chem.scichina.com



# 一种基于液相色谱-质谱技术进行血清代谢组学研究的方法: 从代谢指纹到潜在标志物

陈静 单圆鸿 严沁 赵素敏 尹沛源 赵欣捷 路鑫 席晓薇 万小平 许国旺 ·

中国科学院大连化学物理研究所,中国科学院分离分析重点实验室,大连 116023 上海交通大学附属第一人民医院妇产科,上海 200080

\* 通讯作者, E-mail: xugw@dicp.ac.cn

收稿日期: 2009-07-14; 接受日期: 2009-08-20

摘要 本文描述了一种基于液相色谱-质谱技术(LC-MS)的代谢组学发现疾病潜在标志物的方法. 该方法利用LC-MS获得代谢指纹图谱, 并通过多种统计分析方法对产生的海量数据进行分析, 最终筛选出潜在标志物. 数据分析过程包括: 通过归一化、修正80%规则、数据集分割和数据缩放等方法对数据集进行预处理; 通过正交校正的偏最小二乘(OPLS)模式识别方法对样品进行分型; 根据模型的变量重要性因子(VIP值)、非参数检验结果和 z 值筛选潜在标志物. 以宫颈癌血清样本为例, 应用上述方法, 15 个变量被确认为潜在标志物, 操作者接受曲线(ROC)下的面积为0.667~0.956. 经过相关性分析和结构鉴定, 发现这15个变量来自9个化合物. 其中7个化合物被鉴定为色氨酸、硬脂酸、花生四烯酸、溶血磷脂酰胆碱(0:0/16:0, 16:0/0:0, 18:1/0:0 和 18:0/0:0), 说明在宫颈癌中花生四烯酸和溶血磷脂酰胆碱的代谢发生异常.

关键词 代谢组学 液质联用 宫颈癌 肿瘤标志物

#### 1 引言

代谢组学是通过考察生物体系受刺激或扰动后 其代谢产物的变化,揭示机体生命代谢活动的本质 的科学<sup>[1,2]</sup>.与基因组和蛋白质组学不同,代谢组学的 研究对象是分子质量小于1000的所有内源性代谢物.

据估计,人体中各类内源性小分子代谢物成千上万<sup>[3]</sup>.体液中各种代谢物的浓度范围和物理化学性质差异很大,目前没有一种分析技术能对体液中所有的代谢物进行分析.NMR、质谱与多种分离手段联用成为代谢组学中的重要分析工具,如:液相色谱-质谱联用(LC-MS)<sup>[4~11]</sup>、气相色谱-质谱联用(GC-MS)<sup>[7,9,12]</sup>、毛细管电泳-质谱联用(CE-MS)<sup>[13]</sup>、串联液

相色谱-质谱联用(HPLC+HPLC-MS)<sup>[14]</sup>、全二维气相色谱-质谱联用(GC×GC-MS)<sup>[15,16]</sup>等, 其中最常用的是 RPLC-MS<sup>[4~7,9,17,18]</sup>和 GC-MS<sup>[7,9,12]</sup>.

从代谢组学分析产生的海量数据中发现潜在标志物,需要借助多变量模式识别方法. 无师监督的模式识别方法包括主成分分析(PCA)<sup>[19]</sup>、聚类分析(HCE)<sup>[20]</sup>等; 有师监督的方法包括判别分析(DA)<sup>[21]</sup>、偏最小二乘分析(PLS)<sup>[19]</sup>、偏最小二乘判别分析(PLS-DA)<sup>[19]</sup>, 正交校正的偏最小二乘分析(OPLS)<sup>[19,22]</sup>等. 当变量数量远大于样品数量时, PLS或PLS-DA模型容易过拟合. 根据模式识别模型抽提出对分类有重要贡献的化合物后,还需要进一步验证这些化合物的差异性.

1268

恶性肿瘤是对人类危害最大的一类疾病,如何实现恶性肿瘤的早期诊断一直是一个难题.目前临床上常用的肿瘤标志物多为基因表达产物和蛋白等大分子.随着代谢组学的发展,越来越多的研究开始关注临床医学领域中的内源性小分子代谢物[18].

本文以宫颈癌血清样本为例,以液质联用系统为分析平台,运用统计分析手段,描述利用代谢组学技术从代谢指纹图谱中发现潜在标志物的方法.

# 2 实验部分

#### 2.1 试剂

乙腈(色谱纯)购自Merk公司、甲酸(色谱纯)购自TEDIA公司(TEDIA, OH, USA). 溶血磷脂酰胆碱(12:0/0:0)和溶血磷脂酰胆碱(18:0/0:0)购自 Avanti Polar Lipids(Alabaster, AL, USA); 色氨酸、硬脂酸购自Sigma-Aldrich(MO, USA); 花生四烯酸购自Cayman (Michigan, USA), 标样用水或乙腈稀释或溶解. 去离子水经 Milli-Q 系统(Millipore Co., USA)净化后使用.

# 2.2 样品的预处理

25 例宫颈癌患者和 22 例健康对照血清样本均由上海交通大学附属第一人民医院采集. 宫颈癌患者均通过宫颈组织活检并经病理检查确诊, 患者一经确诊收入院后, 即在第二天清晨空腹采取血清样本; 健康对照组均在清晨空腹采取血清样本. 样本储存于-80 冰箱中, 运输过程中以干冰作为冷链. 宫颈癌患者和健康对照的平均年龄分别为(44.8 ± 6.4)岁和(40.5 ± 6.7)岁. 分析时, 将血清从-80 冰箱中取出解冻, 在 180 μL 血清中加入 720 μL 乙腈和 100 μL 内标液(溶血磷脂酰胆碱(12:0/0:0), 4.5 ng/μL), 涡旋0.5 min, 离心 10 min(4 , 15000 g). 取上清液冻干, 冻干后的固体溶于样品溶液中(水/乙腈, 1/4).

#### 2.3 超高效液相色谱-质谱分析条件

液相色谱为美国 Waters 公司 Acquity 超高效液相色谱系统. 色谱柱为 BEH C18 柱(2.1~mm i.d.  $\times 50~\text{mm}$ ,  $1.7~\mu\text{m}$ ). 液相色谱条件: 流动相 A 为高纯水(0.1%的甲酸溶液), 流动相 B 为乙腈, 流速为 0.35~mL/min. 起始梯度为 2%乙腈并维持 0.5~min, 然后在 24~min 内乙腈线性增加至 100% 并维持 3.5~min, 然后乙腈含量

降回 2%, 平衡色谱柱2 min. 进样量为5 μL, 柱温 35 , 自动进样器温度维持在 4 .

质谱为 Waters 公司 Micromass Q-TOF 串联四极杆-飞行时间质谱仪,配有电喷雾离子源(ESI). 质谱采用正离子和负离子两种模式进行检测,正离子条件为:毛细管电压 3100 V、锥孔电压 30 V、离子源温度 120 、电喷雾气(脱溶剂气)温度 300 、电喷雾气流量(脱溶剂气流量)600 L/h、锥孔气流量 50 L/h、数据采集范围 m/z 100~920. 负离子模式条件为:毛细管电压 3100 V、锥孔电压 30 V、离子源温度 120、电喷雾气(脱溶剂气)温度 300 、电喷雾气流量(脱溶剂气)温度 300 、电喷雾气流量(脱溶剂气流量)600 L/h、锥孔气流量 50 L/h、数据采集范围 m/z 100~920.

# 2.4 潜在标志物筛选

图 1 描述了 LC-MS 分析和潜在标志物筛选的流 程. 采用Markerlynx软件(Waters, USA)对原始数据进 行峰检测和匹配. 每个样本的总峰面积归一化至 10000. 采用主成分分析(PCA)和 OPLS 两种方法对样 品进行模式识别分析. 软件采用SIMCA-P 11.0 (Umetrics AB, Umea°, Sweden), 数据进行 Pareto 缩放 (Scaling). 潜在标志物根据 VIP 值、VIP 值置信区间、 S图和载荷矩阵(loading plot)进行筛选. 用 SPSS 13.0(SPSS, USA)进行非参数检验、判别分析和 ROC 曲线制作. 非参数检验中采用两个独立样本检验的 方法. 聚类分析采用软件 HCE 1.0 (http://www.cs.umd. edu/hcil/hce/)[23], 在这一步中变量峰面积归一化为  $\frac{X-\bar{X}}{\sigma}$ , 皮尔森相关系数大于0.58的样品对被认为具 有相似性. z 值采用 Excel 2003(Microsoft, USA)计算, 公式为  $z = \frac{Xi - \overline{X}\text{control}}{SD\text{control}}$  [9],用 Matlab 7.0(Mathworks, USA)作图.

#### 3 结果与讨论

#### 3.1 标志物的筛选和发现

代谢组学通过研究血清中所有代谢产物在健康和疾病状态下的规律来揭示与疾病发生发展相关的代谢途径. 通过 LC-MS 技术分析血清代谢指纹可以获得海量的数据, 但是这些数据中包含噪声、缺失值

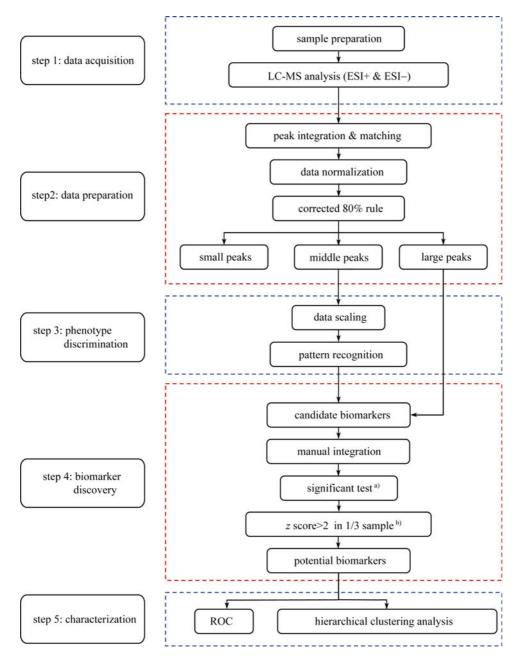


图 1 基于液相色谱质谱联用平台的血清代谢指纹分析和潜在标志物筛选的详细流程

- a) 本文中采用的是非参数检验; b) 通过 z 值检验的变量在至少 33%的宫颈癌样本中 z 值大于 2
- (0)以及少数大峰,导致潜在标志物筛选受到干扰. 因此,潜在标志物筛选过程通常包括:数据采集、数据 预处理、样品分型、潜在标志物发现等几个过程(图 1).
- (1) 代谢组数据采集. 样品预处理和分析技术 在我们之前的工作已有详细描述<sup>[4,6,8]</sup>. 由于血清代谢 物成分复杂, 碱性化合物更适合在质谱正离子模式 (ESI+)下分析, 酸性化合物更适合在质谱负离子模式 (ESI-)下分析, 两种数据采集模式下筛选出来的潜在

标志物会有所区别<sup>[6]</sup>. 为了得到尽可能多的信息,分别用两种离子模式对 25 例宫颈癌患者和 22 例健康对照血清样本进行检测. 图 2 是宫颈癌血清样本经LC-MS 分析得到的典型代谢指纹谱图.

(2) LC-MS 数据预处理. 代谢指纹图谱经过峰检测、峰匹配后得到峰表. 峰表由代谢物碎片的保留时间, 质量数和相应的峰面积组成. 用峰面积归一化的方法对比了47例血清样品的峰组分的相对含量.

1270

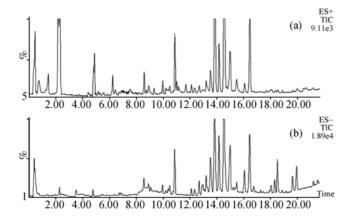


图 2 宫颈癌病人的 LC-MS 分析代谢指纹总离子流图(TIC) (a) 正离子模式下采集的 TIC 图; (b) 负离子模式下的 TIC 图

由于采用了ESI+和ESI-作为检测手段,总峰面积归一化的优点是不会因为某个采集模式响应比较高而在后续的筛选中产生歧视. 合并正负离子模式下得到的峰表,得到一个47样品×5201变量的峰表. 此峰表包含了大量的缺失值. 对缺失值最直观的解释是:在该样本中没有检测到这种物质,即低于检测限了. Bijlsma等人详细描述了产生缺失值的原因<sup>[24]</sup>. 目前常用的处理缺失值的方法有两种: 赋给一个值或者去掉缺失值过多的变量. 采用修正80%规则来去除缺失值<sup>[24]</sup>,即去除在某一组(宫颈癌组或健康对照组)中出现频率(非零值)低于80%的质谱离子. 缺失值去除后,得到一个47样品×958变量的数据矩阵.

将这 958 个变量划分为 3 个部分: 大峰, 小峰和中等强度的峰. 响应大的离子常常在模式识别中占据优势, 从而屏蔽掉一些确实在两组之间差异明显但是响应中等的离子<sup>[6]</sup>. 响应太小的离子由于分析方法的局限性通常测量误差比较大, 欠缺准确性. 即便这些离子能够经过筛选成为潜在标志物, 也要经过改进后续的分析方法来确认. 因此, 在模式识别这一步只关注中等强度的离子. 经归一化后每个样品的总峰面积为 10000, 平均每个样品可以检测到 800 个峰, 所以认为平均峰面积是 12.5(10000/800). 去掉在两组中(健康对照组和宫颈癌组)平均峰面积都大于125 的 19 个变量, 和都小于 1.25 的 283 个变量, 可得到一个 47 样品×656 变量的数据矩阵进行模式识别分析.

(3) 样品分型. 基于检测到的代谢物信息来建

立多变量统计模型,一方面可以表征样品之间的关系,另一方面可筛选潜在标志物. 血清中代谢物的浓度动态范围较大,高浓度的化合物容易掩盖低浓度化合物的变化. 虽然在前一步已经去掉了峰面积大于125 的变量,仍然需要采用一定的方式来减小变量响应值对模型的影响. 常用的方法有: 方差归一化,中心化,pareto方法等[25]. 方差归一化使得每个变量有相同权重,这个方法会放大由于仪器或者其他因素导致的系统偏差. 中心化是将每个变量减去该变量的平均值,对变量的大小没做任何变化,即响应大的变量有较大的权重,这也是要避免的. Pareto是介于方差归一化和中心化之间的缩放方式,即每个变量除以该变量方差的平方根,在一定程度上消除了变量响应的影响又不至于放大系统偏差. 根据代谢组学的特点,在进行样品分型前对数据采用 pareto 缩放.

样品分型常用模式识别方法. PCA 是一种无师监督模式识别方法,可以直观地在多维空间上描述样品间的差异. 从图 3 可知,在 PCA 模型中,两组之间的区别不太明显,模型仅解释了39.6%的原始数据,需要发展更加合适的方法来表征代谢指纹.

PLS-DA 是一种有师监督的模式识别方法,判别能力优于PCA. 用本实验的数据建立的PLS-DA 模型发生了过拟合,不能保证模型的有效性. 此时,应去除与分类信息无关的变量,使判断集中在这些与类别相关的变量上,从而提高分类准确性,消除过拟合,确保模型有效. 为此,采用OPLS方法[22]来区分宫颈癌组和健康对照组、筛选潜在标志物. 将80%的样本作为训练集建立OPLS 模型,剩下20%的样本作为预测集来评价模型的准确性(图4). 根据得分矩

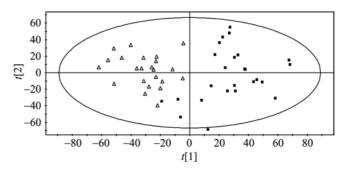


图 3 主成分分析(PCA)模型的得分矩阵,模型有 3 个主成分,用主成分 1 对主成分 2 作图,模型参数  $R^2X = 0.396$ 图中每个点代表一个样品, 代表宫颈癌组, 代表健康对照组

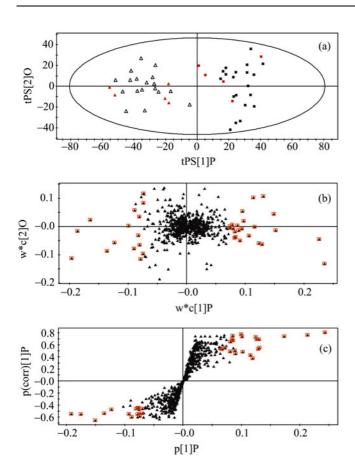


图 4 OPLS 模型,模型有 2 个主成分 模型参数  $R^2Y = 0.918$ , $Q^2 = 0.735$ 

(a) 得分矩阵图,主成分 1 对主成分 2 作图,80%的样品作为训练集建模,20%的样品作为预测集.图中 宫颈癌组, 宫颈癌预测组, 健康对照组, 健康对照预测组;(b) 载荷矩阵图,图中每个点代表一个变量,用 标记的变量为 VIP 值最大的前 40 个变量;(c) S 图(利用主成分 1 作图),图中每个点代表一个变量,S 图体现了变量与模型的相关性.用 标记的变量为 VIP 值最大的前 40 个变量,它们具有与模型良好的相关性

阵主成分1可以区分宫颈癌组与健康对照组,且预测集样本都得到了正确的预测,表明所建立的模型是稳健的.

(4) 潜在标志物的筛选和发现. (i) 根据模式识别模型的 VIP 值筛选潜在标志物. 在 OPLS 模型中提取 VIP值最大的前 40 个变量(VIP 值大于 1.95). 这 40 个变量在载荷矩阵图中距原点都有比较大的距离,表明它们对分类的贡献较大. S 图代表了变量在主成分 1 上的重要性和与主成分 1 的相关系数. 这 40 个变量在 S图上也表现出与主成分 1 良好的相关性. 然而,仅根据 VIP 来筛选潜在标志物是不够准确的. 一方面,还需要进行另外的显著性检验;另一方面,变

量面积都是软件自动积分生成的, 当峰面积较小或 者峰形不好的时候, 用软件积分常常存在一定的误差. (ii) 根据本文提出的方案筛选潜在标志物. 为了消除 积分过程中带来的误差,对 OPLS 模型中 VIP 值最大 的 40 个变量进行手动积分和内标校正. 在样品分型 之前因为峰面积过大排除了19个变量,同样对这些 变量进行手动积分和内标校正. 对这59个变量进行 非参数检验, 验证这些变量在宫颈癌组和健康对照 组中的统计差异. 排除非参数检验结果中 p 值大于 0.05的13个变量. 通过 z 值检验来研究每一个变量在 宫颈癌组样本中与在健康对照组样本中的偏离程度, 图 5 是通过非参数检验的 46 个变量的 z 图. 认为在至 少 33%的宫颈癌样本中 z 值绝对值大于 2 的变量是潜 在标志物. 因此, 15 个变量成为潜在标志物(表 1). 对 目标变量进行手动积分和内标校正, 并综合 VIP 值、 非参数检验和 z 值来筛选标志物. 得到的潜在标志物 更准确地反映了宫颈癌与健康对照组的代谢差异.

潜在标志物表征 这 15 个变量由 5 个强度大的变量和 10 个强度中等的变量组成. 其中 4 个变量在宫颈癌中上调, 11 个变量下调. 将 15 个变量作 ROC曲线, 曲线下面积(AUC)为 0.667~0.956. 在宫颈癌组上调的变量中 AUC 最大的是 neg\_13.74\_476.28, AUC为 0.784. 该变量在负离子模式下检测得到, 保留时间为 13.74 min, 质荷比为 m/z 476.28. 在宫颈癌组下调的变量中 AUC 最大的是 pos\_14.16\_497.53, AUC为 0.956. 该变量在正离子模式下检测得到, 保留时间为 14.16 min, 质荷比为 m/z 497.53, 其定性结果是溶血磷脂酰胆碱(Lysophosphatidylcholine, LPC) (16:0/0:0). 可见经过筛选后的标志物具有很好的诊断前景.

# 3.2 潜在标志物的结构鉴定和生物学解释

只有明确筛选出标志物的分子结构才能进一步研究其功能. 潜在标志物的结构推导主要包括: 精确分子量测定、多级质谱碎片推导结构、数据库比对、保留时间比对以及相关性分析等步骤<sup>[17]</sup>. 常用的代谢组学数据库有<sup>[26]</sup>: NIST、the Golm Metabolite Database、MassBank、METLIN、HMDB等. 经过结构推导,这14个变量归结为9个化合物(表1). LPC (0:0/16:0)与LPC(16:0/0:0)的保留时间不同,为同分

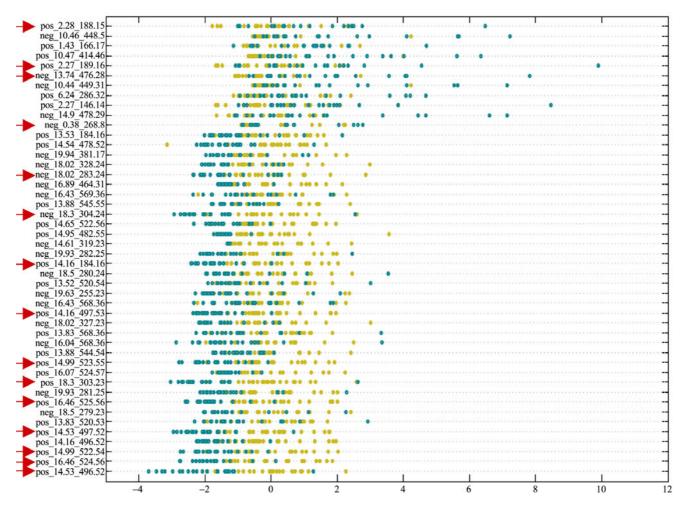


图 5 z 值图, 以非参数检验 P < 0.05 的 46 个变量作图

图中每一行代表一个变量,每个点代表一个样品在这个变量上的 z 值, 代表宫颈癌组, 代表健康对照组. 符号意义: 以变量  $pos_2.28_188.15$  为例,指这个变量是在正离子模式(pos)下分析得到,保留时间是 2.28,荷质比是 188.15. 图中, neg 代表这个变量在负离子模式下分析得到. 图中标有红色箭头的变量至少在 1/3 的宫颈癌样本中 z 值绝对值大于 2,被认为是潜在标志物

异构体.这两个同分异构体的鉴别主要依据 LPC (0:0/16:0)分子中有端羟基,极性比 LPC(16:0/0:0)大,在反相色谱中的保留相对较弱.因此,保留时间 14.16 min 的峰对应LPC(0:0/16:0),保留时间14.52 min 的峰对应 LPC(16:0/0:0).同时二级质谱碎片 m/z 104 丰度大于 m/z 125 的是 LPC(0:0/16:0), m/z 104 丰度小于 m/z 125 的是 LPC(0:0/16:0).

人体中,溶血磷脂酰胆碱在溶血磷脂酶D的作用下转变成溶血磷脂酸(LPA)<sup>[28]</sup>, LPA可以促进细胞增殖,被普遍认为与癌症的发生发展有关<sup>[29]</sup>. 此外,在胶质母细胞癌、甲状腺癌和肾癌中都发现溶血磷脂酶D过表达<sup>[28]</sup>. 因此,推测溶血磷脂酰胆碱(0:0/16:0、

16:0/0:0、18:1/0:0 和 18:0/0:0)等在宫颈癌中下调与溶血磷脂酶 D 过表达有关.

花生四烯酸是合成前列腺素的前体物质,其中的限速酶是环氧合酶(COX). 哺乳动物的 COX有两种异构酶: COX-1和 COX-2. COX-2在宫颈癌中的表达上调与宫颈癌的发生有关<sup>[30]</sup>, COX-2过表达生成前列腺素等生物活性物质从而促进了肿瘤细胞的生长代谢. 硬脂酸在人体内可以合成油酸,而花生四烯酸正是由油酸转变而来. 因此推测, 硬脂酸和花生四烯酸含量下调可能是与宫颈癌中 COX-2 过表达有关.

#### 4 结论

本文提出了一个从血清代谢指纹中筛选潜在标

表 1 潜在标志物

ESI ioniza- tion mode <sup>a)</sup>	t <sub>R</sub> (min)	Mass	average peak area b)		aamnaund	neleted methodes
			cancer group	health control	- compound	related pathway
ESI-	0.38	268.80	1.13±1.01	0.68±0.70	unknown	unknown
ESI+	2.27	189.16	$7.79 \pm 4.72$	4.81±1.83	Tryptophan <sup>c)</sup>	Tryptophan metabolism
ESI+	2.28	188.15	96.12±42.00	65.84±22.49		
ESI-	13.74	476.28	$7.17 \pm 3.42$	$4.82\pm1.81$	unknown	unknown
ESI+	14.16	184.16	$2.92 \pm 1.03$	6.01±1.73	Lysophosphatidylcholine(0:0/16:0)	Glycerophospholipid metabolism
ESI+	14.16	497.53	$3.72\pm1.38$	$7.98\pm2.46$		
ESI+	14.53	496.52	123.14±29.67	$176.35\pm24.97$	Lysophosphatidylcholine(16:0/0:0)	Glycerophospholipid metabolism
ESI+	14.53	497.52	$26.31 \pm 8.00$	$45.60\pm9.60$		
ESI+	14.99	523.55	$5.36\pm2.35$	$10.72\pm3.14$	Lysophosphatidylcholine(18:1/0:0)	Glycerophospholipid metabolism
ESI+	14.99	522.54	$20.36\pm9.47$	42.14±12.85		
ESI+	16.46	524.56	$37.06\pm19.22$	$85.19\pm25.96$	Lysophosphatidylcholine(18:0/0:0) c)	Glycerophospholipid metabolism
ESI+	16.46	525.56	$8.56\pm4.56$	$20.35 \pm 6.64$		
ESI-	18.02	283.24	1.34±0.85	2.56±0.94	stearic acid c)	Biosynthesis of saturated and un- saturated fatty acids
ESI-	18.3	303.23	$5.62\pm4.35$	$13.00\pm3.83$	Arachidonic acid c)	Arachidonic acid metabolism Fc epsilon RI signaling pathway Long-term depression GnRH signaling pathway
ESI-	18.3	304.24	1.34±1.07	3.20±0.98		

a) ESI+代表变量在正离子模式下检测得到, ESI-代表变量在负离子模式下检测得到; b) 峰面积由手动积分得到, 用内标峰面积校正; c) 化合物已得到标样验证, 代谢通路来自 KEGG 和 HMDB 数据库

志物的方法,该方法系统地描述了从数据采集、数据集预处理、模式识别到潜在标志物筛选的过程.通过OPLS 方法建立的模型表明宫颈癌病人与健康人群的代谢存在明显差异,同时该模型能对未知样品进行正确分类. 15个变量(9个化合物)经过筛选被确认为潜在标志物,其中7个化合物被鉴定为:色氨酸、硬脂酸、花生四烯酸和4个溶血磷脂酰胆碱(0:0/16:0、

16:0/0:0、18:1/0:0、18:0/0:0). HPV是导致宫颈癌 发生的诱因<sup>[31]</sup>, 癌组织形成和扩散的过程中必然会 导致一些代谢过程发生改变, 本文从代谢组学的角度发现宫颈癌病人中花生四烯酸和溶血磷脂酰胆碱 的代谢发生了异常. 色氨酸、*m/z* 268、*m/z* 476 等化 合物在宫颈癌中上调, 这些化合物有希望成为临床标志物.

致谢 本工作得到国家高技术研究发展计划(编号: 2006AA02Z342)和蛋白质科学重大研究计划(编号: 2007CB914701)项目资助, 特此一并致谢.

# 参考文献

- 1 Nicholson J K. Global systems biology, personalized medicine and molecular epidemiology. Mol Syst Biol, 2006, 2: article 52
- 2 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. Xenobiotica, 1999, 29(11): 1181 1189
- 3 Seo D, Ginsburg G. S. Genomic medicine: bringing biomarkers to clinical medicine. Curr Opin Chem Biol, 2005, 9(4): 381 386
- 4 Yang J, Xu G W, Zheng Y F, Kong H W, Pang T, Lv S, Yang Q. Diagnosis of liver cancer using HPLC-based metabonomics avoiding false-positive result from hepatitis and hepatocirrhosis diseases. J Chromatogr B, 2004, 813(1-2): 59 65
- 5 De Vos R C H, Moco S, Lommen A, Keurentjes J J B, Bino R J, Hall R D. Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry. Nat Protoc, 2007, 2(4): 778 791
- 6 Yin P Y, Zhao X J, Li Q R, Wang J S, Li J S, Xu G W. Metabonomics study of intestinal fistulas based on ultraperformance liquid chromatography coupled with Q-TOF mass spectrometry (UPLC/Q-TOF MS). J Proteome Res, 2006, 5(9): 2135 2143

1274

- 7 Kind T, Tolstikov V, Fiehn O, Weiss R H. A comprehensive urinary metabolomic approach for identifying kidney cancer. Anal Biochem, 2007, 363(2): 185 195
- 8 Chen J, Wang W Z, Lv S, Yin P Y, Zhao X J, Lu X, Zhang F X, Xu G W. Metabonomics study of liver cancer based on ultra performance liquid chromatography coupled to mass spectrometry with HILIC and RPLC separations. Anal Chim Acta, 2009, doi:10.1016/j.aca.2009.03.039
- 9 Sreekumar A, Poisson L M, Rajendiran T M, Khan A P, Cao Q, Yu J D, Laxman B, Mehra R, Lonigro R J, Li Y, Nyati M K, Ahsan A, Kalyana-Sundaram S, Han B, Cao X H, Byun J, Omenn G S, Ghosh D, Pennathur S, Alexander D C, Berger A, Shuster J R, Wei J T, Varambally S, Beecher C, Chinnaiyan A M. Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. Nature, 2009, 457(7231): 910 914
- 10 Cubbon S, Bradbury T, Wilson J, Thomas-Oates J. Hydrophilic interaction chromatography for mass spectrometric metabonomic studies of urine. Anal Chem, 2007, 79(23): 8911 8918
- 11 Idborg H, Zamani L, Edlund P O, Schuppe-Koistinen I, Jacobsson S P. Metabolic fingerprinting of rat urine by LC/MS Part 1.

  Analysis by hydrophilic interaction liquid chromatography-electrospray ionization mass spectrometry. J Chromatogr B, 2005, 828(1-2):
  9 13
- Jonsson P, Gullberg J, Nordstrom A, Kusano M, Kowalczyk M, Sjostrom M, Moritz T. A strategy for identifying differences in large series of metabolomic samples analyzed by GC/MS. Anal Chem, 2004, 76(6): 1738 1745
- 13 Soga T, Ohashi Y, Ueno Y, Naraoka H, Tomita M, Nishioka T. Quantitative metabolome analysis using capillary electrophoresis mass spectrometry. J Proteome Res, 2003, 2(5): 488 494
- 14 Wang Y, Wang J S, Yao M, Zhao X J, Fritsche J, Schmitt-Kopplin P, Cai Z W, Wan D F, Lu X, Yang S L, Gu J R, Haring H U, Schleicher E D, Lehmann R, Xu G W. Metabonomics study on the effects of the ginsenoside Rg3 in a beta-cyclodextrin-based formulation on tumor-bearing rats by a fully automatic hydrophilic interaction/reversed-phase column-switching HPLC-ESI-MS approach. Anal Chem, 2008, 80(12): 4680 4688
- 15 Li X, Xu Z, Lu X, Yang X, Yin P, Kong H, Yu Y, Xu G W. Comprehensive two-dimensional gas chromatography/time-of-flight mass spectrometry for metabonomics: Biomarker discovery for diabetes mellitus. Anal Chim Acta, 2009, 633(2): 257 262
- Hagan S O, Dunn W B, Knowles J D, Broadhurst D, Williams R, Ashworth J J, Cameron M, Kell D B. Closed-loop, multiobjective optimization of two-dimensional gas chromatography/mass spectrometry for serum metabolomics. Anal Chem, 2007, 79(2): 464 476
- 17 Chen J, Zhao X J, Fritsche J, Yin P Y, Schmitt-Kopplin P, Wang W Z, Lu X, Haring H U, Schleicher E D, Lehmann R, Xu G W. Practical approach for the identification and isomer elucidation of biomarkers detected in a metabonomic study for the discovery of individuals at risk for diabetes by integrating the chromatographic and mass spectrometric information. Anal Chem, 2008, 80(4): 1280 1289
- 18 Griffin J L, Shockcor J P. Metabolic profiles of cancer cells. Nat Rev Cancer, 2004, 4(7): 551 561
- 19 Trygg J, Holmes E, Lundstedt T. Chemometrics in metabonomics. J Proteome Res, 2007, 6(2): 469 479
- 20 Sawada Y, Akiyama K, Sakata A, Kuwahara A, Otsuki H, Sakurai T, Saito K, Hirai M Y. Widely targeted metabolomics based on large-scale MS/MS data for elucidating metabolite accumulation patterns in plants. Plant Cell Physiol, 2009, 50(1): 37 47
- 21 卢纹岱. SPSS for Windows 统计分析. 第三版. 北京: 电子工业出版社, 2006. 448 450
- 22 Gabrielsson J, Jonsson H, Airiau C, Schmidt B, Escott R, Trygg J. OPLS methodology for analysis of pre-processing effects on spectroscopic data. Chemom Intell Lab Syst, 2006, 84(1-2): 153 158
- 23 Seo J, Shneiderman B. Interactively exploring hierarchical clustering result. IEEE Comput, 2002, 35(7): 80 86
- 24 Bijlsma S, Bobeldijk L, Verheij E R, Ramaker R, Kochhar S, Macdonald I A, van Ommen B, Smilde A K. Large-scale human metabolomics studies: A strategy for data (pre-) processing and validation. Anal Chem, 2006, 78(2): 567 574
- 25 User's Guide to SIMCA-P, SIMCA-P+ version 11.0, 2005
- 26 Go E P. Database resources in metabolomics: An overview. J Neuroimmune Pharmacol, DOI 10.1007/s11481-009-9157-3 (2009)
- 27 Han X L, Gross R W. Structural determination of lysophospholipid regioisomers by electrospray ionization tandem mass spectrometry. J Am Chem Soc, 1996, 118(2): 451 457
- 28 Zhao Z W, Xiao Y J, Elson P, Tan H Y, Plummer S J, Berk M, Aung P P, Lavery I C, Achkar J P, Li L, Casey G, Xu Y. Plasma lyso-phosphatidylcholine levels: Potential biomarkers for colorectal cance. J Clin Oncol, 2007, 25(19): 2696 2701
- 29 Mills G B, Moolenaar W H. The emerging role of lysophosphatidic acid in cancer. Nat Rev Cancer, 2003, 3(8): 582 591
- 30 Farley J, Uyehara C, Hashiro G, Belnap C, Birrer M, Salminen E. Cyclooxygenase-2 expression predicts recurrence of cervical dysplasia following loop electrosurgical excision procedure. Gynecol Oncol, 2004, 92(2): 596 602
- 31 Walboomers J M M, Jacobs M V, Manos M M, Bosch F X, Kummer J A, Shah K V, Snijders P J F, Peto J, Meijer C, Munoz N. Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. J Pathol, 1999, 189(1): 12 19

# A metabonomics approach based on liquid chromatography-mass spectrometry: from metabolic profiling to potential biomarker

CHEN Jing<sup>1</sup> SHAN YuanHong<sup>1</sup> YAN Qin<sup>2</sup> ZHAO SuMin<sup>1</sup> YIN PeiYuan<sup>1</sup> ZHAO XinJie<sup>1</sup> LU Xin<sup>1</sup> XI XiaoWei<sup>2</sup> Wan XiaoPing<sup>2</sup> & XU GuoWang<sup>1\*</sup>

1 CAS Key Lab of Separation Science for Analytical Chemistry, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian, 116023, China

2 Department of Obstetrics and Gynaecology, The First People's Hospital, Shanghai Jiaotong University, Shanghai, 200080, China

**Abstract:** A LC-MS based metabonomics approach for potential biomarkers screening is described. Metabolic profiles were acquired by LC-MS technology, and potential biomarkers were filtered by multiple statistical methods. Data pretreatment includes data normalization and scaling, corrected 80% rule, and dataset division. According to metabolic patterns, cervical cancer and health group were separated by OPLS. Potential biomarker screening was performed according to *VIP* value, significant test and z-score. As a result, fifteen variables were considered as potential biomarker, and their AUC were 0.667 0.956. The fifteen variables correspond to nine compounds, seven of which were identified as tryptophan, stearic acid, arachidonic acid, and lysoPC (0:0/16:0, 16:0/0:0, 18:1/0:0, 8:0/0:0). It can be concluded that abnormal metabolism of arachidonic acid and lysoPC happened in cervical cancer.

**Keywords:** metabonomics, metabolomics, LC-MS, cervical cancer, cancer biomarker