[文章编号] 1671-587X(2013)05-0938-06 DOI:10.7694/jldxyxb20130517

应用PCA方法分析拉曼光谱检测结果对乳腺良恶性疾病鉴别诊断的价值

张海鹏[1]，付 彤[2]，张志茹[2]，范志民[2]，郑 超[2]，韩 冰[2]
（1. 吉林大学第一医院产科，吉林 长春 130021；2. 吉林大学第一医院乳腺外科，吉林 长春 130021）

[摘 要] 目的：探讨便携式拉曼光谱在检测新鲜乳腺病灶、正常乳腺组织中应用价值，阐明主成分分析(PCA)方法在处理拉曼光谱检测结果、构建乳腺病变鉴别数学模型和鉴别病变性质中的应用价值。方法：收集2011年5月—2012年5月吉林大学第一医院乳腺外科手术的168例患者(乳腺癌及乳腺良性病变)的新鲜乳腺组织，患者均为女性，年龄22～75岁。其中51例正常组织、66例良性病变组织和51例恶性病变组织，应用便携式拉曼光谱仪进行检测，得出光谱结果；采用PCA方法处理数据，构建病灶鉴别模型，马氏距离法判别数据处理方法的优劣。结果：检测乳腺组织及病灶标本共得到1 800个拉曼光谱，正常组织的特征峰出现在1 078、1 267、1 301、1 440、1 654和1 746 cm^{-1}，而良性和恶性病变组织的特征峰出现在1 281、1 341、1 381、1 417、1 465、1 530和1 637 cm^{-1}，良性和恶性病变组织的主要不同则集中在1 340和1 480 cm^{-1}。PCA方法判别正常组织、良性和恶性病变组织标本的正确率分别是80%、56%和85%。结论：便携式拉曼光谱仪能够检测乳腺组织和病灶，正常组织、良性与恶性病变组织拉曼光谱结果均存在显著差异，PCA方法可以用来构建鉴别模型，但在鉴别良性病变时准确性还不理想。

[关键词] 乳腺肿瘤；拉曼光谱；主成分分析；模型构建
[中图分类号] R730.4 [文献标志码] A

Value of principal component analysis in Raman spectroscopy detection results for differential diagnosis of breast diseases

ZHANG Hai-peng[1], FU Tong[2], ZHANG Zhi-ru[2], FAN Zhi-min[2], ZHENG Chao[2], HAN Bing[2]
(1. Department of Obstetrics Surgery, First Hospital, Jilin University, Changchun 130021, China;
2. Department of Breast Surgery, First Hospital, Jilin University, Changchun 130021, China)

Abstract: Objective To explore the application value of portable Raman spectroscopy in fresh breast lesions and normal breast tissues, and to clarify the application value of principal component analysis (PCA) method in Raman spectrascopy detetion results, construction of the mathematical model and differential diagnosis. **Methods** The fresh tissues of 168 patients (all female, aged 22－75 years) were obtained by routine surgical resection from Department of Breast Surgery, the First Hospital of Jilin University. 51 normal tissues, 66 benign and 51 malignant breast lesions were detected by Raman spectroscopy. The PCA algorithm was used to process the data

[收稿日期] 2013-05-03
[基金项目] 国家自然科学基金青年基金资助课题（81202078）；吉林省科技厅科技发展计划青年基金资助课题（20130522030JH）；吉林省科技厅科研基金资助课题（201015155）
[作者简介] 张海鹏（1980—），女，吉林省长春市人，主治医师，在读医学博士，主要从事生物拉曼光谱的研究。
[通信作者] 韩 冰（Tel：0431-88782550，E-mail：yintian77@126.com）

网络出版时间：2013-08-23 08:18
网络出版地址：http://www.cnki.net/kcms/detail/22.1342.R.20130823.0818.002.html

and build the mathematical model. Mahalanobis distance and spectral residuals were used as discriminating criteria for evaluating this method. **Results** 1 800 Raman spectra were acquired from fresh samples of human breast tissues. Based on spectral profiles, the presence of 1 078, 1 267, 1 301, 1 440, 1 654, and 1 746 cm^{-1} were indicated in normal tissues. And 1 281, 1 341, 1 381, 1 417, 1 465, 1 530, and 1 637 cm^{-1} were found in benign and malignant tissues. The main differences of benign and malignant were the characteristic peaks of 1 340 and 1 480 cm^{-1}. The accuracies of PCA were 80%, 56%, and 85% in discriminating normal, benign and malignant tissues. **Conclusion** Portable Raman spectra can detect the breast tissues and lesions. The Raman spectra of normal, benign and malignant breast tissues have significant differences. PCA method can be used to build identification model, but there is still insufficient in distinguishing benign lesion tissues.

Key words: breast neoplasms; Raman spectroscopy; principal component analysis; model construction

目前对于乳腺检查(查体、彩超、钼靶和MRI等)所发现的异常(肿物、钙化灶等)，只能通过手术活检明确病变性质，但是总体活检后70%～90%患者证实为良性病变[1]。因此，发展一种无创、快速、客观的用于诊断乳腺癌的方法成为目前研究热点。拉曼光谱是一种无损伤、含有信息量非常丰富的光谱技术，可用于固态、液态的生物分子的结构分析，并可直接对生物样品进行检测，已经被广泛应用于针对肿瘤的诊断研究当中[2-5]。在光谱学研究中，困扰研究者的一个巨大障碍就是有机组织的成分多样性造成的数据分析困难。面对光谱特征较为相似的数据，如何将光谱技术和数学分析技术有效结合，构建出准确的检测模型，如何高效快速地提取有效信息，并应用于临床乳腺癌检测及研究当中去，已经成为亟待解决的关键问题。主成分分析(principal component analysis，PCA)是目前生物学领域拉曼光谱结果分析的重要方法，但应用该方法构建鉴别模型并在乳腺疾病鉴别中应用，国内外少有报道。本研究采用便携式拉曼光谱检测新鲜乳腺病灶和正常乳腺组织，通过组织中化学成分的差异(包括钙化灶、蛋白质构成和脂肪构成等)所造成检测结果的不同，采用PCA方法处理结果，构建数学模型，鉴别病变的性质(良性、恶性和癌前病变等)，探讨其准确性和应用前景，为进一步临床应用研究奠定数据处理及数学模型构建基础。

1 资料与方法

1.1 临床资料 收集2011年5月—2012年5月吉林大学第一医院乳腺外科手术切除的168例患者的新鲜乳腺组织，患者均为女性，年龄22～75岁。其中51例正常组织、66例良性病变组织(包括39例纤维腺瘤、17例乳腺腺病和10例乳腺囊肿)和51例恶性病变组织(均为乳腺浸润性导管癌)。所有病例均在吉林大学第一医院病理科进行病理诊断，正常组织取自乳腺癌患者距离乳腺癌组织5 cm以上的正常乳腺组织。所有的患者均知情同意。

1.2 标本处理 采集临床手术活检乳腺正常组织、良性病变及恶性病变新鲜组织标本；修剪组织，去除周围脂肪组织，标本厚度2～5 mm，面积为20～25 mm^2。

1.3 光谱采集 标本切除后30 min内运用便携式光纤拉曼光谱仪(fiber coupled TE cooled CCD array spectrometer，785 nm，～5 mW，BWTEK，分辨率：2～3 cm^{-1})进行检测，扫描每个点的积分时间是60 s。在同一块组织上扫描多个点(扫描原则为乳腺组织标本在位移台上每次移动0.25 mm，以确保检测组织样本的均匀性)，得到10～15张拉曼光谱(同一标本每个检测部位对应一个光谱结果)。将检测后标本送病理科诊断。

1.4 数据分析 将已获得拉曼光谱进行基线及针对δCH2归一化处理；利用PCA对已得到的拉曼信息进行处理：将所得到的不同性质的乳腺组织拉曼光谱(正常组织、良性和恶性病变组织)分别随机抽取150个，通过PCA降低维度，选择信息量最大的2个主成分作为构建模型中使用的特征。设计及编写计算机程序，建立病变模型：为了区分模型中不同分型的样本，使用马氏距离和光谱残基这2个标度来分析PCA获得的模型。使用马氏距离判别法来判别特征抽取这种数据处理方法的优劣。本方法采用Liu's代码识别模式[6]，测量所得数据直接输入模型计算程序，通过计算得出病灶恶性可能性数值，与病理结果对比，采用二元的质量(两类)分类的灵敏度、特异性和马休斯相关系数(MCC)来衡量本研究中实施的准则。灵敏度和特异性反映该方法确定阴性和阳性的效果。由于MCC同时考虑了灵敏度和特异性，因此其被公认为是比较均衡的检测方法。在本研究分析中，这

3条准则利用表达式来计算。灵敏度＝TP/(TP＋FN)；特异度＝TN/(TN＋FP)；MCC＝TP×TN-FP × FN/$\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}$，其中TP为真阳性，TN为真阴性，FP为假阳性，FN为假阴性。对模型检验部分本研究采集新鲜组织标本方式同前，由本组研究负责人对模型计算结果与病理结果进行比较，检测模型的灵敏度和特异性，验证和选择预测病变性质最理想的模型，进一步完善模型。

2 结 果

2.1 正常组织、良性和恶性病变组织的平均光谱

从不同性质的组织(正常组织、良性和恶性病变组织)当中得到1 800个拉曼光谱，根据病理诊断性质将拉曼光谱分类，分别得到恶性、良性和正常组织的平均拉曼光谱(图1)。正常组织的特征峰出现在1 078、1 267、1 301、1 440、1 654和1 746 cm^{-1}，而良性和恶性病变组织的特征峰出现在1 281、1 341、1 381、1 417、1 465、1 530和1 637 cm^{-1}。良性和恶性的主要不同集中在1 340和1 480 cm^{-1}。

2.2 3种组织平均光谱的差谱结果 将3种性质组织的平均光谱分别做差谱得到图2。由图2A(恶性-正常平均光谱)可见比较明显的正向特征峰出现在954、1 047、1 241、1 285、1 341、1 422、1 530和1 640 cm^{-1}。将以上特征峰进行归属后见表1，负向的特征峰则主要出现在1 303和1 443 cm^{-1}，均为脂类的特征振动。图2B(良性-正常平均光谱)可见正向特征峰出现在908、1 047、1 238、1 340、1 387、1 422、1 480和1 640 cm^{-1}处；负向特征峰则主要出现在1 301、1 441和1 657 cm^{-1}。由图2C(恶性-良性平均光谱)可知，正向特征峰为恶性平均光谱的贡献，特征峰峰位分别为1 153、1 256、1 341、1 441、1 534和1 654 cm^{-1}。

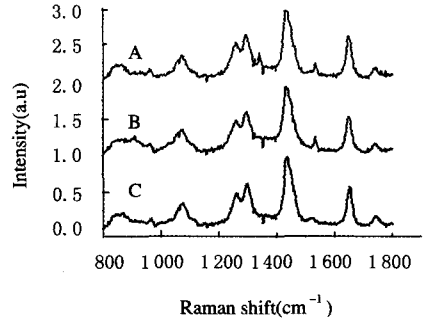

图1 乳腺组织的平均拉曼光谱图

Fig. 1 Mean Raman spectra of breast tissues

A: Malignant tissue; B: Benign tissue; C: Normal tissue.

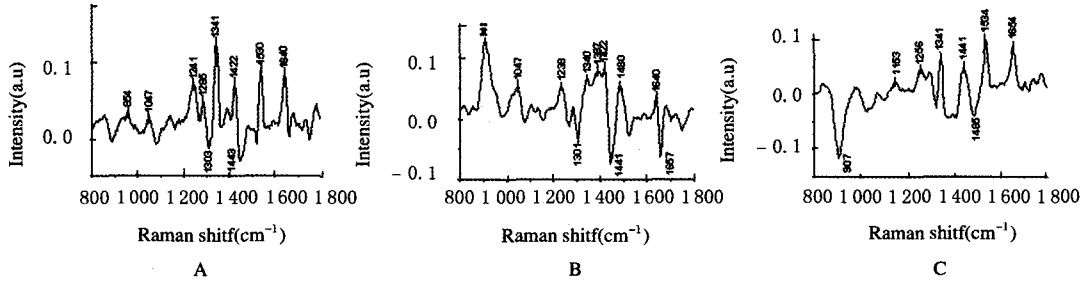

图2 正常、恶性和良性乳腺组织平均光谱的差谱图

Fig. 2 Difference spectra of normal, malignant, and benign breast tissue spectra

A: Malignant-normal tissue; B: Benign-normal tissue; C: Maligant-benign tissue.

2.3 PCA方法的分析结果 利用PCA方法处理光谱数据得出的分析结果见表2和图3A。从表2中可见，PCA方法单纯判别正常和恶性样本的正确率分别为80%和85%，但是判断介于两者之间的良性样本时仅仅获得56%的正确率。而且对于良性样本，PCA方法的灵敏度为0.65、特异度为0.79；对于正常样本其灵敏度和特异度则分别为0.74和0.90；对于恶性样本则分别为0.80和0.92。马氏距离结果表明：不同散点都分布在不同的范围内(图3B)，PCA方法得到的3种表型之间区分明显，并且马氏距离和光谱残基有一定的线性关系(图3C)。

表 1 正常、良性和恶性乳腺组织拉曼光谱特征峰的归属

Tab. 1 Peak assignment of the Raman spectra of normal, benign, and malignant breast tissues

Normal	Benign	Malignant	Assignment
870	873	873	C-C hypro
1 078	1 078	1 078	Lipids
1 155	1 175	1 175	Beta carotene (CH/C-C) in lipids
1 267	1 263	1 263	Amide Ⅲ of proteins
1 301	1 298	1 298	Lipids
—	1 315	1 315	Amide Ⅲ (α-helix) of proteins
—	—	1 341	Nucleic acids
1 363	1 363	1 365	CH_2 and CH_3 symmetric deformation of proteins
1 385	1 381	1 381	CH_2 and CH_3 symmetric deformation of proteins
1 440	1 435	1 435	Lipids
1 461	1 465	1 465	CH_2 and CH_3 symmetric deformation of proteins
1 530	1 530	1 530	Beta carotene
—	1 558	1 558	Tryptophan
1 654	1 650	1 650	Lipids
—	—	1 675	Amide I (collagen)
1 746	1 741	1 745	Lipids

"—": No data.

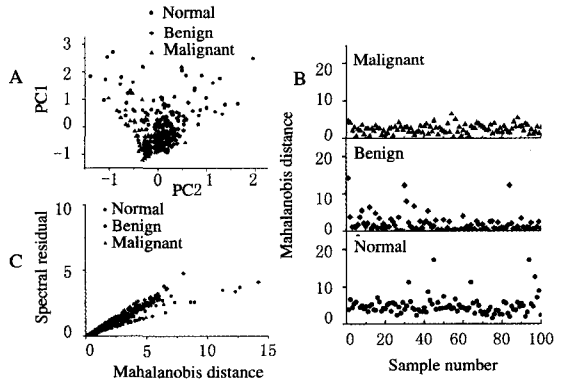

图 3 乳腺组织拉曼光谱的 PCA 结果(A)、马氏距离分析结果(B)和马氏距离与光谱残基的散点图(C)

Fig. 3 PCA results of Raman spectra of breast tissues(A), Mahalanobis distance of different samples(B), and plots of spectral residual *vs* Mahalanobis distance

3 讨 论

光谱学作为一种在医药学活检方面可行的表征方法，可以快速地获得结果。光纤光谱仪对于生物组织研究提供了一种更好的手段。拉曼光谱是一种无损、含有信息量非常丰富的光谱技术，可用于固态、液态的生物分子的结构分析，并可直接对生物样品，如细胞组织、DNA 和蛋白质等进行检测，而无需任何前处理。随着拉曼技术的发展，出现了傅里叶变换拉曼光谱仪、共聚焦显微拉曼光谱技术以及便携式光纤拉曼光谱技术等，其在医学方面的应用也日益受到重视，特别是最近十年来，拉曼技术得到不断的改进，已经被广泛应用于针对肿瘤的诊断研究当中[7-9]。

表 2 利用 PCA 方法处理光谱数据的分析结果

Tab. 2 Analysis results of Raman diagnostic algorithm treated with PCA

Pathological diagnosis	PCA diagnosis			Accuracy rate (η/%)	Sensitivity	Specificity	MCC
	Normal	Benign	Malignant				
Normal	80	19	1	80	0.740 7	0.895 8	0.648 2
Benign	24	56	20	56	0.651 2	0.794 4	0.427 4
Malignant	4	11	85	85	0.801 9	0.922 7	0.734 7

本研究结果显示：正常乳腺组织的平均光谱与病变组织的平均光谱差异较大，正常乳腺组织特征峰通常出现在 1 078、1 267、1 301、1 440、1 654 和1 746 cm^{-1}，这主要归属为脂肪振动模式[3,10]，而良性和恶性乳腺组织谱线轮廓的特征峰则出现在 1 281、1 341、1 381、1 417、1 465、1 530 和

1 637 cm^{-1}，归属为蛋白的特征振动模式[3,10]。因此，正常乳腺组织可能主要包含脂质，而良性和恶性病理组织则主要包含蛋白质。良性和恶性病灶的主要区别集中在1 340和1 480 cm^{-1}，红移的良性光谱显示了宽的δCH_2峰、相对强的酰胺Ⅰ和更强的和更宽的酰胺Ⅲ，这些特征表明恶性组织含有较多脂质而良性组织则含有丰富的基质。这些结果与先前Chowdary等[11]的研究结果一致，同时也与肿瘤的增殖过程相吻合[12]。本研究检测所得的光谱特征可能为探索乳腺组织的生化组成差异提供重要的研究线索。

根据Haka等[3,10]的报道，本研究对3种性质组织的光谱结果进行归属，发现恶性病变组织含有类胡萝卜素、胆固醇、酰胺Ⅲ、胶原蛋白、胞嘧啶、核苷酸琏（DNA）VS（COO-）以及分子间的结合水，而良性病变组织的拉曼光谱中含有比较明显的蛋白（N-H）振动、DNA（C-O）振动、蛋白（酰胺Ⅲ，β-structure，COO-）、CH_3弯曲振动、核酸嘌呤振动及分子间的结合水。与良性病变组织相比较，恶性组织含有更明显的类胡萝卜素、蛋白（酰胺Ⅲ，酰胺Ⅰ）、DNA、脂类（CH_2）、氨基酸残基（tryptophan）和较少的蛋白。这种生物大分子的不同，与肿瘤细胞的特征密切相关，可能为进一步研究提供依据。

虽然3种类型的乳腺组织在平均光谱的光谱学特征中表现出较明确差异，但由于本研究检测的是新鲜病灶乳腺组织，并不能明确定位病灶位置，而拉曼光谱所含有的信息量巨大，敏感性高，单张光谱仍不足以区分这3类组织。本研究将化学计量法引入到对病变组织拉曼光谱的检测中，提取光谱中不能用人工方法所提取的信息。而且，在基于拉曼光谱数据的乳腺癌诊断研究中，特征光谱的选择和选取直接影响着最终的判别结果。因此，本研究中针对乳腺癌数据系统使用了特征光谱的特征抽取法。特征抽取主要是利用当前数据所有特征的信息重新构造新的、能够体现当前数据本质的特征。在本研究中选择了最经典最广泛使用的PCA。PCA是目前广泛采用进行拉曼光谱分析的重要方法，是将多个变量通过线性变换以选出较少个数重要变量的一种多元统计分析方法，又称主分量分析。PCA是一种经典的特征抽取方法，通过奇异值分解过程，从而选取出少数几个能够表示原有数据特性且相互独立的特征向量，达到降维的目的。PCA方法计算简单，数学基础坚固，是一种常用的数据处理方法，具有计算速度快，算法简单等优势，并且该方法获得的结果直接是相互正交的特征，且能够定量的获得每个主成分对原数据的贡献比例。然而针对复杂样本的高维数据，特别是当数据中存在着严重的非线性特征时，往往不能得到正确的结果。PCA已成为一种传统的数据分析方法，被广泛地应用于分析化学信号处理[3,13]。将PCA应用于拉曼光谱的数据处理，不仅可以消除背景和噪音干扰，而且可以消除光谱响应的共线性。

本研究中图3B是PCA模型得到的分型样本和马氏距离的分析结果。从马氏距离结果分析：不同散点都分布在不同的范围内，PCA方法得到的3种表型之间区分明显。从图3C中可以看到在马氏距离和光谱残基这2个标度上，马氏距离和光谱残基有一定的线性关系。所以，模型建立是成功的，而且对于正常和恶性样本的区分，PCA方法得到了较好的结果。

另外，本研究也发现PCA方法对于良性疾病的分析并没有表现出良好结果。从数据方面分析，导致这一结果的原因主要在于数据的噪声，由于PCA方法将所有特征都用于主成分的构建当中，当这些特征中有噪声、错误时，也不可避免的将这些错误和噪声带入构造出的主成分，从而导致无法获得理想的最后结果；另一个原因在于在本数据当中，PCA方法并没有能够将最大量的信息选取包括在使用的2个最大的主成分上，由于前2个主成分只占有大部分信息量，仍有部分信息没有能够被PCA方法捕获，所以在本数据集上PCA方法区分良性样本没有取得好的结果。Abramczyk等[14]使用拉曼光谱检测乳腺组织时，PCA模型区分恶性组织的灵敏度为81%、良性组织的灵敏度为66%，区分正常组织的特异性为88%，这与本研究的结果相似。已有研究[15-16]显示：区分不同类型癌症的PCA算法具有较高的特异性和灵敏度，对于不同类型的组织，特异度为66%～96%。

对于拉曼光谱检测新鲜组织标本结果分析，PCA方法虽然有应用价值，同时存在缺陷，还需要进一步研究，探索算法改良或寻找最适合的分析方法。

综上所述，便携式拉曼光谱仪能够检测乳腺组织及病灶，正常组织、良性与恶性病变组织拉曼光谱结果均存在显著差异，PCA方法可以用来构建鉴别模型，但在鉴别良性病变时准确性还不理想。

[参考文献]

[1] Kolb TM, Lichy J, Newhouse JH. Comparison of the performance of screening mammography, physical examination, and breast US and evaluation of factors that influence them: an analysis of 27, 825 patient evaluations [J]. Radiology, 2002, 225 (1): 165-175.

[2] Krishna CM, Kurien J, Mathew S, et al. Raman spectroscopy of breast tissues [J]. Expert Rev Mol Diagn, 2008, 8 (2): 149-166.

[3] Haka AS, Shafer-Peltier KE, Fitzmaurice M, et al. Diagnosing breast cancer by using Raman spectroscopy [J]. Proc Natl Acad Sci USA, 2005, 102 (35): 12371-12376.

[4] Tu Q, Chang C. Diagnostic applications of Raman spectroscopy [J]. Nanomedicine, 2012, 8 (5): 545-558.

[5] Matousek P, Stone N. Recent advances in the development of Raman spectroscopy for deep non-invasive medical diagnosis [J]. J Biophotonics, 2013, 6 (1): 7-19.

[6] Liu QZ, Sung AH, Chen ZX, et al. Feature mining and pattern classification for steganalysis of LSB matching steganography in grayscale images [J]. Pattern Recognition, 2008, 41 (1): 56-66.

[7] Abramczyk H, Brozek-Pluska B, Surmacki J, et al. Raman optical biopsy of human breast cancer [J]. Prog Biophys Mol Biol, 2012, 108 (1/2): 74-81.

[8] Keller MD, Vargis E, de Matos Granja N, et al. Development of a spatially offset Raman spectroscopy probe for breast tumor surgical margin evaluation [J]. J Biomed Opt, 2011, 16 (7): 077006.

[9] Haka AS, Volynskaya Z, Gardecki JA, et al. Diagnosing breast cancer using Raman spectroscopy: prospective analysis [J]. J Biomed Opt, 2009, 14 (5): 054023.

[10] Haka AS, Shafer-Peltier KE, Fitzmaurice M, et al. Identifying microcalcifications in benign and malignant breast lesions by probing differences in their chemical composition using Raman spectroscopy [J]. Cancer Res, 2002, 62 (18): 5375-5380.

[11] Chowdary MV, Kumar KK, Kurien J, et al. Discrimination of normal, benign, and malignant breast tissues by Raman spectroscopy [J]. Biopolymers, 2006, 83 (5): 556-569.

[12] Hu C, Wang J, Zheng C, et al. Raman spectra exploring breast tissues: Comparison of principal component analysis and support vector machine-recursive feature elimination [J]. Med Phys, 2013, 40 (6): 063501.

[13] Savran CA, Knudsen SM, Ellington AD, et al. Micromechanical detection of proteins using aptamer-based receptor molecules [J]. Ana Chem, 2004, 76 (11): 3194-3198.

[14] Abramczyk H, Placek I, Brozek-Pluska B, et al. Human breast tissue cancer diagnosis by Raman spectroscopy [J]. Spectrosc INT, 2008, 22 (2/3): 113-121.

[15] Stone N, Kendall C, Smith J, et al. Raman spectroscopy for identification of epithelial cancers [J]. Faraday Discuss, 2004, 126: 141-157; discussion : 169-183.

[16] Widjaja E, Zheng W, Huang Z. Classification of colonic tissues using near-infrared Raman spectroscopy and support vector machines [J]. Int J Oncol, 2008, 32 (3): 653-662.