Vol. 30 高等学校化学学报 No. 8
2009 年 8 月 CHEMICAL JOURNAL OF CHINESE UNIVERSITIES 1543 ~ 1547

近红外光谱技术结合主成分分析法用于子宫内膜癌的诊断

徐 可[1], 相玉红[1], 代荫梅[2], 张卓勇[1]
(1. 首都师范大学化学系, 北京 100048; 2. 首都医科大学附属北京妇产医院, 北京 100006)

摘要 应用近红外光谱技术结合化学计量学方法研究了子宫内膜癌组织近红外光谱特征提取和早期诊断的可行性. 测定了 154 例子宫内膜组织切片的近红外光谱, 选取适宜的波段和光谱预处理方法进行主成分分析, 很好地区分了癌变、增生和正常子宫内膜组织切片, 并且分辨出处于不同分化期的组织切片, 为子宫内膜癌的早期诊断提供了可靠依据. 该法快速、简便, 有望发展成为一种新型的肿瘤无创诊断方法.

关键词 子宫内膜癌; 近红外光谱; 主成分分析
中图分类号 O657. 3 **文献标识码** A **文章编号** 0251-0790(2009)08-1543-05

子宫内膜癌(Endometrial carcinoma), 又称宫体癌(Carcinoma corpus of uteri), 是发生于子宫内膜的一组上皮性的恶性肿瘤[1]. 近年来, 其发病率在女性生殖器恶性肿瘤中高居第 3 位, 并且呈现明显的上升趋势. 早期子宫内膜癌的复发转移较少, 5 年生存率可达 90%[2]. 如果能够实现其早期诊断, 将会为众多患者带来治愈的福音.

目前, 子宫内膜癌诊断的传统方法仍是分段诊刮. 但该法存在着不能判断子宫肌层浸润深度及不能明确疾病分期的缺陷. 由于是盲刮, 对小范围的病变有可能会漏诊, 假阳性率也较高[3]. 在病理上有时很难区别分化好的子宫内膜癌与子宫内膜高度不典型增生[4]等病症.

近红外光谱主要是分子从基态向高能级跃迁时非谐振性振动产生的含氢基团 X—H(X = C, N, O)的倍频和合频吸收, 具有丰富的结构和组成信息[5]. 近红外光谱技术可以检测基于细胞形态变化之前的生物大分子的变化, 具有一定的客观性和可重复性. 在癌变发生初期就可以通过分子光谱观察到生物分子结构的细微变化, 从而实现对癌变组织的早期诊断[6]. 同时, 在近红外光谱区域内, 体液和软组织相对透明, 光的穿透力强, 是理想的无创检测光谱段[7]. 该波段可以显示出各种核酸、蛋白质、脂类和细胞膜不同程度的变化, 为诊断患者病情的发展(正常→增生期→恶性期)提供参考依据[8].

近红外光谱检测技术具有无创检测、待测样品无需进行预处理及分析过程快速、简便等特性, 它综合运用化学计量学方法, 可以从测定数据中提取有效信息[9]. 目前, 将近红外光谱检测技术应用于癌症诊断方面的研究比较少, 现有的方法主要是依据癌组织与正常组织近红外光谱特征吸收峰的差异来进行辅助诊断[10~14]. 利用近红外光谱技术研究子宫内膜癌的组织特征和诊断尚未见报道. 本文以子宫内膜组织的病理切片为研究对象, 初步探讨了近红外光谱技术对癌症早期诊断的可行性, 并对相关问题进行了讨论.

1 材料与方法

1.1 样品来源和处理

用于实验的子宫内膜组织石蜡切片均由首都医科大学附属北京妇产医院提供, 共计 154 例. 根据病理诊断结果分类如下: 子宫内膜癌组织切片 58 例, 其中高分化腺癌 26 例, 中分化腺癌 28 例, 低分化腺癌 2 例, 透明细胞腺癌 2 例, 患者年龄 35 ~ 71 岁; 子宫内膜增生组织切片 60 例, 其中单纯性增生

收稿日期: 2009-03-18.
基金项目: 国家自然科学基金(批准号: 20875065, 30772322)资助.
联系人简介: 张卓勇, 男, 博士, 教授, 博士生导师, 主要从事计算机化学与化学信息学研究. E-mail: gusto2008@vip.sina.com

22 例，复杂性增生 36 例，重度非典型增生 2 例，患者年龄 29 ~ 63 岁；正常子宫内膜组织切片 36 例（包含 3 个不同生理期），其中增殖期 22 例，分泌期 12 例，经期 2 例，年龄分布 19 ~ 53 岁[15].

所有组织切片厚度均为 4 μm，均常规取材，4% 甲醛固定，分别经浸蜡、包埋、切片、二甲苯脱蜡、梯度乙醇脱水、粘片及中性树胶封固等一系列技术处理制成.

1.2 实验仪器和光谱采集

实验所用仪器为美国 Thermo 公司的 Nicolet 6700 FTIR 型扩展傅里叶变换近红外光谱分析仪，配有 InGaAs 检测器，分析软件为 Omnic V7.3 和 Matlab V7.1.

室温下以空气作空白，用积分球漫反射法采集样品的近红外光谱，扫描范围 4000 ~ 10000 cm^{-1}，分辨率 4 cm^{-1}，扫描次数 64. 为保证其具有代表性，在每个样品的不同位置进行平行扫描 5 次，应用 Omnic V7.3 软件求取平均光谱.

1.3 主成分分析方法

主成分分析（Principal component analysis，PCA）的主要作用之一是将光谱数据降维，把原变量转换成一组彼此正交的新变量的线性组合，消除了多变量共存中相互重叠的信息. 选取特征值较大的前几个新变量代表主成分表征原变量的数据特征，并尽可能多地保留了原始信息[16~18]. 选择累计方差贡献率大于 85% 时所需的前几个主成分来代表原始光谱信息，从而实现对光谱特征信息的提取. 当前两个主成分（分别称为第一主成分和第二主成分）的累计方差贡献率足够大时，常以每个样品的第一和第二主成分得分值作为横、纵坐标，绘制二维主成分特征投影图，对样品进行分类[19]. 当样品所含光谱信息相似度较大时，其主成分得分值比较接近，样品点会聚集于投影图的某个较小区域内.

影响主成分分析的两个重要因素是测量波段的选取和光谱数据的预处理方法. 本文着重研究以上两方面因素对 PCA 分类效果的影响. 主成分分析采用 Matlab V7.1 程序编写[20,21].

2 结果与讨论

2.1 样品的近红外漫反射光谱

癌变、增生和正常子宫内膜组织切片的原始近红外漫反射光谱图分别如图 1（A）~（C）所示. 比较图 1（A）~（C）可见，每组图的特征吸收峰及其吸收强度都非常接近，仅凭肉眼很难观察出不同组织切片近红外光谱图谱的差异，因此需要借助化学计量学方法来提取光谱的特征信息，应用数学方法最大限度地解析化学数据，以实现对组织切片类别性质的辨识[22].

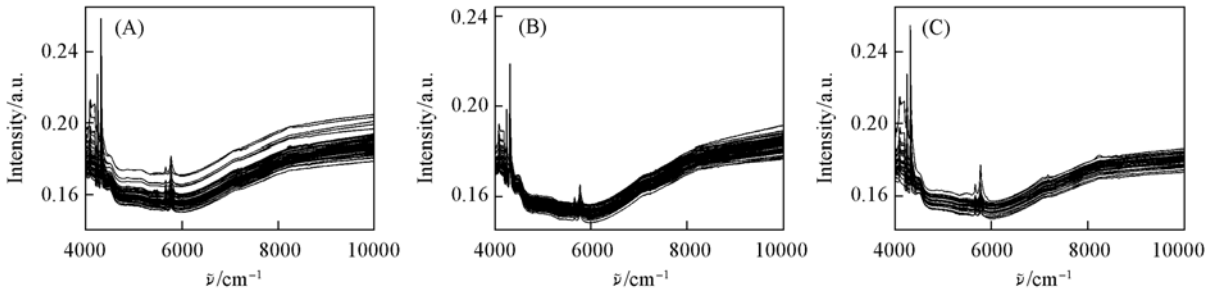

Fig. 1 Original NIR spectra of malignant endometrium（A）, hyperplasia endometrium（B）and normal endometrium（C）

对 4000 ~ 10000 cm^{-1}范围内的原始近红外光谱数据进行主成分分析所得结果表明，前两个主成分的累积方差贡献率达到 98%，因此，采用前两个主成分的得分所绘制的二维主成分特征投影图如图 2 所示（纵坐标括号内分别标注了前两个主成分的方差贡献率及其特征值）. 不同类别的样品用不同的字母表示，聚集在图中相应的区域. 可见，采用未经预处理的原始光谱数据进行分析，3 类样品没有清晰的界线，不能区分开.

2.2 分析波段的选择

利用 Omnic V7.3 软件计算光谱强度在各波数下的标准偏差，以标准偏差值对波数（$\tilde{\nu}$/cm^{-1}）作图，得到如图 3 所示的 154 条谱线的标准偏差图谱. 标准偏差值精确地反映了各波数下样品的光谱数

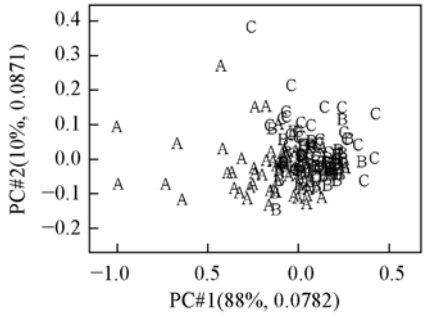

Fig. 2　First two PCs scores plot of samples

A: Malignant; B: hyperplasia; C: normal.

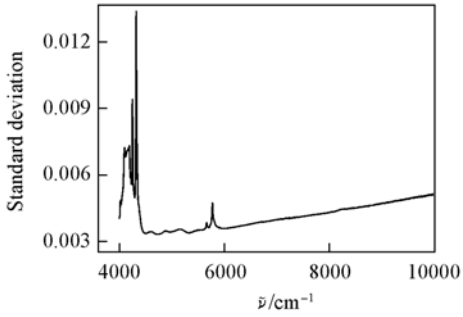

Fig. 3　Standard deviation of intensity at various wavenumbers

据偏离平均值的程度，可用于衡量数据间的差异度. 若标准偏差值较小，则表明这组数据间差异较小，样品性质比较相似；若标准偏差值较大，则表明样品性质的差异较大. 选择标准偏差值较大的波段进行分析，能够有效地提取不同样品光谱中的差异信息. 从图 3 中可以看出，在 4000 ~ 4500 cm^{-1}和 5500 ~ 6000 cm^{-1}区域内样品的近红外光谱差别最大，有用的光谱信息主要集中在 4000 ~ 6000 cm^{-1}波段，因此选择 4000 ~ 6000 cm^{-1}波段并借助主成分分析方法，研究了癌变、增生及正常等 3 类子宫内膜组织切片在分子水平上的细微差异. 相关医学资料报道，在 4000 ~ 6000 cm^{-1}波段范围内包含了癌组织中脂类、蛋白质和 DNA 变化的主要信息[8].

2.3　不同光谱数据预处理方法的选择

为了去除高频随机噪声、基线漂移、样本不均匀以及光散射等因素的影响，需要对光谱进行预处理[23,24]. 本文采用 3 种不同的预处理方式，并对其分类效果的影响进行了比较[25].

方法 1：采用 Savitzky-Golay 方法进行 3 次多项式 5 点平滑，滤除高频噪声；再进行多元散射校正(Multiplicative scatter correction, MSC)处理，消除由样品的不均匀性造成的噪声以及基线漂移的影响.

方法 2：采用 Savitzky-Golay 方法进行 3 次多项式 5 点平滑；再进行多元散射校正；结合标准化法处理.

方法 3：将 Savitzky-Golay 方法应用于求取一阶导数光谱，在进行 3 次多项式 5 点平滑时，拟合窗口中心点的数据用其一阶导数取代后再结合多元散射校正.

分别对用上述 3 种方法所得的吸光度矩阵进行主成分分析，将数据在第二主成分上的得分对在第一主成分上的得分作图，得到二维主成分特征投影图[图 4(A) ~ (C)]，纵坐标的括号内分别标注了前两个主成分的方差贡献率及其特征值.

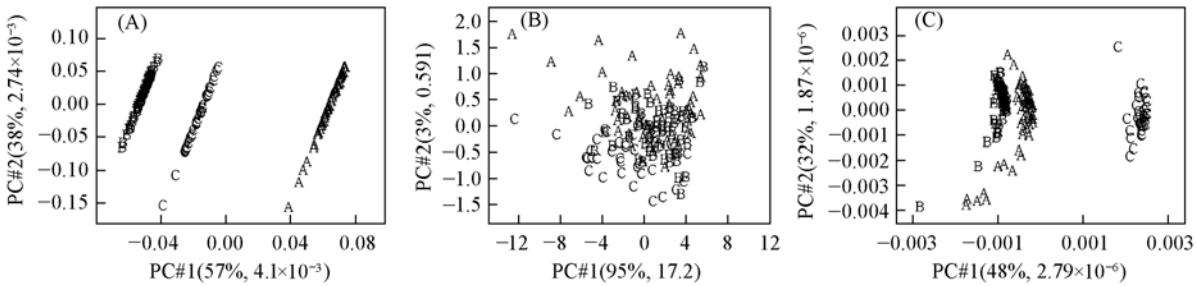

Fig. 4　Comparison of PCs scores plots by method 1(A), method 2(B) and method 3(C)

A: Malignant; B: hyperplasia; C: normal.

比较 3 种预处理方法可知，原始光谱经 Savitzky-Golay 平滑和多元散射校正处理后的分类效果较好，且前两个主成分的累积方差贡献率可达 95%. 而用一阶导数光谱的分类效果欠佳；使用标准化的光谱数据所得的结果最差. 这可能是因为数据的标准化将数据按比例缩放，使之简化并落入一个特定区间，只能消除或减少线性的仪器噪声，但存在非线性噪声时使用该方法来对光谱进行校正会影响有效信息的提取，不利于样品的分辨. 导数光谱可以有效地消除基线和其它背景的干扰，分辨重叠峰，提高分辨率和灵敏度，但它同时会放大噪声，降低信噪比. 分析结果证明，对子宫内膜组织切片近红

外光谱的处理不宜采用一阶求导法和数据标准化法，而 Savitzky-Golay 平滑结合多元散射校正处理方法是可行的选择.

Savitzky-Golay 平滑又称移动窗口多项式最小二乘拟合平滑方法. 在移动窗口运算中采用多项式最小二乘拟合，其实质是一种加权平均法，强调了移动窗口中心点的中心作用. 对于 Savitzky-Golay 平滑，窗口宽度和多项式阶次的优化选择非常重要. 若窗口宽度太小或多项式次数过高均会导致过拟合，使平滑去噪效果欠佳；若窗口宽度太大或多项式次数过低，运算时会平滑掉一些有用信息，造成光谱信号的失真. 经反复实验，将窗口宽度选定为 5，多项式次数选定为 3.

2.4 对每类样品的不同分期进行分类

由于低分化腺癌、透明细胞腺癌、重度非典型增生及经期的样品个数太少，不具有代表性，因此仅取高分化腺癌、中分化腺癌、单纯性增生、复杂性增生、增殖期和分泌期的 6 类样品进行分类实验. 采用方法 1 对光谱数据进行预处理，选择 4000 ~ 6000 cm^{-1} 波段进行主成分分析，所得主成分特征投影图如图 5 所示. 从图 5 中可以看到，6 个不同生理分化期的样品分别聚集在不同区域，分类效果较好. 结果证明，不同分化期组织的生物大分子之间存在一定的分子差异. 借助化学计量学方法进行分辨使细胞在癌变前或癌变早期被准确地识别出来.

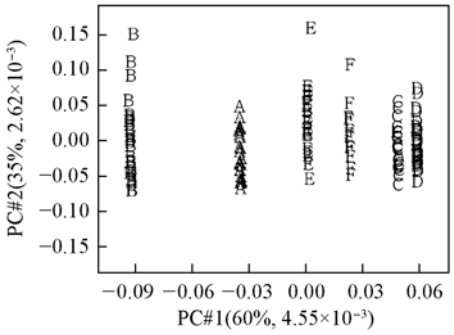

Fig. 5 First two PCs scores plot of samples with endometrium at various differentiation stages

A: High differentiation adenocarcinoma; B: middle differentiation adenocarcinoma; C: simple hyperplasia; D: complex hyperplasia; E: proliferative phase; F: secretory phase.

值得注意的是，单纯性增生和复杂性增生样品在图中的分类界限不明显，彼此间距很小，由此可知这两类样品具有比较接近的近红外光谱性质. 但处于这两个分化期的子宫内膜组织是否具有相似的细胞生理学特性，需要进一步实施相关临床医学实验来深入探讨.

3 结 论

子宫内膜癌组织的近红外光谱特征提取和诊断研究目前尚未见到文献报道. 本文将近红外光谱技术与主成分分析方法相结合，对测量波段和光谱预处理方法进行了优化，实现了对癌变、增生和正常子宫内膜组织病理切片的分辨，对其不同分化期的样品进行分析也获得了较好的分类结果. 对有关方法也进行深入研究，以便提高分类的精确度. 研究结果表明，近红外光谱分析技术结合化学计量学方法可以实现对子宫内膜癌的鉴别诊断，为癌症的早期诊断提供了可靠依据，并有望发展成为一种新型的肿瘤无创诊断方法.

参 考 文 献

[1] PENG Zhi-Lan(彭芝兰). Practical Journal of Clinical Medicine(实用医院临床杂志)[J], 2005, **2**(2): 10—11
[2] Cosson M., Labadie E.. European Journal of Obstetrics & Gynecology and Reproductive Biology[J], 2001, **98**(2): 231—236
[3] SHEN Keng(沈铿). Chinese Journal of Obstetrics and Gynecology(中华妇产科杂志)[J], 2004, **39**(3): 145—147
[4] HE Min-Fu(何民富), LI Hui(李辉), SUN Geng-Tian(孙耕田), *et al*.. China Oncology(中国癌症杂志)[J], 2008, **18**(7): 517—522
[5] Kondepati V. R., Keese M., Mueller R., *et al*.. Vibrational Spectroscopy[J], 2007, **44**(2): 236—242
[6] Nioka S., Chance B.. Technology in Cancer Research & Treatment[J], 2005, **4**(5): 497—512
[7] LIU Rong(刘蓉), XU Ke-Xin(徐可欣), CHEN Wen-Liang(陈文亮), *et al*.. Science in China, Series G: Physics, Mechanics, Astronomy(中国科学, G 辑: 物理学力学天文学)[J], 2007, **37**(Suppl.): 124—131
[8] Kondepati V. R., Heise H. M., Backhaus J.. Analytical and Bioanalytical Chemistry[J], 2008, **390**(1): 125—139
[9] Zou T. T., Dou Y., Mi H., *et al*.. Analytical Biochemistry[J], 2006, **355**(1): 1—7
[10] Cerussi A., Shah N., Hsiang D., *et al*.. Journal of Biomedical Optics[J], 2006, **11**(4): 044005-1—044005-16

[11] Arifler D., Schwarz R. A., Chang S. K., *et al*.. Applied Optics[J], 2005, **44**(20): 4291—4305
[12] Tromberg B. J., Cerussi A., Shah N., *et al*.. Breast Cancer Research[J], 2005, **7**(6): 279—285
[13] Kondepati V. R., Oszinda T., Heise H. M., *et al*.. Analytical and Bioanalytical Chemistry[J], 2001, **387**(5): 1633—1641
[14] McIntosh L. M., Summers R., Jackson M., *et al*.. Journal of Investigative Dermatology[J], 2001, **116**(1): 175—181
[15] Hiroya Y., Shoji K., Ehiichi K., *et al*.. Radiation Medicine[J], 2003, **21**(1): 1—6
[16] Wang X. Y., Garibaldi J. M., Bird B., *et al*.. Applied Intelligence[J], 2007, **27**(3): 237—248
[17] Jolliffe I. T.. Principal Component Analysis[M], New York: Springer-Verlag, 1986: 1580—1584
[18] Jolliffe I. T.. Statistical Methods in Medical Research[J], 1992, **1**(1): 69—95
[19] SU Zhen-Qiang(苏振强), HONG Hui-Xiao, TONG Wei-Da, *et al*.. Chem. J. Chinese Universities(高等学校化学学报)[J], 2007, **28**(9): 1640—1644
[20] Harrington P. D., Vieira N. E., Espinoza J., *et al*.. Analytica Chimica Acta[J], 2005, **544**(1/2): 118—127
[21] Harrington P. D., Vieira N. E., Chen P., *et al*.. Chemometrics and Intelligent Laboratory Systems[J], 2006, **82**(1/2): 283—293
[22] Cohenford M. A., Godwin T. A., Cahn F., *et al*.. Gynecologic Oncology[J], 1997, **66**(1): 59—65
[23] ZHU Xiang-Rong(朱向荣), LI Na(李娜), SHI Xin-Yuan(史新元), *et al*.. Chem. J. Chinese Universities(高等学校化学学报)[J], 2008, **29**(5): 906—911
[24] MENG Qing-fan, TENG Le-sheng, JIANG Chao-jun, *et al*.. Chem. Res. Chinese Universities[J], 2008, **24**(1): 29—31
[25] CHEN Jian(陈建), CHEN Xiao(陈晓), LI Wei(李伟). Spectroscopy and Spectral Analysis(光谱学与光谱分析)[J], 2008, **28**(8): 1806—1809

Near Infrared Spectroscopy Combined with Principal Component Analysis Applied to Diagnosis of Endometrial Carcinoma

XU Ke[1], XIANG Yu-Hong[1], DAI Yin-Mei[2], ZHANG Zhuo-Yong[1*]
(1. *Department of Chemistry*, *Capital Normal University*, *Beijing* 100048, *China*;
2. *Beijing Obstetrics and Gynecology Hospital*, *Capital Medical University*, *Beijing* 100006, *China*)

Abstract The feasibility of near infrared(NIR) spectral feature extraction and early diagnosis of endometrial carcinoma were developed by chemometrics methods. Few papers have been reported about the studies on NIR spectra features of endometrial carcinoma so far. In this study, the NIR spectra of 154 specimens of endometrium were collec-ted, and spectral data were analyzed by principal component analysis. In order to improve the classification, selection of wavelength range and spectral pretreatment methods were discussed. The results suggested that samples of malignant, hyperplasia and normal endometrium were classified correctly. Moreover, endometrium at various differentiation stages could also be identified, which provided reliable evidence for early diagnosis in endometrial carcinoma. This approach was proved to be rapid and convenient, which is able to be developed as a non-invasive diagnosis method for cancer.

Keywords Endometrial carcinoma; Near infrared spectroscopy; Principal component analysis

(Ed.: A, G)