236-1239
8 Winther C, Graem N. *Apmis*, **2011**, 119: 259-262
9 Nakhleh R E. *Arch. Pathol. Lab. Med.*, **2011**, 135: 1394-1397
10 Takáts Z, Wiseman J M, Gologan B, Cooks R G. *Science*, **2004**, 306(5695): 471-473
11 Monge M E, Harris G A, Dwivedi P, Fernández F M. *Chem. Rev.*, **2013**, 113(4): 2269-2308
12 Feider C L, Krieger A, DeHoog R J, Eberlin L S. *Anal. Chem.*, **2019**, 91(7): 4266-4290
13 Swiner D J, Jackson S, Burris B J, Badu-Tawiah A K. *Anal. Chem.*, **2020**, 92 (1): 183-202
14 Chingin K, Liang J, Liu Y, Chen L, Wu X, Hu L, Ouyang Y Z. *RSC Adv.*, **2016**, 6(64): 59749-59752
15 Jia B, Ouyang Y Z, Hu B, Zhang T T, Li J Q, Chen H W. *J. Mass Spectrom.*, **2011**, 46(3): 311-319
16 Banerjee S, Zare R N, Tibshirani R J, Kunder C A, Nolley R, Fan R, Brooks J D, Sonn G A. *Proc. Natl. Acad. Sci. USA*, **2017**, 114(13): 3334-3339
17 Jarmusch A K, Pirro V, Baird Z, Hattab E M, Cohen-Gadol A A, Cooks R G. *Proc. Natl. Acad. Sci. USA*, **2016**, 113(6): 1486-1491
18 Margulis K, Chiou A S, Aasi S Z, Tibshirani R J, Tang J Y, Zare R N. *Proc. Natl. Acad. Sci. USA*, **2018**, 115(25): 6347-6352
19 Sans M, Gharpure K, Tibshirani R, Zhang J L, Liang L, Liu J S, Young J H, Dood R L, Sood A K, Eberlin L S. *Cancer Res.*, **2017**, 77(11): 2903-2913
20 Porcari A M, Zhang J, Garza K Y, Rodrigues-Peres R M, Lin J Q, Young J H, Tibshirani R, Nagi C, Paiva G R, Carter S A, Sarian L O, Eberlin M N, Eberlin L S. *Anal. Chem.*, **2018**, 90(19): 11324-11332
21 Li T G, He J M, Mao X X, Bi Y, Luo Z G, Guo C G, Tang F, Xu X, Wang X, H, Wang M R, Chen J, Abliz Z. *Sci.*

Rep., **2015**, 5: 14089
22 Wei Y, Chen L, Zhou W, Chingin K, Ouyang Y Z, Zhu T G, Wen H, Ding J, Xu J J, Chen H W. *Sci. Rep.*, **2015**, 5: 10077
23 ZHOU Zhi-Quan, ZHANG Ting-Ting, JIA Bin, OUYANG Yong-Zhong, FANG Xiao-Wei, CHEN Huan-Wen. *Chinese J. Anal. Chem.*, **2011**, 39(11): 1665-1669
周志权, 张婷婷, 贾 滨, 欧阳永中, 方小伟, 陈焕文. 分析化学, 2011, 39(11): 1665-1669
24 Ouyang Y Z, Liu J W, Nie B H, Dong N P, Chen X, Chen L F, Wei Y P. *RSC Adv.*, **2017**, 7(88): 56044-56053
25 Breiman L. *Mach. Learn.*, **2001**, 45(1): 5-32
26 Shi T, Horvath S. *J. Comput. Graph. Stat.*, **2006**, 15(1): 118-138
27 Janitza S, Strobl C, Boulesteix A L. *BMC Bioinformatics*, **2013**, 14: 119
28 Gislason P O, Benediktsson J A, Sveinsson J R. *Pattern Recognit. Lett.*, **2006**, 27(4): 294-300
29 Svetnik V, Liaw A, Tong C, Culberson J C, Sheridan R P, Feuston B P. *J. Chem. Inf. Comput. Sci.*, **2003**, 43(6): 1947-1958
30 Rodriguez-Galiano V F, Ghimire B, Rogan J, Chica-Olmo M, Rigol-Sanchez J P. *ISPRS J. Photogramm. Remote Sens.*, **2012**, 67: 93-104
31 Hout M C, Papesh M H, Goldinger S D. *Wiley Interdiscip Rev. Cogn. Sci.*, **2013**, 4(1): 93-103
32 LI Xin-Xin, CHEN Lin-Fei, OUYANG Yong-Zhong, FENG Fang, CHEN Huan-Wen. *Chinese J. Anal. Chem.*, **2016**, 44(1): 25-31
李欣欣, 陈林飞, 欧阳永中, 冯 芳, 陈焕文. 分析化学, 2016, 44(1): 25-31
33 Schapire R E. *Mach. Learn.*, **1990**, 5(2): 197-227
34 Glunde K, Jie C, Bhujwalla Z M. *Cancer Res.*, **2004**, 64(12): 4270-4275

Mass Spectrometric Discrimination of Human Lung Tumors under Ambient Conditions Based on Random Forest Algorithm

OUYANG Yong-Zhong*[1], ZENG Yu-Ting[2], GUO Wei-Qing[1], DENG Jin-Lian[1], WEI Yi-Ping[3]
[1](*School of Environmental and Chemical Engineering, Foshan University, Foshan* 528000, *China*)
[2](*School of Food Science and Engineering, Foshan University, Foshan* 528000, *China*)
[3](*Department of Cardiothoracic Surgery, Second Affiliated Hospital of Nanchang University, Nanchang* 330006, *China*)

Abstract Random forest algorithm (RF) is a machine learning algorithm based on decision trees. Due to the good performance of classification and variables selection, it has been widely used in biomedical high-dimensional data analysis. In order to fast and accurately distinguish human lung cancer from adjacent normal tissues, a model for direct ambient mass spectrometric analysis of lung cancer tissue sections based on random forest algorithm was developed. The purpose of this study was to establish a liquid assisted surface desorption atmospheric pressure chemical ionization mass spectrometry (DAPCI-MS) platform, combined with the random forest algorithm, to directly identify and differentiate the untreated human lung squamous cell carcinoma tissue sections under normal temperature and pressure, as well as obtaining the biomarkers of lung cancer for differentiation from normal tissue. The results showed that when the number of decision trees n_{tree} = 100, the accuracy of distinguishing human lung squamous cell carcinoma from adjacent normal tissues reached 100%. Compared with other methods, this model had higher robustness, better classification effect and stronger generalization ability. This study provided a more accurate and reliable classification model for rapid differentiation of human lung cancer tissues from adjacent normal tissues in complex matrix.

Keywords Random forest algorithm; Surface desorption atmospheric pressure chemical ionization; Lung cancer tissue section; Characteristic biomarkers

(Received 18 February 2020; accepted 6 May 2020)

This work was supported by the National Natural Science Foundation of China (No. 21405013).

支持信息

Supporting Information

基于随机森林算法的原位质谱快速鉴别肺癌的方法研究

欧阳永中*[1]，曾玉庭[2]，郭伟清[1]，邓金连[1]，魏益平[3]

[1]（佛山科学技术学院环境与化学工程学院，佛山 528000）

[2]（佛山科学技术学院食品科学与工程学院，佛山 528000）

[3]（南昌大学附属第二医院胸心外科，南昌 330006）

Mass Spectrometric Discrimination of Human Lung Tumors under Ambient Conditions Based on Random Forest Algorithm

OUYANG Yong-Zhong*[1], ZENG Yu-Ting[2], Guo Wei-Qing[1], DENG Jin-Lian[1], WEI Yi-Ping[3]

[1](School of Environmental and Chemical Engineering, Foshan University, Foshan 528000, *China)*
[2](School of Food Science and Engineering, Foshan University, Foshan 528000, *China)*
[3](Department of Cardiothoracic Surgery, Second Affiliated Hospital of Nanchang University, Nanchang 330006, *China)*

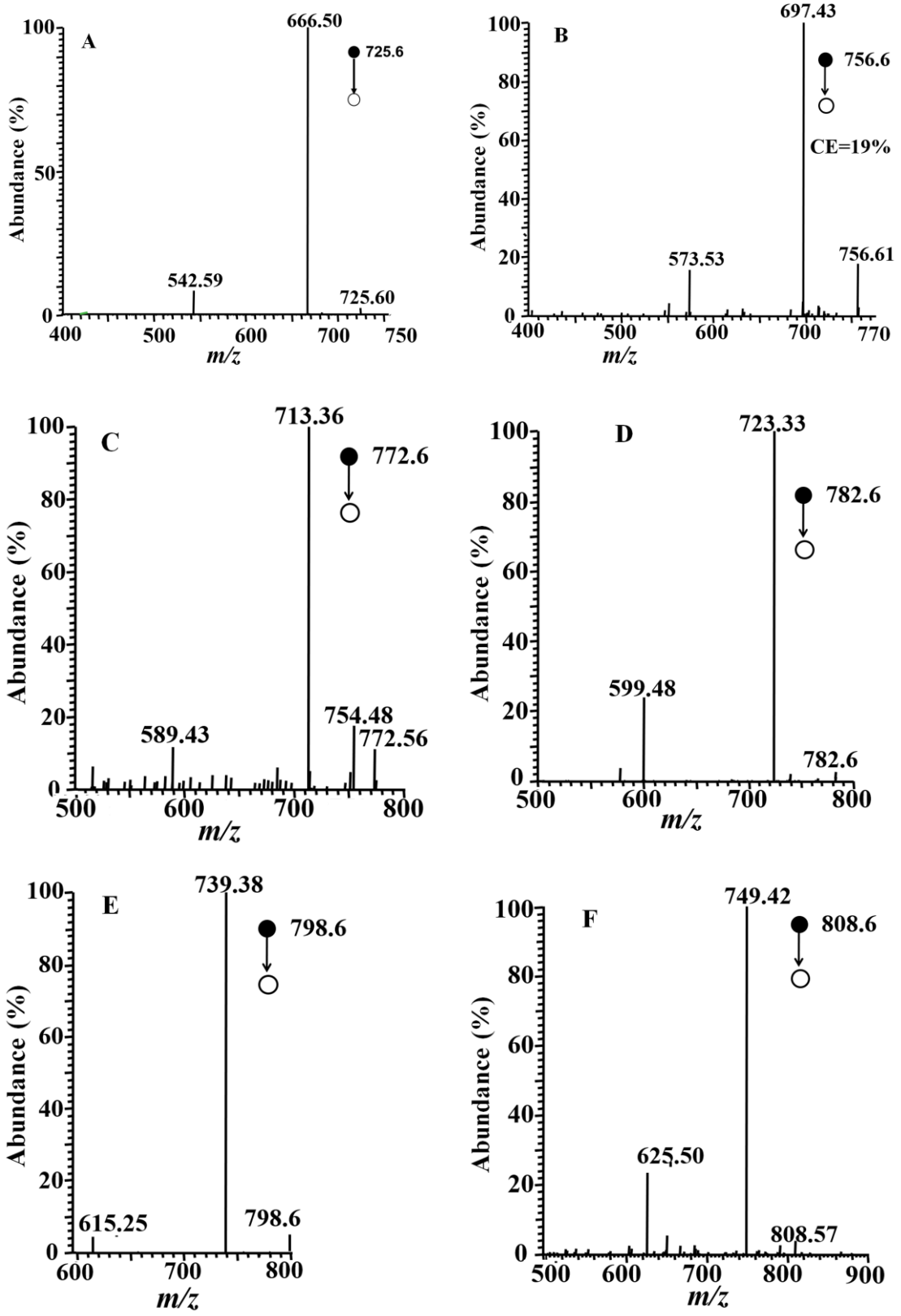
表 S1　20 名肺癌患者的组织学描述情况

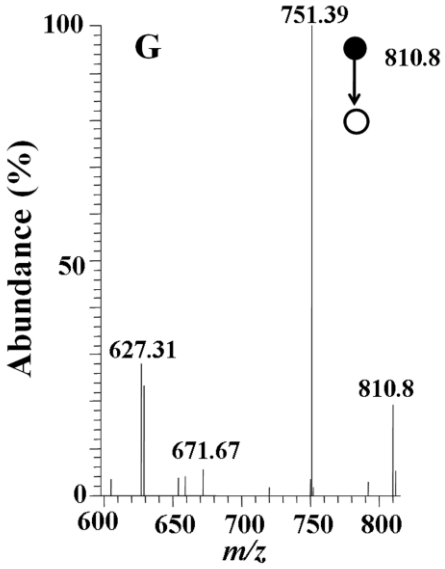
Table S1　Histological description of patients from 20 research subjects

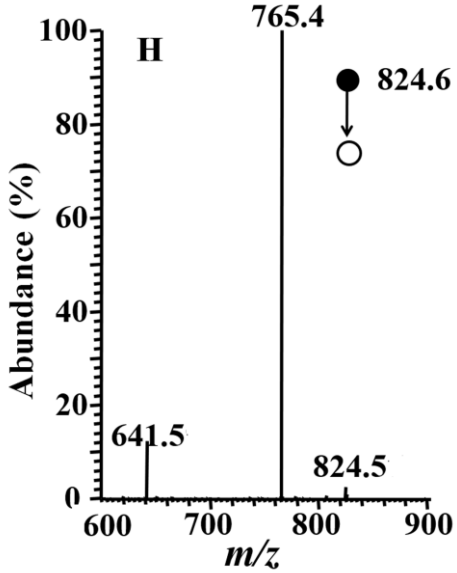
序号 Num	性别 Gender	年龄 Age	肿瘤类型 Tumor type	组织学描述 Histopathological grading
A1	女 Female	62 Sixty-two	鳞状细胞癌 （Squamous cell carcinoma）	中度分化 （Moderate differentiated）
A2	男 Male	47 Forty-seven	鳞状细胞癌 (Squamous cell carcinoma)	低度分化 （Poorly differentiated）
A3	男 Male	48 Forty-eight	鳞状细胞癌 (Squamous cell carcinoma)	中度分化 （Moderate differentiated）
A4	男 Male	57 fifty-Seven	鳞状细胞癌 (Squamous cell carcinoma)	中度分化 （Moderate differentiated）
A5	男 Male	57 fifty-seven	鳞状细胞癌 (Squamous cell carcinoma)	中度分化 （Moderate differentiated）

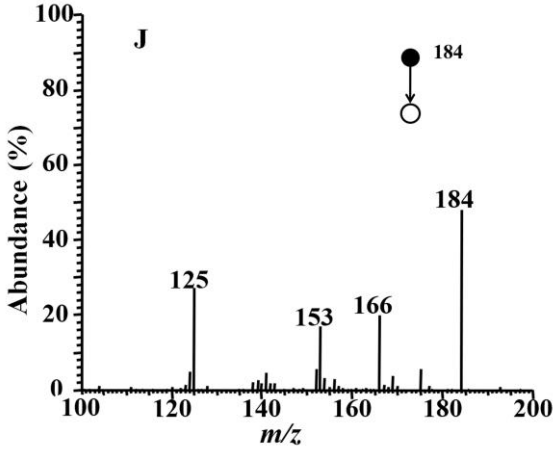
A6	男	59	鳞状细胞癌	中度分化
	Male	fifty-nine	(Squamous cell carcinoma)	（Moderate differentiated）
A7	男	61	鳞状细胞癌	中度分化
	Male	Sixty-one	(Squamous cell carcinoma)	（Moderate differentiated）
A8	男	61	鳞状细胞癌	中度分化
	Male	Sixty-one	(Squamous cell carcinoma)	（Moderate differentiated）
A9	男	61	鳞状细胞癌	低度分化
	Male	Sixty-one	(Squamous cell carcinoma)	（Moderate differentiated）
A10	男	73	鳞状细胞癌	中度分化
	Male	Seventy-three	(Squamous cell carcinoma)	（Moderate differentiated）
A11	女	64	鳞状细胞癌	中度分化
	Female	Sixty-four	(Squamous cell carcinoma)	（Moderate differentiated）
A12	男	55	鳞状细胞癌	中度分化
	Male	fifty-five	(Squamous cell carcinoma)	（Moderate differentiated）
A13	男	58	鳞状细胞癌	中度分化
	Male	fifty-eight	(Squamous cell carcinoma)	（Moderate differentiated）
A14	男	63	鳞状细胞癌	中度分化
	Male	Sixty-three	(Squamous cell carcinoma)	（Moderate differentiated）
A15	男	74	鳞状细胞癌	中度分化
	Male	Seventy-four	(Squamous cell carcinoma)	（Moderate differentiated）
A16	女	70	鳞状细胞癌	中度分化
	Female	Seventy	(Squamous cell carcinoma)	（Moderate differentiated）
A17	男	69	鳞状细胞癌	低度分化
	Male	Sixty-nine	(Squamous cell carcinoma)	（Moderate differentiated）
A18	男	76	鳞状细胞癌	低度分化
	Male	Seventy-six	(Squamous cell carcinoma)	（Moderate differentiated）
A19	女	46	鳞状细胞癌	中度分化
	Female	Forty-six	(Squamous cell carcinoma)	（Moderate differentiated）
A20	女	47	鳞状细胞癌	中度分化

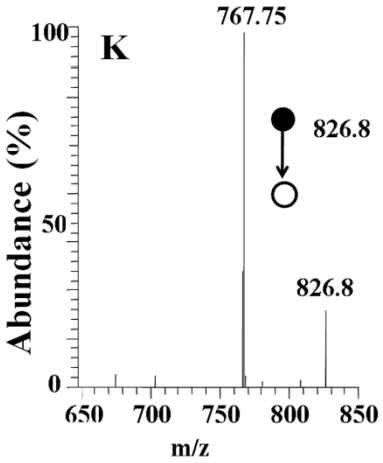
Female	Forty-seven	(Squamous cell carcinoma)	（Moderate differentiated）











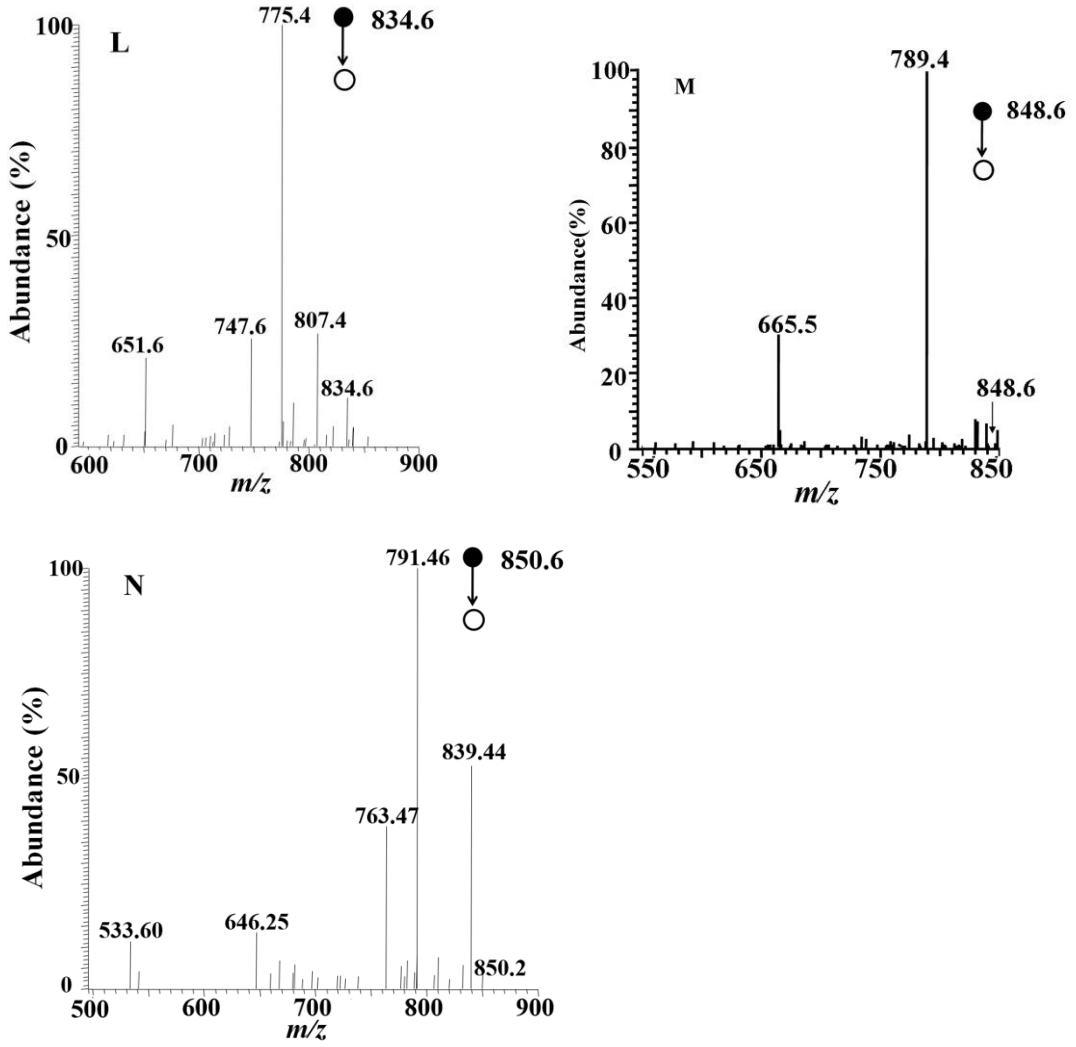


图 S1　区分癌症与邻近正常组织潜在生物标记物的串联质谱图:（A）*m/z* 725.6；（B）*m/z* 756.6；（C）*m/z* 772.6；（D）*m/z* 782.6；（E）*m/z* 798.6 的串联质谱图；（F）*m/z* 806.6；（G）*m/z* 810.6；（H）*m/z* 824.6；（J）*m/z* 184.2；（K）*m/z* 826.6；（L）*m/z* 834.6；（M）*m/z* 846.6；（N）*m/z* 850.6

Fig.S1　Tandem mass spectrometry (MS/MS) spectra of multiple potential biomarkers for distinguishing tumor and adjacent normal tissue: (A) *m/z* 725.6; (B) *m/z* 756.6; (C) *m/z* 772.6; (D) *m/z* 806.6; (E) *m/z* 798.6; (F) *m/z* 806.6; (G) *m/z* 810.6; (H) *m/z* 824.6; (J) *m/z* 184.6; (K) *m/z* 826.6; (L) *m/z* 834.6; (M) *m/z* 846.6; (N) *m/z* 850.6