[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G06F 17/00 (2006.01)
G01N 24/08 (2006.01)

[12] 发明专利说明书

专利号 ZL 200610057865.7

[45] 授权公告日 2009 年 11 月 18 日

[11] 授权公告号 CN 100561463C

[22] 申请日 2006.3.1
[21] 申请号 200610057865.7
[30] 优先权
[32] 2005.3.24 [33] EP [31] 05006476.5
[73] 专利权人 F·霍夫曼-拉·罗奇股份有限公司
地址 瑞士巴塞尔
[72] 发明人 弗兰克·迪特勒 阿尔弗雷德·罗斯格茨·施洛特贝克 汉斯·森
[56] 参考文献
EP1275011A1 2003.1.15
WO03087834A2 2003.10.23
WO02057989A2 2002.7.25
审查员 冯慧萍

[74] 专利代理机构 北京集佳知识产权代理有限公司
代理人 顾晋伟 刘继富

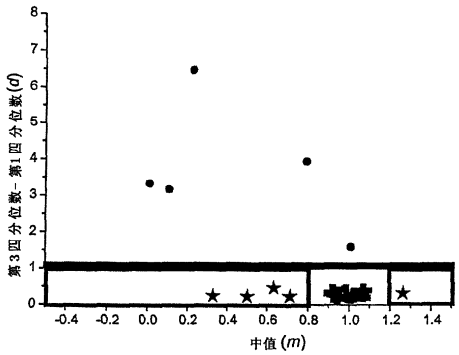
权利要求书 2 页 说明书 23 页 附图 6 页

[54] 发明名称

自动删除系列波谱中的异常值的方法

[57] 摘要

处理系列波谱具体是 NMR 波谱的方法，包括下列步骤：a)选择主波谱范围；b)记录所述主波谱范围内的多个主波谱；c)获得所述主波谱范围内的基准主波谱；d)对每一个所述主波谱进行所述主波谱除以所述基准主波谱的分格除法以获得对应的系列谱商；e)对至少一个所述主波谱计算来自对应系列谱商的相关系列统计量；和 f)对至少一个所述系列统计量进行异常值检测测试。



1. 一种自动删除由波谱法获得的一系列波谱中的异常值的方法，所述波谱法产生含有强度或面积与组分浓度成比例的信号或峰的波谱，所述方法包括下列步骤：

a) 选择主波谱范围；

b) 记录所述主波谱范围内的多个主波谱；

c) 获得所述主波谱范围内的基准主波谱；

d) 对每一个所述主波谱进行所述主波谱除以所述基准主波谱的分格除法以获得对应的系列谱商； 和

e) 对至少一个所述主波谱计算来自对应系列谱商的相关系列统计量，对应于主波谱的所述系列统计量包括其谱商的中值和由其谱商的第三四分位数减去其谱商的第一四分位数而得到的四分位数间差值；

f) 对至少一个所述系列统计量进行异常值检测测试，以确定与其相关的主波谱是否是异常值；和

g) 如果确定所述主波谱是异常值，则删除所述主波谱。

2. 根据权利要求 1 所述的方法，其中所述波谱是核磁共振波谱。

3. 根据权利要求 1 所述的方法，其中所述异常值检测测试包括确定所述四分位数间差值是否超出预定的阈值宽度。

4. 根据权利要求 3 所述的方法，其中所述阈值宽度由取自整个系列主波谱的所述四分位数间差值的分布来确定。

5. 根据权利要求 1 所述的方法，其中所述异常值检测测试包括确定所述谱商的中值与一常数的差是否大于预定的阈值商偏差。

6. 根据权利要求 5 所述的方法，其中所述阈值商偏差由取自整个系列主波谱的所述谱商中值的分布来确定。

7. 根据权利要求 1 所述的方法，还包括在权利要求 1 的步骤 a)之后和在权利要求 1 的步骤 f)之前进行下列步骤：

a) 选择与所述主波谱范围不重叠的辅助波谱范围；

b) 记录与所述主波谱范围中的每一个所述主波谱相关的所述辅助波谱范围中的

辅助波谱；

c) 获得所述辅助波谱范围中的基准辅助波谱；

d) 对每一个所述辅助波谱进行辅助波谱除以所述基准辅助波谱的分格除法以获得对应的系列辅助谱商；和

e) 对每一个所述辅助波谱计算来自对应的系列辅助谱商的相关系列统计量，对应于辅助波谱的所述系列统计量包括其谱商的中值和由其谱商的第三四分位数减去其谱商的第一四分位数而得到的四分位数间差值；

其中所述异常值检测测试包括比较对应于主波谱的所述系列统计量和对应于与所述主波谱相关的所述辅助波谱的所述系列统计量。

8. 根据权利要求 1 所述的方法，其中对每一个所述主波谱实施权利要求 1 的步骤 e) 以获得全部系列的统计量，从该全部系列的统计量得出系列全局统计量，并且其中对所述系列全局统计量进行所述异常值检测测试。

9. 根据权利要求 1~8 中任意一项所述的方法，其中在实施所述分格除法之前，使每一个所述主波谱或辅助波谱经历归一化过程。

10. 根据权利要求 9 所述的方法，其中对任意一个所述主波谱或辅助波谱的所述归一化过程包括下列步骤：

a) 对所述任意一个所述主波谱或辅助波谱进行预处理以获得预处理波谱；

b) 计算所述预处理波谱的积分强度；和

c) 用与所述积分强度成反比的归一化系数乘以所述预处理波谱。

11. 根据权利要求 1-8 中任意一项所述的方法，其中所述基准主或辅助波谱分别作为在对应的主波谱或辅助波谱范围内记录的多个空白或对照波谱的中值而得到。

12. 根据权利要求 1-8 中任意一项所述的方法，其中由所述主波谱或辅助波谱的子系列获得所述基准主或辅助波谱。

自动删除系列波谱中的异常值的方法

技术领域

本发明涉及自动删除系列波谱中的异常值的方法。

背景技术

混合物的分析和比较是分析化学的重要课题，尤其是在环境科学、生物学、食品工业和过程化学中。例如，在代谢组学领域中，利用确定的波谱方法例如液相色谱/质谱法（LC-MS）或核磁共振（NMR）波谱法获得的波谱来表征动物和人的生物流体（biofluids）。通常需要分析和比较整组波谱，例如由一系列样品获得的众多单个波谱。为了区分与样品总浓度变化相关的效应（例如通过稀释样品，改变样品的全部被分析物）及影响样品组成的效应（混合物中组分的相对浓度），需要使用所谓的归一化方法。在不同的实验条件下获得各种样品的数据时，也需要归一化。

至今，在代谢组学研究中－例如研究尿样－的常用方法是使给定 NMR 波谱中的信号归一化，从而获得所述波谱的恒定总积分。这意味着系列波谱中每一个 NMR 波谱按比例绘制成相同的曲线下预定面积。潜在的假设是每一个波谱的积分主要是总尿浓度的函数。假设由于代谢组响应引起的各个被分析物浓度的变化相对于总尿浓度的变化而言相对较小，则总尿浓度的变化分别影响全部波谱和波谱的所述预定面积。但是，代谢组研究中的动物可以排泄极大量的物质例如糖类，其可以在波谱中占主要地位并因此实质性地影响归一化。另外，随尿一起排泄的药物相关化合物也可以通过它们的相应峰的积分来影响归一化，并且因此对波谱的总积分有显著贡献。在比较混合物的其它分析应用中出现同样的问题，其中相对高浓度的未知污染物的出现可显著影响波谱的总积分或者波谱的所述预定面积。

在 US 2003/0111596 A1 中公开了利用质谱法研究化学混合物组分的定量方法。如具体在所述文件的 0040 段中所述，该已知方法基于：

a) 从多个化学样品获得一组样品波谱，每一个波谱包含具有峰强度的峰；

b) 选择基准波谱；

c) 至于所要归一化的所述样品波谱的每一个，对所有峰或对峰总数的一部分计算样品波谱和基准波谱之间的强度比；

d) 用从所述强度比计算得到的归一化系数乘以样品波谱。

上述方法基于这样的事实，即在许多实际环境下，所述强度比的大多数将基本相等，代表在样品和基准波谱之间浓度不变化的组分。然后利用非参数测量由所述强度比计算归一化系数。优选地，归一化系数选择为所述强度比的中值。

如在 US 2003/0111596 A1 的 0031 段中进一步指出的，该已知归一化方法可应用于产生含信号（或峰）的波谱的任何类型的波谱方法或波谱测量方法中，所述信号（或峰）的强度或面积与组分浓度成比例。具体地，它因此可应用于 NMR 波谱学。

然而，在 US 2003/0111596 A1 中公开的方法没有解决识别和消除所谓“异常值”的问题，这些异常值尤其可以是来自人为因素或人为扭曲的单个信号，但也可以是具有某种偏差的整个波谱，例如由于在采集期间的技术故障。这种问题在大量波谱的定量分析中是特别重要的，例如在代谢组研究中。

发明内容

本发明的主要目的是克服现有定量处理波谱、尤其是处理 NMR 波谱方法的局限和缺点。

通过本发明的方法来实现上述和其它目的。

根据本发明的一个实施方案，提供一种自动删除由波谱法获得的一系列波谱中的异常值的方法，所述波谱法产生含有强度或面积与组分浓度成比例的信号或峰的波谱，例如 NMR 波谱，所述方法包括下列步骤：

a) 选择主波谱范围；

b) 记录所述主波谱范围内的多个主波谱；

c) 获得所述主波谱范围内的基准主波谱；

d) 对每一个所述主波谱进行所述主波谱除以所述基准主波谱的分格除法（bin-wise division）以获得对应的系列谱商（spectral quotients）； 和

e) 对至少一个所述主波谱计算来自对应系列谱商的相关系列统计量，对应于主波谱的所述系列统计量包括其谱商的中值和由其谱商的第三四分位数减去其谱商的第一四分位数而得到的四分位数间差值；

f) 对至少一个所述系列统计量进行异常值检测测试，以确定与其相关的主波谱是否是异常值；和

g) 如果确定所述主波谱是异常值，则删除所述主波谱。

虽然本发明方法适用于NMR波谱学并通过在NMR波谱学中的应用来举例说明，但是它还可应用于产生包含信号（或峰）的波谱的其它类型波谱，所述信号（或峰）的强度或面积与组分浓度成比例，例如质谱或各种类型的光谱。

在本发明内容中，视情况而定，术语“波谱范围”将用作波谱中的单一区域或多个不连接的区域。具体地，感兴趣的波谱范围可以是多个波谱区域，每一个波谱区域含有一定数量的信号峰。如本发明的某些实施方案所采用的，定语“主要的” 在本文中用来与“辅助的”进行对比。具体地，“主波谱范围”将被用于表示含有某些样品的一个或多个相关信号峰的波谱范围。应该注意，通常在含有所述主波谱范围和辅助波谱范围（如果可应用的话）的整个波谱范围中获得波谱；而且，整个波谱范围可含有不用于进一步分析的其它波谱范围。例如，代谢组研究中的 ^{1}H NMR 波谱通常记录-8－+14ppm 范围的总波谱范围，从中选择分别由 1-4.5 和 6-9.5ppm 的两个区域构成主波谱范围。如今基本上在所有波谱学应用中，获得的是数字化形式的波谱数据。例如，通常一维NMR波谱可以作为系列强度值而获得，每一个强度值与特定波谱频道（spectral channel）或“格”（bin）相关。因此，术语“格”也可以指强度值的总和。因而，第二波谱除以第一波谱的“分格除法”在本文中应理解为：采用第一波谱的某些格中强度值，用第二波谱的相同格中的强度值除以它，将除法结果分配到所得的系列谱商的相同格中并对感兴趣的波谱范围中所有格重复该程序。应该理解，如果可得的波谱是非数字化的，即模拟态，则根据本发明仍然可以实施分格除法。这仅仅需要利用现有信号处理技术中已知的适当分格（binning）程序首先将所有模拟波谱转换成数字形式。

术语“统计量”是指任意数，其大小表示所关心的某些量的量值，例如关联强度、

偏差量、差值大小和分布形状。例子包括平均值、方差、相关系数等。

术语“异常值”是指任意实体（entity）－具体是信号峰、包含几个单峰或者甚至是整组波谱的其谱图或部分谱图－其对于给定变量的分值基本偏离预定数值范围。因此，“异常值检测测试”应该理解为目的在于测定一个给定实体相对于给定测试标准是否应该被认为是异常值的任意类型的程序。

根据本发明的方法可以容易地在波谱数据的自动处理中实施。如下文将进一步举例说明的，将用于处理系列 NMR 波谱的实际类型的异常值检测测试可适用于所关心的任何应用。本方法可以嵌入整个程序中，其中例如证实为异常值的谱图应该在进一步的分析中被放弃。

以下说明本发明的有利的实施方案。

在根据本发明的一个实施方案中，所述异常值检测测试包括测定所述四分位数间差值是否超出了预定的阈值宽度。该四分位数间的差值大表示谱商的宽分布，并因此意味着横切瞬时波谱的强度相对于基准波谱表现出实质性差异并且不仅仅是简单的比例缩放行为。根据本发明的一个实施方案，阈值宽度由得自整个系列主波谱的所述四分位数间差值的分布来确定。换言之，首先获得在给定系列的全部波谱中发现的四分位数间差值的总印象（impression）从而随后限定用于异常值检测测试的阈值宽度。

在根据本发明的一个实施方案中，所述异常值检测测试包括确定谱商的所述中值与一常数的差是否大于预定阈值商偏差。该大偏差值表示瞬时波谱的总强度相对于给定标准值的偏差。根据本发明的一个实施方案，阈值商偏差由得自整个系列主波谱的谱商的所述中值的分布来确定。换言之，首先获得在给定系列的全部波谱中得到的谱商中值的总印象从而限定用于异常值检测测试的阈值偏差。

根据本发明的一个有利的实施方案，本方法还包括下列步骤：

a) 选择与所述主波谱范围不重叠的辅助波谱范围；

b) 记录与所述主波谱范围中的每一个所述主波谱相关的所述辅助波谱范围中的辅助波谱；

c) 获得所述辅助波谱范围中的基准辅助波谱；

d) 对每一个所述辅助波谱进行辅助波谱除以所述基准辅助波谱的分格除法以获

得对应的系列辅助谱商；和

e) 对每一个所述辅助波谱计算来自对应的系列辅助谱商的相关系列统计量，对应于辅助波谱的所述系列统计量包括其谱商的中值和由其谱商的第三四分位数减去其谱商的第一四分位数而得到的四分位数间差值；

其中所述异常值检测测试包括比较对应于主波谱的所述系列统计量和对应于与所述主波谱相关的所述辅助波谱的所述系列统计量。

具体地，可以在预期可能出现问题或人为现象的区域中选择辅助波谱范围，而在已知不易受问题和人为现象影响的区域中选择主波谱范围。然后，辅助波谱可以用作一种诊断工具。如前所述，本文中所讨论的任意波谱范围可以由单一波谱区域构成，或者由两个或甚至更多的不相连的波谱区域构成。

根据本发明的一个有利的实施方案，对每一个主波谱实施步骤 1e)以获得全部系列的统计量，从该全部系列的统计量中导出系列全局（global）统计量，其中对所述系列全局统计量实施所述异常值检测测试。换言之，利用来自整个系列波谱的统计信息来实施异常值检测测试，这应该使测试尽可能客观。

原则上，可利用未校正的强度波谱数据来实施上文中讨论的方法。但是，在大多数应用中，优选应用在本发明的一个实施方案中限定的方法，根据该方法，在实施分格除法之前，使每一个所述主波谱和每一个所述辅助波谱（如果可应用的话）经历归一化程序。有利的是，这根据本发明的一个实施方案来完成，其中对任意一个所述主波谱或辅助波谱的所述归一化程序包括下列步骤：

a) 对所述波谱进行预处理以获得预处理波谱；

b) 计算所述预处理波谱的积分强度；和

c) 将所述预处理波谱乘以与所述积分强度成反比的归一化系数。

这种预处理程序将通常取决于波谱的类型和质量。在噪声数据的情况下，其可以包括平滑或滤波程序；具体地，其可以包括基线校正或减法程序，这将适合于具有基本平坦或缓慢变化背景成分的波谱。具体对于 NMR 波谱而言，预处理程序可包括填零、定相、应用窗口函数和线性预测。其它的预处理步骤可包括波谱的积分和微分。

实际上，归一化系数包含比例常数，该常数确保任意归一化的波谱具有预定的积

分强度，例如 1 或 100 或任意其它的适当值。

对于将用于处理系列波谱的基准波谱存在几种选择。例如，它可以是计算基准波谱、得自数据库或理论波谱的基准波谱。根据本发明的一个实施方案，基准主或辅助波谱分别作为在对应主或辅助波谱范围中记录的多个空白或对照波谱的中值而获得。作为替代方案，根据本发明的一个实施方案，基准主或辅助波谱可以由所述主或辅助波谱的子系列获得。这种子系列可以仅由单一波谱或多个波谱构成；在由多个波谱构成时，例如，基准波谱可以作为所述多个波谱的中值或均值而获得。最后，应该注意的是，术语“子系列”应该理解为包括子系列等同于整个系列波谱的情况。

附图说明

结合附图参考本发明各种实施方案的下列描述，本发明的上述和其它特征和目的以及实现它们的方法将变得更加清楚并且本发明本身将得到更好地理解，其中：

图 1 表示关于 208 变量（格）代谢系列研究的四个样品相对于相同研究的基准样品的谱商分布；

图 2 表示典型代谢系列研究的第 3 四分位数减去第 1 四分位数的差值（所谓四分位数之间得差值（d））－中值（m）的图；方形表示未偏离的波谱，圆点表示显著损伤的波谱（没有样品、不良接收器获取、错误定相...），而星号表示局部不规则的波谱（尖峰、药物相关化合物、异常量的代谢物...）；为检测异常值设置 m（0.8 和 1.2）和 d(1)的固定阈值；

图 3 表示研究具体阈值而不是固定绝对阈值的来自图 2 数据的图；对于所有阈值，n 设置为 3；

图 4 表示用于系统变化的分格（binned）形式的“金（golden）”^{1}H NMR 波谱；

图 5 表示由于系统变化而不同的四组系列波谱：（A）样品浓度的系统变化；（B）单峰强度的系统变化：（C）样品浓度和单峰强度的系统同时变化；（D）10 格的块（blocks）的系统变化；

图 6 表示在不同时间点（16 小时到 72 小时和 16 小时到 168 小时）的两个动物（30 号和 28 号动物）的原始波谱，同时在右手侧示出图像放大区的波谱；

图 7 表示通过对图 5 的四个数据系列的不同归一化方法而实现的恢复；恢复率为 1 意味着归一化程序将不同样品中具有相同相对浓度的分析物重新按比例缩放至相同的归一化浓度；

图 8 表示利用不同基准波谱的商归一化结果；

图 9 表示 30 号和 28 号动物在不同时间点的含主要信号的波谱区，所述信号没有由于代谢组学响应（1.44-1.84ppm）而变化；顶行表示商归一化波谱，中间行表示积分归一化波谱，底行表示向量长度归一化波谱；

图 10 表示在环孢菌素研究中（左侧）和在罗格列酮（rosiglitazone）（右侧）研究中图解测定异常值的四分位数间差值 d－中值 m 的图；

图 11 表示确认为图 10 右侧图中的异常值的一些波谱：（A）水共振的不良抑制；（B）负基线；（C）空白样品和（D）伴随技术问题例如气泡而获得的波谱；和

图 12 表示对于对应 6.04ppm 化学位移的信号格的谱商对样品编号的图，所述谱商通过将样品信号除以全部商的中值而获得；该值大大偏离 1 表示在对应样品中抑制水共振存在问题。

具体实施方式

下文描述实施本发明方法所需的背景和技术，最主要的是包括各种归一化方法的讨论。虽然这些方法可用于各种类型的波谱学，但是以下讨论将用 NMR 波谱特别是 ^{1}H-NMR 波谱来举例说明。

通常，感兴趣的 NMR 波谱将根据它的作为化学位移 δ 函数的强度 I 来描述。但是，将假设波谱能够以分格数字化形式获得并因此计为 I(i)，其中 i 是指示给定格的变化指数。“信号” I(i)可以理解为通过对第 i 个格所跨越的波谱范围的信号进行积分而获得的结果。在多数情况下，将等距离分格。

1.波谱归一化

本部分描述波谱归一化常用的三种技术，即：积分归一化、肌酸酐归一化和向量长度归一化。接着，导出商归一化。前三种归一化技术可以表达为下列具体的通式：

$$I(i)=\frac{I^{old}(i)}{\sum_{k}\int_{j_k^l}^{j_k^u}(I(x))^n\,dx} \qquad (1)$$

其中 $I^{old}(i)$和 I(i)分别是归一化之前和之后的波谱强度，k 是用于归一化的波谱区指数，j_k^l 和 j_k^u 分别是波谱区 k 的上界和下界，求强度 I(x)的幂次 n 对该波谱区 k 的积分。

1.1 积分归一化

至于积分归一化，假设波谱的积分主要是样品浓度的函数。尿的线性浓度系列应该产生线性系列的对应波谱积分。假设单个分析物的各自浓度变化的影响与尿的总浓度变化相比很小。

积分归一化程序用波谱的积分或部分波谱的积分来除每一个波谱。因此，通式（1）中的幂次 n 取 1。在代谢组学 NMR 测量领域中，通常的方法是选择实际包含两个波谱区的波谱范围，即一个是 9.98-5.98ppm 和另一个是 4.50-0.22ppm。而且，通常进一步将每一个波谱乘以系数 100，因此最终每一个波谱的总积分为 100。

积分归一化的问题是信号的相互依存性。很明显，任意单一强信号将导致归一化程序按比例减小所有其它的信号，因此导致混合物中所有分析物浓度明显减少。

1.2 肌酸酐归一化

对于人和动物尿液的检测，常用的程序是利用肌酸酐的浓度使分析物的浓度和波谱归一化。潜在的假设是进入到尿中的肌酸酐排泄物恒定。对于归一化存在两种可能：可以通过临床化学方法外部测定或通过 NMR 波谱中肌酸酐相关信号的积分内部测定肌酸酐的水平。后一种方法可以表达为特定情况的积分归一化[4]。根据通式（1），采用两个积分区（对应 3.04 和 4.05ppm 的肌酸酐峰）和幂次 1。

但是，肌酸酐归一化的实际应用面临技术和生物性困难。如果通过 NMR 波谱测定肌酸酐浓度，则具有重叠峰的代谢物会干扰肌酸酐浓度的测定（例如在 3.04ppm 的肌酸）。利用 ^{1}H NMR 波谱测定肌酸酐的第二个难题在于肌酸酐在 4.05ppm 附近的

化学位移取决于样品的 pH 值。因此，必需将选峰算法或更大范围的波谱用来进行归一化。

肌酸酐归一化的生物性挑战是由于代谢组学响应而引起的肌酸酐浓度变化，这在一些研究中已被发现。在归一化时，通常还不知道由于代谢组学响应而引起的肌酸酐水平的可能增大。因而，基于肌酸酐的归一化在代谢组学研究中并不常用并且在本文中将不再讨论。因此，肌酸酐峰将被用于研究子系列波谱的肌酸酐水平和用各种方法获得的归一化系数之间的相关性，在该子系列波谱中，已知在浓度水平和肌酸酐之间存在严格的相关性。

1.3 向量长度归一化

在许多科学领域中应用的归一化技术基于将波谱视为向量。换言之，采用序列强度值 I(i)来表示相关向量的分量。进一步假设通过对应样品的浓度来测定这种向量长度；样品的组分将因此确定向量的方向。因此，通过将向量长度设置为 1 来完成不同浓度的调整。注意，这等价于将通式（1）的幂次 n 设置为 2。与积分归一化类似，由于普通向量长度的计算，使得波谱中所有峰相互影响。

1.4 商归一化

商归一化依赖于下面的假设：单个分析物的浓度变化仅影响部分波谱，而样品的总浓度变化影响全部波谱。与使用积分归一化不同，计算给定波谱和基准波谱之间的最可几商，然后将所述商用于获得归一化或换算系数。

对于该方法，实施波谱的分格除法和预先选择基准波谱以获得系列谱商。应该理解，该方法在一些适当选择的波谱范围中进行。理想地，谱商的分布将是窄的；在两个不同浓度的等同样品的限制中，各个谱商将随浓度比而不同。

可以由几种方法测定的最可几谱商表示样品和基准之间的浓度比。图 1 的图面 A 表示代谢组学研究的样品 R14r30h+000 和同一研究的基准样品之间谱商的分布。发现该样品比基准样品略微更浓缩，因为最可几商（近似为直方图的最大值）位于约 1.1。另一方面，图 1 的图面 B 表示同一研究的对比性稀释样品的结果，结果发现与

基准样品相比具有约 0.6 的最可几商。

但是，如果单个分析物由于代谢组学变化而变化，则仅有对应波谱的某些部分将受影响。结果是谱商的分布更宽。图 1 的图面 C 表示样品的谱商分布，该样品由于强代谢组学变化而具有跨越其波谱的极端强度变化。由于分别具有增大和减小强度的部分波谱，这产生了宽的分布，超过值 10 的极端谱商是由该特定样品中排出的极端量的葡萄糖所导致。因此，由于尿的总浓度是基本不变的，因此最可几谱商仍接近 1。相反，图 1 的图面 D 表示具有两种效果的样品直方图，即，由于强代谢组学响应和由于增加尿排泄的样品稀释而引起的特定变化。因此，谱商的分布被扩大并且移动到更低值。

商归一化的一个重要方面是最可几谱商的确定，这是由于这将被用作换算系数。在前文中，最可几谱商通过取谱商直方图的最大值来确定。但是，分布最大值的精确位置取决于分格宽度。因此，目前所提出的直方图的图解分析不能够认为是限定最优谱商的稳定和普遍的方法。采用过粗的分格导致相当大的量化误差（例如 1 和 1.1 的商之间的差对应 10%的量化误差），而太细分格导致直方图没有清晰的最大值，如图 1 的图面 D 所示。实际方法是通过利用商的中值来近似得到最可几商。中值法的优点在于需要用于直方图的商非离散分组（分格），这是任意的。中值法允许非常精细地调整波谱而没有极端值明显影响调整的危害。

用于计算谱商的基准波谱可以是由“金”基准样品获得的单一波谱。作为替代方案，可以使用几个波谱的中值或平均波谱。基准波谱的类型的影响在 4.1.3 部分中讨论。发现基准波谱应该尽可能具有代表性。因此，推荐计算基准波谱作为多个没有给药样品（对照和给药前样品）的中值波谱。

积分归一化（通常积分到 100）可以在商归一化之前进行。这使得利用不同波谱仪测量的比较研究简单化，所述不同波谱仪产生不同绝对标度值的波谱。通常，商归一化方法将由此包括下列步骤：

A1.实施积分归一化（通常采用 100 的恒定积分）。

A2.选择或计算基准波谱（最优方法：计算未给药样品的中值波谱）。

A3.进行样品波谱除以基准波谱的分格除法以获得对应的系列谱商。

A4.计算谱商的中值。

A5.通过将样品波谱乘以所述中值的倒数而重新按比例缩放样品波谱。

如前文所述，步骤 A1 是任选的，但在大多数情况下将是有利的。

2.异常值检测

2.1 背景

作为商归一化基础的方法学得以进一步发展以提供异常值的自动检测。在自动样品制备、测量和数据处理过程中，可以发生许多影响所得数据质量的事情。例如，技术问题如检测器的增益不正确、水共振的抑制不充分、在波谱两个边界处的尖峰，或在数据处理过程中的问题，如基准不正确、基线校正错误和定相不适当均可在测量 NMR 波谱时发生。另外，尿不存在或浓度太低的样品应该被自动检测出来。

在代谢组学研究中（或例如用于产品批次质量控制的所得 NMR 波谱），大多数分析物具有稳定的相对浓度，并因而大多数信号峰将相应地表现出来。相反，有缺陷的波谱通常具有不同的整体形状，这导致了与原波谱相比异常宽的谱商分布。这种特征可用于检测异常值和限定范围以判断对化学组成类似的样品所进行的测量的总体质量。

2.2 异常值检测方法

异常值的离线检测方法（本文中“离线”是指在实验完成后进行的检测）包括下列步骤：

B1.对来自研究的整个系列波谱实施积分归一化。

B2.计算基准波谱（未给药样品的中值波谱）。

B3 进行样品波谱除以基准波谱的分格除法以获得对应的系列谱商。

B4.对每一个波谱，计算谱商的中值（下文中表示为 m）和谱商的第一和第三四分位数之间的差（下文中称为“四分位数间差值”并表示为 d）。中值 m 可用于瞬时波谱的商归一化（实际上利用 m 的倒数作为换算系数）。

B5.进行异常值检测测试。例如，四分位数间差值 d 是瞬时波谱的形状与基准波

谱形状差异程度的量度。因此，异常值的标准可以是超出预选的阈值宽度的 d 值。下文中进一步讨论异常值检测测试。

如上文所述，步骤 B1 是任选的，但是在大多数情况下将是有利的。通常，对测量系列的所有波谱实施步骤 B4，但这不是严格的要求，即，也可以仅对子系列波谱实施步骤 B4。

由于仅有步骤 B2 需要修改，因此上述程序的在线版本的修改是相当温和的，所述在线版本可用于实时控制波谱检测和处理。实际方法是在系列测量开始时测量给药前样品或对照样品的基准系列。然后基于该系列波谱来计算基准中值波谱。采用稳定中值允许一定百分比的有缺陷数据存在基准系列中。上述方法的所有后续步骤均基于波谱/波谱基准（spectrum basis）并由此很好的适合于算法的在线版本。

2.3 发现异常值和损坏数据

由于采集期间的技术故障导致的大部分异常值将得到整体形状不同的波谱（例如，任意形状的线和曲线，而不是真实波谱）。因此，不同形状导致相应波谱的谱商分布非常宽。结果是异常大的四分位数间差值 d。虽然是任意的，但是 d 的固定值 1 被证明对于代谢组学研究的 NMR 波谱是合理的阈值宽度。超出该阈值宽度的 d 值通常表示损坏波谱或影响大部分波谱的问题。除了利用固定和任意阈值的 d 外，也可以利用具有中值 md 的和 d 的四分位数之间差值 dd 的 d 的分布。然后可以根据 $d_{临界}=md+n\cdot dd$ 来设定 d 的阈值，其中 n 是用于调整异常值测定的灵敏度和特异性的参数。

通过察看谱商的中值 m 可以检测第二种异常值。如果四分位数间差值 d 没有指示异常值，但是谱商的中值 m 明显偏离 1，则可能是仅在小波谱范围中的强偏离波谱“愚弄”了积分归一化（步骤 B1）。这可以是由仅影响一小部分波谱的技术问题（例如尖峰或不良的水峰抑制）或由具有异常代谢组学响应的动物所引起。虽然不能在异常动物和技术问题之间做出区分，但是异常值检测指示必须进一步研究该样品。对于与 m 相关的异常值检测的经验方法是与理想值 1 具有±0.15 的偏差。而且，除了采用这些硬性阈值之外，还可以采用研究的特性阈值例如中值 m 的中值 mm 和中值 m 的四分位数间差值 dm。因此，所述中值的临界阈值可以根据 $m_{临界}=mm\pm n\cdot dm$ 来设定，参

数 n 确定灵敏度和特异性。

2.4 发现特定技术问题

如果已知波谱内的区域通常受特定技术问题的影响，则可以选择该区域作为辅助波谱范围。然后可以将在该辅助波谱范围内的谱商与在波谱的不受影响波谱区内即下文中所谓“主波谱范围”的谱商比较，从而检测特定问题。该方法因而包括下列步骤：

C1.选择主波谱范围和与主波谱范围不重叠的辅助波谱范围。

C2.主波谱范围中记录多个主波谱，并且对每一个所述主波谱，记录辅助波谱范围内的相关辅助波谱。

C3.在主波谱范围中获得基准主波谱和在所述辅助波谱范围内获得基准辅助波谱。

C4.对每一个主波谱进行主波谱除以基准主波谱的分格除法以获得对应的系列谱商；并且对每一个辅助波谱进行辅助波谱除以基准辅助波谱的分格除法以获得对应的系列辅助谱商。

C5.对至少一个所述主波谱计算主谱商的中值，称为 mp。

C6.对与所述主波谱相关的每一个辅助波谱计算辅助谱商的中值，称为 ma。

C7.计算 mp 和 ma 的商，称为 qs。

C8.通过比较 qs 和基准值 1 来进行异常值检测测试。

如果该谱商基本偏离 1，则在特定辅助区域中的波谱基本与其余波谱不同，这是特定问题的强烈指示。而且，可以采用固定阈值或基于 qs 分布的软性阈值。用这种方法可以处理的典型问题是水共振抑制的质量。为此，选择特定辅助范围以包含与水共振附近的波谱区域。

2.5 确定整个研究的质量

通常，代谢组学研究通过测量相当数量的样品来进行。这些样品由接受一定药物或物质的“给药”动物和没有接受所述药物和物质的“未给药”动物而获得。由此测量的波谱将表现出变化，这是由于固有的动物间变化、代谢组学响应和测量间变化（技术问题...）。在这些变化中，代谢组学响应通常导致波谱中最大的局部变化。通常由

于代谢组学研究的目标是研究代谢组学响应，因此代谢组学研究的“质量”可以通过观察波谱形状的非局部变化来判断，因为这些很可能不是由代谢组学响应所引起。如果在 2.2 部分中引入的算法通过下列步骤扩展，则可获得研究整体质量的测量：

C9.对所有波谱实施步骤 C5（和相应步骤 C6）并且对每一个波谱计算商的四分位数间差值 d 和商的中值 m。然后对该研究的所有中值 m 计算四分位数间差值 dm。之后计算该研究的全部中值 m 的中值 mm。还计算该研究的所有差值 d 的中值（还称为 md）。最后计算该研究的所有四分位数间差值 d 的四分位数间差值 dd。

如果研究含有许多形状显著不同的波谱，则 md 和 dd 的值将会更高。这可能由于波谱的不良定相、非常低浓度的样品、测量失败等原因而发生。对波谱质量的相应影响给代谢组学响应的数据分析提出了更多挑战。

dm 的高值表示对于几个波谱来说积分归一化和商归一化偏差显著。如果 md 和 dd 较高，则这种偏差可能是由有缺陷的波谱引起。另一方面，dm 高值和同时 md 与 dd 低值意味着研究中的几个波谱中仅有一小部分是不一致的。这意味着由于强代谢组学响应或者由于波谱中的局部缺陷，例如尖峰、污染物或药物相关化合物，使得这几个样品超出范围（outlying）。

2.6 图解检测异常值

检测异常值的图解工具可以通过将 d 对 m 作图来获得。因此，正常波谱将围绕 m＝1 群集，同时具有较低的 d 值。在图 2 中，这些波谱表示为方块。具有广泛损伤的波谱具有比通常设定为 1 的特定阈值更大的 d 值。在图 2 中，这些波谱位于厚水平线之上并由圆圈表示。由于强代谢组学响应、由于污染物或由于波谱局部损坏而超出范围的波谱位于 d 的阈值以下，但明显偏离 m＝1。通常将 m 的阈值设定为 0.80 和 1.20。在图 2 中，这些波谱位于两个箱体（boxes）之内并用星号表示。利用参数 n＝3 用可变阈值实施同样的程序产生的结果示于图 3 中。

3.实施例

在此分析了具有不同背景的四种数据组。第一种数据基于模拟。因此，代谢组学

研究的典型尿 NMR 波谱是系统性变化的以便模拟可能影响归一化的各种影响。模拟范围为从在实际变化直到非实际极端变化。第二实验数据组是基于来自一个代谢组学研究样品的 NMR 测量值。因此，归一化方法受到具有极大量代谢物样品和同时受到尿浓度变化的挑战。对于该数据组，由于技术问题引起的异常值已被滤除。第三数据组是未给药大鼠的超过 4000 个 NMR 测量值的集合。这些样品仅表示正常的生物和分析变化，因此该数据组允许在最低要求的条件下比较各种归一化程序的性能。第四数据组基于来自没有排除任何数据的两个代谢组学研究的测量值。因此，可能遇到包括空白样品测量、具有次最优质量的样品和由于技术问题导致的不良波谱的各种挑战。这些数据将用于说明在实际情况中异常值的验证。

3.1 用于归一化的模拟数据组

对于归一化方法的稳定性的模拟，系统性地改变“金波谱（golden spectrum）”。将金波谱计算为来自未给药大鼠的超过 4000 波谱的中值波谱。因此认为金波谱代表大鼠尿代谢组学领域中典型的波谱。波谱范围（9.96-0.4ppm）被均分成 0.04ppm 的积分格。4.48-6ppm（水和尿）之间范围的格被排除并且发生柠檬酸盐共振的范围（2.72/2.67ppm 和 2.56/2.52ppm）的格被合并到两个格中，从而产生总共 201 个格。使波谱归一化到 100 的总积分。第 201 个格（0.4ppm）的强度被人为设定为 0.5。仅更改该峰以模拟浓度的非特定变化但是不用于特定变化。该格将被用作基准格以判断归一化方法的质量。在图 4 中示出分格的金波谱。

通过以 0.1 的步幅（steps）系统变化样品的非特定浓度，直到双倍浓度从而产生第一组模拟数据。这通过将金波谱的每一个格乘以系数 1.1、1.2、1.3 等来实施。11 个波谱的系列示于图 5 的图面 A 中。

通过系统改变一个单一格而产生第二组的模拟数据。因而，在 2.7ppm 的峰（通常在波谱中可见的柠檬酸盐的两个峰中一个）以 10％的步幅增大总强度积分（总计 10 步幅）。11 个波谱的系列示于图 5 的图面 B 中。显然，该单峰占据了整个波谱。

第三组模拟数据表示第一和第二模拟数据组修改的组合。因此，对于每一步，样品的非特定浓度增大 10％的步幅并且同时金波谱的积分强度的 10％被加到 2.7ppm

处的峰处。对应波谱示于图 5 的图面 C 中。

对于第四组模拟数据，系统更改 10 个格的块，以模拟几个峰的特定变化。因此，对 10 个格的每一个，起始 10 个格的强度增加了 1%的金波谱积分强度。对于第二波谱，前 20 个格被增大。总体而言，产生 20 个波谱，该波谱不断地逐步增大更多的格直到最后波谱的 300%积分强度（与金波谱相比）。系统变化可视为图 5 的图面 D 中的格的块，由此，仅系列波谱的第一个和最后一个波谱是直接可视的。必须注意，对于研究中所有回归一化方法，错误格的位置是不相关的。

3.2 用于归一化的代谢组学研究波谱

实际代谢组学研究波谱被用于利用实验数据来测试各种归一化方法，在该代谢组学研究中，将环孢霉素施加到动物中。对于动物研究，根据在别处（Lindon JC，Nicholson JK，Holmes E，Antti H，Bollard ME，Keun H，Beckonert O，Ebbels TM，Reily MD，Robertson D，Stevens GJ，Luke P，Breau AP，Cantor GH，Bible RH，Niederhauser U，Senn H，Schlotterbeck G，Sidelmann UG，Laursen SM，Tymiak A，Car BD，Lehman-McKeeman L，Cole JM，Loukaci A，Thomas C，Tox Appl Paramacology 187，2003，137-146）描述的 COMET 协议来实施测量和处理。数据组含有通过在不同的时间点对 10 个对照动物、10 个低剂量给药动物和 10 个高剂量给药动物取样而获得的总共 231 个样品。从该数据组中除去已经检测为由于技术问题而导致的异常值的 18 个样品。如 3.1 部分中所述来实施波谱区域的分格和排除。两个高剂量给药动物的所有时间点的非归一化波谱和放大部分示于图 6。

3.3 用于归一化的对照样品

为了使不同归一化方法的实施不仅在强代谢组学响应的困难条件下有效而且在正常条件下也有效，建立了未给药大鼠 NMR 波谱的集合。因此，基于最小极端变化，从来自对照动物和给药前样品的 4521 样品中选择出 4023 样品。该样品集合表示大鼠的代谢组学特征的正常变化。

由于并不是所有的动物都已给药，因此样品的肌酸酐水平不受代谢组学变化的影

响并因此样品的肌酸酐水平代表对样品的总浓度的良好测量。因此，归一化方法的性能可以基于归一化样品的肌酸酐水平的变化来比较。因此，通过 4.02-4.10ppm 之间波谱的积分来测定肌酸酐水平以说明由于 pH 变化导致的肌酸酐峰的移动。

3.4 异常值验证的研究

对于异常值的验证，采用两个研究。第一研究对应于 3.2 部分中描述的环孢霉素研究，但是现在包括人工验证的异常值（与 3.2 部分对比）。第二代谢组学研究使用罗格列酮作为给药化合物并含有 80 个样品，其中 45 个波谱来自给药动物，35 个波谱来自未给药动物。根据 COMET 协议（Lindon JC 等人；Ioc.cit.）实施测量和数据的处理。而且，没有去除人工验证的异常值。用于判断研究的总体质量（参见 2.5 部分）的标准被应用到含有 60 个样品的附加第三研究（30 个样品来自给药动物，另三十个来自未给药动物）。该研究的第一测量面临着关于水共振抑制的问题，水共振导致自动基线校正和定相不良。该研究的数据也进行人工基线校正和定相。另外，利用最优化脉冲系列（Bax 脉冲）再次测量该研究的样品，产生最优视觉质量的波谱。

4.结果

按照下列次序来展示结果。首先，利用模拟数据组来比较积分归一化、向量长度归一化和商归一化。然后对代谢组学研究的波谱进行比较。最后利用几个代谢组学研究来展示采用商归一化检测异常值的可能性。

4.1 归一化方法－模拟

对于模拟数据组的结果（详见 3.1 部分），计算改进波谱和金波谱的 0.4ppm 格的强度的商。仅根据浓度的非特定变化人工修改该峰。因此，通过构建这种基准幅度，商 1 意味着相对峰强度的最优恢复并且通过对应的归一化方法商 1 意味着波谱的最优归一化。

4.1.1 各种归一化方法的性能

在这部分中，比较积分归一化和向量归一化与商归一化。对于商归一化，利用金波谱作为基准波谱。基准波谱的系统变化稍后讨论。在图 7 中，表示了四个数据组的三种归一化方法的结果。对于仅含有非特定变化的总浓度的数据组 1，所有三种方法均表现出最优归一化。恢复率 1 意味着仅随总浓度而变化（例如样品稀释）的峰和分析物被归一化到相同的恒定浓度。象预期一样，全部三种方法均能够足以将稀释样品的实际系列波谱归一化。

仅含有一个单信号和非稀释样品的特定变化的第二数据组示出表现大不相同的三种方法。向量长度归一化对单峰的变化高度灵敏。因而，由于利用二次项计算长度，因此由于波谱单格增大导致的浓度向量长度增加极大。向量长度的重新缩放均布于全部格，这导致低估没有变化的格。对于积分归一化，由于一个格的强度增加的影响均匀分布而不是对全部格取二次项，因此与实际性能的偏离不是那么惊人。例如，当积分归一化分布于所有格时，波谱 10 在一个单格中含有附加 100%的总强度，因此导致所有格缩小两倍。另一方面，商归一化没有受单格变化的影响并因此产生对全部波谱的最优归一化。

含有第一和第二数据系列的组合变化的第三数据系列表现出与第二数据系列很相似的结果。单格的变化强烈影响对向量长度和固定积分的归一化。

第四数据系列模拟几个格的组合变化。对于第一波谱，10 个格（来自 201）的强度被增大（每一个格增大 1%总强度），对于第二波谱，20 个格的强度被增大，依此类推。对于这种情况，向量长度归一化表现出比积分归一化更好的性能，但是这两种方法对第一波谱均表现出偏离最优归一化。另一方面，商归一化一直表现为最优归一化，系统性增大了 201 个格中 100 个格的强度。对于系统性增大甚至更多格的波谱，性能急剧下降。然而，超过半数格在相同方向上系统性变化的这种情况是非常不现实的。对于可以被认为是现实情况的 5-25%的格的系统变化，并且甚至对于已是非常极端情况的 30%-50%的格的系统变化，商归一化表现良好。

4.1.2 噪声的影响

上文列出的四个数据系列的数据分析被重复两次，由此人为噪声被加入到波谱

中。对于第一次重复，均匀噪声被加入到每一个信号，所述均匀噪声具有 0.6%平均强度的标准偏差/信号。这种噪声量估计是来自波谱区域中超过 4000 个波谱的典型波谱仪噪声，在所述波谱区域中没有生物性变化存在。对于第二次重复，噪声量增加了十倍从而接近未处理动物的典型生物噪声。对于两次重复，所有归一化方法证明是对噪声不敏感的。实际上，由于归一化方法考虑到所有的格（平滑效应），因此，归一化系数的变化明显低于每一个格的变化。例如，数据系列 1 的归一化系数对 0.6%噪声的标准偏差为 0.04%-0.1%，对 6%噪声的标准偏差为 0.2%-0.4%。

4.1.3 商归一化基准波谱的影响

与向量长度归一化和积分归一化相比，商归一化需要基准波谱。基准波谱对商归一化性能的影响在本部分中研究。除了利用“金波谱”（1）作为基准波谱之外，还利用下列基准波谱：

（2）3×4 个模拟数据系列的所有波谱的中值波谱，所述所有波谱仅由于非特定变化和由于噪声而不同。利用来自所有候选波谱的每一个信号格的中值来构建中值波谱。

（3）所有波谱的中值波谱，所述所有波谱仅由于非特定变化、由于噪声和由于小于或等于总积分的 20%的变化而不同。

（4）所有波谱的中值波谱，所述所有波谱仅由于非特定变化、由于噪声和由于小于或等于总积分的 100%的变化而不同。

（5）所有波谱的中值波谱（所有 3×4 个模拟数据系列）。

（6）在每一个格中具有恒定值 1 的波谱。

对于上述六个不同的基准波谱，对前述四个无噪声数据系列实施商归一化。对于前三个数据系列，基准波谱之间没有发现显著差异。对于第四数据系列，可以在图 8 中看见明显差异。利用恒定值作为基准波谱表现出非常差的性能。由于波谱和基准波谱之间谱商的分布对应于波谱本身的分布并因此谱商的分布是平坦和宽的，因而这种发现是不言而喻的。因此，几个峰的增大将显著地改变中值。可以看出，如果基准波谱尽可能对应于代表性波谱而没有特定变化（允许非特定变化），则获得最稳定归一

化。还可以看出，总强度的高达 20%的特定变化不会明显地影响归一化，而具有高达 100%积分强度特定变化的波谱的归一化较不稳定。由于完整数据系列 6 包含具有许多格超高特定变化的非现实超大量的波谱，因此在图 8 中表示的所有数据的模拟明显是基准波谱影响的放大。然而，模拟显示出归一化研究的最优方法是利用代表未给药动物例如对照动物和/或给药前时间点的数据作为基准波谱。计算代表性波谱的可行方法是分别使用大量对照波谱或给药前波谱的平均值或中值。计算中值而不是平均值的优点是在波谱中对异常值的较高容限。这在代谢组学研究中经常遇到。

模拟数据系列的目的是测试在现实、极端和一定程度上的非现实条件下的不同归一化方法。信号的特定变化均仅在一个方向上实施（信号增大），这是由于这更需要归一化过程：如果不同信号特定变化成不同方向，则对归一化过程而言变化相互抵消。例如，20 个信号增加 10 强度单位和 15 个信号增加 10 强度单位，则归一化方法仅受 5 个信号变化的影响。因此，对于现实条件，利用数据系列 4 的模拟应该仅被认为适用于最初一些波谱，而不是全部波谱。

当查看不同的模拟时，很明显，对于所有不同现实和极端条件，商归一化均表现得优于普通归一化方法。特别是当单格极端变化时，这可能发生在代谢组学研究中，商归一化仍发现了最优归一化系数，而其它的归一化方法将单格的过量强度的影响分配到全部格上。因此，引入了具有仅影响单格的系数的所有其它格的人为负相关。

此外，在模拟中表明，商归一化的最佳基准波谱是最具代表性的没有特定变化的波谱。因此，对于代谢组学波谱，最优基准波谱应该基于对照动物或在给药前的时间点的动物波谱来计算（例如作为中值波谱计算）。

4.2 归一化方法－环孢霉素研究

在这部分中，利用在 3.2 部分中详细描述的完全代谢组学研究的数据来比较三种归一化方法的性能。通过目测发现，在 1.44ppm-1.84ppm 化学位移之间的所有信号对不同动物和不同时间点非常稳定。这些信号不受该研究中特定代谢组学变化的影响而仅受尿浓度差异的影响。因此，不同样品之间的这部分波谱积分的相对标准偏差被用作归一化方法的质量标准。对于两个高剂量给药的动物 28 和 30，在图 9 中对利用

不同方法归一化的这部分波谱作图。动物 28 在时间点 48h 和 72h 排出极大量的葡萄糖，而动物 30 在所有时间点表现出典型的代谢组学响应。为了比较，两个动物的完整非归一化波谱在图 6 中表示。

最后，将该研究的所有样品用于研究不同方法的归一化系数和在 4.02-4.10ppm 间积分测定的肌酸酐浓度之间的相关性。

对于积分归一化，动物 28 和 30 的波谱在图 9 的中间行表示。明显的是，动物 30 在 1.44ppm-1.84ppm 之间表现出更适合的信号，而对于动物 28，时间点 48h 和 72h 的波谱太低。这些时间点的波谱表现出非常高的葡萄糖峰，由于限制总积分因此这些葡萄糖峰抑制剩余的波谱。该研究所有积分归一化样品在 1.44ppm-1.84ppm 之间的积分表现出 10.3%的相对标准偏差。归一化系数对肌酸酐峰的线性回归表现为 0.87 的相关系数。

向量长度归一化图（图 9，底行）表现出较差的归一化。两个动物的几个波谱在 1.44ppm-1.84ppm 之间表现出太低或太高的信号，并且两个葡萄糖样品（动物 28，48h 和 72h）偏离最大。该区域中信号在所有样品中不同，这可以由 15.0%的相对标准偏差看出。而且与肌酸的相关性差（r＝0.62）。

从图 9 的顶行可以看出，商归一化对本文中绘图的样品是较好的。1.44ppm-1.84ppm 之间的信号非常适合于所有样品。该区域中所有样品信号的 4.5%的低相对标准偏差和与肌酸酐峰极佳的相关性（r＝0.99）表明商归一化是对全部研究最一致的归一化。

4.3 归一化方法－正常样品

一个令人感兴趣的问题是验证不同归一化方法对对照动物和给药前动物不仅在困难条件下而且在“正常条件下”的性能。所选的 4023 未给药样品（详见 3.3 部分）不含有强代谢组学响应或药品相关化合物。因此，预计所有三种归一化方法应该表现出类似的性能。由于未给药动物应该具有很稳定的相对肌酸酐水平，因此本文中通过肌酸酐峰的相对标准偏差来评估三种归一化方法的性能。

结果很值得关注：向量长度归一化具有不可接受的 12.2%的肌酸酐峰高相对标准

偏差，而积分归一化具有 7.6%的低相对标准偏差，并且商归一化性能最佳，具有 6.7%的低相对标准偏差。这意味着甚至当观察对照动物时，由于代谢组学变动引起的特定变化是如此之高使得在归一化方法之间存在显著差异，由此商归一化表现出最佳的性能。

4.4 异常值验证

用商归一化方法将包括通过目测波谱检测的异常值的环孢霉素研究归一化（详见 3.4 部分数据系列）。对于异常值的自动验证，对每一个样品，除了商的中值 m 外，还计算谱商的第三和第一四分位数之间的四分位数间差值 d。在图 10 的左图面中，对该研究的全部样品用 d 对 m 来绘图。通过目测波谱发现的异常值表示为圆点、三角形和菱形。很明显，所有非远离中心的样品群集在非常低的 d 值和约为 1 的中值处。具有极端代谢组学响应的样品在这种情况下是极大量的葡萄糖，该样品位于低 d 值和低 m 值处。由于技术问题、空白样品和具有水共振抑制不良的样品所引起的异常值均位于高 d 值处（高于 3）。这意味着 $d > 1$ 的样品阈值检测出由于非代谢组学相关问题引起的所有异常值。而且，极端葡萄糖样品可以用 $d < 1$ 和 $m < 0.8$ 的简单阈值来检测为极端代谢组学响应。第二研究的质量图（参见 3.4 部分）在图 10 的右图面中表示。而且，$d > 1$ 的样品阈值检测出先前通过目测所有波谱验证的全部异常值。由于典型问题导致的一些远离中心的波谱示于图 11 中。

在图 12 中，证实了对于罗格列酮研究而言，商归一化是如何被用于检测特定问题的。在该实施例中，监测了水共振抑制的质量。首先，对该研究实施商归一化。然后，计算 6.04ppm 处波谱除以所有商的中值得到的谱商。如果对应值强烈偏离 1，则波谱明显不同于水共振附近的基准。很明显，四个样品具有较差的水抑制。这四个样品也已经被人工验证并且在图 10 的右图面中表示为三角形▲。

表 1：三个数据系列的不同质量特征。质量特征值越低表示全部研究中的波谱的形状越类似。

数据系列	md	dd
第一次测量	0.38	0.60

人工重新处理	0.33	0.38
第二次测量	0.27	0.29

对于第三数据系列（参见 3.4 部分），商归一化的另一个应用示于表 1 中。其中，示出了四分位数间差值的中值 md 和四分位数间差值的差值 dd（详见 2.5 部分）。第一次测量对几个样品具有较差的水抑制，较差的水抑制对自动定相和基线校正有不利的影响。人工基线校正和定相可以改善这些样品的视觉质量。然而，具有最优脉冲次序的样品的第二次测量明显地改善了水抑制和波谱质量（如目测所证实）。从表 1 明显可知，全部三个异常值标准均与波谱质量的视觉印象一致。因此 md 表达了波谱内的平均不均匀性。由于仅有一些波谱受不良水抑制的影响，因此 md 仅适度减少。在另一方面，由于样品的再处理和再测量主要改善具有较差水抑制的样品，因此描述不同波谱之间这种不均匀性的差异的 dd 急剧减少。两种质量特征均允许评估波谱内和波谱之间的均匀性而不需要检验波谱。

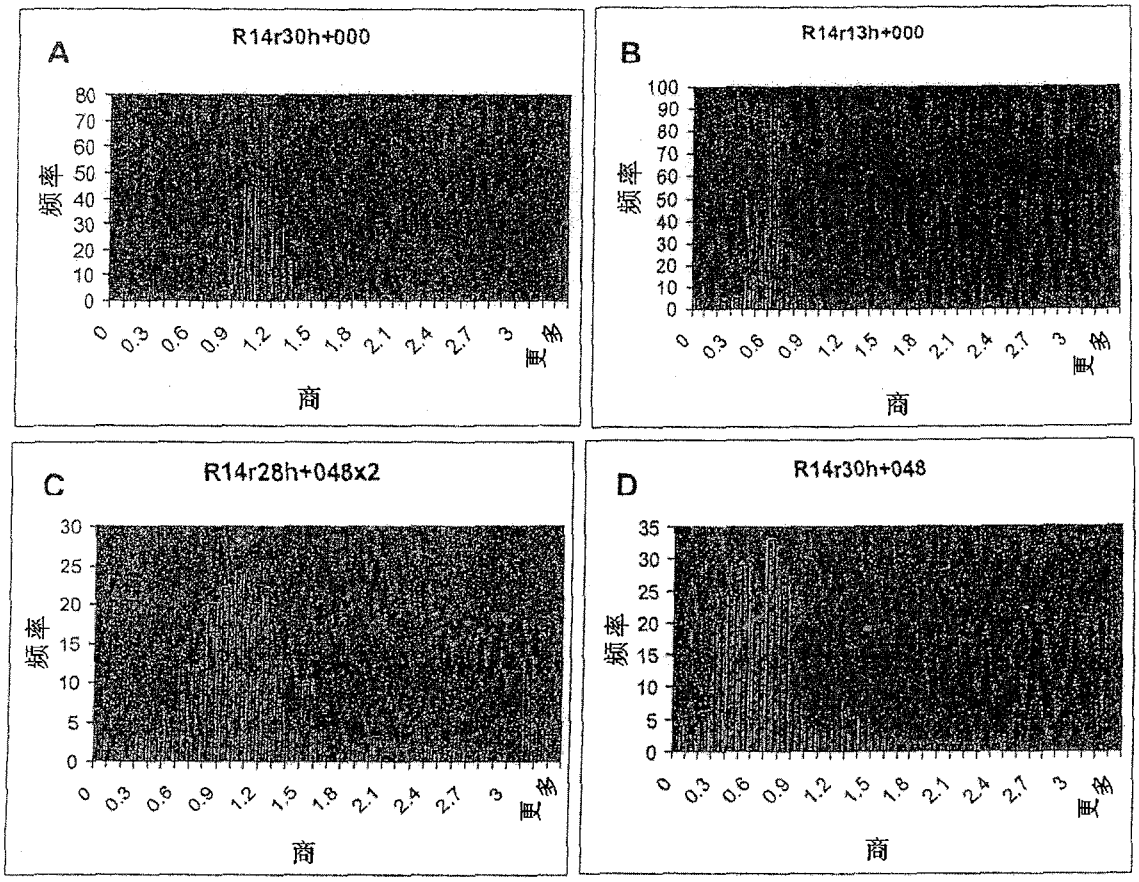

A
R14r30h+000
频率
商
B
R14r13h+000
频率
商
C
R14r28h+048x2
频率
商
D
R14r30h+048
频率
商
更多

图1

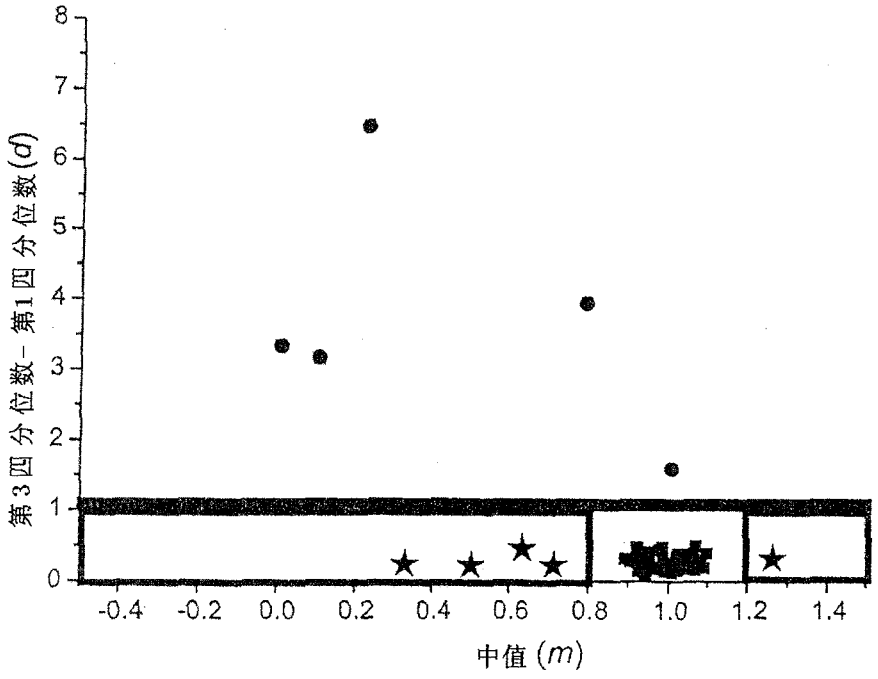

第3四分位数-第1四分位数(d)
中值(m)

图2

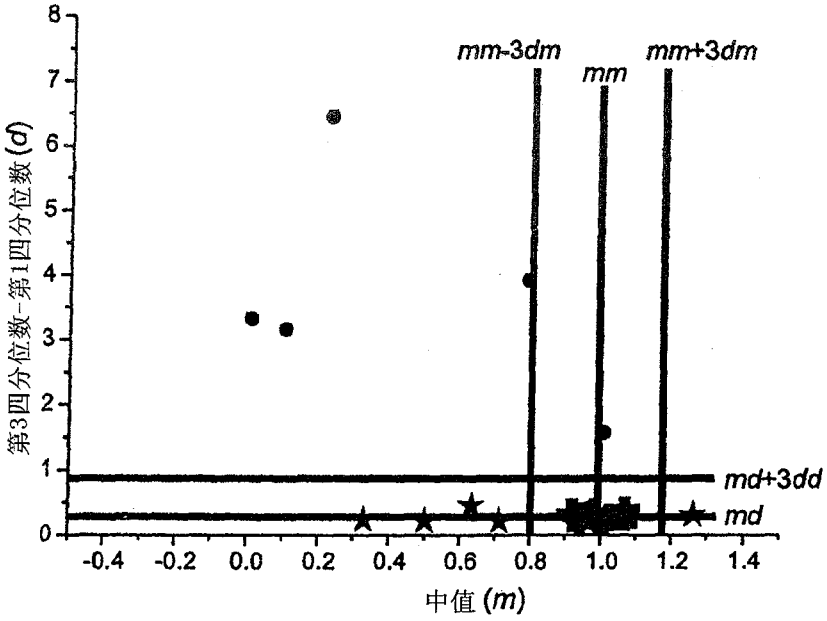

mm-3dm
mm
mm+3dm
md+3dd
md
第3四分位数-第1四分位数(d)
中值(m)
8
7
6
5
4
3
2
1
0
-0.4
-0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4

图 3

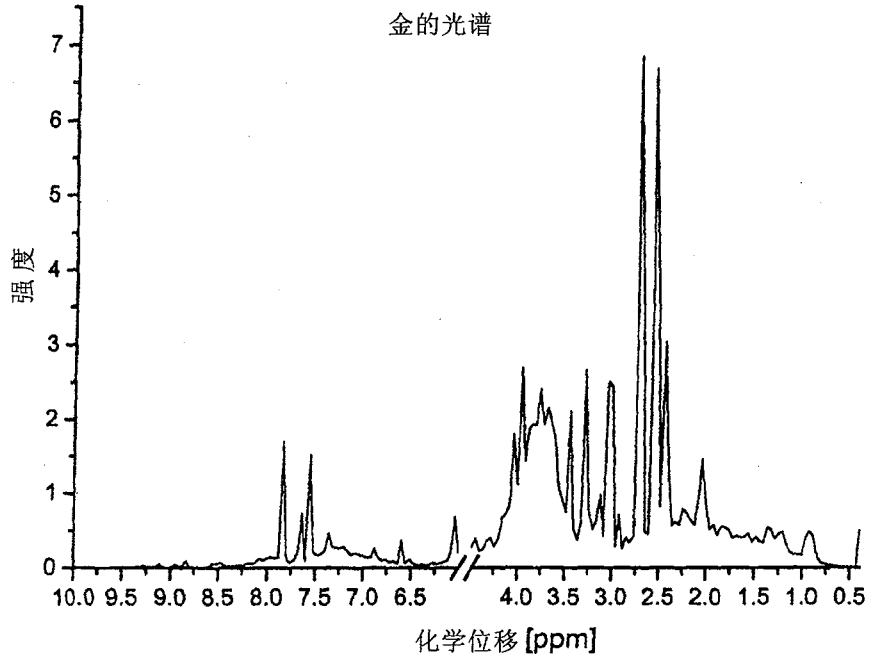

金的光谱
强度
化学位移 [ppm]
7
6
5
4
3
2
1
0
10.0 9.5 9.0 8.5 8.0 7.5 7.0 6.5
4.0 3.5 3.0 2.5 2.0 1.5 1.0 0.5

图 4

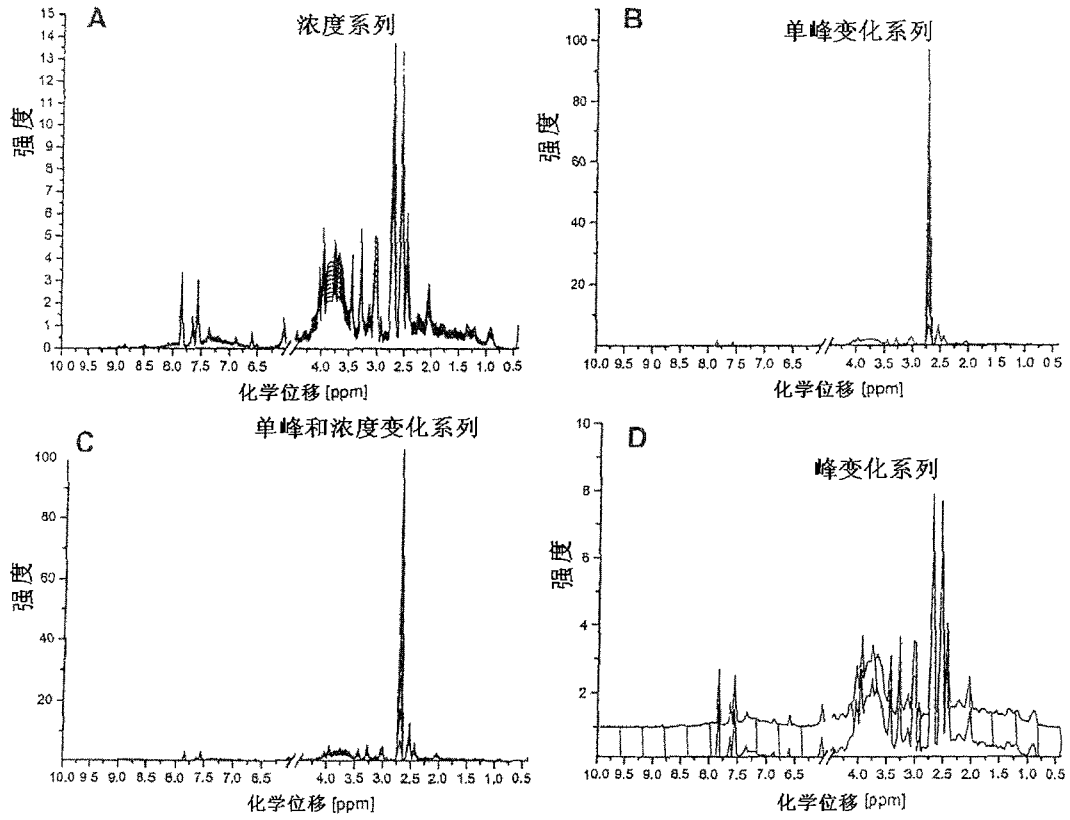

A
浓度系列
强度
化学位移 [ppm]
B
单峰变化系列
强度
化学位移 [ppm]
C
单峰和浓度变化系列
强度
化学位移 [ppm]
D
峰变化系列
强度
化学位移 [ppm]

图5

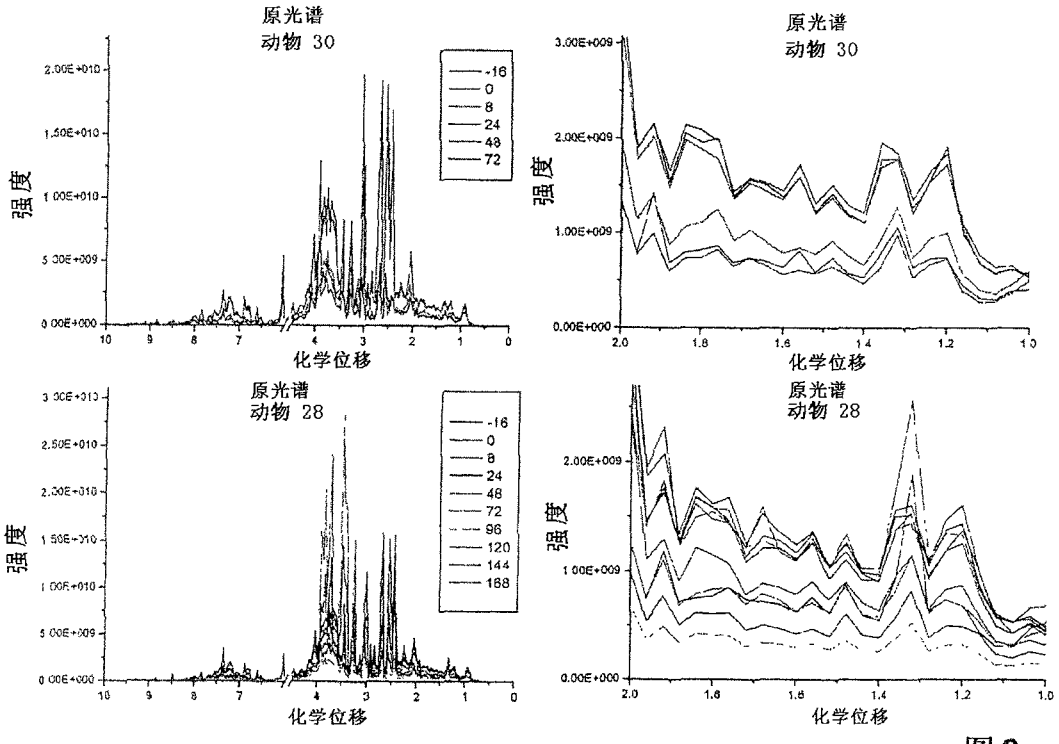

原光谱
动物 30
-16
0
8
24
48
72
强度
化学位移
原光谱
动物 30
强度
化学位移
原光谱
动物 28
-16
0
8
24
48
72
96
120
144
168
强度
化学位移
原光谱
动物 28
强度
化学位移

图6

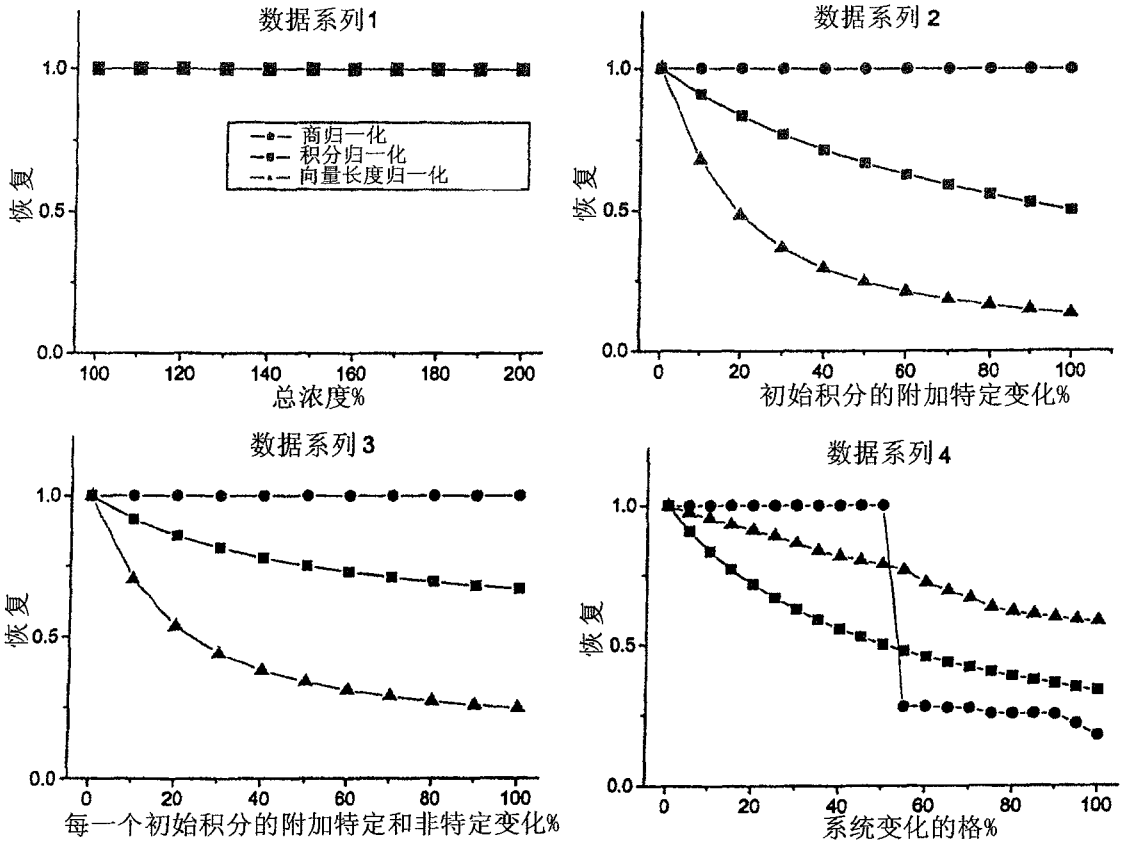

数据系列1
商归一化
积分归一化
向量长度归一化
恢复
总浓度%
数据系列2
初始积分的附加特定变化%
数据系列3
每一个初始积分的附加特定和非特定变化%
数据系列4
系统变化的格%

图7

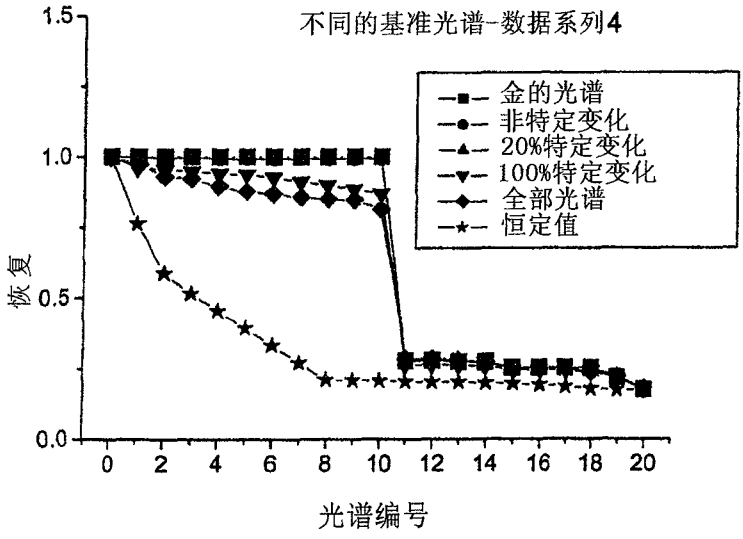

不同的基准光谱-数据系列4
金的光谱
非特定变化
20%特定变化
100%特定变化
全部光谱
恒定值
恢复
光谱编号

图8

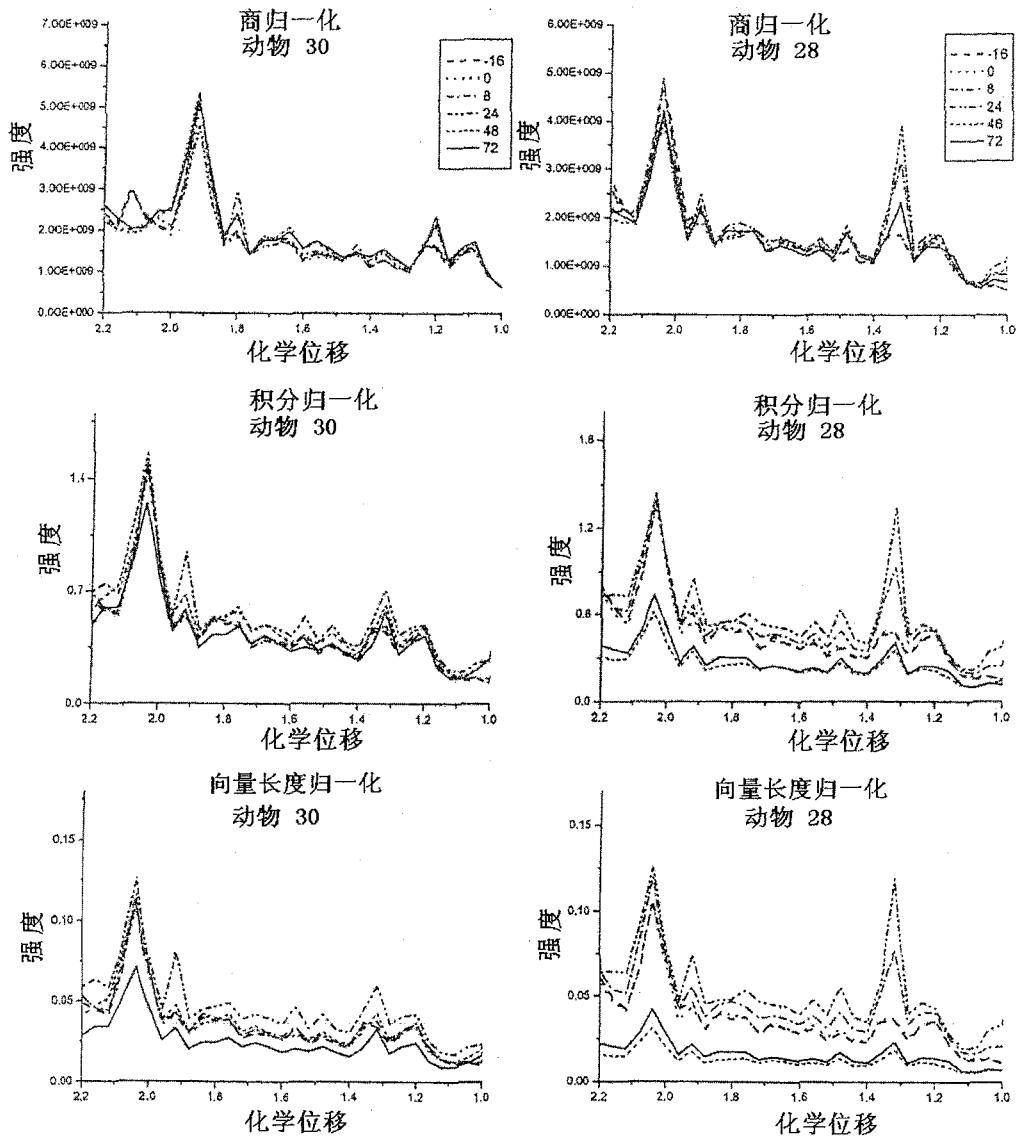

商归一化
动物 30
强度
化学位移
-16
0
8
24
48
72
商归一化
动物 28
积分归一化
动物 30
积分归一化
动物 28
向量长度归一化
动物 30
向量长度归一化
动物 28

图9

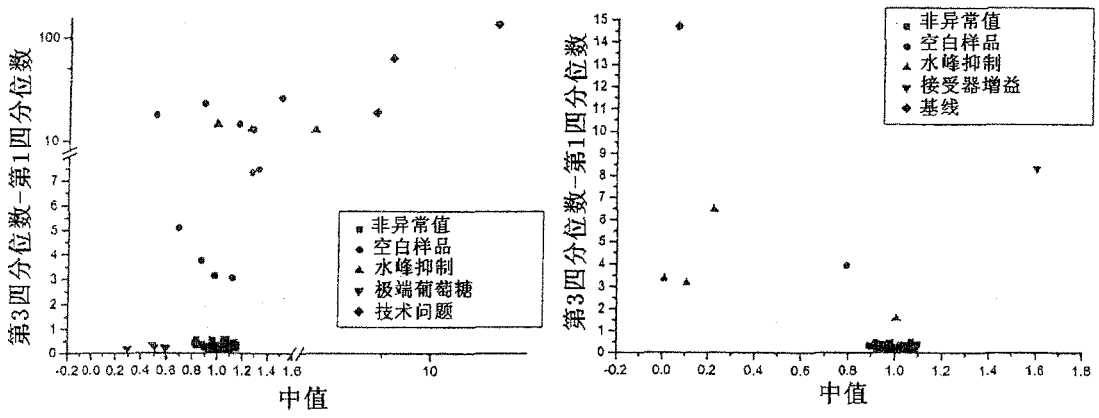

第3四分位数-第1四分位数
中值
非异常值
空白样品
水峰抑制
极端葡萄糖
技术问题
接受器增益
基线

图10

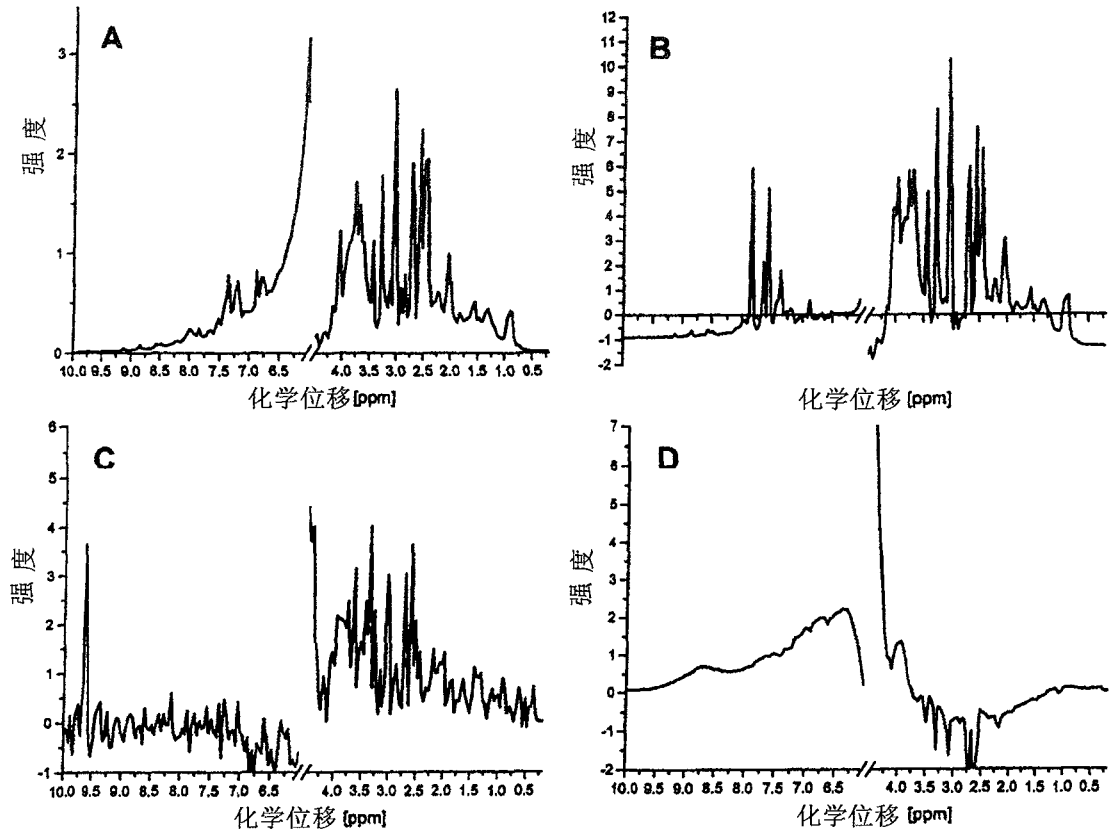

A
强度
化学位移[ppm]
B
强度
化学位移[ppm]
C
强度
化学位移[ppm]
D
强度
化学位移[ppm]

图11

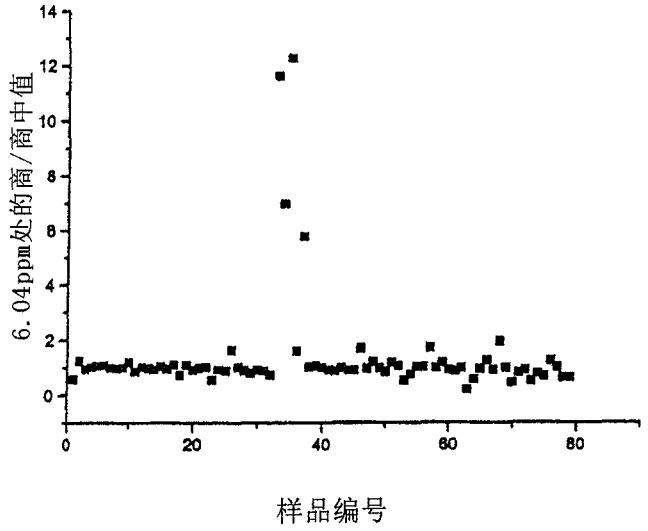

6.04ppm处的商/商中值
样品编号

图12