



(12)发明专利

(10)授权公告号 CN 104380311 B

(45)授权公告日 2018.06.26

(21)申请号 201380030388.3

(22)申请日 2013.04.10

(65)同一申请的已公布的文献号
申请公布号 CN 104380311 A

(43)申请公布日 2015.02.25

(30)优先权数据
2008620 2012.04.10 NL
2009015 2012.06.15 NL

(85)PCT国际申请进入国家阶段日
2014.12.09

(86)PCT国际申请的申请数据
PCT/NL2013/050260 2013.04.10

(87)PCT国际申请的公布数据
W02013/154425 EN 2013.10.17

(73)专利权人 拜斯帕尔有限公司
地址 荷兰莱顿

(72)发明人 雷内·雷蒙德·帕钦
阿尔扬·劳伦·范乌伊杰克胡吉斯
阿德里安努斯·博斯

(74)专利代理机构 北京安信方达知识产权代理有限公司 11262

代理人 宁晓 郑霞

(51)Int.Cl.
G06K 9/00(2006.01)

(56)对比文件
US 2007184455 A1,2007.08.09,
US 2007184455 A1,2007.08.09,
CN 1500000 A,2004.05.26,
JP 2001322816 A,2001.11.20,
Yamit Sharaabi等.Two-state analysis
of single-molecule Raman spectra of
crystal violet.《Chemical Physics》.2005,
(续)

审查员 姚晓斌

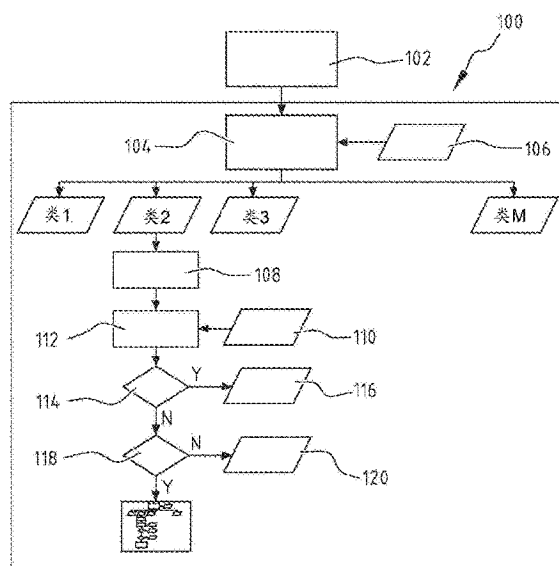
权利要求书2页 说明书20页 附图13页

(54)发明名称

依据谱数据的样本分类方法、创建数据库的方法和使用该数据库的方法、以及相应的系统

(57)摘要

本发明涉及以谱数据为基础的样本分类方法,创建数据库的方法和使用该数据库的方法,以及相应的计算机程序,数据存储介质和系统。方法包括:a)获得用于作为参考谱的至少两组第一谱,每组包括属于同一类的参考样本谱;b)确定每个参考谱的涉及谱特征的相同的至少一个量的值;c)在确定的值的基础上,关联概率与量的不同的值;d)从样本中获得谱并确定该谱的相同的至少一个量的值;和e)以获得的所述至少两类中的每类的谱的概率和量的值为基础,计算样本属于该类的概率。



[转续页]

[接上页]

(56)对比文件

R.Dieckmann等.Rapid identification and characterization of Vibrio species using whole-cell MALDI-TOF mass spectrometry.《Journal of Applied

Microbiology》.2010,

蔡飞等.傅立叶变换红外光谱结合化学计量学在微生物判别、分类、鉴定中的应用.《The Journal of Pharmaceutical Practice》.2002,

1. 一种用于在谱数据的基础上对至少两类中的一类中的样本进行分类的方法, 所述谱数据包括喇曼谱、近红外谱、FT-IR谱、MALDI MS谱或MALDI TOF MS谱, 所述方法包括:

a) 获得用于用作参考谱的至少两组第一谱, 每个组包括属于同一类的多个参考样本的谱;

b) 通过将预定的函数或操作应用到每个所述参考谱来确定每个所述参考谱的与包括在所选择的一个所述谱数据中的谱特征相关的相同的至少一个量的值;

c) 对于每组参考谱, 在步骤b) 中所确定的值的基础上关联概率与所述至少一个量的不同值, 并且从所述至少一个量的所述不同值和关联的概率中构造概率密度函数;

d) 从所述样本获得谱, 并确定这个谱的所述相同的至少一个量的值; 以及

e) 基于在步骤c) 中确定的所述概率密度函数和所获得的所述至少两类中的每类的谱的所述量的值, 计算所述样本属于所述类的概率,

其中所述参考谱和获得自所述样本的所述谱是单粒子谱。

2. 根据权利要求1所述的方法, 其中, 所述样本是生物样本, 并且其中所述生物样本包括微生物, 且分类包括微生物的分类, 和/或

在参考类的特有谱特征的基础上选择所述至少一个量。

3. 根据权利要求1或2所述的方法, 其包括, 针对获得与谱特征量 Q_i 相关的值 I_i 的样本谱, 计算给定值 I_i 时所述样本属于参考类 A^j 的概率 $P(A^j | I_i)$, 根据的公式为:

$$P(A^j | I_i) = \frac{P(I_i | A^j) P(A^j)}{P(I_i | A^j) P(A^j) + \sum_{k \neq j} P(I_i | A^k) P(A^k)}$$

其中 $P(I_i | A^j)$ 是参考类 A^j 与值 I_i 关联的概率, 且 $k \neq j$ 时的 $P(I_i | A^j)$ 是不同于参考类 A^j 的至少一个参考类与值 I_i 相关联的概率。

4. 根据权利要求1-2中的任一项所述的方法, 其中步骤b) 包括确定每个所述参考谱的与谱特征相关的相同的至少两个量的值, 且步骤e) 包括结合获得的关于全部量的概率与所述样本属于各个类的整体概率。

5. 根据权利要求3所述的方法, 其中步骤b) 包括确定每个所述参考谱的与谱特征相关的相同的至少两个量的值, 且步骤e) 包括结合获得的关于全部量的概率与所述样本属于各个类的整体概率。

6. 根据权利要求1-2、5中的任一项所述的方法, 其中所述值与在至少一个预定的谱值处的强度或在预定的谱值范围内的强度相关, 和/或

所述值被归一化。

7. 根据权利要求1-2、5中的任一项所述的方法, 其中, 通过将在预定谱值范围上的各个谱的强度乘以权重函数来确定所述值,

其中所述权重函数基于同一类谱的累积谱。

8. 根据权利要求1-2、5中的任一项所述的方法, 其中所述值涉及在预定谱值处的强度之间的比值或预定谱范围内的强度之间的比值, 和/或, 其中所述值涉及在预定谱值范围内的峰值的位置, 和/或

所述值是向量。

9. 根据权利要求1-2、5中的任一项所述的方法, 其中, 针对第一组参考类且随后针对第

二组参考类执行步骤d)和e),其中在所述第一组参考类中的一个参考类中的样本分类的基础上选择所述第二组,其中:

所述样本是生物样本,所述生物样本包括微生物,且分类包括微生物的分类;以及在生物学分类层次的基础上选择所述第一组参考类和所述第二组参考类。

10.根据权利要求1-2、5中的任一项所述的方法,步骤d)包括从所述样本获得至少两个谱并确定这至少两个谱的所述相同的至少一个量的值,且步骤e)包括结合分类在同一类中的样本谱与累积谱,并通过结合各个类的所述参考谱来比较该累积谱和累积参考谱。

11.一种用于创建用于在谱数据的基础上对在至少两类中的一类中的样本进行分类的数据库的方法,所述谱数据包括喇曼谱、近红外谱、FT-IR谱、MALDI MS谱或MALDI TOF MS谱,所述方法包括:

- a)获得用作参考谱的至少两组第一谱,每组谱包括属于同一类的参考样本的谱;
- b)通过将预定的函数或操作应用到每个所述参考谱来确定每个所述参考谱的与包括在所选择的一个所述谱数据中的谱特征相关的相同的至少一个量的值;以及
- c)对于每组参考谱,在所确定的值的基础上关联概率与所述至少一个量的不同值,并且从所述至少一个量的所述不同值和关联的概率中构造概率密度函数,

其中,所述参考谱是单粒子谱。

12.一种使用数据库在谱数据的基础上对在至少两类中的一类中的样本进行分类的方法,所述数据库根据权利要求11所述的方法而获得,所述谱数据包括喇曼谱、近红外谱、FT-IR谱、MALDI M谱或MALDI TOF MS谱,所述方法包括:

- d)从所述样本获得单粒子谱,并确定该谱的所述相同的至少一个量的值;以及
- e)在从所述数据库获得的概率密度函数和所获得的所述至少两类中的每类的谱的所述量的值的基础上,计算所述样本属于所述类的概率。

13.一种计算机实现系统,其被配置为执行根据前述权利要求中的任意一项所述的方法的步骤。

14.一种用于在谱数据的基础上分类样本的系统,所述系统包括:

- 设置为从样本获得单粒子谱的装置;
- 设置为执行根据权利要求1-12中的任一项所述的方法的分析装置。

依据谱数据的样本分类方法、创建数据库的方法和使用该数据库的方法、以及相应的系统

技术领域

[0001] 本发明涉及用于在谱数据的基础上对属于至少两类中的一类的样本进行分类的方法。例如,谱数据包括喇曼(Raman)谱、近红外谱、FT-IR 谱、频谱、MALDI MS谱、MALDI TOF-MS谱。用这些技术中的一个获得的谱可与样本成分(即呈现在样本中的分子)相关。

背景技术

[0002] 然而,不考虑用于获得谱的技术,在属于同一类的样本的谱之间发生了变化。这些变化可归因于随机效应。这些随机效应可涉及测量技术或在调查下的样本,或二者兼而有之。

[0003] 因此,常规的分析谱数据的方法要求被分类的谱获得自具有大量粒子(诸如分子)的样本。通过要求大量粒子,由于随机过程的谱中的特征相比于由稳定过程引起的特征将被抑制。然而,这使得常规方法不适合用于分析具有较少数量粒子的样本。当随机效应主要是由于测量技术本身时,常规方法要求待被分类的谱为所谓的总谱,也被称为累积谱,通过增加相同样本的大量谱而获得。通过这种方法,由于测量技术的随机过程将在总谱中得到抑制。然而,当不能获得总谱时,这使得常规方法不适合用于分类样本。

[0004] 以上问题将针对其中使用质谱分析法(MS)获得谱数据的示例性例子进行说明。

[0005] 传统MS谱,且特别是飞行时间(TOF)质谱在质量上显示大的离子强度,该质量对应于通过电离形成的离子流中的高度丰富的离子质量。通常,这些高度丰富的离子来自于出现在分析物中的高度丰富的分子。

[0006] 然而,因为在电离分子的情况下容易存在大的可变性,故高度丰富的分子的出现将不会自动地产生大的离子强度。

[0007] 此外,电离的效率依赖于大量因素,其可在电离事件到电离时间之间强有力地改变。实例为存在的分析物数量、基质和分析物的比值、基质的晶体结构、及在基质晶体和分析物的位置处的激光束的强度。因此,通过各单独电离事件产生的离子强度会显著变化。

[0008] 最重要的是,在电离期间,考虑的分子会分裂成更小的部分,有些分子可能获得两倍或三倍的电荷,分子可聚集到流当中,等。因此,除了发生在对应于分析物分子质量的质量上的高离子强度,单独电离事件谱将在不与分析物分子质量相关的位置处显示峰值。然而,由于分裂、充电和聚集的过程本质上是高度随机的,故非相关的峰值的位置将在电离事件和电离事件之间强烈地变化。

[0009] 为产生可靠和稳定的谱,在关于MALDI MS的传统的分类方法中,累加了记录在靶板地点上不同位置处的单独电离事件的数量。

[0010] 因此,对比于由稳定过程引发的离子强度,电离期间将抑制由随机过程引发的离子强度。稳定过程将产生总谱中的高强度的窄区域(峰值),而带更多随机性质的过程将产生总谱中的宽的低强度响应(基准)。

[0011] 在总谱出现的峰值可与在分析物中容易地(有效地)电离的、丰富的分子相关联。

通常这些峰值被用于谱的分类,且因此被认为描述谱特征的特征。

[0012] 因为合成谱来源于记录在靶板地点上不同位置处的大量单独电离事件,故这些常规MALDI谱是样本的许多粒子的谱的总和。例如,在样本包括微生物的一些情况下,常规方法依赖于获得自许多微生物的谱。

[0013] 当对单粒子谱应用常规方法时,出现相似的问题,单粒子谱诸如单气溶胶粒子谱。在常规的方法中,为产生可靠和稳定的谱,获得自来源于单个独立种群的微生物的大量的单气溶胶谱被结合到总谱中。这个总谱以有限数量的不同谱特征(峰值)叠加在光滑的基准上为特征。

[0014] 建立累积谱的各个单粒子谱以大量峰值和不存在基准为特点。一些峰值发生在与累积谱中峰值位置相一致的位置上,其它峰值发生在不与累积谱中峰值位置相一致的位置上。此外,看起来单粒子谱呈现大量的点对点变化。

[0015] 图1示出累积自大量单粒子谱的谱(P)的实例,其中单粒子谱获得自源于单独独立种群的样本,其中还包括一些典型的基本单粒子谱的实例(Q、R、S、T)。

[0016] 由于大量的点对点变化和在单粒子峰值位置和累积谱峰值位置之间的有限关联,某一位置(质量过载)的谱的振幅和带相关的质量的离子形成的物质的存在之间不存在确定性关系。

[0017] 以上实例说明分析谱数据的常规方法的问题。这些方法不能直接地处理单粒子谱,因为他们没有考虑以上点对点的变化。此外他们不能处理由于测量技术的变化,诸如,如以上描述的在MALDI MS中由于电离的变化。

发明内容

[0018] 本发明的目的是提供用于在谱数据的基础上对属于至少两类中的一类的样本进行分类的改进方法,当属于同一类的谱显示出变化时其是有效的和可靠的。

[0019] 根据本发明,通过在谱数据的基础上对在至少两类中的一类中的样本进行分类的方法来完成该目的,该方法包括:

[0020] a) 获得至少两组第一谱用于作为参考谱使用,每组包括属于同一类的参考谱样本;

[0021] b) 为每个参考谱确定与谱特征有关的相同的至少一个量的值;

[0022] c) 在确定值的基础上,关联概率与量的不同的值;

[0023] d) 从样本获得谱,并确定这个谱的相同的至少一个量的值;以及

[0024] e) 在关于获得的所述至少两类中的每类的谱的量的概率和值的基础上计算样本属于该类的概率。

[0025] 步骤a)-c)描述了参考库的创建。步骤d)和e)描述了对比样本和这个参考库。

[0026] 在步骤b),针对每个参考谱确定了相同的至少一个量的值。例如,对谱数据执行数学操作,以获得该值。换句话说,步骤b)描述通过对每个参考谱应用预定的函数或操作而获得分值(值)。

[0027] 步骤b)的结果是每个参考谱已与其至少一个值关联,每个值对应于预定的量,即对谱的预定操作。下一步,对于每一组参考谱,在针对该组中谱确定的值的基础上,概率与至少一个量的不同值相关联。这可被认为是构造概率密度函数(PDF)。例如,PDF是离散概率

密度函数,例如其可表示为数组或直方图。

[0028] 步骤a)-c)的最终结果是每组参考谱已针对至少一个量中的每个量关联的各个量的值的PDF。例如,在MS谱数据情况下,选择的一个量是谱的强度,也称为谱振幅,在 $M/Z=1000$ Dalton。在这个实例中,对于每组参考谱PDF为强度在 $M/Z=1000$ Dalton时创建。值得注意的是,通常对于不同组的参考谱PDF将显著不同,因为不同组的参考谱对应于不同的类。

[0029] 在步骤d),获得样本的谱,且此外关于这个谱的至少一个量的值被确定。在先前构造的PDF的基础上,在步骤e)中计算样本属于每一个参考类的概率。

[0030] 因此根据本发明的方法使用每个类的量(与预定的操作/函数有关)的值(分值)的PDF,故该方法考虑同一类的谱中的变化。特别地,值得注意的是,通过至少一个量的值表示的特征的存在和不存在都在根据本发明的方法中被考虑。

[0031] 根据本发明的方法的另一优点是在步骤e)中的最终结果是概率。因为概率是归一化的量,即概率是0到1之间的值,故样本谱属于第一参考类的概率可以与谱属于第二类的概率相比较。这是超越常规方法(其中分值被计算而不是归一化)的一个优点。

[0032] 此外,本发明使得定义用于接受或拒绝分类的标准成为可能。通常,谱将被分配给类,其为样本谱属于该类的概率最高的类。然而,这个概率可能是相对低的,例如0.2。该发明使得定义表示关于可接受的分类的最小值的阈值成为可能。关于所有类的最大概率低于阈值时不分类样本谱,即它们被分类到未知类中。当如在常规方法中的非归一化分值被使用时这种过程是不可能,因为在那种情况下阈值将是任意的。

[0033] 根据本发明的方法的优选方案中,谱数据包括喇曼谱、近红外谱、FT-IR 谱、频谱、MALDI MS谱或MALDI TOF-MS谱,优选地为MALDI TOF-MS 谱。

[0034] 如以上指出的,特别地在MALDI MS谱和MALDI TOF-MS谱中,属于同一类粒子的谱之间的变化是相当大的。因此,根据本发明的方法对于这些测量技术是特别有利的。

[0035] 在本发明的优选的实施方案中,获得自样本的参考谱和谱是单粒子的谱。

[0036] 这代表小数量粒子的极端情况。如上所述,根据本发明的方法能够说明在属于同一类的谱中的变化,其将特别地存在于单粒子谱中。

[0037] 如以上提及的,可在单粒子和两个或多于两个粒子以及甚至许多粒子下应用根据本发明的方法。

[0038] 优选地,该方法包括使用MALDI MS从样本获得单粒子谱,优选地使用MALDI TOF-MS。这使得粒子混合物的分析成为可能。取代分析获得自混合物的多种粒子的谱,即复合谱,获得的关于混合物的单粒子和各个粒子的谱在单粒子谱的基础上进行分类。

[0039] 例如,单粒子的MALDI质谱,诸如MALDI TOF MS谱,其通过将样本分成包含至多一个粒子的液滴而获得。通过这些形成了液滴气溶胶,其随后使用MALDI MS进行电离和分析。参考WO 2010/021548,其通过参考以其整体被合并,其描述了产生包含用于通过MALDI MS后续分析的至多一个细胞的液滴的系统和方法。通过使用这种设备,获得混合物的各个细胞的谱。这些谱可使用根据本发明的方法分类,以获得样本成分的分类。

[0040] 在优选的实施方案中,样本为生物样本。

[0041] 生物样本包括有机体(诸如微生物)分子,或由这种有机体产生的分子,如蛋白。特别地,获得自样本的参考谱和谱为生物样本的单粒子谱。属于同一类的生物样本的谱的变化是特别明显的,使得本发明的方法对于生物样本是特别有利的。这些变化例如来源于有

机体的生命周期和同一类（例如种）的有机体的个体之间的差异。

[0042] 根据本发明的另外优选的实施方案中，生物样本包括微生物且分类包括微生物分类。

[0043] 优选地，获得自样本的参考谱和谱为单微生物谱。这使能分析微生物的混合物。例如，微生物的混合物被稀释并随后被分为包含至多一个粒子的液滴，且随后使用MALDI MS从液滴获得谱，诸如以上描述。

[0044] 在根据本发明的优选的实施方案中，在参考类的特有谱特征的基础上，至少一个量被选择。

[0045] 例如，在步骤a)到c)中获得了类A、B和C的涉及类A的特有谱特征的量的值的PDF。例如，类A的谱示出了在位置X的作为特有谱特征的峰值。相应地，选择在位置X的谱强度作为值（分值）。计算所有类A、类B和类C的在位置X的这种强度。因为在位置X的峰值对于类A是特有的，故获得的类A的PDF将示出在X的高强度的高概率。通常，类B和类C将示出在位置X的高强度的较低概率。所以通过选择某一类的特有谱特征作为至少一个量中的一个，在特有的谱特征存在或不存在的基础上，该方法能够确定谱是否属于这个类。

[0046] 优选地，在优选地多于一个参考类的优选地多于一个的特有谱特征的基础上，选择多于一个的量。例如，对于类A、B和C，计算分值 I_1 和 I_2 ，其涉及参考类A的两种不同的谱特征。在另一实例中，计算分值 I_1 、 I_2 和 I_3 ，其中 I_1 涉及类A的谱特征， I_2 涉及类B的谱特征且 I_3 涉及类C的谱特征。

[0047] 优选地，在参考类的累积谱的基础上，特有谱特征被确定。例如，可通过求和属于同一类的各个参考样本的谱而获得参考类的累积谱。如上所述，这些累积谱将示出特征，其对于给定类是特有的，虽然它们可能不存在于类的每个单个谱中。值得注意的是，如步骤a)中所示，通过增加每组的参考谱和确定累积谱的特有谱特征，可使用多组参考谱确定量。

[0048] 优选地，特有谱特征包括在参考类的累积谱中的峰值。例如，将峰值寻找算法应用到累积谱以找到特有峰值的位置。然后，选择的量为在这些峰值位置或在包括这些位置的预定间隔处的强度。

[0049] 在本发明的优选的实施方案中，该方法包括针对样本谱进行计算，其中获得值 I_i ，其涉及谱特征量 Q_i ，根据下式获得给定值 I_i 时样本属于参考类 A^j 的概率 $P(A^j | I_i)$ ：

$$[0050] \quad P(A^j | I_i) = \frac{P(I_i | A^j)P(A^j)}{P(I_i | A^j)P(A^j) + \sum_{k \neq j} P(I_i | A^k)P(A^k)}$$

[0051] 其中 $P(I_i | A^j)$ 为参考类 A^j 与值 I_i 相关联的概率，且 $k \neq j$ 时的 $P(I_i | A^k)$ 是至少一个不同于参考类 A^j 的参考类与值 I_i 相关联的概率。

[0052] 特征 Q_i 优选地涉及特定类的特有特征。例如，量 Q_1 、 Q_2 与类 A^1 的特征相关联，且量 Q_3 、 Q_4 和 Q_5 与类 A^2 的特征相关联。

[0053] 值的注意的是，在量 Q_i 与指定类 A^j 相关联的情况下，也可使用这个量以获得不同类 $A^{k \neq j}$ 的概率。用以上实例说明的，量 Q_2 的值 I_2 是用于类 A^1 的特有特征的测量，然而，其可被使用以计算 $P(A^2 | I_1)$ ，即在给定 I_1 的测量结果时谱属于类 A^2 的概率。

[0054] 当类非常不同时，这将导致非常低的概率。然而，对于特征更相似的类这种计算可变得更重要。

[0055] 在方法的步骤c)中，每个参考类 A^j 的值 I_i 的PDF已经被计算。使用这个PDF值计算获

得的样本的值的概率 $P(I_i|A^j)$ 。这个概率代表如果样本属于参考类 A^j 时样本将产生测量值 I_i 的概率。根据贝叶斯 (Bayes) 定理,这可与概率相关,该概率为给定值 I_i 的测量结果时样本属于类 A^j 的概率。

[0056] 如在公式中所示,这种计算也需要在所有其它参考类内的 $P(I_i|A^k)P(A^k)$ 的总和。此外,使用在方法的步骤c)获得的PDF来获得每个k的 $P(I_i|A^k)$ 的值。

[0057] 优选地,假定概率 $P(A^j)$ 对于所有j(包括 $j=k$)是相等的。可以看出因此项 $P(A^j)$ 和 $P(A^k)$ 退出方程。如果我们假设样本谱属于参考类的概率对库的每个参考类是相等的,则这种方法被证明是合理的。当没有可获得的关于样本的内容的先验知识时,这种方法是特别有用的。然而,如果在属于特定类的样本的先验概率比属于其它类的样本的先验概率更高或更低的基础上,上述信息是可获得,那么根据本发明可对 $P(A^j)$ 使用不同的值。

[0058] 在根据本发明的优选的实施方案中,步骤b)包括针对每个参考谱确定与谱特征相关联的相同的至少两个量的值,且步骤e)包括将获得的关于所有量的概率结合到样本属于各个类的整体概率。

[0059] 如果在使用如上所述的本发明的方法的分类谱中仅使用了一个量,那么将获得关于每个参考类的对应于所述一个量的单概率。然而,如果谱的多于一个量被确定,即至少两个不同的量被确定,那么针对每个参考谱获得了对应于每个量的一些概率。根据本发明的这个实施方案,结合了这些概率与整体概率,例如通过逻辑OR和/或AND。这转换为各个量的各个概率的总和相加。

[0060] 这可被表示为关于每个参考类 A^j 的函数 $F(P(A^j|I_1), P(A^j|I_2), \dots, P(A^j|I_n))$ 。

[0061] 对于函数F,可用公式表示两个极端形式:

[0062] 所有n个特征存在于类 A^j

$$[0063] \quad F(P(A^j|I_1), P(A^j|I_2), P(A^j|I_3), \dots, P(A^j|I_n)) = \prod_{i=1}^{i=n} P(A^j|I_i)$$

[0064] n个特征中的任意一个存在于类 A^j

$$[0065] \quad F(P(A^j|I_1), P(A^j|I_2), P(A^j|I_3), \dots, P(A^j|I_n)) = 1 - \prod_{i=1}^{i=n} P(\neg A^j|I_i)$$

[0066] 其中

$$[0067] \quad P(\neg A^j|I_i) = \frac{P(I_i|\neg A^j)P(\neg A^j)}{P(I_i|A^j)P(A^j) + P(I_i|\neg A^j)P(\neg A^j)}$$

[0068] 除了以上两种极端形式,可使用任何其它多变量(自学习的)分类方法,诸如主要成分分析或支持向量机。

[0069] 我们注意到关于 $P(\neg A^j|I_i)$ 的表达式可用与如上述关于 $P(A^j|I_i)$ 相似的方法获得。

在这种情况下, $P(\neg A^j|I_i)$ 的PDF在步骤a)-c)中从参考类的组中获得。事实上,单PDF在所有不属于类 A^j (即 $k \neq j$ 的所有参考类 A^k)的参考谱的值 I_i 的基础上获得。例如,当选择三个参考类A、B和C时,在获得的关于参考类A和B的谱的值的的基础上,获得用于计算 $P(\neg C|I_i)$ 的PDF。

[0070] 在根据本发明的优选的实施方案中,该值涉及至少一个预定谱值的强度或在谱值

的预定范围内的强度。

[0071] 例如,值是其等于预定谱值的强度的标量,即沿谱的X轴的给定位置处的振幅。在另一个实例中,值等于谱值的预定范围的强度的总和。

[0072] 例如,在质谱中,该量为预定质量与电荷比值的峰值强度或预定范围内的质量与电荷比值的峰值强度。

[0073] 优选地,在参考类的特有谱特征的基础上选择预定谱值或预定谱范围。例如,选择谱值或谱值的范围作为其中参考类的总谱示出了谱中的峰值的值或范围。

[0074] 在优选的实施方案中,获得的值被归一化。

[0075] 归一化校正不同谱的强度中的变化,例如由于MS谱中的全部离子场的变化。例如,这种变化可由于随机效应或由于测量系统的漂移而发生。此外,当使用两个或多于两个的测量系统以获得用于该方法的谱数据时,预期不同系统获得的谱的强度存在差异。

[0076] 因此,归一化导致更可靠的算法。例如,通过将值除以整个谱或部分谱上的强度的总和而归一化该值。然后,根据方法的步骤c)使用归一化的谱以获得PDF。另外,归一化从样本谱获得的值以对比每个参考类的相应的量的PDF。

[0077] 在另外的优选的实施方案中,通过将预定的谱值范围上的各个谱乘以权重函数而确定值。

[0078] 例如,如果谱表示为向量,即表示谱的振幅的预定数量的标量的数组,则权重函数采用权重向量的形式。在这种情况下权重函数乘以谱对应于进行权重向量和谱向量的点乘。

[0079] 在实践中,对于给定的量,谱的某一区域比其它区域更相关。这可通过与权重函数相乘来解释。例如,整个谱乘以权重函数。然而,在大多数情况下仅选择谱的感兴趣的某一区域并乘以权重函数。我们注意到这等价于整个谱乘权重函数,其中权重函数在感兴趣区域之外具有0值。

[0080] 优选地,方法包括归一化值和与权重函数相乘。我们注意到可通过选择适当的权重函数结合这些步骤。

[0081] 如上所述,权重函数对谱的重要部分给出更大的权重。实际构成重要部分的谱部分可以基于各个类的累积谱。例如,将峰值寻找算法应用到累积谱以找到在总的谱中的峰值位置,并选择权重函数使得其相对于谱的其它部分,对在峰值处的和/或围绕峰值的位置处的强度设置更大的权重。例如,确定了在累积谱中的峰值并确定该峰值的宽度。然后,选择权重函数作为谱的拷贝,其中将在峰值宽度之外的权重函数的值设为零。可选地,权重函数还针对累积谱的基准进行修正。优选地,权重函数被归一化,使得其值的总和等于1,或使谱曲线下的面积等于1。

[0082] 固有地,以上实例的权重函数将把最大权重放在峰值的最大值处,且把更小的权重放在远离峰值的强度处。

[0083] 在根据本发明的另外的优选的实施方案中,该值涉及在至少两个预定谱值之间的强度或在至少两个预定谱范围内的强度之间的比值。

[0084] 这具有的优势是,该值对谱的某些区域间的关联是敏感性的。

[0085] 在可替换的实施方案中,该值涉及在预定的谱值的范围内的峰值的位置。

[0086] 如以上所提及的,该值可以是标量值。然而,在根据本发明的优选的实施方案中,

该值是向量。在这种情况下,与向量相关联的概率将是多变量概率密度函数 $P(I_1, I_2, \dots | A^j)$ 。例如,向量的元素涉及不同的量,即涉及不同谱特征的值。在极端情况下,仅一个向量被计算,其中向量的元素对应于所有被选择的谱特征。在这种情况下,使用逻辑“与”和/或逻辑“或”结合不同量的概率,例如结合使用如上所述的函数 $F(\dots)$ 的这些概率,是不必要的。

[0087] 在根据本发明的优选的实施方案中,针对第一组参考类且随后针对第二组参考类执行步骤d)和e),其中在第一组的参考类中的一个中的样本的分类基础上选择第二组。

[0088] 这定义了参考库的层次结构。首先在主类(第一组参考类)中分类样本,且随后在主参考类的子类中(第二组参考类)进行分类。这在计算上比使用包括全部最低等级类的参考库的分类更不精细。因此,根据本发明该方法有效地。

[0089] 此外,对于每个分类步骤,样本与包含相对较少的参考类的参考库相比较。因此,可选择各组参考类使得每组参考类示出少量重叠且分类质量提高。

[0090] 优选地,在有机体分类层次结构的基础上选择第一组和第二组。

[0091] 例如,当样本包括生物样本,使用可由用于样本分类的生物分类层次结构构成。例如分类组的选择是基于微生物的顺序、家族、基因、种类或菌株。

[0092] 在优选的实施方案中,步骤d)包括从样本中获得至少两个谱并确定这些至少两个谱的相同的至少一个量的值,且步骤e)包括将分类在同一类中的样本谱结合到累积谱,并对比该累积谱与通过结合各个类的参考谱获得的累积参考谱。

[0093] 此步骤提供了分类的可选的最终检查。从样本中获得多个谱。如上所述分类这些样本谱。随后,结合(例如增加)分类在同一类中的谱以获得累积样本谱。此累积样本谱与各个类的参考谱的累积谱比较。

[0094] 我们注意到,样本可包含成分的混合物,例如不同的分子或微生物的混合物。因此,这样的混合物的谱可被分类在不同的类中。在此情况下,组合每类谱。

[0095] 例如,样本包含粒子A和B。从该样本中,获得100个单粒子谱。使用本发明的方法,89个谱被分类为属于类A且11个被分类为属于类B。随后,对于类A,累积样本谱通过结合89个样本谱中的各个样本谱获得。另外,对于类B,累积样本谱使用11个谱中的各个谱获得。然后,类A和类B的累积样本谱与类A和类B的累积参考谱对比。这提供了最终检查。

[0096] 优选地,累积样本谱和累积参考谱的比较,将更多的权重放在不同于步骤b)的至少一个量的量上。换句话说,最终检查将更多的权重放在还没有被用于分类样本谱的谱特征上。优选地,仅考虑这些特征,即步骤b)的至少一个量没有在比较中使用。

[0097] 这确保了最终检查是基本上或者完全独立于分类步骤。

[0098] 优选地,使用第一组参考类执行方法,且随后使用第二组参考类执行该方法,其中在累积样本谱和累积参考谱之间比较的基础上选择第二组。

[0099] 定义了层次分类程序。该样本首先使用第一参考类的库的分类。该分类使用累积样本谱和累积参考谱证实。如果该比较导致可进一步在子类中划分该样本的结论,则样本随后与第二库比较。

[0100] 例如,比较揭示了相比于累积参考谱,附加峰值存在于样本的累积谱。这表明样本可含有粒子的混合物,例如属于不同的类的分子或微生物。因此,样本随后与第二组参考类比较。

[0101] 本发明还涉及创建用于在谱数据的基础上对在至少两类中的一类中的样本分类的数据库的方法,所述方法包括:

[0102] a) 获得至少两组用作参考谱的第一谱,每个组包括属于同一类的参考样本的谱;

[0103] b) 确定关于每个参考谱的相同的至少一个涉及谱特征的量的值;以及

[0104] c) 在确定值的基础上,关联概率到量的不同的值。

[0105] 此外,本发明涉及在使用以上方法获得的数据库的谱数据的基础上,用于对在至少两类中的一类中的样本分类的方法,该方法包括:

[0106] d) 从样本获得谱,并确定此谱的相同的至少一个量的值;以及

[0107] e) 基于从数据库获得的概率和获得的谱的量的值,针对所述至少两类中的每类计算样本属于该类的概率。

[0108] 本发明还涉及一种计算机程序,当在计算机上执行所述计算机程序时,执行上面描述的方法之一的步骤,以及还涉及包括这样的计算机程序的数据存储介质。

[0109] 此外,本发明涉及一种系统,其用于在谱数据的基础上分类样本,所述系统包括:

[0110] -装置,其设置为从样本获得谱;

[0111] -分析装置,其设置为执行如上所述的根据本发明的方法。

[0112] 如所描述的用于在谱数据的基础上分类样本的方法的相同优点和影响适用于创建数据库的方法、使用根据本发明的数据库、计算机程序、存储介质和系统的方法。特别地,结合用于在谱数据的基础上分类样本的方法的所描述特征可与创建数据库的方法、使用根据本发明的数据库、计算机程序、存储介质和系统结合的方法。

[0113] 在根据本发明的系统的优选的实施方案中,该系统包括设置为从样本获得单粒子谱的装置。例如,该系统包括设置为创建样本的液滴使得液滴包含至多一个粒子的装置,以及设置为获得谱的装置被设置为从各个液滴获得谱。

附图说明

[0114] 在本发明的优选的实施方案的基础上阐述本发明的另外的优点、特征和细节,其中参照附图,在附图中:

[0115] -图1示出了谱(P),其从获得自源于单独立种群的样本的大量单粒子谱累积得到,同时示出了基本的单粒子谱的一些典型实例(Q、R、S、T);

[0116] -图2示出了根据本发明的包括其子系统的系统的示意性概略图;

[0117] -图3图示了原始谱的峰形的实例(左图)和归一化的且基准校正后的峰形(右图);

[0118] -图4示出了谱特征的分值的PDF的实例,其分别针对包含特征诱导物质($P(I|A)$)的粒子集合(种类A)和不包含这种物质($P(I|\neg A)$)的集合(种类B);

[0119] -图5基于在图4中给出的概率密度函数,示出了特征诱导物质存在/不存在的概率,其作为单电离事件谱特征强度的函数;

[0120] -图6示出了包含粒子的种类A的参考,其通过累积用于估算在 $M/Z=5689$ 和 $M/Z=8339$ 的峰值的PDF的单粒子谱进行汇集,通过竖线表示;

[0121] -图7示出了包含粒子的种类B的参考,其通过累积用于估算在 $M/Z=2187$ 和 $M/Z=3558$ 的峰值的PDF的单粒子谱进行汇集,通过竖线表示;

[0122] -图8示出了含有两种有机体(种类A和种类B)的混合物的样本中记录的单电离事

件谱的序列；

[0123] -图9示出了单粒子谱混合物的累积谱(顶部)、分配到种类A类的单粒子谱(中间)和分配到种类B类的单粒子谱(底部)；

[0124] -图10示出了分配给种类A类的粒子累积谱和源于独立的种类A的粒子累积谱之间的比较；

[0125] -图11示出了分配给种类B类的粒子累积谱和源于独立的种类B的粒子累积谱之间的比较；

[0126] -图12示出了基于这些菌株中记录的MALDI质谱的95个金黄色葡萄球菌菌株的集合的簇聚图(clustergram)的实例；

[0127] -图13示出了根据本发明的层次分类方案。

具体实施方式

[0128] 在示例性实施方案中,根据本发明的系统2(图2)被设置为由用于生物材料检测的气溶胶创建谱,诸如在空气中使用MALDI TOF MS的细菌。本系统和经典的MALDI仪器之间的差异是气溶胶束发生器4、12、14的入口和样本制备,其中基质成行加入。其以这种方式被构造,这种方式使各个气载粒子在气溶胶束6输入系统。这开启了分析各个粒子的可能性,各个粒子诸如细菌、病毒或一定大小的其它生物材料的气溶胶的混合物。例如,该系统的医疗应用是可能的。系统控制器子系统8被用于从子系统设置、控制、测量、记录和监测参数。

[0129] 针对样本的气溶胶的制备,系统2还包括样本制备单元10,其尤其用于样本的稀释,第一部分12、空气动力学透镜14以及喷嘴和撇取器4的部件。

[0130] 激光子系统16包括由触发电子设备22控制的电离激光18和电离光学器件20。触发电子设备22连接到系统控制器8,和连接到设置有光电倍增管光学器件26的光电倍增管检测器24。此外,在子系统16中提供检测激光28和检测光学器件30。

[0131] 质谱分析子系统32包括离子源34,其是气溶胶使用来自激光18的激光束进行电离的位置。MS子系统32还包括离子偏转器36和MS检测器38。在容器和飞行管40中包含离子源34、离子偏转器36和MS检测器38。借助于压力表42、涡轮分子泵44、46和前级泵48的组合保持此管真空。离子源34和离子偏转器36连接到离子源控制50,其被连接到触发电子设备22和信号处理和数据管理子系统52。

[0132] 子系统52包括连接到MS检测器38的数字转换器54。经由数据管理模块56,数字转换器54连接到系统控制器8。系统控制器8连接到分析器58。

[0133] 系统2还包括电源单元60。虽然在该图中没有连接被绘制,但该单元为不同组件提供电力。此外,系统机架62和空气/水冷却装置64提供用于壳体系统2。

[0134] 在系统2上用气溶胶化的细菌和蛋白粒子实现原理证明。在这个示例性实施方案中,使用分析器58执行根据本发明的方法步骤。

[0135] 确定单粒子物质的存在

[0136] 由于在MS中的单电离事件谱的高度随机性质,在对应于分析物分子的质量的位置处的单粒子谱强度 I 只能被认为是对于该分析物分子 A 存在于粒子中的概率 $P(A|I)$ 的度量。

[0137] 依赖于贝叶斯定理和概率密度函数 $P(I|A)$ 量化这个概率,其中概率密度函数 $P(I|A)$ 指定给出分析物分子 A 存在时特征强度 I 的概率,被量化的概率如下:

$$[0138] \quad P(A|I) = \frac{P(I|A)P(A)}{P(I|A)P(A) + P(I|\neg A)P(\neg A)}$$

[0139] 其中

[0140] $\neg P(A)$ 是包含物质A的粒子的部分；

[0141] $\neg P(I|\neg A)$ 是不包含物质A的粒子的强度I的概率密度函数；以及

[0142] $\neg P(\neg A)$ 不包含物质A的粒子的部分；

[0143] 因此，只要函数 $P(I|A)$ 和 $P(I|\neg A)$ 是已知的，则该物质A存在于粒子中的概率可由测量的由物质A的存在导致的谱特征的强度推导出。

[0144] 因此，分类单粒子谱所需的参考信息必须包含关于所有预期特征PDF。

[0145] 基于其单粒子质谱确定单细胞的源

[0146] 微生物(诸如细菌)以存在许多产生MALDI-MS的可区分的特征的物质为特征。依赖于有机体(细菌、病毒等)和有机体的状态(植物细胞、孢子)，这个数目可在少至5到多达50或甚至更多之间变化。

[0147] 其中一些特征对于手边的有机体所属的属是代表性的，其中一些是关于种类，其中一些是关于菌株。

[0148] 对于每个特征 Q_i ，定义概率密度函数 $P(I_i|A^j)$ ，其表示测量包括物质 A^j 的样本的谱的特征 Q_i 的强度 I_i 的概率。

[0149] 此外，pdf $P(I_i|\neg A^j)$ 被定义表示测量当样本不包括物质 A^j 时的特征 Q_i 的值 I_i 的概率。

[0150] 根据贝叶斯定理，这两个pdf与 $P(A^j)$ 和 $P(\neg A^j)$ 相结合以获得给定测得的特征强度 I_i 时特征诱导的物质 A^j 存在的概率 $P(A^j|I_i)$ 。

[0151] 值得注意的是，在该实例中，参考了“物质 A^j ”而不是如上所述的“类 A^j ”。事实上，该实例代表极端情况，其中类 A^j 包括单个物质。例如，该物质包括特定蛋白。

[0152] 为估算单细胞MALDI-MS谱源自属于属、种或菌株的有机体的概率，关于各个特征的概率必须结合到关于特征组合的概率，这些特征对于属、种或菌株是代表性的，且存在于谱中，使用函数 $F(P(A^j|I_1), P(A^j|I_2), P(A^j|I_3), \dots, P(A^j|I_n))$ ：

[0153] $P(A^j|I_1, I_2, I_3, \dots, I_n) = F(P(A^j|I_1), P(A^j|I_2), P(A^j|I_3), \dots, P(A^j|I_n))$ 假设物质 A^j 存在于样本中的概率。

[0154] 此外，对于每个物质 A^j 的概率可任选地组合。例如，结合关于一组蛋白中的每个蛋白的概率以获得给定微生物的总概率。

[0155] 然而，不是所有的在从大量单细胞谱累积的谱中能够确定的特征都需要存在于每个单细胞谱中，见图1。

[0156] 例如，在其生命周期的不同阶段(裂变之前不久、刚裂变之后等)，该细胞可表达不同的蛋白。因此，当累积谱呈现出在其生命周期的各个阶段由有机体产生的所有的可电离的蛋白时，单个有机体的谱只能呈现那些在被分析的特定细胞的特定阶段中产生的被表达的蛋白。

[0157] 严格来说，当关于在生命周期中的阶段的蛋白表达的相关性的知识是可用的时，

有可能制定函数 $F(\dots)$ 。可惜的是,一般这种信息是不可用的。

[0158] 然而,如上面所指出的,除了在生命的阶段中由差异诱导的差异之外,存在许多更多的关于单细胞谱之间的易变性的原因。因此,即使当上述信息将是可用的时,在其基础上的函数 $F(\dots)$,将是一个最好的估算值。

[0159] 忽略在生命阶段和单细胞蛋白表达之间的关系的信息的概率,函数 $F(\dots)$ 的两个极端的形式可用公式表示为:

[0160] -所有特征存在

$$[0161] \quad F\left(P(A^j | I_1), P(A^j | I_2), P(A^j | I_3), \dots, P(A^j | I_n)\right) = \prod_{i=1}^{i=n} P(A^j | I_i)$$

[0162] -任何特征存在

$$[0163] \quad F\left(P(A^j | I_1), P(A^j | I_2), P(A^j | I_3), \dots, P(A^j | I_n)\right) = 1 - \prod_{i=1}^{i=n} P(\neg A^j | I_i)$$

[0164] 其中

$$[0165] \quad P(\neg A^j | I_i) = \frac{P(I_i | \neg A^j)P(\neg A^j)}{P(I_i | A^j)P(A^j) + P(I_i | \neg A^j)P(\neg A^j)}$$

[0166] 确定概率密度函数 $P(I | A)$ 和 $P(I | \neg A)$

[0167] 引言

[0168] $P(I | A^j)$ 和 $P(I | \neg A^j)$ 的PDF可通过分别包含物质 A^j 和不包含物质 A^j 的足够大的组的参考粒子的估算来近似。

[0169] 为了确定所有特征 Q_i 的 $P(I_i | A^j)$,一组单粒子谱必须使用已知含有产生特征 Q_i 的物质 A^j 的粒子。

[0170] 特征 Q_i 的特点是,在与此特征相关联的质量处的强度的期望值,其比相邻质量区域处强度的期望值更大。因此,如果有大量的单粒子谱相加,则累积谱将以峰值(或在聚合物的情况下的峰值的组合)。

[0171] 以累积谱中的峰值为特征表征的PDF,可简单地通过以下进行估算,记录在关于每个单粒子谱的关联质量处的振幅,将这些振幅合并到离散振幅容器中,以及将在每个容器中的分值除以记录的单粒子谱的总数。

[0172] 给定足够大数量的单粒子谱和足够精细的容器分配,这将获得对概率密度函数的充分的估算。

[0173] 提取特征

[0174] 对于任何实际的质谱仪,在累积谱中的峰值将具有有限的宽度。因此,当记录了每个特征的单粒子振幅时,必须考虑到特征可能发生的质量区间的有限宽度。

[0175] 这里,在累积谱中的峰值的形状是用来考虑有限峰值宽度的效果。为此目的,峰值可能发生的质量区间的累积谱的形状被复制,在区间边缘针对基准振幅校正并进行归一化,使得 $AUC = \int I(MZ) dMZ = 1$ 。

[0176] 图3图示了峰值形状的实例。

[0177] 因此,可定义特征形状函数 $S_{\text{特征}}$ 使得:

$$[0178] \quad S_{\text{特征}}(MZ) = \begin{cases} 0 & \text{for } MZ < MZ_{\min} \vee MZ > MZ_{\max} \\ F(I_{\text{谱}}(MZ)) & \text{for } MZ_{\min} \leq MZ \leq MZ_{\max} \end{cases}$$

[0179] 其中, $F(I_{\text{谱}}(MZ))$ 的定义方式为:

$$[0180] \quad \int_{MZ_{\min}}^{MZ_{\max}} F(I_{\text{谱}}(MZ)) dMZ = 1 \quad F(I_{\text{谱}}(MZ_{\min})) = 0 \quad F(I_{\text{谱}}(MZ_{\max})) = 0$$

[0181] 最后, 可因此定义单粒子谱的强度分值 IS 为形状函数和单粒子离子强度谱的乘积在整个质谱仪的质量范围上的积分:

$$[0182] \quad IS_{\text{特征}} = \int_{MZ_{\min} MS}^{MZ_{\max} MS} S_{\text{特征}}(MZ) I_{\text{单粒子}}(MZ) dMZ$$

[0183] 由于以有限的采样率记录离子强度, 上述积分被替换为离散的总和:

$$[0184] \quad \hat{IS}_{\text{特征}} = \sum_{i=i(MZ_{\min} MS)}^{i=i(MZ_{\max} MS)} S_{\text{特征}}(MZ_i) I_{\text{单粒子}}(MZ_i) \Delta MZ_i$$

[0185] 因此强度分值可通过两个向量的内积来表示:

$$[0186] \quad IS_{\text{特征}} = \bar{S}_{\text{特征}} \cdot \bar{I}_{\text{单粒子}}$$

[0187] 其中

$$[0188] \quad \bar{S}_{\text{特征}} = S_{\text{特征}}(MZ_i) \quad \text{对于 } i=i(MZ_{\max} MS) \text{ 到 } i=i(MZ_{\min} MS)$$

[0189] 是特征选择向量, 且

$$[0190] \quad \bar{I}_{\text{单粒子}} = I_{\text{单粒子}}(MZ_i) \Delta MZ_i \quad \text{对于 } i=i(MZ_{\max} MS) \text{ 到 } i=i(MZ_{\min} MS)$$

[0191] 是单粒子谱能量向量。

[0192] 估算 $P(I|A)$

[0193] 通常, 微生物的MALDI谱的特点是有限数目的峰值, 在2000Da和 20000Da之间的质量区域中通常是10到50个。这些峰值表示诸如蛋白、肽等物质的存在。

[0194] 这些物质中的一些对于大群体有机体(属)是代表性的, 它们中的一些对于小群体(种类)是代表性的, 且它们中的一些可能甚至针对单个有机体(菌株)。

[0195] 假设细菌的参考集合具有完全特有的峰值图案, 则可构造概率密度函数库。这个库将实现依照特有物质存在于粒子当中的概率对含有单个细菌的未知粒子的分类。

[0196] 需要在参考库中表示的特征的整个集合是所有有机体的所有特征的联合:

$$[0197] \quad S_n^m \text{ 其中 } \begin{cases} m = 1 \dots M_{\text{隔离}} \\ n = 1 \dots N_{\text{特征}}^m \end{cases}$$

[0198] (值的注意的是, 每个独立特征的数量都依赖于所涉及的有机体)

[0199] 当全部特征集合被转换成一组特征选择向量 S_n^m 时, 这些向量可被转换成特征选择矩阵:

$$[0200] \quad \overline{\overline{S}} = \left(\left(\overline{S}_1^1 \right) \left(\overline{S}_2^1 \right) \cdots \left(\overline{S}_1^m \right) \left(\overline{S}_2^m \right) \cdots \left(\overline{S}_N^M \right) \right)$$

[0201] 其中, NS列等于出现在特征集合的总特征数量, 且 M_{MZ} 行等于在单粒子质谱中的时间(质量)样本的数量。

[0202] 将此特征选择矩阵与单粒子谱能量向量相乘, 产生强度分值IS向量, 其保存所有特征的强度分值:

$$[0203] \quad \overline{IS}_{\text{单粒子}} = \overline{I}_{\text{单粒子}} \cdot \overline{\overline{S}}$$

[0204] 如上面所指出的, 通过处理足够大组的单粒子谱, 其由源于单个独立种群的微生物产生, 有可能创建概率密度函数。利用上述特征选择矩阵, 与特定独立种群 m 关联的概率密度函数 $P^{m, 1 \cdots NS}$ 可一次性被确定。

[0205] 估算 $P(I|\neg A)$

[0206] 涉及存在于考虑的有机体中的物质的特征的概率密度函数, 将与不涉及存在于有机体中的物质的那些特征的概率密度函数显著不同。作为示例, 图4示出来自包含特征诱导物质的粒子集合和不包含这种物质的粒子集合的特征的PDF实例。

[0207] 图4示出了包含该物质的集合的PDF的振幅超出不包含该物质的集合的PDF的振幅, 特征强度高出约0.08, 反之亦然。

[0208] 在这个只有两个可能粒子类型的简单的实例中, 一种包含诱导特征 Q_1 的物质 A^1 , 且一种包含物质 A^2 , 不包含 A^1 即 $\neg A^1$ 的集合的PDF等于包含 A^2 的粒子的PDF:

$$[0209] \quad P(I|\neg A^1) = P(I|A^2)$$

[0210] 如上面所指出的, 给定概率密度函数 $P(I|A)$ 和 $P(I|\neg A)$, 能够确定单个粒子包含特征诱导物质的概率。

[0211] 为了计算物质A存在的概率, 除了PDF以外还必须提供遇到A的概率 $P(A)$ 。在这个简单的实例中, 只有两种类型的粒子被考虑, 包含A的粒子和不包含A的粒子。因此遇到A的概率假设等于 $P(A) = P(\neg A) = 0.5$ 。

[0212] 复杂粒子集合

[0213] 如果更复杂粒子集合被考虑, 即带 n 种可能的粒子类型, 其中粒子型 1 中含有物质 A^j , 乘积 $P(I|\neg A^j)P(\neg A^j)$ 等于

$$[0214] \quad P(I|\neg A^j)P(\neg A^j) = \sum_k P(I|A^k)P(A^k) \text{ 其中 } k \in \{1, 2, 3, \dots, n\} \setminus j$$

[0215] 因此, 如在二元混合物的情况下, 在复杂的混合物中遇到的概率 $P(A^j)$ 等于 $P(A^i) = 1/n$ 。

[0216] 二元混合物的分析

[0217] 基于图4中呈现的特征PDF和贝叶斯定理, 作为(单粒子谱)特征强度的函数的特征

诱导物质A存在(或不存在)的概率可被推导。图5示出了这个概率。

[0218] 根据该图的临界强度(由图5中在大约0.08的细的竖线表示)之上,特征诱导物质存在的概率始终大于该物质不存在的概率。

[0219] 因此,高于这个强度,这个特征是这种物质的存在的可靠的测量。

[0220] 低于这个强度,概率和特征强度之间的关系是模糊的。因此,没有针对特征诱导物质的存在的判断可被执行,且必须标记 $P(A|I)$ 为未定义的。

[0221] 为了说明上面描述的方法的功能,预先记录的单粒子谱的(二元)混合物可被汇集。为拆散此混合物,需要定义两种粒子类型、特征且需要估算相应的PDF需要。

[0222] 图6和图7示出了用于估算这些特征PDF的粒子的累积谱。在这种情况下,选择两种粒子类型的两个清晰的峰值为特征(用在图6和图7中的竖线表示)。

[0223] 图8示出了从两个批次的预先记录的单粒子谱随机选择的单粒子谱的序列(值得注意的是,针对混合物使用与估算PDF使用的不同批次)。此外,图8示出了在单粒子谱中的极端易变性和参考谱与单粒子谱的相关性的明显缺乏。

[0224] 在这个实例中,单粒子谱根据以下规则进行分类:

[0225] -根据如以上描述的“任何特征存在”规则,源自于参考独立种群中的一个的单粒子谱 $P_{\text{总}}$ 由两个独立种群中的每个的两个单特征概率推导出。

[0226] -对于明确的特征概率(见图5),只有当特征强度 I 超过了临界特征时,才分配关于这个特征的特征概率,否则,特征概率为“未定义”。

[0227] 谱被分配给类(独立种群)1,此时

$$[0228] \quad P_{\text{总}}^1 > P_{\text{总}}^2 \wedge P_{\text{总}}^1 > P_{\text{min}}$$

[0229] 且分配给类2,此时

$$[0230] \quad P_{\text{总}}^2 > P_{\text{总}}^1 \wedge P_{\text{总}}^2 > P_{\text{min}}$$

[0231] 其中 P_{min} 是需要超过的阈值概率以便分类谱。在这个实例中使用阈值概率 $P_{\text{min}} = 0.9$ 。

[0232] 这些不超过阈值概率的谱(例如因为两个特征概率为未定义的)被分配到“未知”类。

[0233] 当分配给类中的一个的谱被累积时,看起来两个类的合成谱彼此明显不同,且与混合物的累积谱明显不同,参见图9。

[0234] 图9中的检查示出了在峰值位置(由竖线表示),该峰值仅保留对应于参考独立种群的谱。这表明该分类算法在某种意义上说是有效的,它能够选择那些促进特征的谱,且能够忽略那些没有促进特征的谱。

[0235] 因此,该算法能够选择促进特征的单粒子谱,且属于不同参考类的、源于基于特征的选择的累积谱源彼此显著不同。然而,在某些特定情况下,这将不能确切证明由算法产生的两个类实际对应于该参考类。

[0236] 因此,作为最终检查,根据不是用于分类的其它信息,比较可在样本谱和参考谱之间进行。

[0237] 分类结果的确认

[0238] 为确认被分配到不同的类的单粒子谱确实对应于参考谱,每类中的累积单粒子谱

可与参考谱进行比较。

[0239] 图10和图11示出了两个粒子类的这种比较。这些附图示出了除了用于选择的峰值外,还有存在于汇集自经分类的粒子的谱中的对应于参考谱中的峰值的其它峰值。

[0240] 另外,从经分类的粒子累积的谱不包含不存在于参考谱中的峰值。

[0241] 这两个属性都清楚地表明,经分类的谱的汇集谱实际上对应于参考谱。因此,产生的经分类的谱中的粒子实际上包含对应于包含在用于参考谱的独立种群中的细胞的细胞(的片段)。

[0242] 层次分类程序

[0243] 一般地,细菌的累积谱包含在介于10至40之间的某处的可区分的峰值。可能所有峰值的位置(按照它们的质量)可被用作单粒子分类的特征。

[0244] 峰值的适用性的必要条件是单粒子峰值强度的概率密度函数,其与在相同质量处的参考库中的其它独立种群的概率密度函数显著不同。

[0245] 对两个分布之间的差异/相等的方便的测量是Kolmogorov-Smirnov统计。Kolmogorov-Smirnov统计量化两个样本的两个经验累积分布函数之间的距离。如果该统计信息是足够大,这两个基本的概率密度函数可被认为是可区分的。

[0246] 由于质谱仪的分辨率有限,故参考库越拥挤,不同独立种群(部分地)重叠的峰值的概率越大。显然,不同的独立种群的峰值的重叠使得它们不适合用于分类。

[0247] 因此,根据参考库的库内容,仅在谱中出现的峰值的一部分可有效地用于单粒子分类。

[0248] 可区分的类的数量由所使用的峰值的数量来确定。可区分的类的数量用 2^N 来度量,其中N是存在于参考库中不重叠的峰值的数量。

[0249] 最初,当必须考虑所有可能的有机体时,参考库将十分拥挤。因此,不重叠的峰值的数量将比较低,且可被区分的类的数量将比较低。因此,当分析包含混合源的有机体的样本时,单一类包含多于一种种类的风险相对较高。

[0250] 然而,当分别保存单粒子谱时,可第二次分类它们。因为现在仅需要考虑参考,其属于这些谱被分配给的类,类特定的参考库可被构造,其明显没有原来的参考库拥挤。

[0251] 由于类特定的参考库较不拥挤,峰值之间重叠的概率降低。因此,一些在原始库中不适合的峰值,将适合于在类特定的库。

[0252] 因此,可进一步区分包含样本的混合物。

[0253] 因此,通过递归地分类混合物和重新定义参考库,包含混合源的有机体的样本可被区别为大程度的特异性。

[0254] 峰值存在和分类学之间存在关联。在微生物学的基础上,通过特定于微生物的顺序、家族、属、种或菌株的峰值使用,可激励峰值的选择。在这种方式中,实现层次分类方案。

[0255] 在图12中,层次分类方案的簇聚图实例被示出。在簇聚图中,在金黄色葡萄球菌的不同菌株的质谱中的峰值被描述。x轴表示m/z。不同菌株沿y轴作图。黑色区域对应在各个菌株的谱中的峰值。

[0256] 该簇聚图示出了包含特定种类的峰值的所有菌株。然而,在区域B和C中它们可使用菌株特定峰值区分。

[0257] 在示例性实施方案中,根据本发明的层次分类方案100(图13)接收如在步骤102中

获得的输入单粒子。随后,单粒子谱在步骤104中使用PDF 的第一参考组106被分类。在此分类的基础上,在类1、类2、…、类M 中分类样本谱。

[0258] 对于每组样本谱,在组中的谱和各个类的参考谱之间进行比较。其被示出为类2。在第一步骤108中,累积样本谱,即创建累积样本谱。在步骤110,计算累积参考谱。通过模式匹配的手段在步骤112比较累积样本谱和累积参考谱。

[0259] 在步骤114中,累积谱之间的匹配被计算。如果匹配高于最小阈值,则分类被认为是正确的,且在步骤116中该算法返回该分类。如果匹配低于阈值,即不被接受,则在步骤118中该算法检查累积样本谱是否是混合谱。例如,如果另外的峰值存在于累积样本谱中,其不存在于累积参考谱中,则累积样本谱被认为是混合谱。

[0260] 当谱不是混合谱时,该算法返回该信息,并在步骤120计算谱到最接近的相关有机体的分类学距离。换句话说,该有机体不属于库106的类。然而,在随后的步骤,谱可与不同的库比较。

[0261] 如果在步骤118中确定谱为混合谱,则该算法返回到步骤104,其中使用不同的库。这个新库包含混合谱的类的子类,在这种情况下为类2的子类。

[0262] 换句话说,在混合谱被观察的情况下,通过观察各个类的子类,该算法向下进入更深的层次。当在114中匹配被找到,这个程序停止。如果没有发现匹配,且没有观察到混合谱,则有机体不属于源类106。

[0263] 下面,根据算法示出了方法的步骤的伪代码。

概率密度函数的生成和选择

%生成特征选择向量%

遍历独立种群

[0264] **遍历单粒子谱文件**

读取单粒子谱为强度数组

计算总离子数量

用总离子数量归一化谱

将归一化的谱加入到总谱

结束循环

显示总谱

标记谱特征（或手动或自动）

存储特征/独立种群的识别

遍历特征

提取特征形状函数

存储特征形状函数为特征选择向量

结束循环

结束循环

%%%

%生成特征概率密度函数%

[0265]

遍历独立种群

遍历单粒子文件

读取单粒子谱为强度数组

提取属于所有独立种群的所有特征的特征强度

存储对应于特征强度数组的所有特征的特征强度

结束循环

遍历特征

将特征强度数组的元素放入适当间隔的容器中

返回特征强度频率数组中的每个容器的元素数量

将特征强度频率数组除以处理的单粒子谱的数量

结束循环

结束循环

%%%%%%%%%

%选择特征概率密度函数%

遍历特征概率密度函数

生成累积概率密度函数

结束循环

遍历独立种群

根据特征/独立种群识别表，为独立种群选择特征“天然的”

遍历“天然的”特征

遍历“其它的”特征

确定并存储关于天然的特征和其它特征的每个组合的

Kolmogorov-Smirnov 统计量

结束循环

[0266] 确定关于每个“天然的”特征的最小 Kolmogorov-Smirnov 统计量，并存储为 *minKS*

结束循环

选择 N 个（可由用户指定但至少为 1 的数量）具有最大 *minKS* 的特征

存储关于每个独立种群的特征选择

结束循环

%%%%%%%%%

混合分析

%单粒子分类%

等待循环

读取单粒子质谱

归一化单粒子质谱

确定（选择的）特征强度

遍历（选择的）特征

遍历独立种群

确定 $P(I|A_n^m)P(A)$ 和 $P(I|\neg A_n^m)P(\neg A)$ 的贡献

结束循环

确定 $P(A_n^m|I)$

结束循环

遍历独立种群

使用“天然的”特征列表，确定每个独立种群的 P

结束循环

分配谱到对应于参考的类中的一个或到未知类

结束循环

[0267]

遍历类

如果谱的数量大于最小数量

对分配给类的所有谱求和

平滑累积谱

重采样平滑谱

重采样谱减去基准

提取峰值

将峰值列表分类为属于所考虑的类（如属）的子类（如种）

如果分类成功

报告身份

否则

检查在类内的种的混合物是否可能存在

[0268]

如果混合物存在

产生用于混合物的进一步区分的新特征组

针对分配到被考虑的组的这些粒子执行新的单粒子分类

否则

确定累积谱到参考谱的分类学距离

结束

结束

结束

结束循环

%%%%%%%%%

[0269] 因此,通过优选实施方案描述了本发明。然而,应当理解的是,本公开内容仅仅是说明性的。提出结构和函数的各种细节,但通过所附权利要求中表示的、在本发明的原则内应理解的术语的通用含义在其中做出的变化延伸至完整范围。本说明书和附图应用于解释权利要求书。权利要求书不应解释为意指寻求保护的范围,应理解为权利要求书、说明书和附图中使用的词的严格的字面意义限定的仅用于解析权利要求书中发现的不明确。对于确定由权利要求书寻求保护的范围,应严格遵守等同于其中指定的元素的任何要素。

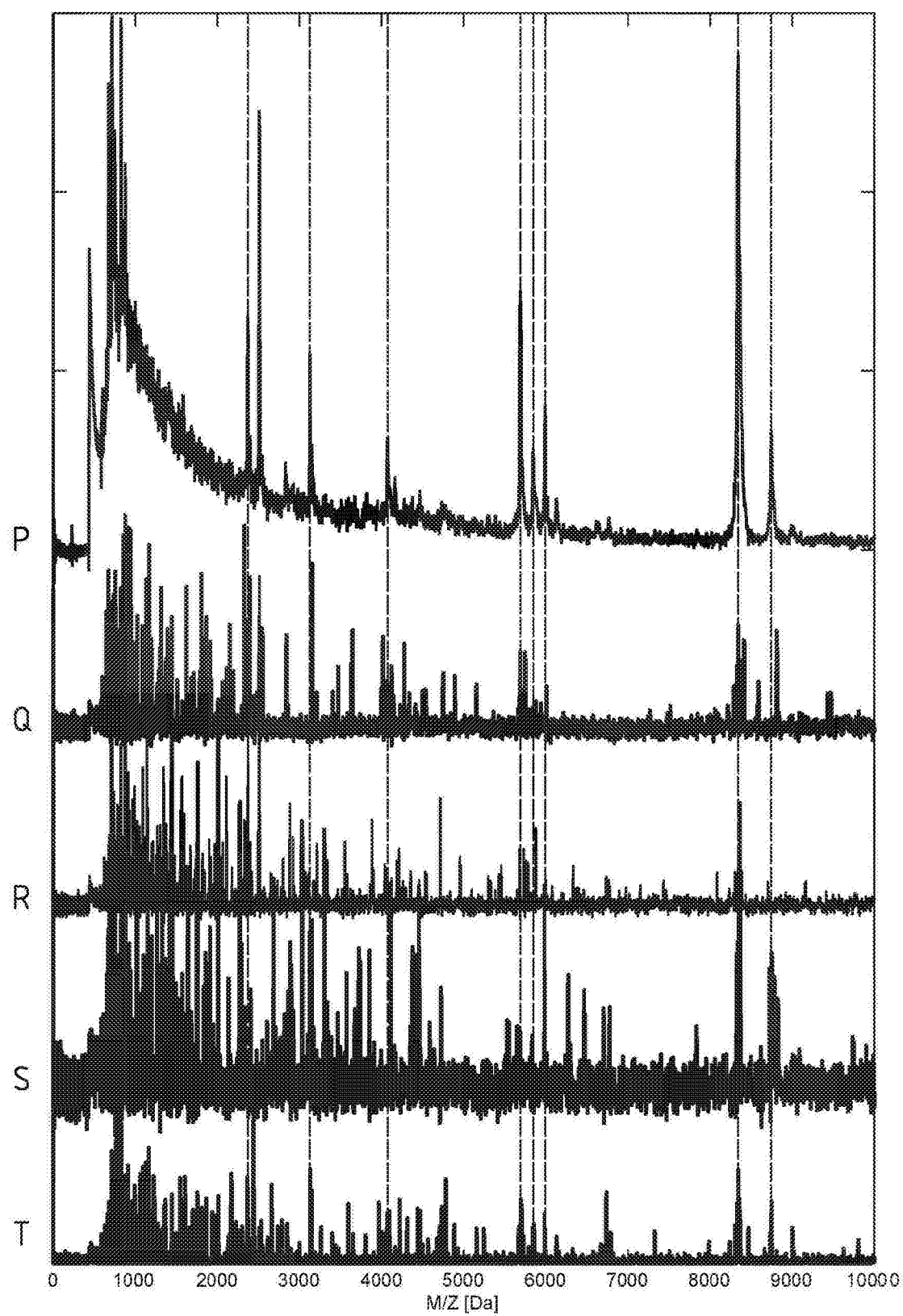


图1

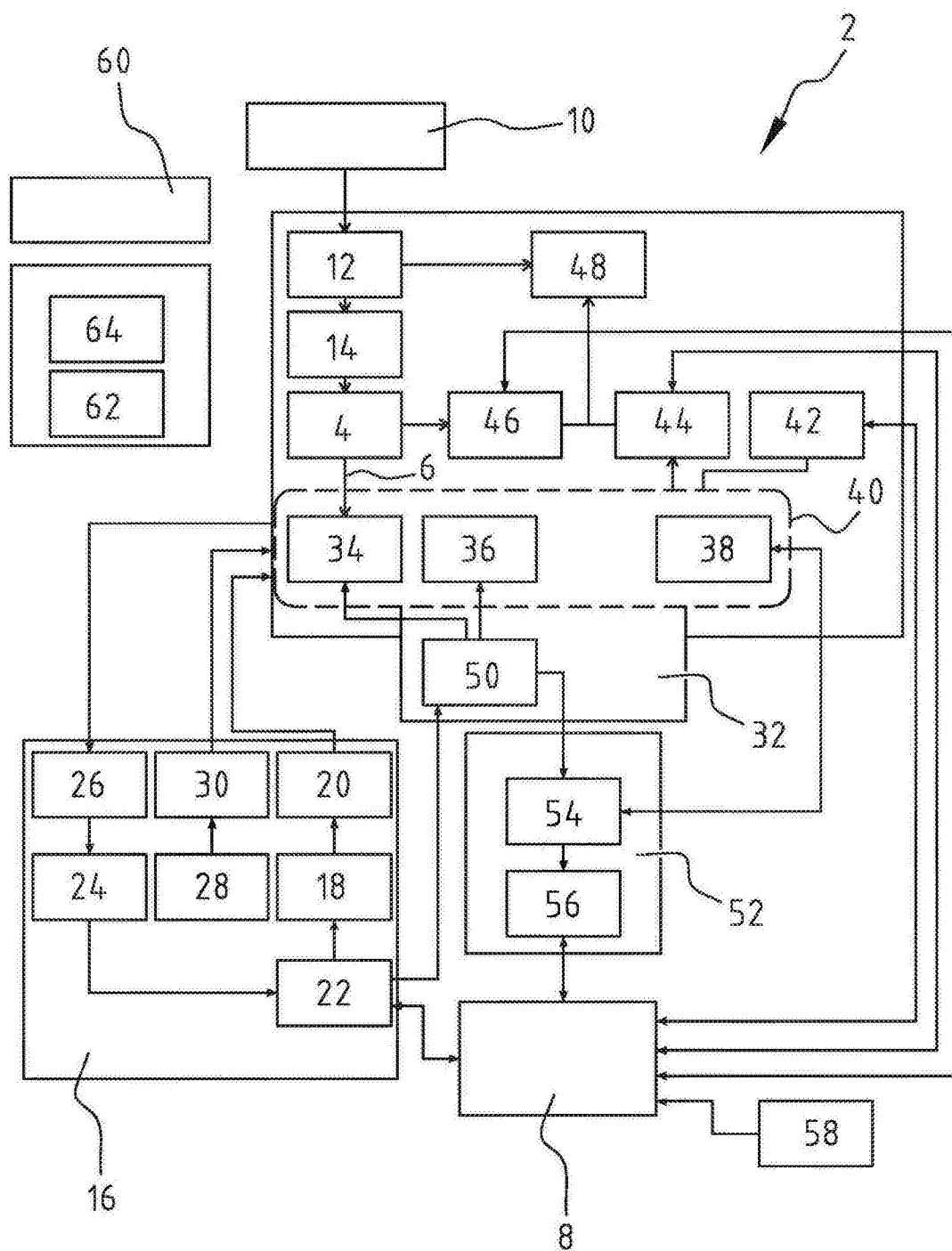


图 2

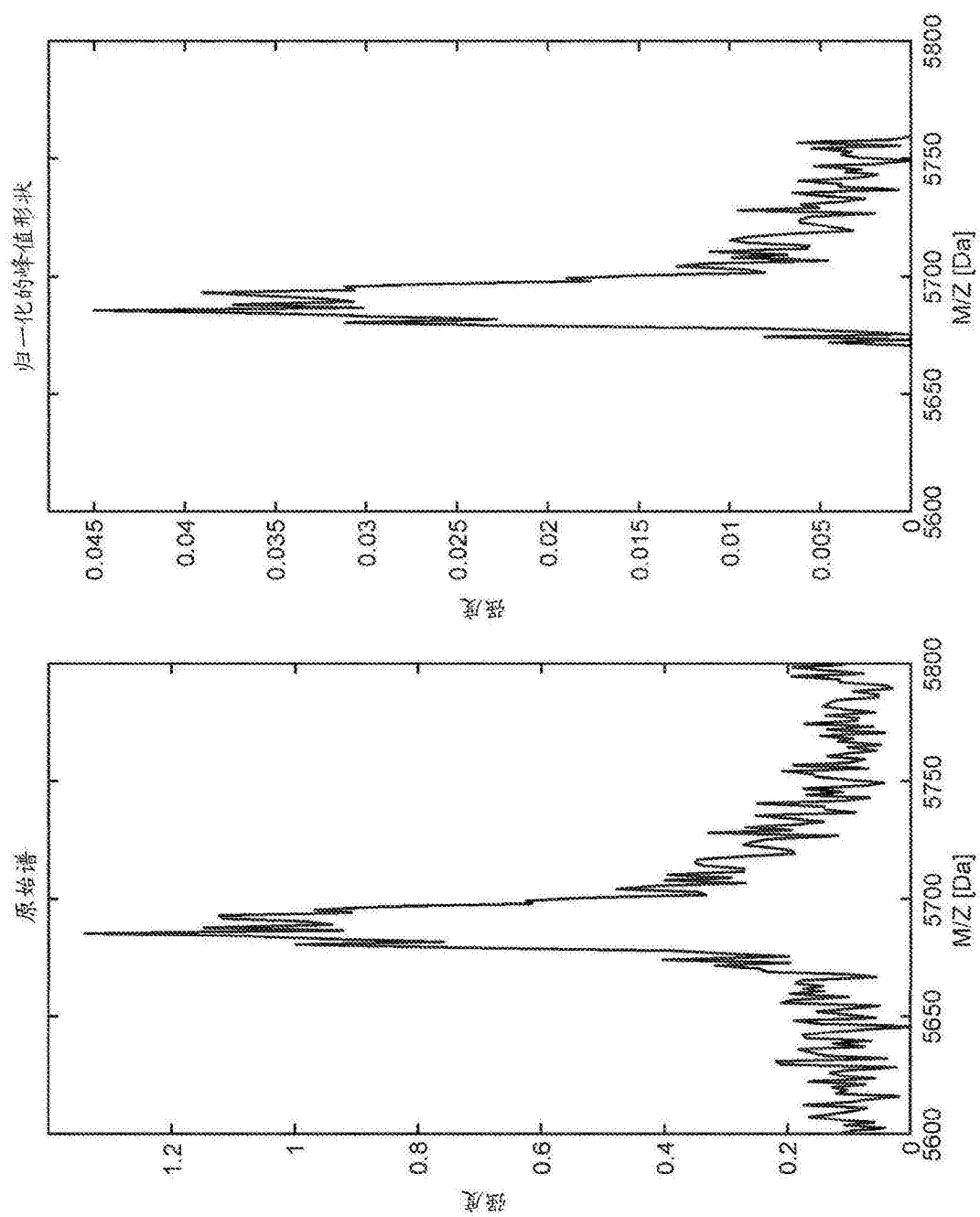


图3

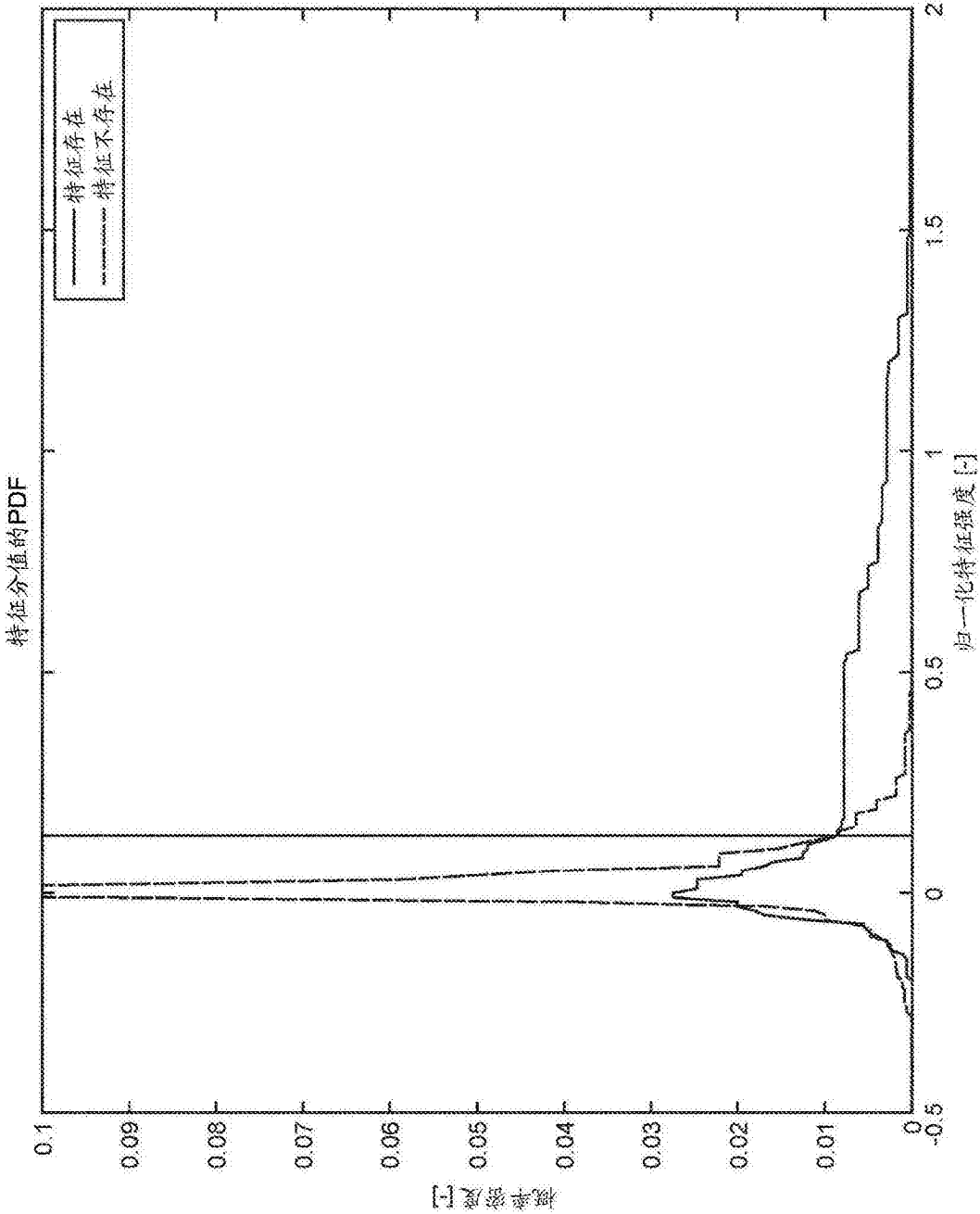


图4

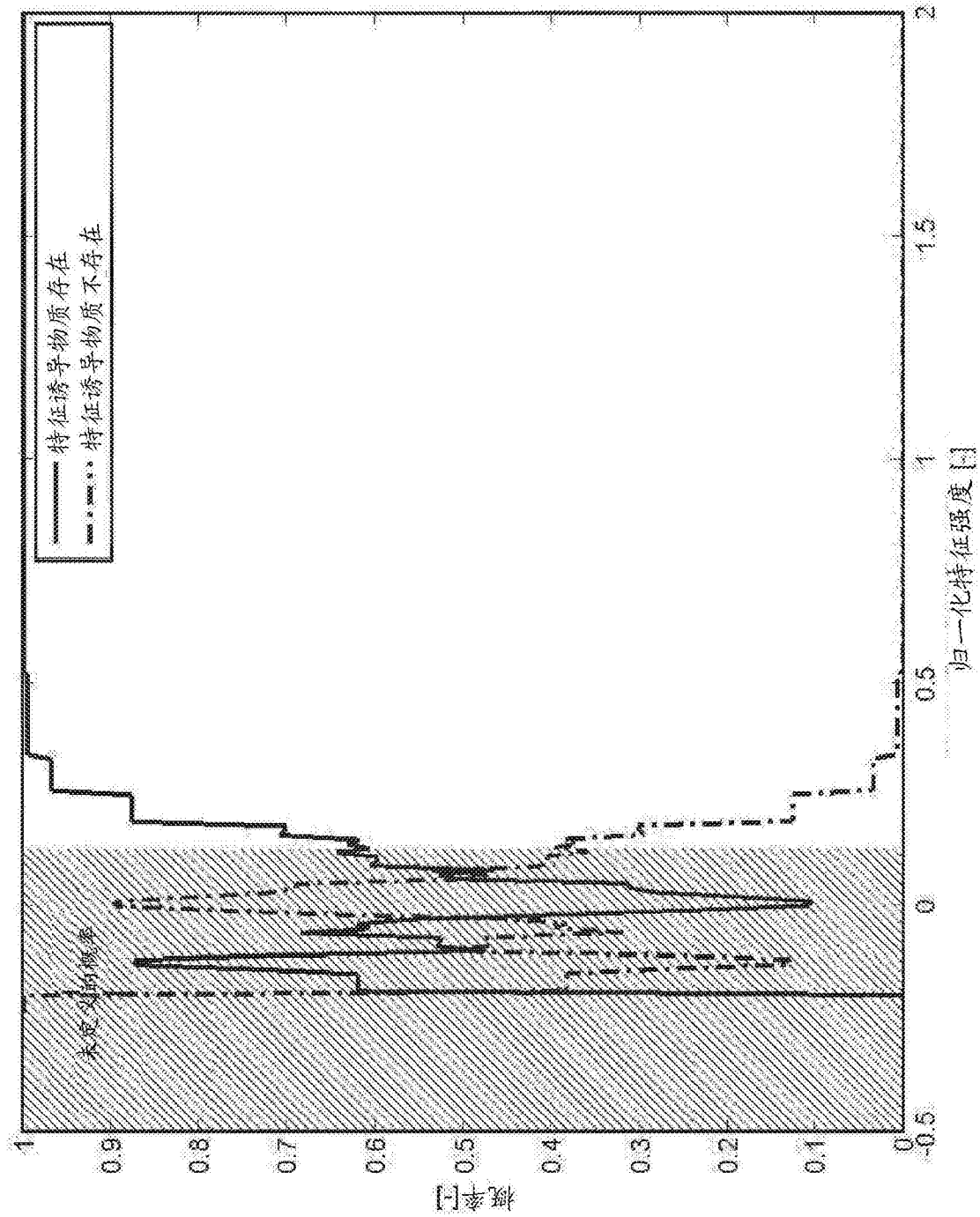


图5

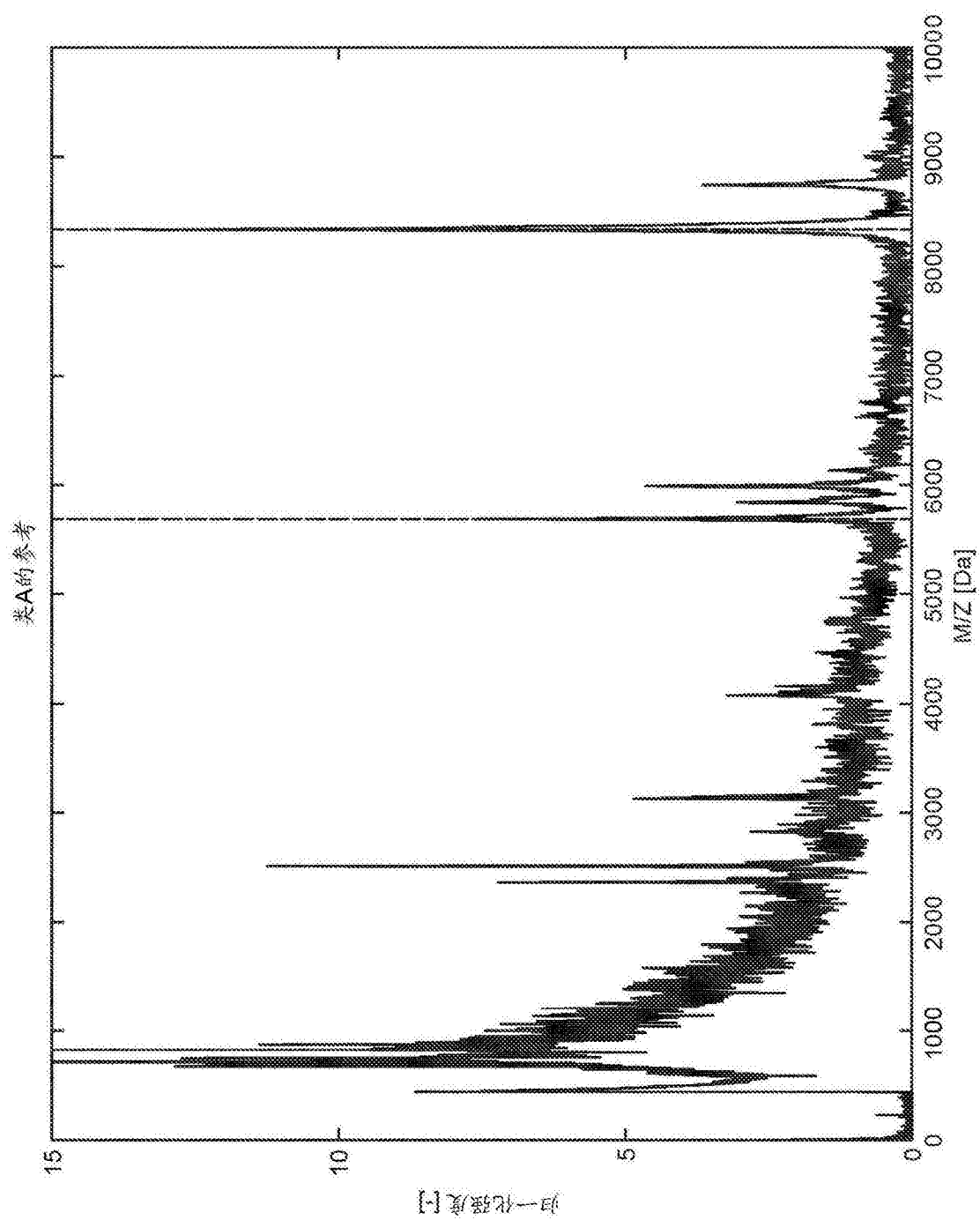


图6

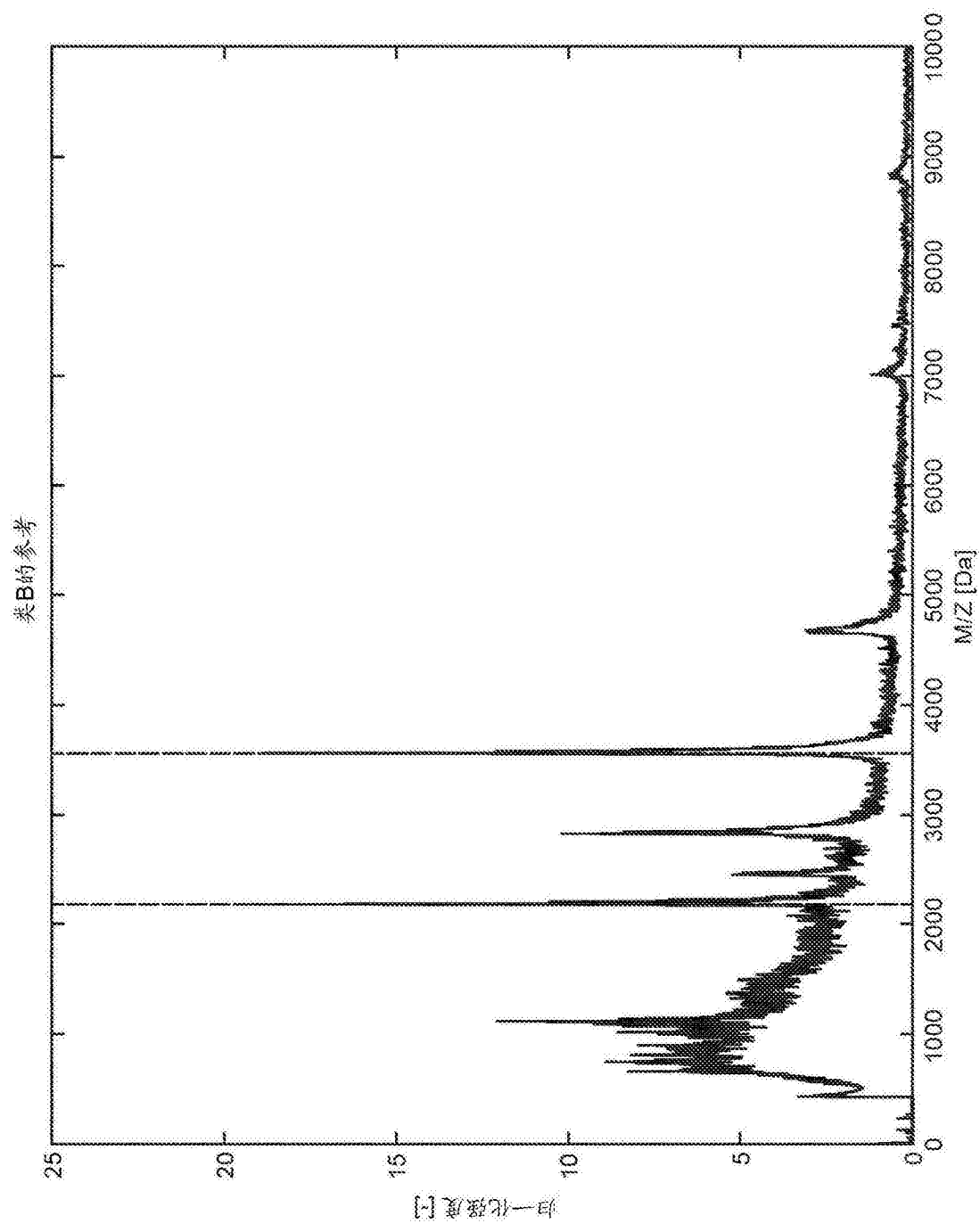


图7

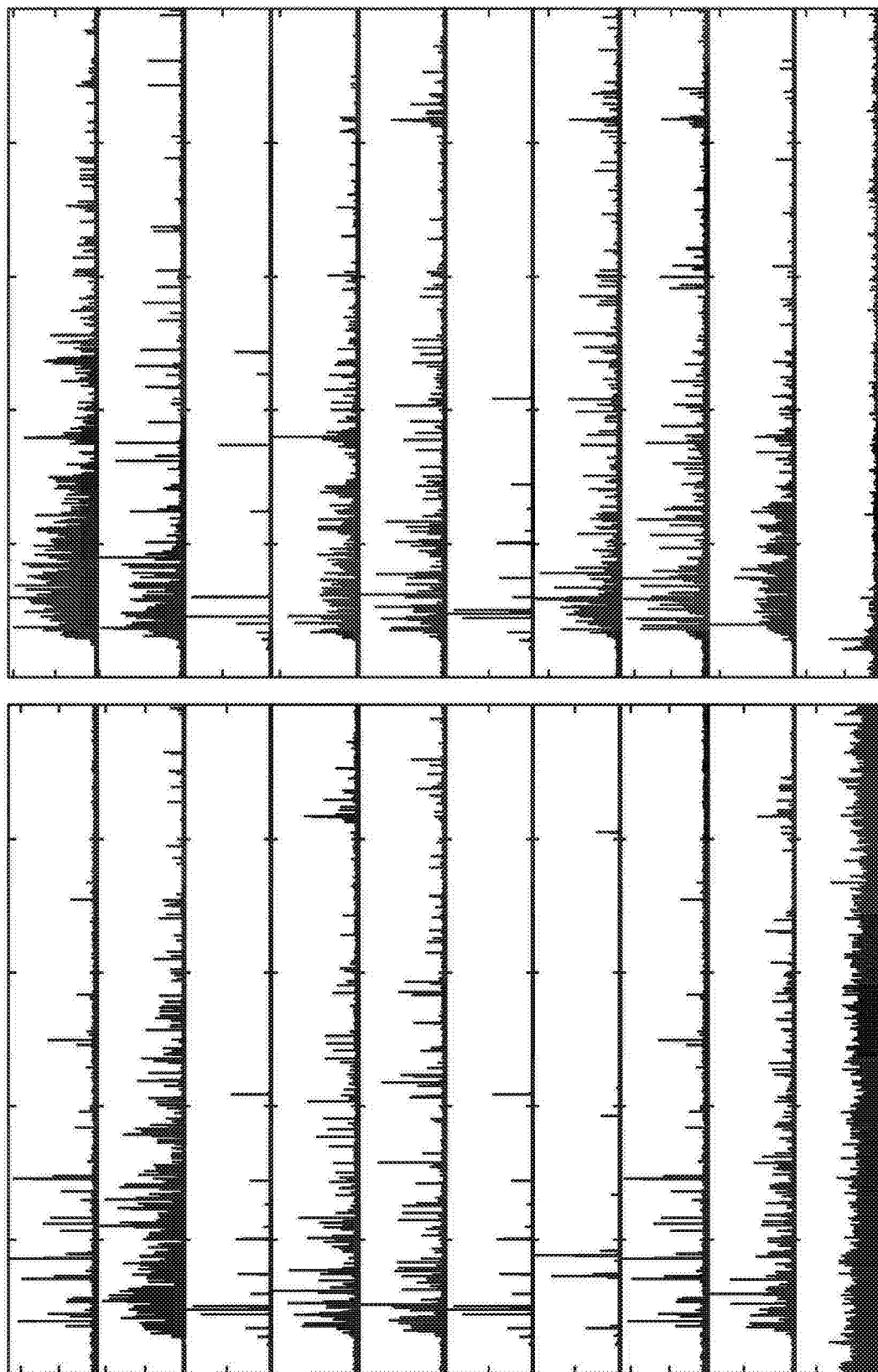


图8

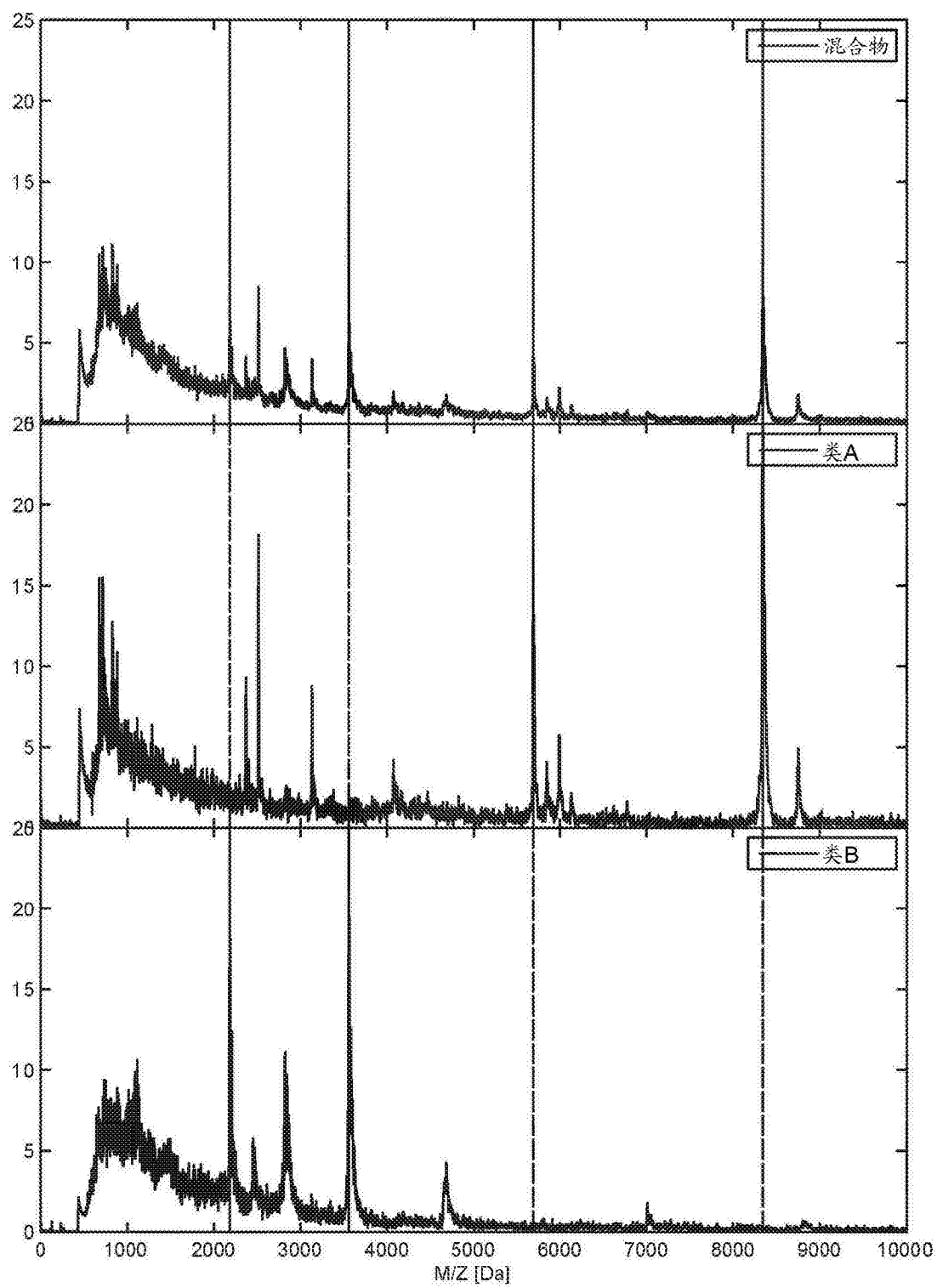


图9

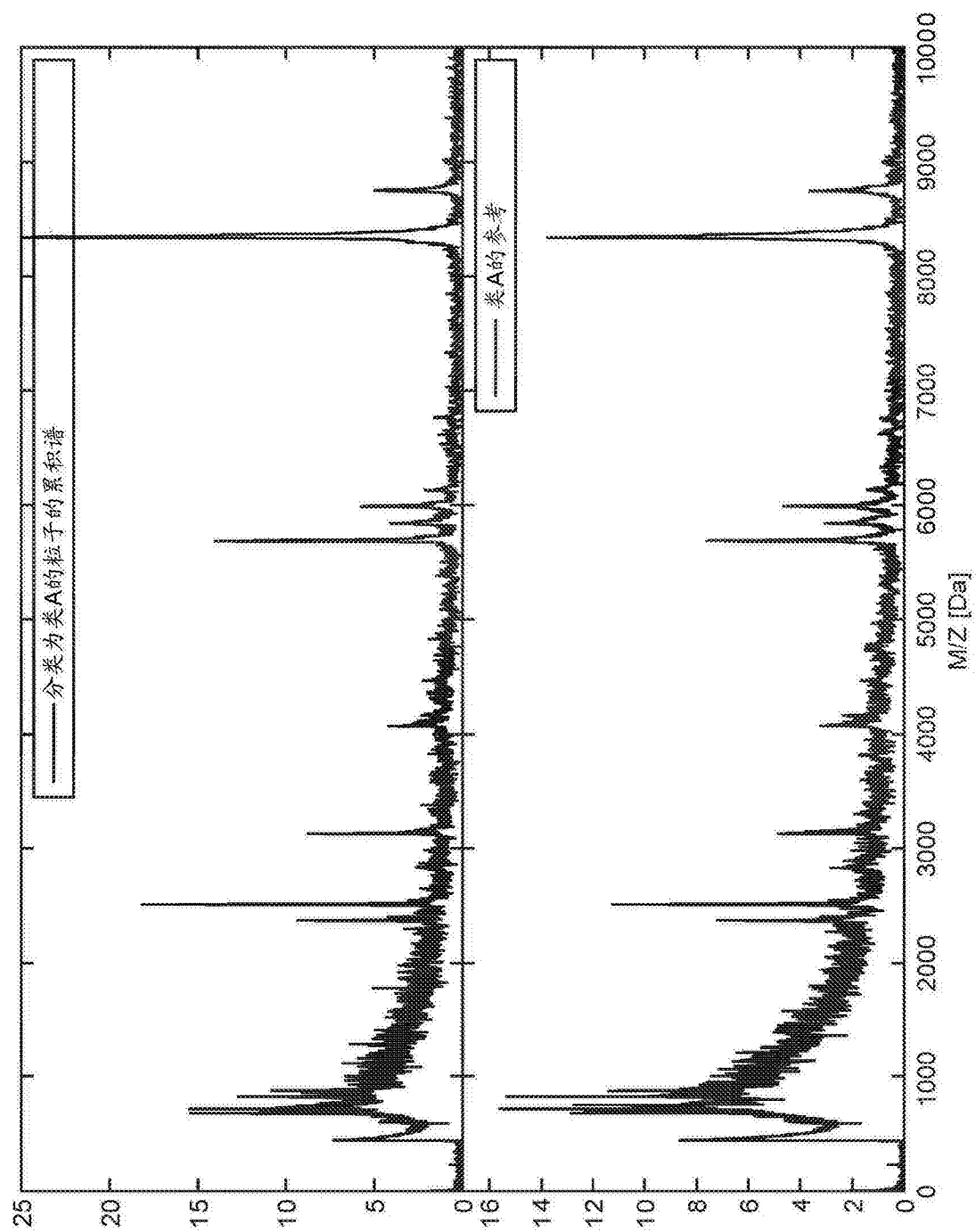


图10

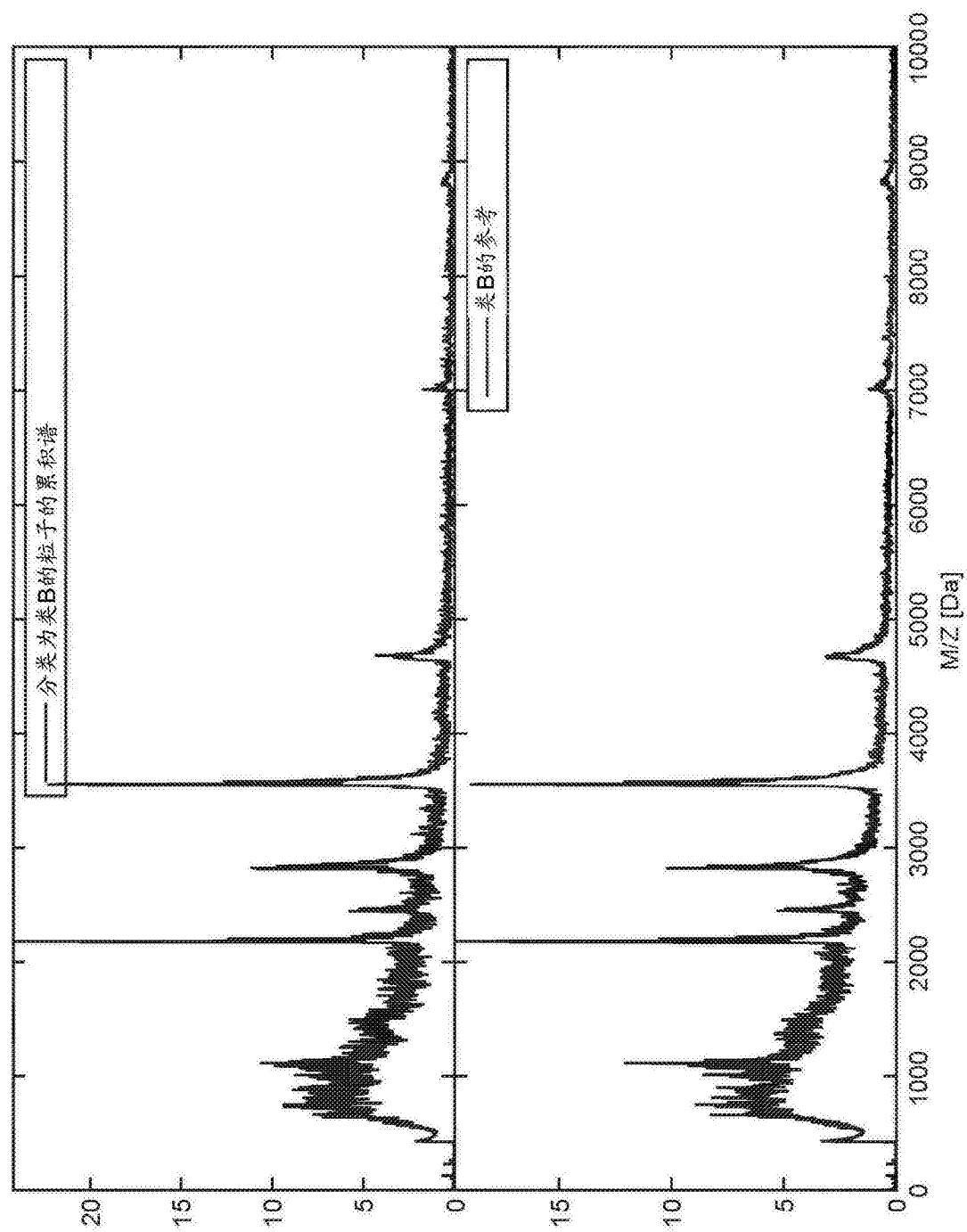


图11

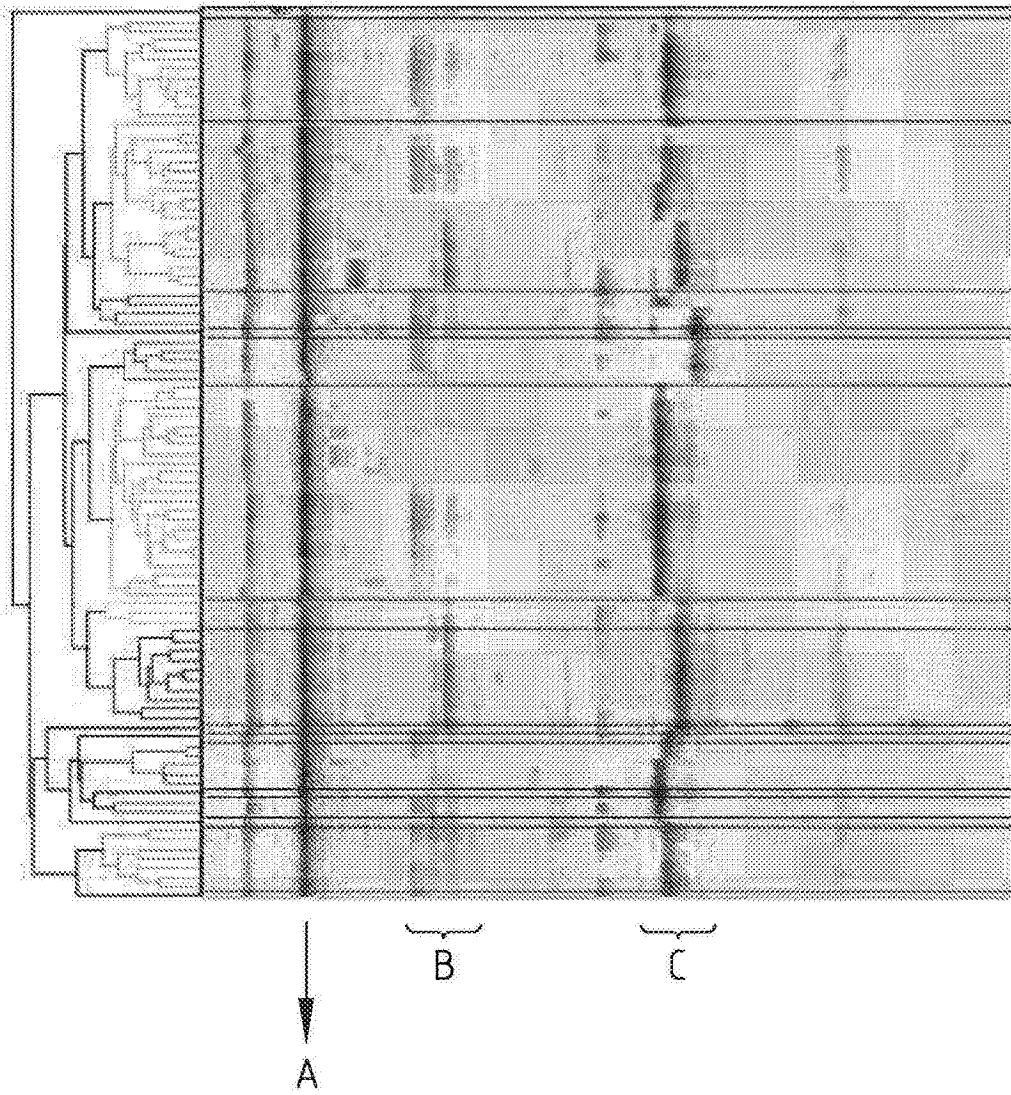


图12

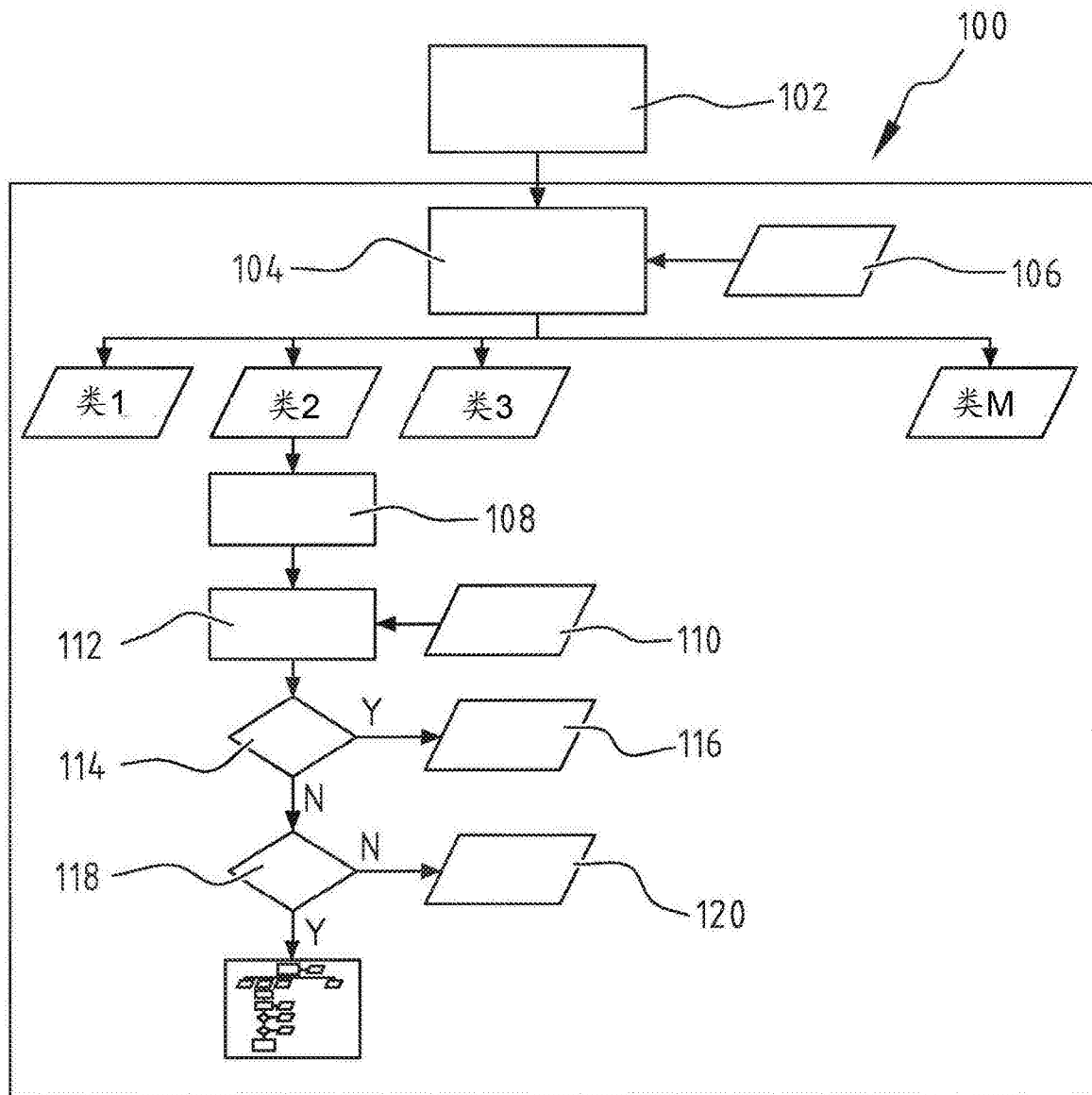


图13