波 谱 学 杂 志
第 24 卷第 4 期 Chinese Journal of Magnetic Resonance Vol. 24 No. 4
2007 年 12 月 Dec. 2007

文章编号: 1000-4556(2007)04-0381-13

代谢组学数据分析方法及在糖尿病研究中的应用

董继扬[1], 徐 乐[1], 曹红婷[1], 戴晓侠[2],
李学军[3], 杨叔禹[3], 陈 忠[1]*
(1. 厦门大学 物理系, 福建 厦门 361005; 2. 厦门大学 医学院, 福建 厦门 361005;
3. 厦门市第一医院, 福建 厦门 361005)

摘 要: 对 NMR 波谱数据的统计分析是基于 NMR 代谢组学研究的关键问题之一. 鉴于 NMR 波谱信号可以近似为样品中各种成分谱信号的线性叠加, 本文将非负矩阵分解(NMF)方法引入基于 NMR 代谢组学的数据处理中, 并与代谢组学中常用的统计方法 ——主成分分析(PCA)进行比较. 通过 NMF 和 PCA 两种方法对健康志愿者与 2 型糖尿患者血液和尿液的 NMR 谱图的统计分析, 对所获取的特征代谢物进行比较和验证, 并探讨了 PCA 方法可能存在的不足之处及其原因; 阐明了 NMF 方法是基于 NMR 的代谢组学研究中较理想的数据分析方法. 最后, 讨论了基于 NMR 代谢组学在糖尿病研究中的前景.

关键词: 基于 NMR 的代谢组学; 2 型糖尿病; 非负矩阵分解; 主成分分析
中图分类号: O591 **文献标识码**: A

引言

核磁共振(NMR)技术作为一种非侵入和高效的检测手段已成为当今化学、生物学和医学研究等领域的一种强有力的工具. 近年来, 随着 NMR 技术的提高和系统生物学的发展, 出现了基于 NMR 的代谢组学这一新兴学科[1], 它主要利用生物体液的 NMR 谱所提供的生物体内小分子代谢物的丰富信息, 研究相关生物体在功能基因组学、病理生理学、药理毒理学等方面的状况及动态变化, 以及它们所揭示的生物学意义, 并从分子水平上认识生命运动的规律[2]. 目前代谢组学已应用于包括疾病的诊断、药物作用机

收稿日期: 2007-08-28
基金项目: 福建省自然科学基金(T0750015)和厦门市重大疾病攻关研究基金(3502Z20051027)资助项目.
作者简介: 董继扬(1974-),男,博士,福建安侯人,副教授,从事生物医学信号处理. *通讯联系人: 陈忠, 电话: 0592-2181712, E-mail :chemz @xmu. edu. cn

制研究、药物研发、分子生理学、分子病理学、基因功能组学、营养学、环境科学等重要领域[3-6].

糖尿病是一种以糖代谢失常为主的内分泌代谢性疾病通常表现为整体的代谢紊乱,因此代谢组学的方法非常适合于糖尿病的研究. 事实上,早在 20 多年前,Nicholson 等人[7]就尝试性地将 NMR 技术用于糖尿病人和正常人的尿液和血液的分析,并发现糖尿病人样品中某些代谢物与正常人样品中的浓度相比有显著的差异. 这一发现预示了 NMR 技术在糖尿病的诊断方面的巨大潜力,激起了人们利用 NMR 技术研究糖尿病的兴趣. 近些年,越来越多的研究者开始了基于 NMR 代谢组学的方法对糖尿病的研究[8,9]. 例如: Mäkinen 等人[10]应用^{1}H NMR 技术,结合多变量数据分析方法,检测 1 型糖尿病人的血清样品中代谢物的变化,结果表明该方法可以清楚地辨别糖尿病人和正常人的特征代谢物. Jin 等人[11]分析了 Zucker 糖尿病肥胖鼠的血液的^{1}H NMR 谱和^{13}C NMR 谱,发现了在患病状态下葡萄糖的代谢途径,从而加深了对糖尿病发病机理的认识. 我们研究小组也从糖尿病人和正常人血液的^{1}H NMR 指纹图谱的统计分析中得到了相关的特征代谢物[12-15].

本文主要研究代谢组学的数据处理. 首先,对收集到的 2 型糖尿病志愿者和健康志愿者的血清和尿液样本进行 NMR 检测,获取高分辨的^{1}H NMR 指纹图谱. 其次,根据 NMR 谱数据特点,将非负矩阵分解(Non-negative Matrix Factorization, NMF)方法引入代谢组学数据处理中,分别利用 NMF 和主成分分析(PCA)方法对正常人与 2 型糖尿病人的血液和尿液的 NMR 谱数据进行统计分析和代谢组学分析,标记 2 型糖尿病的特征代谢物. 本文研究结果表明:样本不仅在 NMF 得分图中的可分性比在 PCA 得分图中更好、更容易解释,而且从 NMF 的负载图上所得到的标记代谢物比 PCA 负载图上所获得的标记代谢物更准确可靠. 最后,对糖尿病的代谢组学研究进行讨论与展望.

1 实验数据采集

实验所用的生物体液样品均由厦门市第一医院专职医务人员采集,其中糖尿病组的血液样品采自厦门市第一医院的住院糖尿病志愿者,糖尿病组的尿液样品采自厦门市第一医院门诊部的糖尿病志愿者,正常对照组的血液和尿液样品均采自厦门市第一医院体检部的健康自愿者.

1.1 血液样品的采集及制谱

实验样品包括 14 个 2 型糖尿病患者和 13 个正常人的血液,采用静脉穿刺抽取法,其中正常对照样品来自健康的、年龄与糖尿病患者相配的志愿者. 血液抽取之后经 3000 r/min 离心 10 min,取上层血清,放入 -80 ℃冰箱储存,2 个月内实验.

配置血清样品时,取一份血清与两倍体积的磷酸盐缓冲液(溶质: 0.2 mol/L Na_2HPO_4/0.2 mol/L NaH_2PO_4,溶剂为 80 % H_2O/20 % D_2O)混合,以消除 pH 对化学位移的影响,配好后滴入 5 mm 样品管. 在 Varian Unity plus 500 MHz 谱仪上采集样品的一维^{1}H NMR 谱图. 实验采用 5 mm HCN 三共振探头,温度 300 K,谱宽 5.12 kHz,信号累加 256 次,采样点数 16 k,采样 NOEPR-CPMG 脉冲序列,其中预饱和脉冲序列(NOESYPRESAT)抑制水峰信号,CPMG 自旋回波序列用来抑制由蛋白和脂蛋白等产生的宽峰而得到窄的谱线. 谱图采用 DSS 为内标.

糖尿病人和正常人血清的典型^{1}H NMR 谱图如图 1 所示.

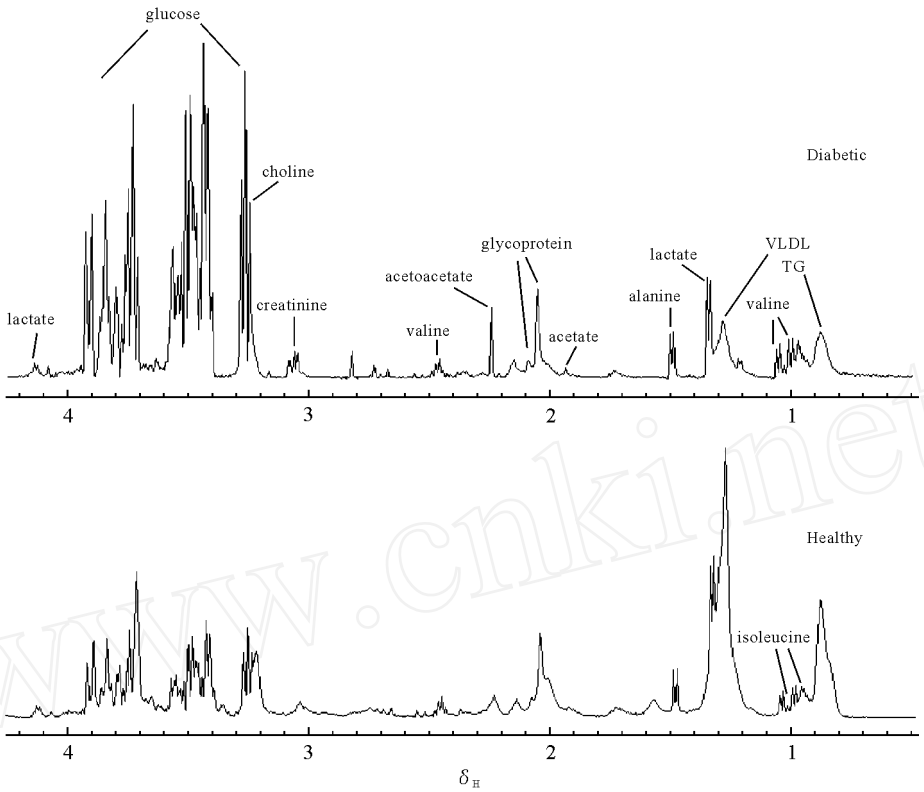

图 1　2 型糖尿病人和正常人血清的典型^{1}H NMR 谱图

Fig. 1　Typical ^{1}H NMR spectra of serum from a type 2 diabetic and a normal individuals

谱图先进行手工调相和基线校正, 并将^{1}H 化学位移 δ为 4.6～5.0 区间的谱峰强度置为零, 以消除残留水信号对分析结果的影响. 此外, 为了减少噪声和化学位移漂移的影响, 采用分段积分的方法, 积分间隔 $\Delta\delta$为 0.04. 通过分段积分, 得到一个 256 ×27 的观测数据矩阵, 每列代表一个样本.

1.2　尿液样品的采集及制谱

实验样品包括 14 个 2 型糖尿病患者和 14 个正常人的尿液, 其中正常对照样品来自健康的、年龄与糖尿病患者相配的志愿者. 样品收集后经 3000 r/ min 离心 10 min, 取上清液, 放入 - 80 ℃冰箱储存, 2 个月内实验.

NMR 制谱前尿液样品的准备: 取溶解后的尿样 500 μL, 尿样加入 250 μL 缓冲液 (0.2 mol/ L Na_2HPO_4/ NaH_2PO_4, pH = 7.4), 将 60 μL 的 D_2O 以及 10 μL 的 DSS 溶液转入 5mm 样品管中. 在 Varian Unity plus 500 MHz 谱仪上采集样品的一维^{1}H NMR 谱图. 实验采用 5 mm HCN 三共振探头, 温度 300 K, 谱宽 5 kHz, 信号累加 128 次, 采样点数 8 k, 用预饱和脉冲序列(NOESYPRESAT) 采样并抑制水峰信号. 糖尿病人和正常人尿液的典型^{1}H NMR 谱图如图 2 所示.

从图 2 可见, 人体尿液中的主要代谢物包含: 肌氨酸酐、乳酸、柠檬酸、丙氨酸、二甲胺、氨基乙酸等, 这与 Zuppi 等[16]的研究结果相同.

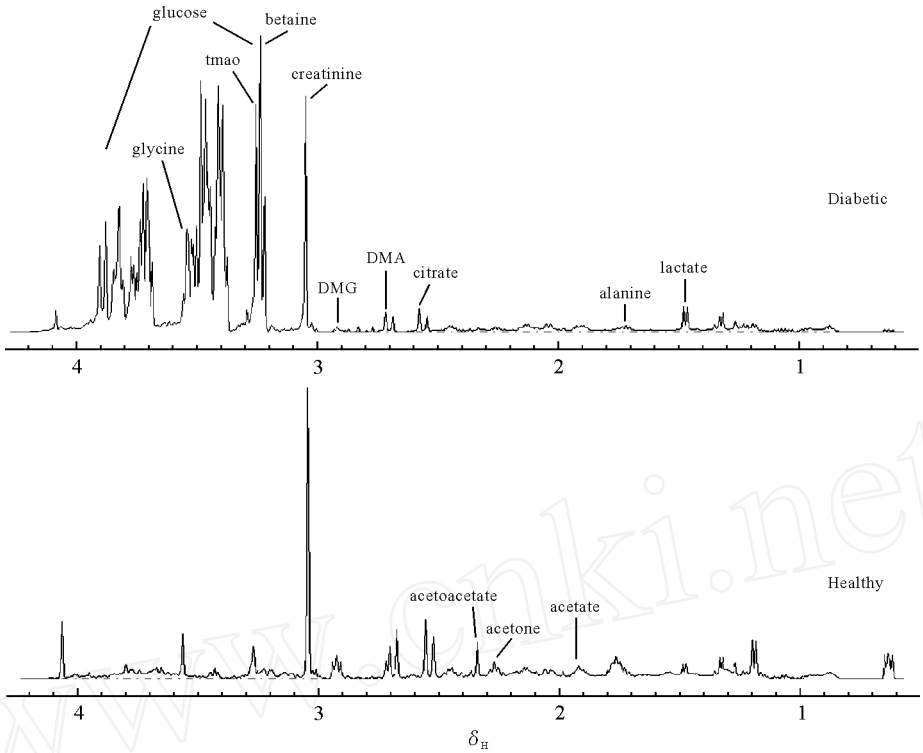

图 2　2 型糖尿病人和正常人尿液的典型^{1}H NMR 谱图

Fig. 2　Typical ^{1}H NMR spectra of urine from a type 2 diabetic and a normal individuals

谱图先进行手工调相和基线校正，并将^{1}H 化学位移 δ为 4.6～5.0 区间的谱峰强度置为零，以消除残留的水信号对分析结果的影响. 此外，为了消除噪声和化学位移漂移的影响，采用分段积分的方法，积分间隔 $\Delta\delta$为 0.04. 通过分段积分，得到一个 256 ×27 的观测数据矩阵，每列代表一个样本.

2　非负矩阵分解算法

NMF 算法[17, 18]是 Lee 和 Seung 于 1999 年提出的. 该算法可描述为：给定一个 $n \times m$ 非负观测矩阵 V，根据适当的目标函数，寻找一个 $n \times r$ 非负矩阵 W 和一个 $r \times m$ 非负矩阵 H，使

$$V = WH \tag{1}$$

其中 W 称为特征向量矩阵，H 称为权重矩阵.

通常采用欧氏距离作为 NMF 问题的目标函数，将 NMF 问题转化为如下最优化问题：

$$\begin{aligned} &\min \quad F = \sum_{i=1}^{n}\sum_{u=1}^{m}[V_{iu} - (WH)_{iu}]^2 \\ &\text{s.t.} \quad W, H \geq 0, \quad \sum_{i} W_{ij} = 1 \end{aligned} \tag{2}$$

该问题可以用迭代方法求解 W 和 H，求解过程如下[19]：

Step 1: 初始化 W, H 矩阵为非负随机矩阵;

Step 2: 按公式(3)对 W, H 进行同步迭代运算:

$$\begin{cases} W_{ia} = W_{ia} \sum_j \dfrac{V_{ij}}{(WH)_{ij}} H_{aj} \\ W_{ia} = \dfrac{W_{ia}}{\sum_j W_{ja}} \\ H_{aj} = H_{aj} \sum_i W_{ia} \dfrac{V_{ij}}{(WH)_{ij}} \end{cases} \tag{3}$$

Step 3: 根据公式(2)计算 V 和 WH 之间的距离, 如果大于预定值, 返回 Step 2 继续计算, 否则停止, 运算结束.

应用 NMF 分析代谢组学数据时, 可以先将样品的^{1}H NMR 谱数据经过一定的预处理后, 分段积分成 n 个数据点, 然后将 m 个样品的^{1}H NMR 谱数据分别作为观测矩阵 V 中的一个列向量, 组成一个 $n \times m$ 的观测矩阵 V. 经 NMF 算法分解后, V 中每一个列向量(即一个样本)就可以表示为 W 中的 r 个列向量的线性组合, H 中的行向量则为相应的权值系数. 这样 W 中的 r 个列向量就可认为是 V 空间中的 r 个特征基, 用这 r 个特征基所张成的空间代替原始的 m 维空间.

特征基数目 r 的选择是一个值得探讨的问题, 它与观测矩阵的性质有关, 不同的观测矩阵通常应该选择不同的 r 值, 目前尚无理论可循, 通常是根据多次试验的经验来选择. 若 r 取值太小, 则由新特征基所张成的空间不足以表达原始数据, 将产生严重的信息损失, 导致"欠拟合(underfitting)"现象, 使新特征基的泛化能力不足. 当然, r 取值也不能太大, 否则将导致"过拟合(overfitting)"现象. 此时虽然原始信息得到了较好的保留, 但新特征基的泛化能力却下降了. 在二分类问题中, r 值的选择通常是在保证新特征基泛化能力的前提下, 以获得较好的类可分性为原则. 但总说来, 通常 $r = m$, 因此 NMF 具有较好的降维作用.

与 PCA 等常用的统计方法相比, NMF 方法的优点主要体现在:

(1) 特征矩阵和权重矩阵的"非负"性限制, 与常见的物理信号(如图像信号、语音信号以及生物医学信号等)的非负性相吻合, 使分析得到的特征具有实际的物理意义, 而不是数学意义上的特征. 同样, 由于 NMR 谱数据都是非负的, 是由不同代谢物的共振峰线性叠加而成的, 这些性质与 NMF 的出发点完全一致, 因此谱数据的 NMF 分析结果将更加具有实际意义.

(2) 由于 NMF 是用非负性约束来获取数据表示的一种方法, 也即所获取的数据只允许是原始数据的加性组合, 而不允许减运算, 这一约束导致了 NMF 的基于局部表示的性质. 它可以克服整体表示的方法(如 PCA)可能由于能量集中而导致的分类困难等缺点, 更适于小浓度标记代谢物的检测.

在 NMR 制谱时, 样本之间的个体差异、实验条件的细微差别以及谱仪的不稳定性等各种因素都将引起 NMR 谱图的变化. 这类变化是由外界因素引起的, 它往往引起谱图的整体变化, 例如谱峰强度成比例下降或升高等. 这类变化与样品中代谢物成分不同所引起谱峰间的差异是两种性质完全不同的变化. 大家只关心由代谢物成分不同所引起的变化, 而把外界因素引起的变化当成一种干扰噪声. 由于 PCA 是对数据进行整体表

示的一种方法，其本身无法区分这两种变化，因此用 PCA 分析时需要先对谱图进行一定的归一化(通常采用面归一或线归一)，尽量消除外界因素对谱图的影响. 然而，归一化可能带来另外一个负面影响，即那些总体信号较弱的谱图，其噪声在谱图放大的过程中也随着放大，从而影响样本在得分图中的位置，甚至造成两类样本在得分图中交混在一起的现象. 而对于 NMF，由于它是对数据进行局部表示的方法，样本间的个体差异主要降低了 NMF 对观测数据的重构精度(重构精度可以通过增大 r 值来提高)，对分析结果产生的影响较小，因此只要个体差异不是太悬殊，便无需对谱图进行归一化.

3 实验分析与结果讨论

本文采用 Visual C^{++} 编程实现了 PCA 算法和 NMF 算法，分别对糖尿病人和正常人的血清样本及尿液样本的 NMR 谱进行分析. 为了消除外界因素的影响，在进行 PCA 分析之前，一律先对谱数据矩阵 V 进行面归一，即

$$v_{\alpha\mu} = \frac{v_{\alpha\mu}}{\sum_{\alpha=1}^{n} | v_{\alpha\mu} |}, \quad \alpha = 1, L, n; \ \mu = 1, L, m \tag{4}$$

而在 NMF 分析时，没有对谱图进行归一化处理.

此外，为了定量衡量所获得的新特征基对分类的贡献，本文引入如下一些变量.

设 X，Y 表示两个新特征基，例如 PCA 中的 PC1 和 PC2，NMF 中的 W1 和 W2. P_k^i 表示第 i 类的第 k 个样本，P_k^i 在特征基 X，Y 中的坐标为 $P_k^i = (x_{ik}, y_{ik})$. $C^i = (\overline{x_i}, \overline{y_i})$ 表示第 i 类的类中心，

$$\overline{x_i} = \frac{1}{n_i}\left(\sum_{k=1}^{n_i} x_{ik}\right), \quad \overline{y_i} = \frac{1}{n_i}\left(\sum_{k=1}^{n_i} y_{ik}\right) \tag{5}$$

式中 n_i 为第 i 类的样本数.

用类中心的距离表示两个类的类间距，即第 i 类和第 j 类的类间距 Inter(i, j) 为，

$$\mathrm{Inter}(i,j) = D(C^i, C^j) = \sqrt{(\overline{x_i} - \overline{x_j})^2 + (\overline{y_i} - \overline{y_j})^2} \tag{6}$$

用样本到类中心的平均距离加上 3 倍标准差表示该类的类内距，即第 i 类的类内距 Inner(i) 为，

$$\mathrm{Inner}(i) = \overline{D(P^i, C^i)} + 3\sigma \tag{7}$$

其中 $\overline{D(P^i, C^i)} = \frac{1}{n_i}\sum_{k=1}^{n} D(P_k^i, C^i)$，$\sigma = \sqrt{\frac{1}{n_i - 1}\sum_{k=1}^{n_i}[D(P_k^i, C^i), - \overline{D(P^i, C^i)}]^2}$.

在 X，Y 空间中，类间距越大表示不同类的样本在 X，Y 空间中的距离越大，类内距越小表示同一类样本在得分图中越集中，此时特征基 X，Y 越能代表两个类别之间的差异. 通过 X，Y 的负载图，我们便容易找到区分这两类样本的标记代谢物. 对于两类别的情况，可以定义一个综合参数 R 来评估新特征基 X，Y 的信息表征能力，

$$R = \frac{\mathrm{Inter}(1,2)}{\mathrm{Inner}(1) + \mathrm{Inner}(2)} \tag{8}$$

R 值越大，表明这两类的可分性越好；R 值越小，表明两类的可分性越差.

实验一、糖尿病组和正常对照组血清样品的分析结果

PCA 和 NMF 两个特征基的得分图和负载图如图 3 所示, 其中 NMF 分析中的参数 $r = 2$.

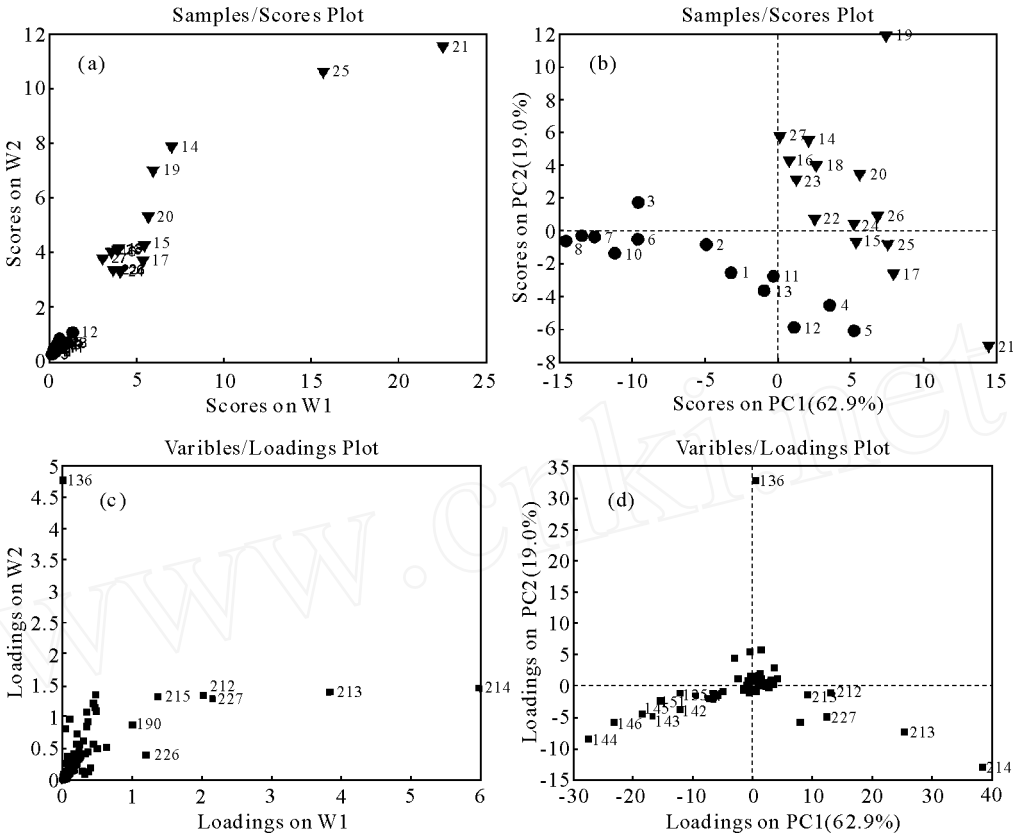

图 3　不同分析方法对血液 ^{1}H NMR 谱的分析结果

(a) NMF 得分图; (b) PCA 得分图; (c) NMF 负载图; (d) PCA 负载图

符号　·= 正常人样本; ▼= 2 型糖尿病人样本; ■= 谱积分段

Fig. 3　Analysis results of the plasma ^{1}H NMR data with different approach

(a) NMF scores plot, (b) PCA scores plot, (c) NMF loading plot, (d) PCA loading plot

Key: ·= healthy samples, ▼= type 2 diabetes samples, ■= spectra ingredients

在 NMF 和 PCA 的各个特征基上, 糖尿病组和对照组间的类可分性测度如表 1 所示.

表 1　NMF 和 PCA 分析方法在各特征基上类的可分性测度

Table 1　Class Separability on each basis of NMF and PCA

Basis	NMF			PCA		
	W1	W2	W1-W2	PC1	PC2	PC1-PC2
Separability	0.36	0.68	0.46	0.42	0.24	0.38

综合 PCA 负载图和表 1, 可以将最有可能存在特征代谢物的前 10 个谱峰段排列为: 214、213、136、212、227、215、226、138、152、137. 而综合 NMF 负载图和表 1 来看, 排列次序基本相同: 214、213、136、227、212、215、152、228、190、226. 我们将这些谱峰段中与糖尿病相关的可能代谢物列在表 2 中.

表 2 谱图中对类可分性贡献较大的区域及所包含的可能的代谢物

Table 2 Potential metabolites in the spectra slices with better class separability contribution

No.	δ_H	Potential Metabolites
136～138	3.65～3.75	葡萄糖(glucose)、山梨醇(sorbitol)
212～215	1.22～1.37	胆固醇(cholesterin)，低密度脂蛋白(LDL)、乳酸(lactate)
226～228	0.82～0.94	胆固醇(cholesterin)、低密度脂蛋白(LDL)、缬氨酸(valine)、异亮氨酸(isoleucine)、丁酸(Butyric)
152	3.21～3.25	胆碱(choline)
190	2.02～2.06	醣蛋白(glycoprotein)

糖尿病的病理生理为绝对或相对胰岛素分泌不足所引起的代谢紊乱，包括糖、蛋白质、脂肪、水及电解质等，严重时常可导致酸碱平衡失常. 上述分析结果与国际上最新的研究结果相符[20,21,13,22].

实验二、糖尿病组和正常对照组的尿液样品的分析结果

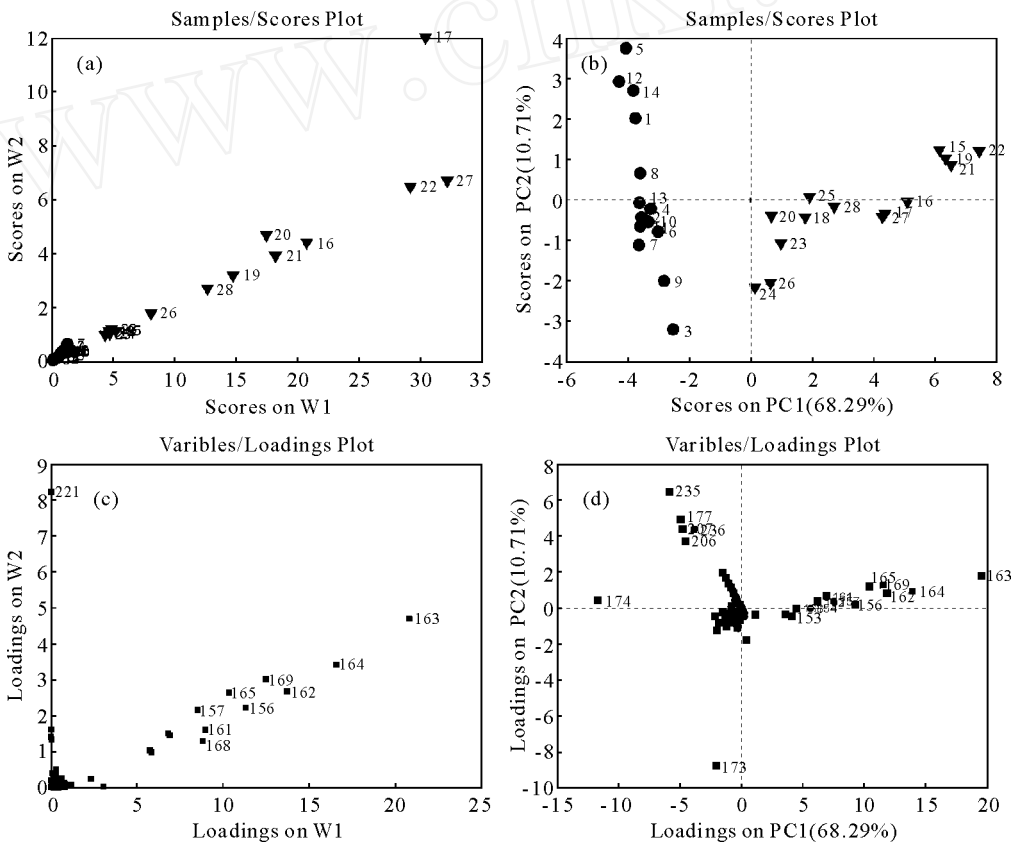

图 4 不同分析方法对血液 ^{1}H NMR 谱的分析结果

(a) NMF 得分图；(b) PCA 得分图；(c) NMF 负载图；(d) PCA 负载图

符号 ·= 正常人样本；▼= 2 型糖尿病人样本；■= 谱积分段

Fig. 4 Analysis results of the plasma ^{1}H NMR data with different approach

(a) NMF scores plot, (b) PCA scores plot, (c) NMF loading plot, (d) PCA loading plot

Key: ·= healthy samples, ▼= type 2 diabetes samples, ■= spectra ingredients

PCA 和 NMF 两个特征基的得分图和负载图如图 4 所示, 其中 NMF 分析中的参数 $r = 3$.

在 NMF 和 PCA 的各个特征基上, 糖尿病组和对照组间的类可分性测度如表 3 所示.

表 3　NMF 和 PCA 分析方法在各特征基上类的可分性测度

Table 3　Class Separability on each basis of NMF and PCA

Basis	NMF			PCA		
	W1	W2	W1-W2	PC1	PC2	PC1-PC2
Separability	0.55	0.39	0.54	1.08	0.05	0.64

综合 PCA 负载图和表 3, 可以将最有可能存在特征代谢物的前 10 个谱峰段排列为: 163、164、162、169、165、156、157、161、152、154. 而综合 NMF 负载图和表 3 来看, 排列次序基本相同: 163、164、162、169、156、165、221、161、168、157. 我们将与糖尿病相关的可能代谢物列在表 4 中.

表 4　谱图中对类可分性贡献较大的区域及所包含的可能的代谢物

Table 4　Potential metabolites in the spectra slices with better class separability contribution

No.	δ_H	Potential Metabolites
161～165	3.35～3.56	葡萄糖(glucose)、氨基乙酸(glycine)
168～169	3.19～3.29	葡萄糖(glucose)、氧化三甲胺(tmao)、三甲铵乙内酯(betaine)
156～157	3.66～3.76	葡萄糖(glucose)
152	3.86～3.91	葡萄糖(glucose)
154	3.78～3.84	葡萄糖(glucose)
173～174	3.01～3.08	肌氨酸(creatine)

上述分析结果与国际上的最新研究结果[23,24]一致. 这些成果可望为糖尿病诊断提供理论依据.

NMF 和 PCA 两种分析方法的比较:

3.1　NMF 得分图中样本的分布更符合实际情况

在 NMF 得分图中, 正常对照组样本比较集中, 而糖尿病组的样本比较分散, 如图 3(a)和图 4(a)所示. 这种分布更接近于实际样本的性质. 因为实验中的正常对照组样品来自一些健康的志愿者, 他们的年龄差异比较小, 而糖尿病组样品来自一些晚期糖尿病患者, 他们大部分都患有其它不同的并发症, 而且年龄差异也比较大. 而在 PCA 得分图中, 两类样本的离散度都比较大, 如图 3(b)和图 4(b)所示, 样本分布不符合实际. 进一步的研究表明, 正如本文第二节所指出的, 导致正常组样本离散的原因是由于在 PCA 分析之前, 对所有样本做了面归一化. 样本归一化使得在对信号较弱的样本进行放大时, 其噪声也得到了放大, 从而引入了一些额外的信息, 加大了同类样本间的离散度.

3.2 NMF 的负载图的物理意义更明确、更容易解释

NMF 是数据的局部表示，其特征(基)矩阵和权重矩阵都仅包含非负的元素，即观测样本只允许是特征基的加性的和非负的组合，这使得原始数据空间的每一维变量在 NMF 的负载图上，都处于新的特征基所张成的坐标系的第一象限中，如图 3(c)和图 4(c)．因此，NMF 的特征基与类之间关系的相关性往往比较小，即 NMF 负载图上每一点(即原始数据空间中的一维)对类的可分性贡献(如区分糖尿病和非糖尿病)，可以只从其在负载图某一坐标的大小来判断．例如，根据血液样本的 W1-W2 负载图(图 3(c))，给定一个新的未知样品，若该样品的^{1}H NMR 谱在 214、213、136、227、212、215、152 等谱峰段的值较大，那么该样品是糖尿病样品的可能性就很大．而 PCA 的主成分与类之间的关系往往具有较大的相关性．

3.3 在 PCA 的主成分中，个体差异信息和代谢组分差异信息往往相互交错，影响分析结果的准确性

PCA 是对数据的一种整体表示，是基于信息量保留最大的数据处理原则．PCA 不仅保留了样本内代谢物差异的信息，而且保留了样本间的个体差异信息，而个体差异信息会造成同一类样本在 PCA 得分图上距离较远，如图 3(b)和图 4(b)，甚至使不同类之间有相互的交叉，难以达到正确的分类．而且，如果个体差异信息较大，那么这种差异将掩盖代谢物组分差异的信息，从而使得 PCA 方法的分析结果误差较大．此外，当存在个体差异时，由于浓度较大的代谢物，其在样本间的差异信息往往比浓度较小的组分大很多，使得 PCA 负载图上距离原点较远的点主要是一些浓度较大的代谢组分，浓度小的代谢组分的影响通常体现不出来，而这些小浓度组分往往具有重要的生物学意义．如实验一，在 PCA 负载图中对类可分性贡献较大的前 10 个谱峰段大多是浓度较高的组分(即峰值较高)，峰值较低的谱峰段 228 和 190 的作用则体现不出来，而实际上 228 和 190 谱峰段所包含的代谢物缬氨酸、醣蛋白以及低密度脂蛋白(LDL)等确实与糖尿病有很大的关系．NMF 方法则保留了这些谱峰段．因此，可以说 NMF 比 PCA 更适合作为检测低浓度生物标记物的手段．

4 总结与展望

糖尿病是继心血管疾病、癌症之后的人类第三大顽疾，有现代文明病之称．随着人们生活方式的改变，糖尿病的患病率急剧增加，且日趋年轻化．专家预计到 2010 年，世界糖尿病患病人数将达到 2.2 亿，到 2025 年将超过 3 亿．目前在临床医学上尚没有根治糖尿病的方法．虽然通过早期发现可以对其进行有效的控制，但由于糖尿病早期一般不表现典型的“三多一少”症状，只表现为轻微的疲乏，甚至没有任何症状，故很容易被人们忽视．而且当前常用的医学检测手段主要是一些生物化学方法，如尿糖检查、血糖检查、血清胰岛素、24 h 尿 C 肽、相关抗体检测以及肝功、肾功和血脂等等，这些检测方法通常只是检测单个或少数几个成分的变化来获得疾病的相关信息，具有一定的局限性．因此，发展高效的早期检测和早期诊断技术一直是糖尿病研究中的一个热点．

基于 NMR 的代谢组学是探索糖尿病早期检测和早期诊断技术的一种充满希望和挑战的方法．国际上在这方面的研究已经开始起步[25-30]，并显示出很好的发展势头，但离实际应用还有很长一段距离．目前，这方面的研究还存在以下几点困难：一是样品收集

困难. 由于在医学上没有对早期糖尿病的有效诊断方法, 大部分已有的研究结果都是从已确诊的糖尿病病人和正常人的对照中得到的, 而糖尿病确诊病人往往同时患有不同的并发症, 这给糖尿病的生物标记物的确定及代谢途径的理解等研究带来了很大的困难. 虽然有些研究者在动物模型的研究上取得了一些成果, 但动物模型与真实的糖尿病发病过程仍有较大的差别. 因此, 糖尿病的研究是一个需要长期积累的过程. 二是数据处理困难. 生物样品的组成非常复杂, 通常高分辨率一维^{1}H NMR 谱含有数以千计的共振峰, 这些共振峰又存在随机的漂移、重叠和噪声等现象, 虽然可借助模式识别方法来挖掘谱图数据中所蕴涵的信息, 但不同的数据处理方法侧重点不同, 仅能获得数据的某一个或几个方面的信息. 例如最常用的 PCA 方法侧重的是保留原始数据中的二阶矩信息, 但其未能利用原始数据中的类别信息, 使得降维后的数据有时反而不利于模式分类. 因此必须探索更多的有效的统计分析方法, 将它们有机地结合起来, 以获得更多的有效信息. 三是生物学分析困难. 生物体液中的代谢物是生物体各个组织代谢物的综合, 同一种生理或病理变化在不同时期对代谢产物的影响不同, 不同的生理病变对某些代谢物的影响则可能相同. 从体液代谢产物的变化来推断生物组织生理或病理变化是一个异常复杂的逆过程. 以上这些困难不仅仅是糖尿病代谢组学研究所面临的, 也是所有代谢组学研究所面临的, 需要不同学科的研究者一起携手攻克.

本文的工作是对 2 型糖尿病初步的代谢组学研究. 我们收集了 2 型糖尿病人的血液及尿液样品, 并利用 NMR 技术获取样品的一维^{1}H NMR 谱, 分别与健康人的血液及尿液样品的^{1}H NMR 谱进行统计分析, 获得了一些与型糖尿病相关的特征代谢物. 此外, 我们将非负矩阵分解方法引入基于 NMR 代谢组学的数据处理中, 并与常用的统计方法 PCA 比较. 通过对比两种方法的分析结果, 阐述了 PCA 方法在 NMR 谱数据处理中的一些不足, 证明了 NMF 方法更适用于代谢组学数据处理.

参考文献:

[1] Nicholson J K, Lindon J C, Holmes E. Metabonomics: understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data [J]. Xenobiotica, 1999, 29: 1 181 - 1 189.

[2] Lindon J, Holmes E, Nicholoson J K. Pattern recognition methods and applications in biomedical magnetic resonance [J]. Prog Nucl Mag Res Spe, 2001, 39: 1 - 40.

[3] Lindon J C, Holmes E, Nicholoson J K, *et al*. Metabonomics technologies and their applications in physiological monitoring, drug safety assessment and disease diagnosis [J]. Biomarkers, 2004, 9: 1 - 31.

[4] Clayton T A, Lindon J C, Cloarec O, *et al*. Pharmaco-metabonomic phenotyping and personalized drug treatment [J]. Nature, 2006, 440: 1 073 - 1 077.

[5] Hector C K. Metabonomic modeling of drug toxicity [J]. Pharmacology & Therapeutics, 2006, 109: 92 - 106.

[6] Griffin J L, Nicholls A W. Metabolomics as a functional genomic tool for understanding lipid dysfunction in diabetes, obesity and related disorders [J]. Pharmacogenomics, 2006, 7(7): 1 095 - 1 107.

[7] Nicholson J K, Oflynn M P, Sadler P J. Proton-nuclear-magnetic-resonance studies of serum, plasma and urine from fasting normal and diabetic subjects [J]. Biochem J, 1984, 217: 365 - 375.

[8] Wang C, Kong H W, Guan Y F, *et al*. Plasma phospholipid metabolic profiling and biomarkers of type 2 diabetes mellitus based on high-performance liquid chromatography/electrospray mass spectrometry and multivariate

statistical analysis [J]. Anal Chem, 2005, 77(13): 4 108 - 4 116.

[9] Yuan K L, Kong H W, Guan Y F, *et al*. A GC-based metabonomics investigation of type 2 diabetes by organic acids metabolic profile [J]. J Chromatogr B, 2007, 850(1 - 2): 236 - 240.

[10] Mäkinen V-P, Soininen P, Forsblom C, *et al*. Diagnosing diabetic nephropathy by ^{1}H NMR metabonomics of serum [J]. Mag Res Mat Phy, 2006, 19: 281 - 296.

[11] Jin E S, Burgess S C, Merritt M, *et al*. Differing mechanisms of hepatic glucose overproduction in triiodothyroninetreated rats vs. Zucker diabetic fatty rats by NMR analysis of plasma glucose [J]. Am J Physiol Endocrinol Metab, 2005, 288: E654 - E662.

[12] Xu L, Dong J Y, Dai X X, *et al*. Non-negative matrix factorization for diabetes II metabolic profiling analysis [C]. ICBBE2007, 2007, 1(2): 651 - 653.

[13] Wen J B, Xiao X, Dai X X, *et al*. Data normalization for diabetes II metabonomics analysis [C]. ICBBE2007, 2007, 2(2): 694 - 697.

[14] Wen J B(温锦波). NMR based metabonomics's data preprocess methods and its application on diabetes mellitus study(基于 NMR 的代谢组学的数据预处理方法及其在糖尿病研究的应用) [D]. Xiamen(厦门): Xiamen University(厦门大学), 2007.

[15] Xiao X(肖娴), Yang S Y(杨叔禹), Dong J Y(董继扬), *et al*. Optimization of NMR pulse sequences for metabonomics(基于 NMR 代谢组学的脉冲序列优化) [J]. J Fuzhou Univ(福州学报), 2007, 35(S): 37 - 40.

[16] Zuppi C, Messana I, Forni F, *et al*. ^{1}H NMR spectra of normal urines reference ranges of the major metabolites [J]. Clinica Chimica Acta, 1997, 265: 85-97.

[17] Lee D D, Seung H S. Learning the parts of objects by non-negative matrix factorization [J]. Nature, 1999, 401: 788 - 791.

[18] Rao N N, Shepherd S J. Clustering gene expression data for periodic genes based on INMF [J]. Comp Sci, 2006, 4 115: 412 - 423.

[19] Lee D D, Seung H S. Algorithms for non-negative matrix factorization [J]. Adv Neural Info Proc Syst, 2001, 13: 556 - 562.

[20] Garvey W T, Kwon S, Zheng D, *et al*. Effects of insulin resistance and type 2 diabetes on lipoprotein subclass particle size and concentration determined by nuclear magnetic resonance [J]. Diabetes, 2003, 52: 453 - 462.

[21] Soedamah-Muthu S S, Colhoun H M, Thomason M J, *et al*. The effect of atorvastatin on serum lipids, lipoproteins and NMR spectroscopy defined lipoprotein subclasses in type 2 diabetic patients with ischaemic heart disease [J]. Atherosclerosis, 2003, 167: 243 - 255.

[22] Ding S Y, Xenia T, Hansen B. Nuclear magnetic resonance-determined lipoprotein abnormalities in nonhuman primates with the metabolic syndrome and type 2 diabetes mellitus [J]. Metabolism, 2007, 56(6): 838 - 846.

[23] Messana I, Forni F, Ferrari F, *et al*. Proton nuclear magnetic resonance spectral profiles of urine in type II diabetic patients [J]. Clin Chem, 1998, 44(7): 1 529 - 1 534.

[24] Ciurtin C, Nicolescu A, *et al*. Metabolic profiling of urine by H-1-NMR spectroscopy - A critical assessment of interpreting metabolite concentrations for normal and diabetes groups [J]. Revista de Chimie, 2007, 58 (1): 51 - 55.

[25] Williams R E, Lenz E M, Evans J A, *et al*. A combined 1H NMR and HPLC-MS-based metabonomic study of urine from obese (fa/fa) Zucker and normal Wistar-derived rats [J]. J Pharmaceut Biomed, 2005, 38: 465 - 471.

[26] Williams R E, Lenz E M, Evans J A, *et al*. The comparative metabonomics of age-related changes in the urinary composition of male Wistar-derived and Zucker (fa/fa) obese rats [J]. Mol Biosys, 2006, 2(3 - 4): 193 - 202.

[27] Hodavance M S, Ralston S L, Pelczer I, *et al*. Beyond blood sugar: the potential of NMR-based metabonomics for type 2 human diabetes, and the horse as a possible model [J]. Anal Bioanal Chem, 2007, 387: 533 - 537.

[28] Lauridsen M, Hansen S H, Jaroszewski J W, *et al*. Human urine as test material in ^{1}H NMR-based metabo-

nomics: recommendations for sample preparation and storage [J]. Anal Chem, 2007, 79: 1 181 - 1 186.

[29] Cloarec O, Dumas M E, Andrew C, *et al*. Statistical total correlation spectroscopy: an exploratory approach for latent biomarker identification from metabolic ^{1}H NMR data sets [J]. Anal Chem, 2005, 77: 1 282 - 1 289.

[30] Gipson G T, Tatsuoka K S, Brian C S, *et al*. Weighted least-squares deconvolution method for discovery of group differences between complex biofluid ^{1}H NMR spectra [J]. J Mag Res, 2006, 183: 269 - 277.

A New Data Processing Method for Metabonomic and Its Application in a Study of Diabetes

DONG Ji-yang[1], *XU Le*[1], *CAO Hong-ting*[1], *DAI Xiao-xia*[2], *LI Xue-jun*[3], *YANG Shu-yu*[3], *CHEN Zhong*[1*]

(1. Department of Physics, Xiamen University, Xiamen 361005, China;
2. School of Medicine, Xiamen University, Xiamen 361005, China;
3. The Xiamen First Hospital, Xiamen 361005, China)

Abstract: Multivariate statistical methods are frequently used in nuclear magnetic resonance (NMR)-based metabonomic researches to analyze NMR spectra of biofluids. Based on the fact that the NMR spectrum of a given sample are a sum of the NMR signals from all constituting ingredients, we developed a non-negative matrix factorization (NMF) method, capable of finding parts-based and linear representations of non-negative data, for analyzing the data acquired in NMR-based metabonomic studies. Detail comparisons were made between the NMF method and the commonly use principal component analysis (PCA) method by employing the two methods to discriminate the urine and serum spectra of type-2 diabetic patients from those of the healthy controls. It was shown that, compared to the PCA method, the NMF method is a more effective and accurate method for processing NMR spectra acquired in the metabonomic studies, partially due to its unique features such as the non-negative constraints and part-based representation. The disadvantages of the PCA method were also analyzed and discussed.

Key words: NMR, metabonomics, type 2 diabetes, non-negative matrix factorization, principle component analysis

* Corresponding author: Chen Zhong, Tel: 0592-2181712, E-mail: chenz@xmu.edu.cn.