• 方法学研究 •

# 变步长积分：代谢组学数据积分预处理新方法

李晓红[1,2]，卢 姗[3]，虞明阳[4]，蔡 爽[1,2*]

1. 中国医科大学附属第一医院 药学部，辽宁 沈阳 110001
2. 中国医科大学药学院，辽宁 沈阳 110001
3. 中国石油天然气集团公司中心医院 药学部，河北 廊坊 065000
4. 安捷伦科技（中国）有限公司，北京 100000

**摘 要：目的** 以核磁共振（NMR）分析不同年龄 SD 大鼠尿样中内源性代谢物的改变实验数据为基础，提出新的积分方法。**方法** 通过在原有固定步长积分的基础上引入步长变动区间，使数据积分区间可以根据峰的位置进行一定范围的调整，形成变步长积分方法。以固定步长和变步长积分方法，对实际实验数据进行比较研究。**结果** 变步长积分方法既能够增强样品聚类能力，能够减少差异代谢物指认缺失现象的发生。**结论** 变步长积分方法克服了固定步长积分方法存在的不足，解决了固定步长积分方法不能够同时兼顾图谱分辨率和减少由于环境引起的化学位移变化的矛盾。

**关键词：**代谢组学；核磁共振；变步长积分；多元数据分析；代谢物指认

**中图分类号：**R969.1 **文献标志码：**A **文章编号：**1674 - 6376 (2015) 01 - 0023 - 06

**DOI:** 10.7501/j.issn.1674-6376.2015.01.004

# Variable step integrating algorithm: A new algorithm for integral pretreatment of metabolomic data

LI Xiao-hong[1,2], LU Shan[3], YU Ming-yang[4], CAI Shuang[1,2]

1. Department of Pharmacy, The First Hospital of China Medical University, Shenyang 110001, China
2. College of Pharmacy, China Medical University, Shenyang 110001, China
3. Department of Pharmacy, Central Hospital of China National Petroleum Corporation, Langfang 065000, China
4. Agilent Technologies Co., Ltd., Beijing 100000, China

**Abstract: Objective** To present a new integrating algorithm for NMR analysis based on the data of endogenous metabolites in urine which changed by the age of SD rats. **Methods** This new algorithm—variable step integrating algorithm (VSIA) is by adding variable step range on fixed step, thus the integrating range can be adjusted by different peak widths. The integral method of VSIA included four procedures such as denoising of sample data, recognition of peak-valley point, domain integral based on input parameters, and output data of integral domain. In this paper, we compared VSIA and fixed step algorithm, using pattern recognitions, principal component analysis (PCA), and partial least squares-discriminate analysis (PLS-DA), to analyze the integral data from the NMR spectra. **Results** VSIA can significantly enhance the aggregative of different group samples, and can also reduce the loss of different metabolites caused by naive data preprocessing. By integrating effectively according to the signal peak position, VSIA can both enhance the spectrum resolution and reduce the chemical shift changes caused by environment, thus the loss of fixed step algorithm could be made up. **Conclusion** This study suggests that VSIA could be applied to metabonomic studies, and also could be extended to the other multi-dimensional data processing analysis.

**Key words:** metabonomics; nuclear magnetic resonance; variable step integrating algorithm; multivariate data analysis; metabolites identification

**收稿日期：**2014-01-05

**基金项目：**辽宁省自然科学基金项目（20072113）；辽宁省科学技术项目（2012225107）；沈阳市科技计划项目（F12-277-1-79）

**作者简介：**李晓红（1982—），女，沈阳人，助教，硕士，研究方向为药物分析及医院药学。Tel: 15640205479 E-mail: shelling23951037@163.com

***通信作者** 蔡 爽，女，博士，教授，硕士生导师，主要从事药物分析及医院药学。Tel: 15840418555 E-mail: caishuang1972@126.com

代谢组学被定义为定量分析与时间相关的由于生命体病态生理刺激或者基因改变后产生的代谢物，是系统生物学中自上而下研究方法的一个分支[1]，正活跃于公共或者私人的研究机构中，尤其是在植物代谢组研究、药物开发以及临床医学等领域[2-9]。其分析方法有 GC-MS、（U）HPLC-MS、NMR、（U）HPLC-NMR-MS 等多种手段[10-12]。其中 NMR 分析方法，由于对于样品分析的无损伤性和无偏向性成为代谢组学分析的重要手段。在应用这些分析方法对经实验设计的从动物或者植物中得到的生物样品进行分析后，得到的原始数据一般要经过数据预处理、模式识别、生物样品定性/定量、代谢异常途径指认等过程，才能达到基本研究目的。

在代谢组学数据预处理阶段，从图谱中得到的数据需经过积分、归一化、峰对齐等步骤，才能将由于不同仪器或者样品受外界影响甚至操作所产生的谱图差异减小到最低，以减少由于非实验设计因素所产生的差异对实验研究结果造成的影响。其中对结果影响最大的是积分步骤。积分是指将 NMR 或者 LC（GC）-MS 所得到的数据，根据一定的区间（如 NMR 一般为 $\delta$0.04）或者指定一定积分数目（如 300 个区间），将区间内的或者所有数据根据指定数目，将峰强度积分，积分后的个数作为下一阶段数据处理中的自变量个数，峰强度作为自变量的数值。这样做的目的是防止由于样品受到 pH、测试时温度等外界影响对于仪器的响应位移所发生的改变，如 NMR 中某个物质由于 pH 的微小差异引起化学位移的偏移；GC-MS 或者（U）HPLC-MS 中保留时间的微小变化，减少变量数量便于统计分析。但是，如果样品间的差异过大，现有的积分方法就不会达到所期望的效果，单纯的增加区间范围或者减少积分个数会降低谱图的分辨率，将两个峰积分为一个整体，从而导致代谢物质定性缺失；而减少区间范围或者增多积分个数虽然提高了谱图分辨率，但是却没有起到积分的应用目的[13]。本文以 NMR 分析 SD 大鼠随年龄增长尿样中内源性代谢物的改变实验数据为基础，提出了新的积分方法。此方法是将积分区间在一定范围内根据峰的位置变动（如 $\delta$0.04±0.02）而改变，既能够不降低谱图的分辨率，又能够合理地将化学偏移影响减少到最低，完整地对信号峰进行积分，确保差异代谢物指认的准确性。

## 1 材料与方法

### 1.1 仪器与试剂

$^1$H-NMR 样品测试仪器为 Bruker DRX 600 NMR（Reinstetten，Germany），Norell 标准核磁管（ST500—75 mm o.d.，Norell，Inc.，Landisville，NJ，USA），三（三甲基硅烷）磷酸酯（TMSP，98 atom% D，Cambridge Isotope Laboratories，Inc）作为 NMR 化学位移 $\delta$ 0.0 校正与内标试剂，重水（$D_2O$，99.8 atom % D，Schweres Wefsser，Norell，Inc），磷酸一氢钠与磷酸二氢钠（分析纯，汕头陇西化工厂）。

### 1.2 动物实验与样品收集

成熟雄性 SD 大鼠（购自沈阳药科大学动物实验中心，动物许可证号 SYXK( 辽) 2003-0013）8 只，初始体质量为 200 g。在正式实验前动物适应实验室与代谢笼 2 周。实验动物在室温 22±2 ℃，12 h 昼夜交替，相对湿度为 45%～65%条件下饲养。自由饮水与饮食。分别于第 1 周、第 12 周收集大鼠 24 h 尿样（4 ℃左右），记为“UZ”和“UB”。3 000 r/min 离心 10 min，分取上清液，于−80 ℃保存。

### 1.3 样品制备与数据采集

取−80 ℃冷藏尿样，在室温（20 ℃）下自然融化，于 13 000 r/min 离心 10 min 后，上清液用 0.45 μm 滤膜滤过。取续滤液 400 μL 与重水磷酸缓冲液（PBS 缓冲液调节 pH＝7.4，50 μL $D_2O$，10 mmol/L TMSP）250 μL，混合。取 600 μL 涡旋后样品溶液转移到核磁管中。$^1$H 测试频率为 600.129 MHz，298.16 K。水峰抑制脉冲应用 Bruker “noesypr1d”脉冲序列（RD-90°-$t_1$-90°-$t_m$-90°-acquire FID），延时时间 $t_1$＝3 s，混合时间 $t_m$＝100 ms。每个样品扫描 64 次，数据采集点为 131，072（128 K），谱图宽度为 12 019.230 Hz，采集时间为 2.73 s。

### 1.4 NMR 数据模式识别

所有尿样 $^1$H 核磁数据使用 Topspin（version 2.1pl1，Bruker Biospin，Rheinstetten，Germany）手动进行基线与相位校正，化学位移校正 TMSP 于 $\delta$ 0.0。化学位移区间 $\delta$ 4.7～5.0、$\delta$ 5.5～6.0 因水峰抑制和尿素峰的影响而去除。积分区间 $\delta$ 0.2～4.7、$\delta$ 5.0～5.5 和 $\delta$ 6.0～10.0 使用软件 ProMetab Suite version 1.4，应用两种不同的积分方法进行积分：①以 $\delta$0.04 为间距固定步长积分方法积分；②应用变步长积分（adaptive binning）方法进行积分。积分后数据由 ProMetab Suite 根据每个图谱峰总和进行

归一化，导出为 TXT 文本。数据矩阵用 Pareto scaling（$1/\sqrt{\text{standard deviation}}$）方法做数据预处理。主成分分析（PCA）和偏最小二乘法-判别分析（PLS-DA）使用 SIMCA-P 软件包（version 11.5，UMETRICS AB，Sweden）进行分析。PCA 分析用以可视化显示样品分类情况，溢出点判别，以及趋势判断，PLS-DA 分析应用于聚类后生物标志物发现。

### 1.5 算法

变步长积分方法包含以下步骤：①样本数据滤噪；②峰谷点辨识；③根据输入参数进行区间积分；④输出积分区间数值。

**1.5.1** 样本数据滤噪 在没有信号的区间，如 $\delta$ −1.0～3.0，计算此区间内噪音信号的均值与 SD 值，根据 Bruker AMIX 报道，图谱噪音 noisylevel＝mean（$\delta$ range）＋noisyfactor×SD，其中 noisyfactor 一般为 3.5～5.0。根据图谱噪音水平对图谱进行滤噪，选择高于噪音信号值的数据点为信号数据，或者采用小波分析的方法对全部谱图进行滤噪。

**1.5.2** 峰谷点辨识 将所有谱图数据加和的均值作为标准谱图，用于寻找峰谷点。这样能够满足不同组分差异峰的峰谷点辨识，全面而又不丢失所有的峰数据。spectra_matrix_ref 为输入标准图谱数值矩阵。对 YS 求两次差分，算法如下：

SlopeSign = diff (spectra_matrix_ref) ＞ 0；

SlopeSignChange = diff (slopeSign) ＞ 0；%过滤掉局部最低点微小的数据波动

h = find (slopeSignChange) + 1；%标记差分后最低点峰谷索引地址值

peakvalleyfind ( : , 2) = spectra_matrix（h）；%输出峰谷点数值

**1.5.3** 变积分方法积分 初始化参数，max_i_bin = bin_stepsize.×（1＋mean_range）；min_i_bin = bin_stepsize.×（1－mean_range），其中 max_i_bin 为设置区间中最大值（上边界），min_i_bin 为设置区间中最小值（下边界），bin_stepsize 为积分区间（一般为 $\delta$0.04），mean_range 为区间变动百分比（一般为 50%），matrix 为计算得出峰谷点数值。

具体算法实现如下（Appendix I:MATLAB 算法实现）：

```
For 峰谷点数值长度，将数值分段处理
  res = 上段积分后残差与新数值段之和
  If res ＞ max_i_bin
      将 res 根据 bin_stepsize 分段
      res 作为 max_i_bin 分段后残差
      If res 小于 min_i_bin 大于 0
          将 res 加入新数值矩阵段
          res = 0
      Elseif res 大于 min_i_bin，小于 bin_stepsize
          保存 res 值
      End
  Elseif res ＞ = min_i_bin & res＜ = max_i_bin
      res = 0；
  Else % res＜min_i_bin
      保存 res 值
  End
End
If res～ = 0 % res 最后剩余值不大于最小值
    新矩阵加入 res
End
new_matrix = 新矩阵为分段区间边界点数值
NMR 样品根据分段区间边界点，将区间内的峰强度积分。
```

**1.5.4** 输出积分区间数值

## 2 结果与讨论

### 2.1 模式识别分析比较

以 $\delta$0.04 为区间固定步长积分方法和变步长积分方法积分后的数据经过 PCA、PLS-DA 方法进行模式识别分析，得到结果见表 1。

**表 1 两种积分方法数据经模式识别分析的结果比较**

**Table 1 Comparison on PCA and PLS-DA results between two methods**

| 方法 | 模式识别 | A | R2*X* | R2*Y* | Q2 |
|---|---|---|---|---|---|
| 固定步长积分 | PCA | 1 | 0.492 | — | 0.384 |
| | | 2 | 0.152 | — | 0.124 |
| | PLS-DA | 1 | 0.492 | 0.98 | 0.968 |
| | | 2 | 0.086 6 | 0.013 5 | −0.062 1 |
| 变步长积分 | PCA | 1 | 0.559 | — | 0.478 |
| | | 2 | 0.121 | — | 0.093 5 |
| | PLS-DA | 1 | 0.559 | 0.989 | 0.983 |
| | | 2 | 0.048 | 0.009 03 | 0.007 3 |

表中结果表明，变步长积分方法在 PCA、PLS-DA 对于 *X*（自变量）的解释能力（R2*X* 值）要比固定步长积分方法更强，对于 PLS-DA 建立预测模型的预测能力（对因变量 *Y* 的解释能力，R2*Y* 值）也要更好。这说明变步长积分方法能够有效地判断差异信号峰所引起的变化。

对于两种不同的积分方法采用 PCA 分析后计

算 $R$ 距离来考察其组内聚合、组间分离能力，其中 $R$ 定义为 $R$＝Distance（Between Groups）/∑[Distance（Within Group）]。$R$ 值反映了样品聚类分散性的大小，$R$ 值越大说明其样品聚类性越好，也就是说组内间距小而组间间距大。通过比较计算后组内距离和组间距离以及 $R$ 值可知，变步长积分方法能够有效的减少组内距离，增加组间距离，有效地将两个不同的类进行聚类（表 2）。

## 2.2　误差原因分析

数据预处理的目的是减少主观因素引起的差异，放大试验设计所要考察的差异。数据预处理阶段所引入的不准确的差异数据会直接影响数据处理后的差异代谢物指认，导致由差异代谢物在代谢网络中的不全面而引起的解释错误。两种积分方法得到的潜在生物标志物是有所不同的（表 3），这种不同很可能是因为数据预处理阶段人为因素引入了误差。

**表 2　组内距离和组间距离以及 *R* 值**

**Table 2　Distances of samples within group and between groups and *R* value**

| 方法 | ∑ (Distance) UB | ∑ (Distance) UZ | Distance | *R* (UB/UZ) |
|---|---|---|---|---|
| 固定步长积分 | 55.754 209 | 39.368 279 | 185.689 | 3.330 492 9/4.716 716 2 |
| 变步长积分 | 49.335 239 | 36.950 897 | 248.98 | 5.046 696 9/6.738 131 4 |

**表 3　两种积分方法 PLS-DA 中的 VIP**

**Table 3　VIP of PLS-DA between two methods**

| 序号 | VIP | | | |
|---|---|---|---|---|
| | 固定步长积分 | | 变步长积分 | |
| 1 | 3.28～3.24 (−) | 2.92～2.88 (+) | 3.278 5～3.250 2 (−) | 2.922 9～2.892 8 (+) |
| 2 | 3.04～3.00 (−) | 1.96～1.92 (+) | 5.417 9～5.378 8 (−) | 4.008 9～3.984 9 (+) |
| 3 | 4.08～4.04 (−) | 4.00～3.96 (+) | 3.062 8～3.033 3 (−) | 5.378 8～5.338 8 (+) |
| 4 | 2.48～2.44 (−) | 3.64～3.60 (+) | 2.467 4～2.441 2 (−) | 1.959 1～1.911 4 (+) |
| 5 | 5.40～5.36 (−) | 7.52～7.48 (+) | 7.859 3～7.820 2 (−) | 3.641 6～3.612 5 (+) |

(+)表示正相关；(−)表示负相关

(+) means positive relationship, (−) means negative relationship

下面以 $\delta$ 4.008 9～3.96 区间为例分析造成这种误差的原因。综合固定步长积分方法与变步长积分方法两方面，将这个区间分为两部分 $\delta$ 4.008 9～3.984 9 和 $\delta$ 3.984 9～3.96。从图 1 中可以看到，在固定步长积分方法中，$\delta$ 4.00～3.96 为正相关（图 1A）；而在变步长积分方法中 $\delta$ 4.008 9～3.984 9 为正相关，$\delta$ 3.984 9～3.964 4 为负相关（图 1B）。从图 2 中可以看出，在 $\delta$ 4.008 9～3.96 区间，“UB”

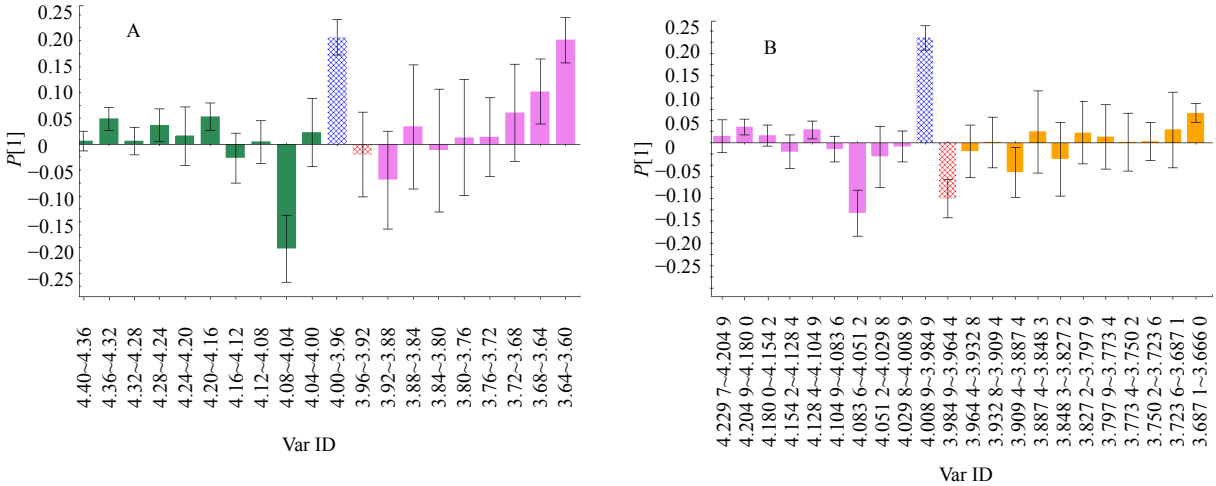


**图 1　固定步长（A）和变步长（B）积分方法的载荷图**

**Fig. 1　Loading plots of normal binning method (A) and adaptive binning method (B)**

组样品存在两个单峰（s），化学位移分别为 δ 3.99，δ 3.97；“UZ”组样品在 δ 3.975 处存在一个双峰（dd）。固定步长积分方法不能够有效地区分这两组不同的峰的影响，而统一将其认为是一个弱的正相关（由于存在负相关的部分抵消，正相关性减弱）；变步长积分方法却能够有效地发现这两组差异峰的存在，正确地判断相关性，从而减少由于数据预处理所引起的潜在生物标志物丢失，减少了由于正负相关性的抵消所产生的相关性值的减小，以及使显著的潜在生物标志物成为非显著生物标志物的现象。

由于固定步长积分方式没有考虑到峰位置，所以会将差异峰分裂为两部分，减弱差异峰对于分组影响的权重。以 δ 7.68 处单峰（s）为例（图 3）。δ7.68 处单峰在固定步长积分方法中被分为两部分分别积分，而变步长积分方法能够识别峰的位置将信号峰作为一个整体进行积分。PCA 载荷图中显示了这个差异对数据分析的影响（图 4）。

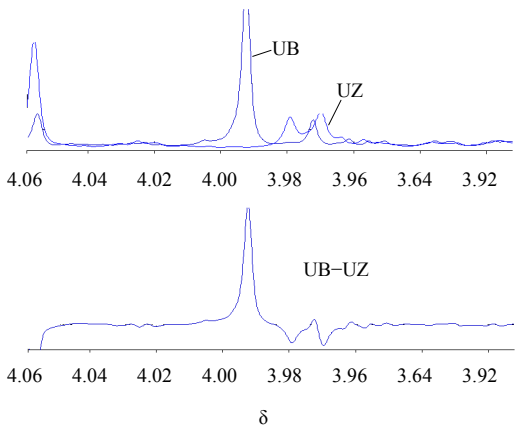


**图 2　在 δ 4.008 9～3.96 两组样品谱图的区别**

**Fig. 2　Differences between two spectra from two different groups at region δ 4.008 9—3.96**

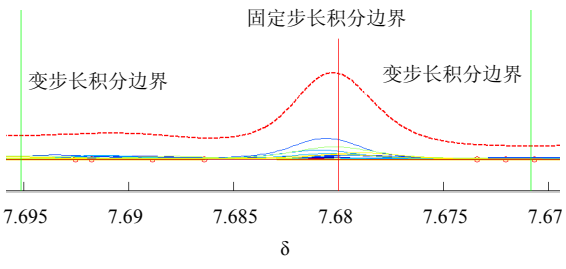


**图 3　固定步长积分（步长 δ0.04）和变步长积分（δ0.04 ± 0.02）方法在 δ7.68 处的边界**

**Fig. 3　Marked borders of normal binning method (fix range δ0.04) and adaptive binning method (δ0.04 ± 0.02) at region δ7.68**

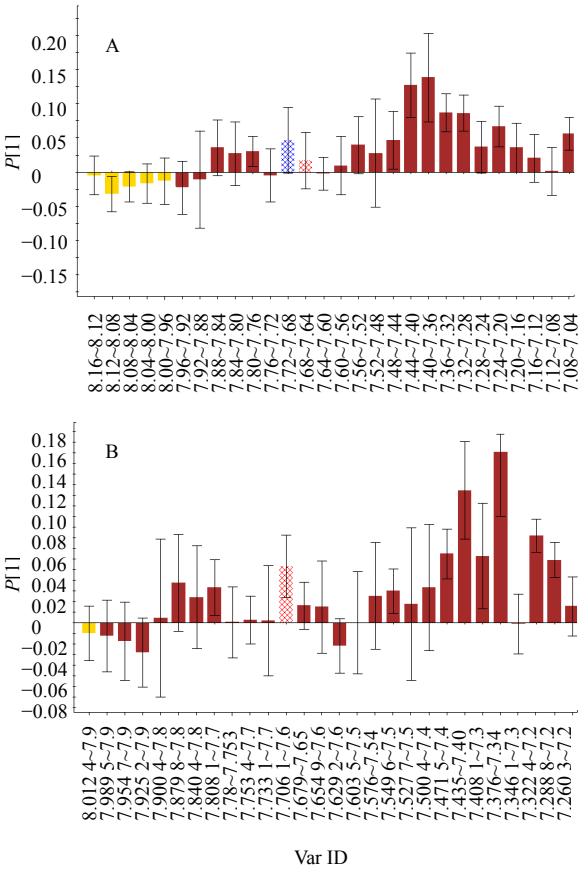


**图 4　固定步长（A）和可变步长（B）积分方法载荷图**

**Fig. 4　Loading plots of normal binning method (A) and adaptive binning method (B)**

## 3　结论

综上分析，变步长积分方法能够有效地根据信号峰的位置对其进行积分，减少了不同信号峰被合并积分或者同一信号峰被拆分积分的情况。由于这一特点，变步长积分方法能够有效地增强判断样品聚类能力同时减少由于积分方法产生的差异代谢物丢失，解决固定步长积分方法不能够同时兼顾图谱分辨率和减少由于环境引起的化学位移变化的矛盾。

变步长积分方法也可以扩展应用到三维数据中（如 GC-MS、HPLC-MS 采集的数据），增强数据预处理的可靠性，避免由于误差引起的差异代谢物、生物标志物指认错误。

致谢：感谢李发美实验室各位老师和研究生郑姝宁、唐静同学在实验过程中给予的帮助。

### 参考文献

[1] Keun H C. Metabonomic modeling of drug toxicity [J]. *Pharmacol The*, 2006, 109(1-2): 92-106.

[2] Schripsema J. Application of NMR in plant metabolomics: techniques, problems and prospects [J]. *Phytochem Anal*, 2010, 21(1): 14-21.

[3] Ala-Korpela M. Potential role of body fluid 1H NMR metabonomics as a prognostic and diagnostic tool [J]. *Expert Rev*. *Mol Diagn*, 2007, 7(6): 761-773.

[4] Aich P, Babiuk L A, Potter A A, *et al*. Biomarkers for prediction of bovine respiratory disease outcome [J]. *OMICS*, 2009, 13(3): 199-209.

[5] Ala-Korpela M. Critical evaluation of $^{1}$H NMR metabonomics of serum as a methodology for disease risk assessment and diagnostics [J]. *Clin Chem Lab Med*, 2008, 46(1): 27-42.

[6] Lindon J C, Holmes E, Nicholson J K. Metabonomics in pharmaceutical R&D [J]. *FEBS J*, 2007, 274(5): 1140-1151.

[7] Coen M, Holmes E, Lindon J C, *et al*. NMR-based metabolic profiling and metabonomic approaches to problems in molecular toxicology [J]. *Chem Res Toxicol*, 2008, 21(1): 9-27.

[8] Duarte I F, Diaz S O, Gil A M. NMR metabolomics of human blood and urine in disease research [J]. *J Pharm Biomed Anal*, 2014, 93C: 17-26.

[9] Beger R D, Sun J, Schnackenberg L K. Metabolomics approaches for discovering biomarkers of drug-induced hepatotoxicity and nephrotoxicity [J]. *Toxicol Appl Pharmacol*, 2010, 243(2): 154-166.

[10] Zhang M, Deng M, Ma J, *et al*. An evaluation of acute hydrogen sulfide poisoning in rats through serum metabolomics based on gas chromatography-mass spectrometry [J]. *Chem Pharm Bull*, 2014, 62(6): 505-507.

[11] Lee D K, Lim D K, Um J A, *et al*. Evaluation of four different analytical tools to determine the regional origin of Gastrodia elata and Rehmannia glutinosa on the basis of metabolomics study [J]. *Mol*, 2014, 19(5): 6294-6308.

[12] Sarker S D, Nahar L. Hyphenated techniques and their applications in natural products analysis [J]. *Methods Mol Biol*, 2012, 864: 301-340.

[13] Jacob D, Deborde C, Moing A. An efficient spectra processing method for metabolite identification from 1H-NMR metabolomics data [J]. *Anal Bioanal Chem*, 2013, 405(15): 5049-5061.