基于支持向量机的大肠黏液腺癌术前血清蛋白质标志物检测及分析

徐文鸿[1] 陈益定[2] 胡 跃[2] 余捷凯[3] 王连聪[1] 郑 树[3] 张苏展[3]*

(浙江大学医学院附属第二医院,[1] 肿瘤放疗科,[2] 肿瘤外科,[3] 肿瘤研究所, 杭州 310009)

摘要 应用表面加强激光解吸电离 - 飞行时间质谱(SELDI-TOF-MS)技术和CM10蛋白质芯片从大肠黏液腺癌和非黏液腺癌患者中成功地筛选出了大肠黏液腺癌患者血清特异性相关蛋白。应用美国CipherGen公司CM10蛋白质芯片和PBS II型蛋白质芯片阅读仪检测53例大肠癌患者(黏液腺癌12例,非黏液腺癌41例)患者血清蛋白质指纹图谱。采用ZUCI-Protein Chip Data Analyze System分析软件包进行分析,离散小波去噪音,结合支持向量机筛选肿瘤标志物,建立大肠黏液腺癌的术前诊断模型。12例大肠黏液腺癌患者与41例大肠非黏液腺癌患者的血清蛋白质有12个蛋白质峰强度有显著差异。其中质荷比为24 297和23 434 *m/z*处的蛋白质峰强度统计学 *P* 值分别为0.0067和0.0092,差异有极显著统计学意义。支持向量机筛选出24 297、3 322、3 822和4 353 *m/z* 蛋白质峰作为生物标志物进行检测和预测准确率,其中12例大肠黏液腺癌患者中有10例患者被正确识别,41例大肠癌非黏液腺癌患者中有39例被正确识别,准确率为92.45% (49/53)。该方法可以较好地应用于区别大肠黏液腺癌和非黏液腺癌,进行术前病理鉴别,指导进行大肠黏液腺癌的手术和综合治疗。

关键词 大肠黏液腺癌; 表面加强激光解吸电离 - 飞行时间质谱; 生物信息学

随着对大肠癌的生物学行为及病理学特征的不断深入研究,认识到病理类型与疾病预后关系密切。国内外文献及资料均报道大肠黏液腺癌的预后不良,大肠黏液腺癌患者更易发生远处转移和更具有侵袭性[1-4]。故对大肠黏液腺癌的手术应强调扩大根治术的原则,研究还表明,术中腹腔灌注化疗可以显著提高大肠黏液腺癌的生存期[5,6]。因此大肠黏液腺癌的术前确诊,对于肿瘤学家术中实施更彻底手术以及术中腹腔灌注化疗等,从而改善大肠黏液腺癌患者的预后具有重大意义。

但是,迄今为止,大肠黏液腺癌的术前活检诊断准确率很低,原因主要是因为黏液腺癌中也可有高分化或者中分化的腺癌成分,而黏液腺癌的黏液成分所处的病变位置也较深,活检不易取到[7]。而且目前没有用于大肠黏液腺癌诊断的肿瘤标志物。

表面加强激光解吸电离 - 飞行时间质谱(surface-enhanced laser desorption/ionization time-of-flight mass spectrometry, SELDI-TOF-MS)是近年来发展起来的一种全新的蛋白质组学研究手段,应用该技术实现了质谱技术用于临床检测各种疾病相关标志物[8~10]。本研究旨在通过分析大肠黏液腺癌患者和非黏液腺癌患者的术前血清蛋白质指纹图谱,以期发现一些大肠黏液腺癌的肿瘤标志物, 应用于大肠黏液腺癌的术前诊断。

1 材料与方法

1.1 病例资料

2004年11月至2006年3月间我院收治的初诊术前大肠癌患者共53例,其中黏液腺癌患者12例,非黏液腺癌患者42例(包括6例乳头状腺癌和36例管状腺癌)。上述诊断均经术后病理证实,由我院病理科根据WHO黏液腺癌的判断标准进行分类。两组的年龄和性别比例匹配,无统计学差异。所有患者术前血样均于清晨空腹采集,4 ℃静置离心分离血清,−80 ℃低温冰箱保存。

1.2 主要仪器、软件及试剂

PBS-II表面加强激光解吸电离 - 飞行时间质谱

收稿日期: 2008-05-20 接受日期: 2008-07-03

国家自然科学基金(No.30471987)和国家高技术研究发展计划(863计划) (No.2006AA02Z341)资助项目

* 通讯作者。Tel: 0571-87783956, Fax: 0571-87784404, E-mail: zhangscy@tom.com

仪(SELDI-TOF-MS)、能量吸收分子 SPA、CM10 型蛋白质芯片为美国 Ciphergen 公司产品; 支持向量机(support vector machines, SVM)软件运用美国 MathWorks 公司的 Matlab6.5 数据处理软件; ZUCI-Protein Chip Data Analyze System 分析软件包由浙江大学肿瘤研究所开发; 分析纯尿素、乙酸钠、乙腈、DTT 和 CHAPS 缓冲盐为 Sigma 公司产品。

1.3 方法

血清标本冰浴中解冻, 10 000 r/min 离心 5 min, 取 5 μl 与 10 μl U9 处理液(9 mol/L 尿素, 0.2% CHAPS, 0.1% DTT)混合振荡处理 30 min 后, 加 185 μl 50 mmol/L 乙酸钠溶液(pH=4.0)混匀。CM10 芯片置于 96 孔支架中, 50 mmol/L 乙酸钠溶液(pH=4.0)平衡芯片 2 次, 每次 5 min, 取经过处理、稀释的血清 100 μl 加至芯片上, 4 ℃摇床反应 1 h。反应结束甩干芯片表面液体, 依次以 50 mmol/L 乙酸钠溶液(pH=4.0)清洗芯片 3 次, 每次 5 min; 去离子水快速清洗 2 次。芯片自然干燥后点加 1 μl SPA 一次, 待干燥后行质谱仪检测。

1.4 数据收集和统计学分析

质谱仪参数设定: 激光强度 190, 灵敏度 7, 数据收集范围 2 000~30 000 *m/z* (蛋白质质量和电荷比值, 即质荷比), 每次收集数据前以标准蛋白质芯片校正分子量。以质控血清作重复性检测。原始数据先以 Proteinchip Software 3.2.0 软件校正, 使总离子强度及分子量达到均一, 并过滤噪音, 初始噪音过滤值 5, 二次噪音过滤值 2。

1.5 生物信息学分析

ZUCI-Protein Chip Data Analyze System 分析软件包进行数据处理,通过离散小波分析去除噪音, 用局部极值的方法找出样本质荷峰, 以 10% 为最小阈值对质荷峰进行聚类。质谱原始数据经过滤噪音, 聚类分析处理后, 对初步筛选出的质荷比峰数据做Wilconxon 秩和检验, Wilcoxon 秩和检验分析根据 *P* 值评价各个峰对区分两类样本的相对重要性。将差异显著的质荷峰随机组合输入SVM, SVM [11]: 使用线性的SVM 分类器, 具体设置: 采用径向基核函数(radial based kernel), Gamma 值设为 0.6, 罚分函数(C)设为 19。特征向量的选取采用统计过滤结合模型依赖性筛选的方法, 建立判别模型, 用留一法交叉验证评估模型的判别效果。

2 结果

设定参数分别对 12 例大肠黏液腺癌患者与 41 例大肠非黏液腺癌患者的血清蛋白质指纹图谱进行相对含量的分析, 得到两者间表达差异有统计学意义的蛋白质峰共 12 个(*P*<0.05), 其中 8 个蛋白质峰(24 297, 23 434, 3 322, 4 477, 3 822, 4 391, 4 353 和 6 644 *m/z*)在大肠黏液腺癌中的表达低于大肠非黏液腺癌(*P*<0.05), 其中质荷比为 2 4297 和 2 3434 *m/z* 处的蛋白质峰强度统计学 *P* 值分别为 0.0067 和 0.0092, 差异有极显著统计学意义(*P*<0.01)。4 个蛋白质峰(9 126, 9 144, 4 616 和 4 132 *m/z*)在大肠黏液腺癌中的表达高于大肠非黏液腺癌(*P*<0.05) (表 1)。

表 1 大肠黏液腺癌患者与大肠非黏液腺癌患者 12 个表达有统计学差异的蛋白质峰比较($\bar{x}\pm s$)

蛋白质峰 *m/z*	非黏液腺癌 ($\bar{x}\pm s$)	黏液腺癌 ($\bar{x}\pm s$)	*P* 值
24 297	114.69±53.42	87.22±54.27	0.0067
23 434	376.92±204.48	268.45±202.76	0.0092
3 322	2 151.71±755.24	1 620.10±618.28	0.0236
4 477	1 372.55±1 183.58	1 223.68±1 043.51	0.0293
3 822	599.51±755.16	300.07±127.26	0.0310
4 391	660.57±190.59	523.80±144.40	0.0310
9 126	459.66±231.72	627.87±261.86	0.0363
9 144	607.21±439.15	990.01±666.84	0.0382
4 616	799.94±840.47	1 430.60±1 195.82	0.0402
4 353	2 883.45±1 009.79	2 215.09±747.67	0.0446
4 132	716.96±343.68	978.46±612.94	0.0469
6 644	97.46±134.47	42.83±66.82	0.0469

表 2 黏液腺癌患者与非黏液腺癌患者留一法交叉验证结果(例数)

分组	非黏液腺癌	黏液腺癌	例数
非黏液腺癌	39	2	41
黏液腺癌	2	10	12
共计	41	12	53

SVM 留一法筛选出 24 297、3 322、3 822 和 4 353 *m/z* 四个蛋白质峰组成的模型作为生物标志物进行检测和预测准确率, 其中 12 例大肠黏液腺癌患者中有 10 例患者被正确识别, 41 例大肠癌非黏液腺癌患者中有 39 例被正确识别, 准确率为 92.45% (49/53) (表 2, 表 3)。图 1 显示了这个模型的四个蛋白质峰在大肠黏液腺癌和大肠非黏液腺癌患者中的不同强度表达的差异。

3 讨论

黏液腺癌, 是腺癌的一个亚型, 只要黏液样物质占瘤体体积的50%或者以上定义为黏液腺癌, 黏液腺癌和非黏液腺癌在临床和生物学行为及预后上都有

表3 大肠黏液腺癌患者和非黏液腺癌患者的血清蛋白质指纹图谱筛选构建的模型的4个蛋白质峰($\bar{x}\pm s$)

项目	例数	蛋白质峰(*m/z*)			
		24 292	3 3223	8224	353
黏液腺癌	12	87.22±54.24	1 620.10±618.28	300.07±127.26	2 215.0±747.67
非黏液腺癌	41	114.69±53.42	2 151.71±755.24	599.51±755.16	2 883.0±1 009.79
P 值		0.0067	0.0236	0.0310	0.0446

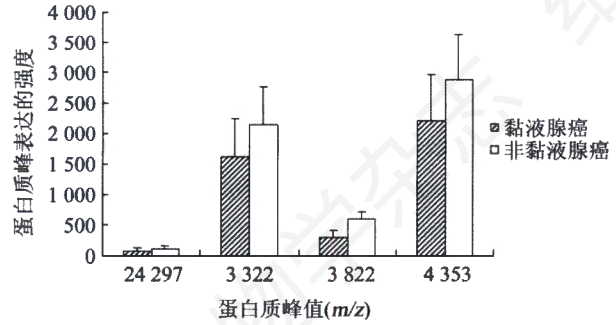

图1 4个蛋白质峰在大肠黏液腺癌和大肠非黏液腺癌患者中的不同强度表达的差异

很大不同, 黏液腺癌在确诊时通常比非黏液腺癌侵犯范围更加广泛, 瘤细胞分泌黏液的水平与其预后密切相关, 因为黏液样物质对周围组织的机械压力使瘤细胞更容易向四周侵犯, 更易发生腹膜转移及远处的淋巴结转移, 大肠黏液腺癌患者的生存率在很多研究中证明均低于非黏液腺癌[3,4,12,13], 大肠黏液腺癌更易于扩散转移, 其病理学特征显示了治愈率低, 复发率高, 预后更差。

因此对于大肠黏液腺癌病例应早期作出诊断, 由于黏液腺癌更容易发生深层侵润及淋巴转移, 因此术中应尽可能行根治或扩大根治性切除, 扩大淋巴结的清扫范围, 对较晚期病例手术时应广泛切除周围组织, 对已浸润到周围脏器者应考虑行联合脏器切除[14], 有条件的患者可考虑术前新辅助化疗, 术前冲击性放疗, 术中腹腔温热灌注化疗, 以改善黏液腺癌患者的生存率[7,8,15]。

然而, 因为黏液腺癌组织中也包含分化不同的腺癌组织, 而且黏液成分常在癌组织深部, 大肠黏液腺癌的术前活检病理诊断准确率比较低, 有报道只有20.6% [9]。而且目前没有用于大肠黏液腺癌诊断的肿瘤标志物。

SELDI-TOF-MS技术是2002年开发的新型蛋白质组学技术, 应用基因芯片的设计原理, 把层析、质谱等技术合理的与蛋白质芯片结合, 可检测出传统方法很难鉴定的蛋白质和多肽, 由于高灵敏度和高通量地反映检测样本中蛋白质全貌的特点[16], 被用于肿瘤标记物的筛选在大肠癌诊断和分期等有重大的发现[17,18], 但目前尚未见有大肠黏液腺癌方面的报道。本实验使用的SVM是Vapnik[19]等提出的一种分类技术, 是在统计学理论基础上发展起来的新型机器学习方法, 模式识别中小样本模型的推广性、模型选择、过拟和、纬数灾难等问题在SVM中得到了很好的解决。 用留一法交叉验证评估模型, 即当一个样本作为测试, 其余样本训练, 反复测试直至结果稳定。因每次的测试集都是独立于用来训练的样本, 完全做到盲法测试。经过以上多步骤、综合多种方法处理数据, 确保了所建模型的推广性和预测的准确性。应用ZUCI-Protein Chip Data Analyze System分析软件包和SVM留一法, 已经成功应用分析了很多疾病血清蛋白质指纹图谱[20,21]。

本实验中, 我们探讨使用SELDI-TOF-MS技术结合生物信息学方法建立基于SVM的血清蛋白质指纹图谱模型,成功地区分大肠黏液腺癌患者和非黏液腺癌患者, 准确率较高。基于SVM的血清蛋白质指纹图谱模型为早期筛查及诊断大肠黏液腺癌提供了一种新方法。

参考文献(References)

[1] 郑 淼等。*中国癌症杂志*, 2005, **15**: 383
[2] Kanemitsu Y *et al. Dis Colon Rectum*, 2003, **46**: 160
[3] Manne U *et al. Clin Cancer Res*, 2000, **6**: 4017
[4] Du W *et al. Dis Colon Rectum*, 2004, **47**: 78
[5] Sugarbaker PH. *Langenbecks Arch Surg*, 1999, **384**: 576
[6] Pestieau SR *et al. Dis Colon Rectum*, 2000, **43**: 1341
[7] Okuyama T *et al. Jpn J Clin Oncol*, 2002, **32**: 412
[8] Issaq HJ *et al. Biochem Biophys Res Commun*, 2002, **292**: 587
[9] Han KQ *et al. Am J Clin Oncol*, 2008, **31**:133
[10] 王 良等。*细胞生物学杂志*, 2007, **29**: 163
[11] Liu Y. *Chem Inf Comput Sci*, 2004, **44**: 1936
[12] Iyomasa S *et al. Jpn J Gastroenterol Surg*, 1988, **21**: 75
[13] Wendum D *et al. Virchows Arch*, 2003, **442** : 111
[14] Okuno M *et al. Am Surg*, 1988, **54**: 681
[15] Pihl E *et al. Pathology*, 1980, **12**: 439
[16] Wright GL Jr. *Expert Rev Mol Diagn*, 2002, **2**: 549
[17] 徐文鸿等。*中华肿瘤杂志*, 2006, **28**: 753
[18] Chen YD *et al*. *Clin Cancer Res*, 2004, **10**: 8380
[19] Vapnik VN. *IEEE Trans Neural Netw*, 1999, **10**: 988
[20] Wang JX *et al. Pediatrics*, 2007, **120**: e56
[21] Wang JX *et al. Proteomics*, 2006, **6**: 5344

Biomarker Discovery of Colorectal Mucinous Adenocarcinoma by Fingerprint and Support Vector Machine

Wen-Hong Xu[1], Yi-Ding Chen[2], Yue Hu[2], Jie-Kai Yu[3], Lian-Cong Wang[1], Shu Zheng[3], Su-Zhan Zhang[3]*

([1]*Department of Radiation Oncology,* [2]*Department of Surgical Oncology,* [3]*Cancer Institute, the Second Affiliated Hospital of Zhejiang University School of Medicine, Hangzhou 310009, China*)

Abstract To discover special serum protein of colorectal mucinous adenocarcinoma (MA) and non-mucinous adenocarcinoma (nMA) preoperatively by using the technology of surface enhanced laser desorption/ionization-time of flight-mass spectrometry (SELDI-TOF-MS) and support vector machine. The potential tumor biomarkers in serum from 12 MA patients and 41 nMA patients were screened by the technology of SELDI-TOF-MS and CM10 Protein Chip (CipherGen Company, USA). The CM10 protein chips was analyzed by PBS II protein chip reader and the protein information was transformed into the form of spectra. The ZUCI-Protein Chip Data Analyze System software package was used to analyze the results. Discrete wavelength analysis was used to eliminate noise and subtract the baseline. A linear support vector machine (SVM) classifier was used to identify peaks. MA was compared with nMA in order to search for proteomic difference between different pathological types. The intensity of 12 proteins in the two groups was significantly different. Among them the *P* value of the 24 297 and 23 434 *m/z* were 0.0067 and 0.0092, respectively. The model formed by four protein peaks of 24 297, 3 322, 3 822 and 4 353 *m/z* was able to distinguish MA from nMA patients with high accuracy. Ten cases in 12 MA patients and 39 cases in 41 nMA patients were classified correctly. The accuracy was 92.45% (49/53). The specific serum protein in MA patients can be preoperatively diagnosed by SELDI-TOF-MS with high accuracy. This method possesses a potential in clinical application.

Key words colorectal mucinous adenocarcinoma; SELDI-TOF-MS; bio-informatics

Received: May 20, 2008 Accepted: July 3, 2008

This work was supported by the National Natural Foundation of China (No.30471987) and the National High Technology Research and Development Program of China (863 Program) (No.2006AA02Z341)

*Corresponding author. Tel: 86-571-87783956, Fax: 86-571-87784404, E-mail: zhangscy@tom.com