2007 年第 65 卷 第 2 期, 152～158
化 学 学 报 ACTA CHIMICA SINICA
Vol. 65, 2007 No. 2, 152～158

• 研究论文 •

正交信号校正在正常成人血清 ^{1}H NMR 谱的代谢组分析中的滤噪作用评价

冒海蕾[a,c] 徐 旻[b] 王 斌[c] 王惠民[c] 邓小明*,[a] 林东海*,[b]

([a] 第二军医大学长海医院 上海 200433)

([b] 中国科学院上海药物研究所 上海 201203)

([c] 南通大学附属医院 南通 226001)

摘要 观察、比较正交信号校正(OSC)滤噪前后, 用不同的模式识别方法对正常成人血清代谢组 ^{1}H NMR 谱进行分析的效果, 以探讨 NMR 代谢组学技术应用于临床研究和疾病早期诊断的可行性. 78 例正常成人在采血前按常规要求禁食 8 h, 记录血清一维 600 MHz 氢谱后, 分别采用主成分分析(PCA)、偏最小二乘法-判别分析(PLS-DA)以及簇类的独立软模式法(SIMCA)对氢谱进行模式识别分析. 结果表明: 虽然采血前并无其它诸如饮食、生活方式、生理周期等方面的严格限制, 采用 OSC 滤噪后, PLS-DA 能够完全区分不同性别的血清氢谱, 其判别能力优于 PCA 和 SIMCA. 而且采用 OSC 滤噪与文献报道的未经 OSC 处理的 PLS-DA 法获得的与性别分类有关的主要 NMR 积分区段基本相同. 从 OSC 去除不同数目的隐变量后所致的 PLS-DA 模型的性能改变可见: OSC 去除两个隐变量时, 前两个隐变量的特征值明显比后面的大; 剩余残差为 20.82%, 即去除了 79.18%的 *X* 变量中与反应变量 *Y* 不相关的系统变异. 此时 PLS-DA 计算所得的隐变量个数为 1; 而不使用 OSC 或用 OSC 去除一个隐变量时, PLS-DA 所得的隐变量个数分别为 3 和 2. 作为 PLS-DA 模型质量的评价指标, R2X 表示 PLS-DA 模型计算所获得的隐变量反映自变量 *X* 的变异的百分比, R2Y 则表示隐变量反映因变量 *Y* 的变异的百分比, Q2 (cum)为交叉验证后 PLS-DA 模型所获隐变量能够预测 *X* 和 *Y* 变异的累计百分比. R2X 在 OSC 去除两个隐变量时达到最低值, 表明此时 PLS-DA 计算模型包含的系统变异最少; R2Y 与 Q2 (cum) 都达到 80%以上并趋于稳定, 说明 OSC 去除两个隐变量时 PLS-DA 模型的质量优良. 显然, OSC 可去除饮食、环境等因素的影响, 降低临床样本的不均一性, 这对于 NMR 代谢组学技术应用于临床研究至关重要. OSC 滤噪去除的隐变量个数应根据剩余残差、去除隐变量的特征值大小、PLS-DA 模型计算所得的隐变量个数和反映模型质量的相关指标加以判断.

关键词 NMR; 代谢组学; 模式识别; 正交信号校正; 血清

Evaluation of Filtering Effects of Orthogonal Signal Correction on Metabonomic Analysis of Healthy Human Serum ^{1}H NMR Spectra

MAO, Hai-Lei[a,c] XU, Min[b] WANG, Bin[c] WANG, Hui-Min[c]

DENG, Xiao-Ming*,[a] LIN, Dong-Hai*,[b]

([a] *Changhai Hospital, Secondary Military Medical University, Shanghai 200433*)

([b] *Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203*)

([c] *Affiliated Hospital, Nantong University, Nantong 226001*)

Abstract Three different pattern recognition methods before and after orthogonal signal correction (OSC) were employed to perform the metabonomic analysis of ^{1}H NMR spectra recorded on healthy human sera, in order to explore the potential of applying ^{1}H NMR-based metabonomics to clinical research. At first, 78 healthy human sera were collected after a routine fasting for 8 h, and the corresponding 1D ^{1}H NMR spectra

* E-mail: dhlin@mail.shcnc.ac.cn; xmdeng@anesthesia.org.cn

Received April 3, 2006; revised June 19, 2006; accepted September 6, 2006.

国家自然科学基金(Nos. 30470351, 30570352)和中国科学院“百人计划”资助项目.

were recorded on a Varian Unity INOVA-600 spectrometer, and then three pattern recognition analyses, PCA (principal component analysis), PLS-DA (partial least squares-discriminant analysis), and SIMCA (soft independent modeling of class analogy), were performed, respectively. In spite of no specific sample-collecting restriction on foods, life styles, and physiological cycles, the PLS-DA method after OSC is able to distinguish the NMR metabonomic profiles of male sera from those of female sera, more perfectly than both the PCA and SIMCA. Furthermore, the major NMR integral regions relevant to gender classification from PLS-DA after OSC were identical with those from PLS-DA without OSC filter in the literature. In the figure of displaying the variation of PLS-DA model before OSC and after removing different OSC latent variables (LVs), the eigenvalues of the first and second OSC-removed LVs were much greater than others. After removing two LVs by OSC, the remaining sum of square (RSS) in the *X* block was 20.82%, that is, 79.18% information unrelated to *Y* was removed from the PLS-DA model. Meanwhile, the LV number of PLS-DA model attained to one; while the LV number was two for the model with the first LV being removed by OSC, and three for the model without OSC. R2X, R2Y, and Q2 (cum) are usually used to evaluate the quality of PLS-DA model. R2X and R2Y are the fraction of the sum of square of the entire *X*'s and *Y*'s explained by the current LV of PLS-DA, and represent the variance of *X* and *Y* variables, respectively; while Q2 is cross validated R2. Q2 (cum) reflects the cumulative cross-validated percent of the total variation of the *X*'s and *Y*'s that can be predicted by the current LV of PLS-DA model. In our study, after OSC filtering the first two LVs, R2X reached the minimum, suggesting that the least systematic variance should be present in the current PLS-DA model. Meanwhile, both R2Y and Q2 (cum) were always higher than 80%, indicative of the good quality of the PLS-DA model. Obviously, OSC is capable of eliminating the influence of dietary and environmental factors, and decreasing the heterogeneity of samples, which is fairly useful and important for clinical investigations. Additionally, the appropriate number of OSC-removed LVs should be determined on the basis of RSS in the *X* block, eigenvalue of OSC-removed latent variables, LV number and the qualitative indicators of the PLS-DA model.

Keywords NMR; metabonomics; pattern recognition; orthogonal signal correction; sera

核磁共振代谢组学(NMR metabonomics)概念是1999年Nicholson等[1]在长期研究生物体液的基础上首先提出的, 主要利用NMR谱学技术和模式识别方法对生物体液和组织进行系统的测量和分析, 对完整生物体(而不是单个细胞)内随时间改变的代谢物进行动态检测、定量和分类, 然后联系生物体的病理生理变化, 确定导致这些代谢改变的靶器官和作用位点, 寻找相关的生物标志物[2].

模式识别技术在NMR代谢组学研究中被广泛地用来从复杂的代谢组NMR波谱数据中尽可能多地获取重要生物学信息. 人们已经发展了多种多样的模式识别技术, 如: 主成分分析法(principal component analysis, PCA), 偏最小二乘法(partial least squares, PLS), 偏最小二乘法-判别分析(PLS-discriminant analysis, PLS-DA), 以及簇类的独立软模式分类法(soft independent modeling of class analogy, SIMCA)等. 其中, PCA方法是最简单、最常用且比较有效的无监督方法, 它只解释自变量(*X*)矩阵中的最大变量信息. PLS方法则还要考虑一个反应变量(*Y*)矩阵, 通过寻找类似于PC的隐变量, 使自变量(*X*)向反应变量(*Y*)回归的程度达到最大. PLS-DA方法应用PLS算法对一个*Y*哑变量矩阵(dummy matrix)进行分类, 这个哑变量矩阵由正交的每个类矢量构成. PLS-DA属于有监督的模式识别方法, 选择已知不同类别的样本数据作为训练集, 构建PLS-DA模型. 一旦模型经计算并确认生效, 就可以用来预测未知样本的类别[3~5]. SIMCA方法是建立在PCA基础上较为通用的有监督的简易分类法, 通过对每类样本建立一个PCA模型, 并限定该类模型的空间范围, 再用这些模型对未知样本进行判别分析.

NMR代谢组学技术用于临床研究目前尚处于探索阶段. 已有的研究报道往往对于研究对象设定了较严格的入选标准[4,6~8], 包括饮食、生活方式、生理周期等诸多方面. 但在临床实践中, 无论对于住院患者还是门诊患者, 都很难真正达到要求. 近年来发展的多种谱学信号校正技术, 能够明显地降低噪声, 提高波谱信号与所要观察的变量的相关性. 正交信号校正(orthogonal sig-

nal correction, OSC)是目前广泛应用的信号校正技术, 以样本的类别作为反应变量(*Y*)建立一个 PLS 模型, 获得几个最长向量表示与 *Y* 不相关的自变量(*X*)信息, 其中主要是 *X* 矩阵中的系统误差, 并将其去除, 这样通过 OSC 滤掉与类别判断正交(不相关)的变量信息, 提高了模式识别方法的判定能力[9,10].

为了观察 OSC 的滤噪作用, 本文研究的体检正常成人, 在收集血清标本之前, 仅采用一般体检时 8 h 禁食的要求, 并无其它诸如饮食、生活方式、生理周期等方面的严格限制; 并比较了 OSC 滤噪前后, PCA, PLS-DA, SIMCA 等方法对血清 600 MHz 一维 ^{1}H NMR 谱进行性别判别分析的效果. 通过 OSC 去除不同隐变量个数后所导致的 PLS-DA 模型的性能改变, 探讨合理使用 OSC 滤噪的判断方法. 为 NMR 代谢组学技术应用于临床研究和疾病诊断提供理论依据.

1 材料与方法

1.1 研究对象及相关的资料

128 例血清标本均来自 2005 年 5～6 月在南通大学附属医院体检的正常健康成人, 其中男性 63 例, 女性 65 例. 体检及生化指标均在正常范围, 包括血尿粪三大常规、肝肾功能、血糖、血脂等项目. 具体生化指标有: 血糖(Glu)、总胆固醇(TC)、总甘油三酯(TG)、谷丙转氨酶(GPT)、总胆红素(TBi)、尿素氮 (BUN)、肌酐(Cr)、尿酸(UA)等. 取其中 78 例(男性 38 例, 女性 40 例)血清作为建立模式识别训练集的标本; 其余 50 例则作为验证模型效果的测试集标本.

1.2 NMR 样本的采集

采血前仅要求 8 h 禁食, 并无其它有关饮食、生活方式、生理周期等的严格限制. 所采血液室温下自然凝固 1 h, 收集血清, 4 ℃ 10000 g 离心 10 min, 收集上层血清, 置－80 ℃冷冻备用.

1.3 NMR 代谢组学技术的流程

1.3.1 ^{1}H NMR 谱的记录

进行 NMR 实验前, 取出冷冻血清, 室温下解冻, 4 ℃ 10000 g 离心 10 min, 取上层血清 500 μL 加至 5 mm 的 NMR 试管中, 加入 50 μL D_2O (用于锁场), 振荡混匀.

在 Varian Unity Inova 600 MHz 超导核磁共振谱仪上记录 ^{1}H NMR 谱. 实验温度为 298 K, 预饱和法压制水峰, 采用 Carr-Purcell-Meiboom-Gill (CPMG)脉冲序列抑制血清中蛋白质及脂蛋白较宽的共振信号[6]. 为了使采样后的磁化矢量完全恢复到热平衡态, 采用 4 s 的弛豫延迟时间. 每张 NMR 谱记录 48 个 FID, 每个 FID 收集 16 K 个数据点, 充零到 64 K 个数据点. 采用线宽为 0.3 Hz 的指数窗函数, 进行傅立叶变换, 对所获谱图进行相位和基线校正后, 参照乳酸的甲基共振峰(δ 1.33)对 ^{1}H NMR 谱的化学位移进行定标.

1.3.2 ^{1}H NMR 谱的分段积分与数据预处理

为了减少 pH 和离子强度等因素产生的化学位移变化, 便于进行模式识别分析, 以 δ 0.04 为单位将 ^{1}H NMR 谱(δ 10～0.2)划分成 245 个等宽的区域, 并对各个区域进行分段积分. 为了消除饱和水峰时引起的谱线差异, 以及由于尿素与溶剂交换质子后发生的部分饱和转移所造成的尿素信号差异, 将区域 δ 6.0～4.5 设为 0 积分段. 由所有积分区间的平均化学位移和强度积分值构成 NMR 分段积分谱, 再将这些分段积分谱的数据转化为 Excel 数据格式, 用于模式识别分析.

由于这些分段积分值大小差异较大, 进行模式识别分析前应对分段积分值进行预处理: OSC 前使用标准化处理, 即将变量减去其列均值后再除以标准差; OSC 后则采用中心化处理, 即将变量减去其列均值. 经标准化处理后的每个变量权重相同, 均值为 0, 方差为 1, 有利于分析低含量的代谢物; 中心化处理可使新坐标系的原点与群点的中心重合, 从而在保持原有数据间相关性的同时减少数据的动态范围.

1.3.3 NMR 数据的模式识别分析

采用 SIMCA-P 10.5 (Umetrics AB, Umea, Sweden)分析软件, 在 OSC 滤噪的前后, 分别用 PCA, PLS-DA, SIMCA 等模式识别方法处理和分析 NMR 数据. 245 个分段积分值作为自变量(*X*), 样本的性别类别作为因变量(*Y*).

2 结果与讨论

2.1 血生化分析

所有研究对象的血生化指标均在正常范围. 采用 STATA 6.0 统计分析软件, 训练集的 78 例研究对象的各指标以均数±标准差表示, *t* 检验结果为: 男女两组年龄, Glu, TC 及 BUN 之间差异无显著意义($P>0.05$); TBi 两组差异有显著意义($P<0.05$); TG, Cr, UA 与 GPT 男女两组差异有显著意义($P<0.01$), 具体数据见表 1.

2.2 OSC 去除两个隐变量滤噪前后用不同模式识别方法分析血清 ^{1}H NMR 谱

NMR 作为当前代谢组学研究所使用的最主要技术之一, 能对化学组成知之甚少的复杂样品(如尿液、血液、组织等)实现非破坏性、无偏向性、重复性好的快速分析[6,11,12], 在临床疾病的研究中具有独特的优势.

表 1　作为训练集的男女两组年龄及临床生化指标的比较

Table 1　Comparison of ages and clinical parameters for male and female training-set subjects

Item	Male group (n=38)	Female group (n=40)	P
Age/year	42.00±12.67	38.95±9.88	0.238
c (Glu)/(mmol•L^{-1})	5.01±0.40	4.9±0.41	0.225
c (TC)/(mmol•L^{-1})	4.52±0.63	4.26±0.71	0.0828
c (TG)/(mmol•L^{-1})	0.98±0.32	0.82±0.28	0.0182^a
c (BUN)/(mmol•L^{-1})	5.12±0.88	4.81±1.29	0.212
c (Cr)/(μmol•L^{-1})	105.87±9.16	88.73±8.22	4.84×10^{-13b}
c (UA)/(μmol•L^{-1})	338.05±48.08	230.93±41.46	1.50×10^{-16b}
c (GPT)/(U•L^{-1})	20.45±7.69	12.55±5.40	1.23×10^{-6b}
c (TBi)/(μmol•L^{-1})	15.10±3.50	13.08±4.35	0.0275^a

^{a} P<0.05; ^{b} P<0.01.

NMR 代谢组学用于疾病诊断的研究目前尚处于探索阶段. 虽然临床患者的血液、尿液标本的获取是比较容易的, 但对样本均一性的控制却有相当的难度. 通常可以通过选用纯种、同一窝、同龄、体重相近等措施减少实验动物的个体差异. 因此, 目前 NMR 代谢组学技术主要应用于研究药物毒性的动物实验[2,3,11]. 在获取临床样本时, 为了保持样本的均一性, 人们对研究对象的饮食、生活方式、生理周期等方面制定了较严格的入选标准[4,6~8], 但在临床实践中很难真正达到要求. OSC 作为一种较为有效的 NMR 波谱信号校正技术, 可以去除饮食环境等因素的影响, 降低临床样本的不均一性. 图 1 为我们使用不同的模式识别方法分析正常男女血清一维 600 MHz ^{1}H NMR 谱的结果. OSC 去除两个隐变量(latent variable, LV)滤噪后, PCA 分析所获得的主成分(principal component, PC) t_1/t_2 得分图可以较好地显示不同性别的丛集分布; PLS-DA 分析所获得的隐变量 t_1/t_2 得分图可以完全正确地区分不同的性别; SIMCA 分析所获得的 Coomans 图可见左下区内难以明确区分性别的样本数明显减少. 结果表明: 采用 OSC 滤噪可明显改善模式识别的分析效果; 有监督的模式识别方法效果好于无监督的方法; 三种方法中 PLS-DA 的判别分析能力最强. 由 PLS-DA 模型得到的性别判别效果与 Brindle 等[6]在文献中报道的研究结果基本一致. 另外, 观察各种模式识别方法显示的样本分布可见: 男性样本分布较为分散, 女性则相对集中, 表明正常男性血清内代谢物质的变化较大, 这可能与男性血清生化指标差异较大、其值相对较高有关.

图 2 为 OSC 去除两个隐变量前后 PLS-DA 回归系数图与 PCA 载荷图. 该图表明: 与男女分类相关的 NMR 谱图积分区段完全相同. 载荷或回归系数为正值表示对应于 NMR 谱图积分区段的物质, 男性组的值比女性组的值高; 负值表示相反. 比较 OSC 前后 PLS-DA 回归系数 CoeffCS[1]的柱形图(图 2a 和图 2b), 可以发现: OSC 滤噪前的 PLS-DA CoeffCS[1]的柱形图包含着更多的 NMR 谱图积分区段—除了与男女分类相关的区段外, 还有噪声以及与性别分类无关的区段. 这表明使用 OSC 滤噪能够滤掉与男女分类不相关的 NMR 谱图积分区段, 明显地提高 PLS-DA 分类的效果.

我们根据文献[7,13]提供的血清代谢组 ^{1}H NMR 谱化学位移的信息, 结合我们的血清 1D ^{1}H NMR 谱, 对图 2(d)与性别分类有关的主要 NMR 积分区段进行质子共振信号指认. 图 2(d)中网格标记的积分区段序数(bin number)与化学位移 δ 的对应关系为(从右到左): 229 对应 δ 0.86 (0.84～0.86), 219 对应 δ 1.26 (1.24～1.26), 201 对应 δ 1.98, 170 与 169 分别对应 δ 3.22 与 δ 3.26 (3.22～3.24). 括号内为文献[3]报道的数据. 这些积分区段与 Ramadan 等[7]报道的未经 OSC 处理的 PCA 性别分析结果揭示的与性别分类有关的主要 NMR 积分区段基本相同, 但回归系数的正负符号却相反. 另外, 还有多个区段与 Ramadan 等[7]报道的未经 OSC 处理的 PLS-DA 分析得到的男女有明显差异的主要区段吻合, 回归系数的正负符号也相同. 回归系数为正值的区段包括: δ 0.98 (0.98, 0.96), 2.18 (2.14), 2.38 (2.36), 2.50 (2.46), 2.94 (2.90), 3.14 与 4.10 (4.08). 回归系数为负值的区段包括: δ 1.18, 1.46 (1.50), 2.10 (2.08), 4.18 与 4.42. 其中部分区段与文献[7]报道的化学位移值(括号所列之值)稍有差别.

OSC 滤噪后 PLS-DA 分析所得到的与性别分类有关的主要区段回归系数多为正值, 表明这些区段对应的男性组血清代谢物浓度要高于女性组, 包括: δ 0.86(总胆固醇的 CH_3), 0.98 与 0.96(主要为游离胆固醇的 CH_3), δ 3.22～3.24(主要为胆碱的$[N(CH_3)_3]^+$, 大多来自脂蛋白 HDL 中的磷酸卵磷酯), δ 4.10(为磷酸卵磷酯中甘油主链骨架的 CH_2), 这些成分均与主要含胆固醇的

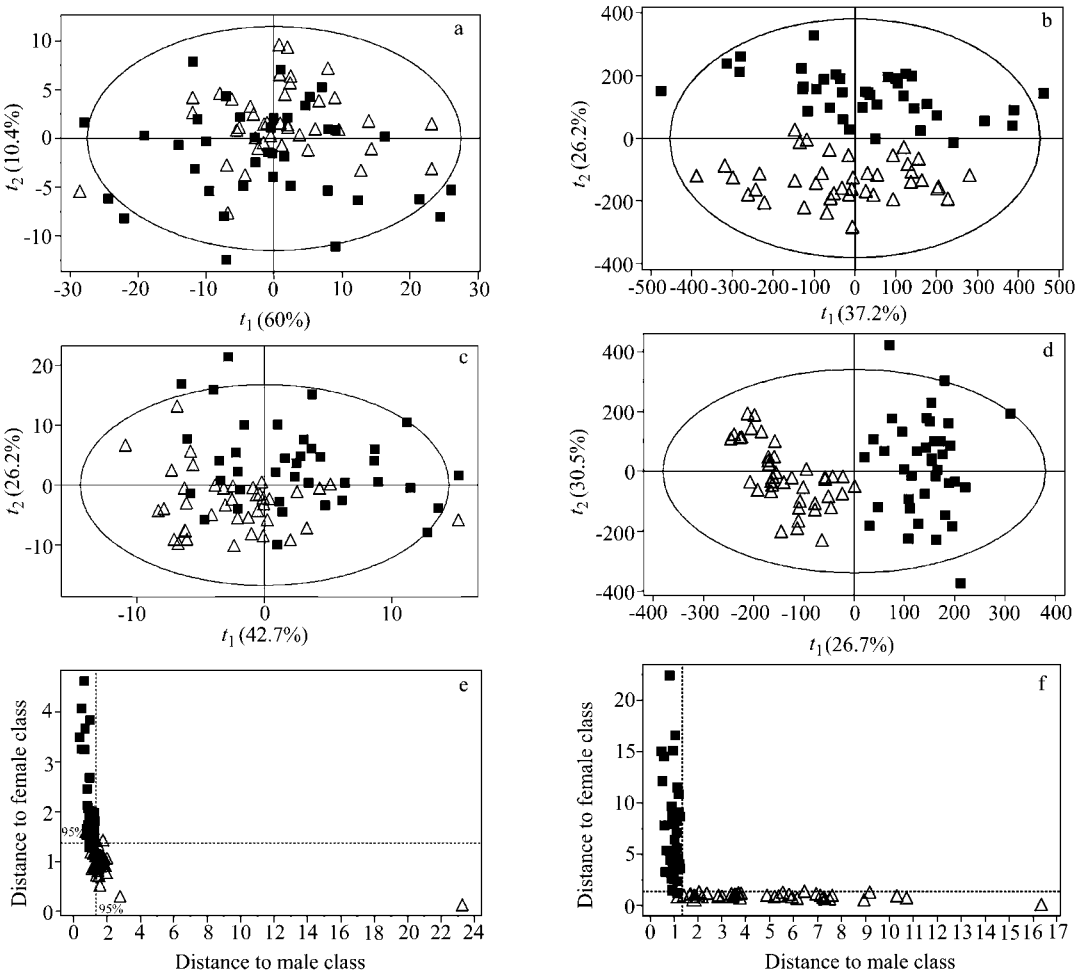

图 1 用不同的模式识别方法分析正常男女血清 600 MHz ^{1}H NMR 谱(■ 男性, △ 女性; 得分图的椭圆形区域和 Coomans 图的虚线以下区域表示 95 %的可信区间)

(a) OSC 滤噪前, PCA 的主成分得分图; (b) OSC 滤噪后, PCA 的主成分得分图; (c) OSC 滤噪前, PLS-DA 的隐变量得分图; (d) OSC 滤噪后, PLS-DA 的隐变量得分图; (e) OSC 滤噪前, SIMCA 的 Coomans 图; (f) OSC 滤噪后, SIMCA 的 Coomans 图

Figure 1 The gender discriminative performance to serum 600 MHz ^{1}H NMR profiling using three pattern recognition methods before OSC and after removing two OSC latent variables (■ male, △ female; the ellipse in the score plot and the area under the short-dash line in Coomans' plot means 95% confidence area)

(a) PC score plot of PCA before OSC; (b) PC score plot of PCA after OSC; (c) LV score plot of PLS-DA before OSC; (d) LV score plot of PLS-DA after OSC; (e) Coomans' plot of SIMCA before OSC; (f) Coomans' plot of SIMCA after OSC

高密度脂蛋白(HDL)有关. δ 1.98(主要为脂肪酸链的 $CH_2C{=}C$)与 δ 2.38(对应脂肪酸链的 CO—CH_2)也与血脂水平有关[7,13]. 可见 PLS-DA 对 ^{1}H NMR 谱进行性别分类结果与血脂水平存在着一定的关系, 这与男女两组血脂的生化分析结果相吻合.

结合相应的得分图, PCA 载荷与 PLS-DA 回归系数的正负符号可以指示与分类相关的 NMR 谱图积分区段内代谢物质含量的差别; 其大小代表着对应的 NMR 谱图积分区段在采用判别分析中的相对重要性. 因此, 有可能利用这些模式识别技术进一步寻找、指认与疾病相关的生物标志物, 探索生物体内的代谢途径. 图 2 中与性别分类有关的区段与 Ramadan 等[7] 报道的 PCA 分析得到的载荷图相比, 指示的区段基本相同, 但正负符号却相反; 这些积分区段内包含的代谢物主要与血脂有关[7,13]. 本文与 Ramadan 的论文[7]所采用的男女两组血脂水平的差异正好相反, Ramadan 等人的研究中[7]女性组血 TC 水平稍高于男性组; 而本文中血 TG 和 TC 水平男性组均高于女性组, 虽然 TG 水平的性别差异并无统计学意义, 但其 P 值为 0.0828, 已接近 0.05. 可见, PCA 分析所获得的主成分代表数据中最大的变量信息, 仅对 $\boldsymbol{X}$ 矩阵起着数据降维的作用; 但 PCA 分析时没有考虑反应变量 Y, 因而 PCA 主要用来反映样本的从集分布情况, 不适于直接用于分类判别.

与性别分类有关的 PLS-DA 回归系数揭示的 NMR

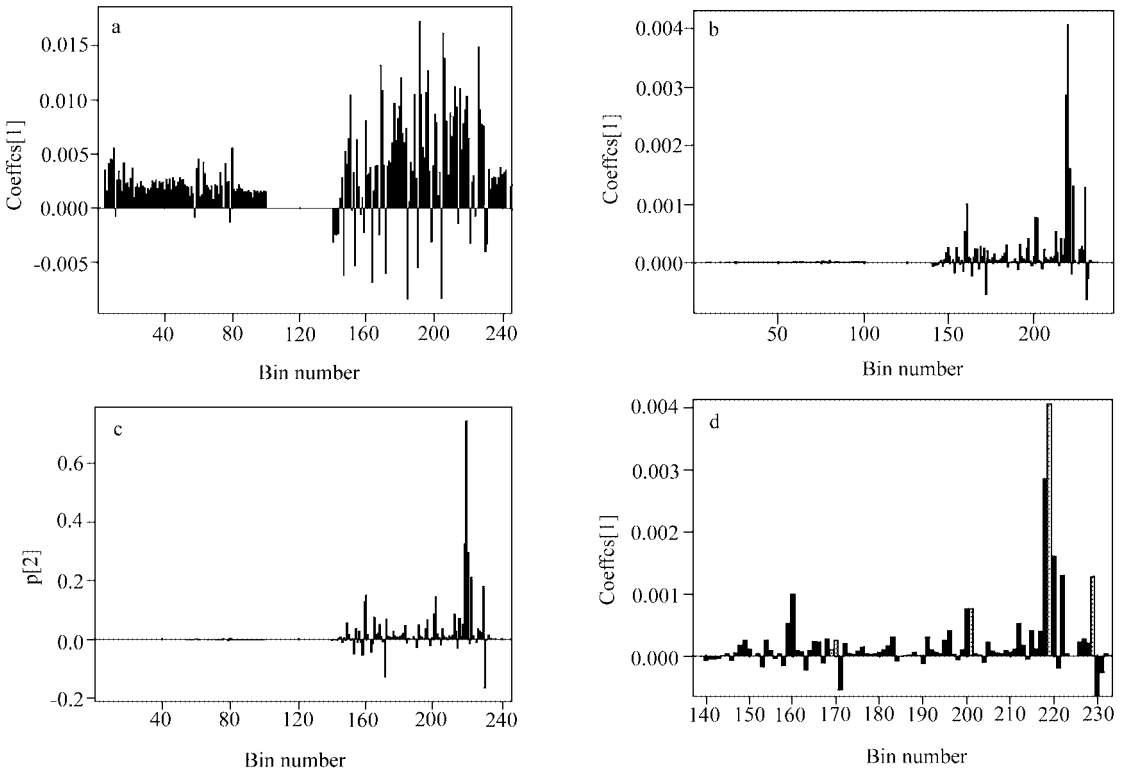

图 2　OSC 去除两个隐变量前后 PLS-DA 回归系数图和 PCA 载荷图

(a) OSC 滤噪前 PLS-DA 的回归系数图; (b) OSC 滤噪后 PLS-DA 的回归系数图; (c) OSC 滤噪后 PCA 的载荷图; (d) OSC 滤噪后 PLS-DA 所获得的与性别分类有关的回归系数图的局部放大

Figure 2　Coefficient plots of PLS-DA and loading plots of PCA before OSC and after removing two OSC latent variables

(a) Coefficient plot of PLS-DA before OSC; (b) Coefficient plot of PLS-DA after OSC; (c) Loading plot of PCA after OSC; (d) Zoomed coefficient plot of PLS-DA after OSC

分段积分区域与 Ramadan 的论文[7]采用遗传算法(genetic algorithm, GA)对变量进行选择后再进行 PLS-DA 分析的结果较为一致, 男女差别的主要积分区段的正负符号也相同, 说明采用 PLS-DA 方法分类临床血清样本性别是依据基本相同的血清代谢物谱峰. PLS-DA 不同于 PCA, 在分析 X 变量的同时还要考虑一个反应变量 Y 矩阵, 并使自变量(X)向反应变量(Y)回归的程度达到最大, 因而其分类判别能力明显增强. 另外, 图 2d 中部分区段与括号内 Ramadan 的论文[7]报道的数值稍有差别, 这可能与分段积分采用的方法不同有关. 他们以 δ 0.02 为单位将 ^{1}H NMR 谱划分成 493 个积分区段, 而我们则以 δ 0.04 为单位将 ^{1}H NMR 谱(δ 10～0.2)划分为 245 个积分区段. 而且, 本文与 Ramadan 的论文[7]设置的 0 积分段也有较大差异. 此外他们采用的实验温度和 pH 与我们的也略有差别, 这些因素可能会对结果产生一定的影响. 另外, Ramadan 的论文[7]收集血清标本时排除了研究对象存在感冒、发热等急性疾病以及服用药物的情况, 对女性还有月经周期的限制; 但他们在进行模式识别分析前没有进行信号校正处理, 而采用遗传算法对变量进行选择, 其 PLS-DA 分类效果虽然不是十分理想, 但却能优化对分类更重要的 X 变量. 本文仅要求受试者采血前 8 h 禁食, PLS-DA 经 OSC 去除两个隐变量时, 就能完全区分不同性别. 尽管 OSC 能够明显有效地改善分类判别效果, 但却不能去除那些与反应变量 Y 高度相关的系统误差, 并容易导致所构建的模型出现过拟合现象[14]. Ramadan 的论文[7]在 PLS-DA 模型计算前采用 Pareto 法进行数据预处理; 而本文在 OSC 前用标准化处理, OSC 后则用中心化处理. 以上差别可以解释本文与 Ramadan 的论文[7]中少数与性别分类有关的积分区段不一致的原因.

2.3　验证 OSC 去除两个隐变量后 PLS-DA 模型的预测性能

将 50 例测试集的标本分成 5 组, 每组 10 例, 其中男女各半; 男性年龄为(48.12±12.20)岁, 女性为(32.04±8.51)岁. 应用已建立的 OSC 去除两个隐变量后的 PLS-DA 模型, 仅根据一个有统计学意义的隐变量, 分别将这 5 组正常人的 NMR 数据作为测试集, 验证模型的性别预测判别效果. 仅根据一个有统计学意义的隐变

量, 计算验证该模型的平均成功率为(82±17.9)%; 而再纳入无统计学意义的第二个隐变量后, 验证的平均成功率则达到(90±10)%, 而且这两个值与对应的训练集计算所得的 PLS-DA 模型的 Q2 (cum)值(分别为 81%与81.7%)相近, 进一步说明该模型的预测性能稳定可靠, 质量优良.

2.4 OSC 去除不同隐变量个数后 PLS-DA 模型的变化

OSC 滤噪时去除隐变量的个数应根据剩余残差、去除隐变量的特征值大小、PLS-DA 模型计算所得隐变量个数以及模型质量改善情况等指标加以判断, 这一点在现有的 NMR 代谢组学研究中似乎并没有引起足够的重视[4,6,8,9]. 图 3 显示 OSC 滤噪时去除不同隐变量个数后 PLS-DA 模型变化. 该图表明: 当 OSC 去除两个隐变量时, 剩余残差为 20.82%, 此后剩余残差值变化明显减少并有稳定趋势, 说明此时 79.18%的 X 变量中与反应变量 Y 不相关的系统变异被去除. 第一、第二个隐变量的特征值分别为 37.21, 24.54, 明显大于后面隐变量的特征值. 此时 PLS-DA 计算所得隐变量个数为 1; 当用 OSC 去除一个隐变量时, PLS-DA 所得隐变量个数为 2; 而不使用 OSC 时, PLS-DA 所得隐变量个数则为 3. 对于只有一个反应变量的 PLS-DA 分析模型, 正确的计算结果应只得到一个隐变量; 如果有两个或更多的隐变量, 说明在 $\boldsymbol{X}$ 矩阵中仍存在较多的系统误差.

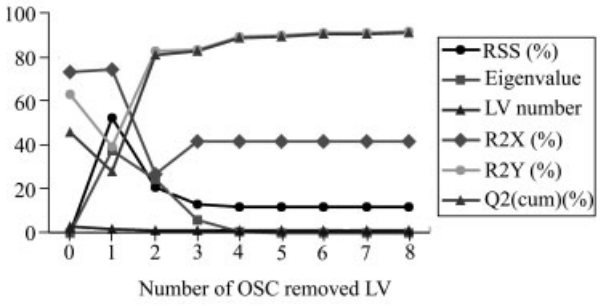

图 3 OSC 滤噪时去除不同隐变量个数后 PLS-DA 模型的变化
Figure 3 Variation of PLS-DA model before OSC and after removing different OSC latent variables

R2X, R2Y 和 Q2 (cum)为所建立的 PLS-DA 模型质量的评价指标. R2X 表示 PLS-DA 模型计算所获得的隐变量反映自变量 X 的变异的百分比; R2Y 则表示隐变量反映因变量 Y 的变异的百分比; Q2 (cum)为交叉验证后 PLS-DA 模型所获得的隐变量能够预测 X 和 Y 变异的累计百分比. 一般说来, 这三个值越高表明所建立的 PLS-DA 模型越能反映原始变量的信息, 其预测能力也越强. 图 3 中 R2X 在 OSC 去除两个隐变量时达到最低值, 表明此时 PLS-DA 计算模型包含的系统变异最少. 同时, R2Y (83%)和 Q2 (cum) (81%)也都达到 80%以上, 并趋于稳定, 说明 OSC 去除两个隐变量时 PLS-DA 计算模型的质量优良, 能够解释 83% (R2Y 值)的 Y 变量的变化, 并具有 81% (Q2(cum) 值)的模型正确预测能力.

NMR 代谢组学技术虽然起步不久, 目前在临床研究和疾病诊断等方面的应用尚处于探索阶段, 却已经受到国际生物医学界的广泛关注和重视. NMR 技术具有快速、无创、无偏向性、重复性好的独特优势; 而 PLS-DA 等有监督的模式识别方法结合合适的信号校正技术, 能够较好地消除各种复杂因素的干扰, 缩小临床样本的个体差异, 减少系统误差, 极大地推动着 NMR 代谢组学技术在临床研究和疾病诊断等方面的应用. 我们完全有理由相信, NMR 代谢组学技术在临床疾病的诊断、监测及治疗等方面的研究中具有广泛的应用前景, 并将带来巨大的经济效益.

References

1 Nicholson, J. K.; Lindon, J. C.; Holmes, E. *Xenobiotica* **1999**, *29*, 1181.
2 Nicholson, J. K.; Connelly, J.; Lindon, J. C.; Holmes, E. *Nat. Rev. Drug. Disc*. **2002**, *1*, 153.
3 Holmes, E.; Antti, H. *Analyst* **2002**, *127*, 1549.
4 Brindle, J. T.; Nicholson, J. K.; Schofield, P. M.; Grainger, D. J. ; Holmes, E. *Analyst* **2003**, *128*, 32.
5 Eriksson, L.; Antti, H.; Gottfries, J.; Holmes, E.; Johansson, E.; Lindgren, F.; Long, I.; Lundstedt, T.; Trygg, J.; Wold, S. *Anal. Bioanal. Chem*. **2004**, *380*, 419.
6 Brindle, J. T.; Anni, H.; Holmes, E.; Tranter, G.; Nicholson, J. K.; Bethell, H. W.; Clarke, S.; Schofield, P. M.; McKilligin, E.; Mosedale, D. E.; Grainger, D. J. *Nat. Med*. **2002**, *8*, 1439.
7 Ramadan, Z.; Jacobs, D.; Grigorov, M.; Kochhar, S. *Talanta* **2005**, *68*, 1683.
8 Odunsi, K.; Wollman, R. M.; Ambrosone, C. B.; Hutson, A.; McCann, S. E.; Tammela, J.; Geisler, J. P.; Miller, G.; Sellers, T.; Cliby, W.; Qian, F.; Keitz, B.; Intengan, M.; Lele, S.; Alderfer, J. L. *Int. J. Cancer* **2005**, *113*, 782.
9 Beckwith-Hall, B. M.; Brindle, J. T.; Barton, R. H.; Coen, M.; Holmes, E.; Nicholson, J. K.; Antti, H. *Analyst* **2002**, *127*, 1283.
10 Gavaghan, C. L.; Wilson, I. D.; Nicholson, J. K. *FEBS Lett.* **2002**, *530*, 191.
11 Bollard, M. E.; Stanley, E. G.; Lindon, J. C.; Nicholson, J. K.; Holmes, E. *NMR Biomed.* **2004**, *18*, 143.
12 Lindon, J. C.; Holmes, E.; Nicholson, J. K. *Curr. Opin. Mol. Ther*. **2004**, *6*, 265.
13 Mika, Ala-Korpela. *Prog. NMR Spectrosc.* **1995**, *27*, 475.
14 Keun, H. C.; Ebbels, T. M.; Antti, H.; Bollard, M. E.; Beckonert, O.; Schlotterbeck, G.; Senn, H.; Niederhauser, U.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. *Anal. Chim. Acta* **2003**, *490*, 265.

(A0604032 YANG, Y.; LING, J.)