(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 104965020 A
(43)申请公布日 2015.10.07

(21)申请号 201510290301.7

(22)申请日 2015.05.29

(71)申请人 中国科学院计算技术研究所
地址 100190 北京市海淀区中关村科学院南路6号
申请人 中国科学院生物物理研究所

(72)发明人 孙世伟 王耀君 卜东波 李岩 黄纯翠 刘亚名 杨飞 武红梅 陈润生

(74)专利代理机构 北京泛华伟业知识产权代理有限公司 11280
代理人 王勇 李科

(51)Int.Cl.
G01N 27/62(2006.01)

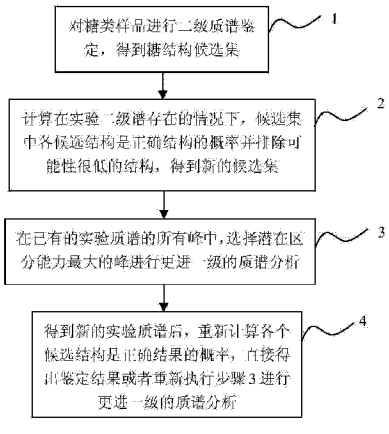
权利要求书2页 说明书10页 附图3页

(54)发明名称

多级质谱生物大分子结构鉴定方法

(57)摘要

本发明提供一种多级质谱生物大分子结构鉴定方法，包括：1)获得样品的二级质谱作为当前质谱；2)在当前质谱中，选择产生下一级质谱的离子，基于所选择的离子进行质谱实验获得所述下一级质谱；3)使用层次贝叶斯模型将所述当前质谱的后验概率以先验概率的方式代入所述下一级质谱，进而对每个候选结构对应的理论质谱进行谱谱比对打分；4)如果无法根据当前的谱谱比对打分结果得出唯一的匹配结构，则将所述下一级质谱作为当前质谱，重复步骤2)进行下一级的质谱分析，直至得出唯一的匹配结构。本发明能够提升生物大分子结构尤其是糖结构同分异构体的鉴定准确度；能够显著降低多级质谱鉴定的开销，例如花费较少的样品量及鉴定时间。



CN 104965020 A

1. 一种多级质谱生物大分子结构鉴定方法，包括下列步骤：

1) 对生物大分子样品进行二级质谱分析，将所获得的二级质谱作为当前质谱；

2) 在当前质谱中，选择产生下一级质谱的离子，基于所选择的离子进行质谱实验获得所述下一级质谱；

3) 使用层次贝叶斯模型将所述当前质谱的后验概率以先验概率的方式代入所述下一级质谱，进而对每个候选结构对应的理论质谱进行谱谱比对打分，其中，所述候选结构根据当前质谱的母离子质量进行结构库搜索得出；

4) 如果无法根据当前的谱谱比对打分结果得出唯一的匹配结构，则将所述下一级质谱作为当前质谱，重复步骤 2) 进行下一级的质谱分析，直至得出唯一的匹配结构。

2. 根据权利要求 1 所述的多级质谱生物大分子结构鉴定方法，其特征在于，所述步骤 2) 中，所述产生下一级质谱的离子根据各候选离子对应的信息熵选出。

3. 根据权利要求 2 所述的多级质谱生物大分子结构鉴定方法，其特征在于，所述步骤 2) 中，所述产生下一级质谱的离子的选择方法如下：

21) 根据当前的质谱的谱峰，选择产生下一级质谱的候选离子；

22) 计算各候选离子对应的信息熵；其中，对于每一个候选离子，模拟生成其所有可能形成的下一级可能质谱；然后每个可能质谱存在的条件下，更新每个候选结构的后验概率，计算得到后验概率的信息熵作为该可能质谱的信息熵；对于每一个候选离子，计算该候选离子所有可能质谱的平均信息熵，将这个平均信息熵作为该候选离子对应的信息熵；

23) 根据所述各候选离子对应的信息熵选择产生下一级质谱的离子。

4. 根据权利要求 3 所述的多级质谱生物大分子结构鉴定方法，其特征在于，所述步骤 23) 还包括：选择各候选离子中对应信息熵最小的作为产生下一级质谱的离子。

5. 根据权利要求 3 所述的多级质谱生物大分子结构鉴定方法，其特征在于，所述步骤 23) 还包括：结合各候选离子的丰度和的对应信息熵选择产生下一级质谱的离子；其中，根据丰度筛除部分候选离子，然后在剩余的候选离子中，选择对应信息熵最小的作为产生下一级质谱的离子。

6. 根据权利要求 3 所述的多级质谱生物大分子结构鉴定方法，其特征在于，所述步骤 21) 还包括：从候选谱峰离子中，排除在所有的候选结构中对应于相同的子结构的离子。

7. 根据权利要求 3 所述的多级质谱生物大分子结构鉴定方法，其特征在于，所述步骤 22) 中，候选离子对应的信息熵的计算方法包括下列步骤：

221) 对于当前实验质谱中的候选离子 i，分析其再次碎裂后可能产生的各个碎片离子，得到每个碎片离子所对应的理论峰，每个理论峰在谱中出现或者不出现构成所有理论谱的集合 $\{s_{i,1}, s_{i,2}, \cdots, s_{i,j}, \cdots, s_{i,n}\}$，其中 $s_{i,j}$ 表示第 i 个离子产生的第 j 种可能的质谱；

222) 然后计算每一个候选结构在下一级质谱实验中产生质谱 $s_{i,j}$ 的概率 $P(s_{i,j}|G_k, s^1, \cdots s^M)$，其中 M 表示当前已得到的实验谱的个数；进一步地，计算各个可能质谱 $s_{i,j}$ 分别存在的情况下，各个候选结构是正确结构的概率，并计算该概率集合的信息熵 $H(s_{i,j})$；进而计算第 i 个离子产生的所有可能质谱 $s_{i,j}$ 得到的信息熵的均值 $H(S_i)$。

8. 根据权利要求 1 所述的多级质谱生物大分子结构鉴定方法，其特征在于，所述步骤 3) 还包括：根据概率 $P(G_i|S^1)$ 和理论谱中的每个峰在实验谱中出现的概率，计算获

得二级实验质谱的条件下，候选结构G_i对应理论谱的谱谱比对打分；其中，$P(G_i|S^1)$表示在一级质谱S^1的条件下，当候选结构数目为m时，各候选结构G_i是正确结构的概率，$P(G_i|S^1)=\frac{1}{m}$。

9. 根据权利要求8所述的多级质谱生物大分子结构鉴定方法，其特征在于，所述步骤3)还包括：基于获得二级实验质谱的条件下，候选结构G_i对应理论谱的谱谱比对打分，利用层次贝叶斯模型，通过将上一级质谱的后验概率以先验概率的方式代入当前所获得的下一级质谱的方式，依次递推地计算出获得更高级别的实验质谱的条件下，候选结构G_i对应理论谱的谱谱比对打分。

10. 根据权利要求1至9中任意一项所述的多级质谱生物大分子结构鉴定方法，其特征在于，所述生物大分子结构为糖结构，所述生物大分子样品为糖样品。

多级质谱生物大分子结构鉴定方法

技术领域

[0001] 本发明涉及生物信息技术和计算技术领域，具体地说，本发明涉及一种使用多级质谱技术进行生物大分子结构鉴定的方法。

背景技术

[0002] 本发明中，生物大分子主要是指核酸、蛋白质、脂类和糖类化合物等作为细胞的主要成分的大分子。在生物信息领域，对生物大分子结构的鉴定在细胞周期调控、凋亡衰老、细胞表面的相互作用等生命过程的研究中起到十分重要的作用。

[0003] 在各种生物大分子的鉴定中，由于糖类化合物以多种结构构型存在于细胞中，因此其复杂性通常也相对较高。例如：对蛋白质数据库 SWISS-PROT 的分析表明：超过一半的蛋白质以糖基化形式存在。此外，糖类还常常与脂类化合物连接形成糖脂。另外，糖类化合物通常由多个单糖通过糖苷键连接而成，并呈现出树形的分枝结构。因此，糖类化合物结构鉴定包括对糖类分子组成、单糖的连接顺序与分枝位置等信息的分析。图 1 示出了 N 糖 GlcNAc2Man9 糖结构的多种经典表示方式，虽然表示方式不同，但它们都是二维的树形结构，其中根节点位于结构的最右边，子节点向左逐步延伸；每一个节点代表一个单糖；每一条边代表连接两个单糖的糖苷键。图 2 示出了糖结构中典型的糖苷键的表示方式。在上述图中，一个单糖可以通过糖苷键跟另外一个或多个单糖相连接，这种分枝结构的复杂性导致了糖类化合物结构构型的多样性。

[0004] 现有的基于质谱的糖类化合物鉴定大多是基于一级质谱或二级质谱数据的策略。一级质谱只能得到母离子质量，不能给出糖类化合物的详细信息，例如支链信息、结构信息和链接位点信息等。二级质谱是将糖在质谱仪中进一步打碎成碎片进行分析，现有技术中的基于二级质谱的糖结构鉴定策略主要包括以下几种：

[0005] (1) 结构库搜索策略：对糖结构库中的每个已知糖结构，首先预测其理论质谱；然后将待鉴定质谱与理论质谱逐一进行比较，返回相似度最高的理论谱所对应的糖结构作为鉴定结果。其缺陷是：糖结构库搜索策略依赖于理论质谱的预测，而目前对于质谱形成机制认识仍然有局限，导致理论质谱预测的精度不高，从而影响了鉴定结果的准确性。

[0006] (2)De Novo 结构鉴定策略：它的基本思想是通过谱图中谱峰间的 m/z 差值，来推断可能的糖结构。与糖结构库搜索策略不同，De Novo 结构鉴定策略不依赖于已知的糖结构数据库，而是直接对质谱数据进行分析，然而De Novo 策略的鉴定准确度严重依赖于质谱谱图数据质量。在高质量的质谱数据中，每一个糖苷键都会有相应的碎裂离子出现；而在低质量的质谱中，部分离子的缺失导致De Novo 策略无法得出准确的鉴定结果。此外，由于要枚举所有可能的糖结构，De Novo 策略通常速度较慢。

[0007] (3) 谱库搜索策略：这种策略的基本思想是将已知鉴定结果的二级质谱以“质谱-结构对”的映射形式收录于数据库中，然后将待鉴定质谱与质谱数据库中的真实谱进行比较，返回相似度最高的糖结构作为鉴定结果。这种策略的优点是：与糖结构库搜索策略相比，谱库搜索策略使用已知的真实谱，而不是预测出的理论谱进行比较，从而使得鉴定结果

更可信。然而，原则上来说，谱库搜索策略仅适用于已鉴定过的谱，对于未收录的质谱则无法鉴定；此外，在收录“质谱－结构对”数据库时的误差，将会影响候选鉴定结果的准确度。

[0008] 总而言之，相对于糖类化合物复杂的结构来说，二级质谱仅提供了有限的信息，从而导致基于二级质谱的糖类化合物鉴定策略准确度不高。基于此，有些研究者使用多种质谱联合的方式，以期提高鉴定结果的准确度。其中一种方案是：使用多种质谱仪对同一样品进行多次实验，从而获得多种类型的质谱，期望利用更多的碎裂信息实现糖结构的鉴定。例如，对同一糖样品分别使用CID质谱仪和ETD质谱仪进行碎裂，产生两张包含不同碎裂信息的质谱，然后综合利用两张质谱的碎裂信息进行糖结构鉴定。然而，多糖化合物结构往往十分复杂，两次实验所提供的碎裂信息，通常远远达不到准确鉴定各种同分异构体的需求；而如果对同一糖样品进行大量的实验，虽然能够提供足够的碎裂信息，但这种鉴定方案的开销过大，尤其是需要借助多台不同型号的质谱仪，导致鉴定过程复杂同时也增加了鉴定成本。

[0009] 有研究者进一步提出另一种方案：多级质谱搜索策略，即使用超过二级质谱的多级质谱能够提供更多的断裂信息，来实现同分异构体的逐步区分。一种已有的使用多级质谱鉴定方法是将多级质谱与谱库搜索策略相结合，使用已知糖结构的二级和三级质谱构建“质谱－结构对”数据库。此方法在一定程度上提高了鉴定准确度，但是由于其谱库中谱数据量少、物种单一，且多级质谱数据仅局限于三级质谱，所以对于某些糖类化合物的同分异构体依然无法准确区分。

[0010] 综上所述，当前迫切需要一种能够以较小的开销实现高准确度的多级质谱鉴定的解决方案。

发明内容

[0011] 因此，本发明的任务是克服现有技术的上述缺陷，提供一种多级质谱鉴定的解决方案。

[0012] 本发明提供了一种多级质谱生物大分子结构鉴定方法，包括下列步骤：

[0013] 1) 对生物大分子样品进行二级质谱分析，将所获得的二级质谱作为当前质谱；

[0014] 2) 在当前质谱中，选择产生下一级质谱的离子，基于所选择的离子进行质谱实验获得所述下一级质谱；

[0015] 3) 使用层次贝叶斯模型将所述当前质谱的后验概率以先验概率的方式代入所述下一级质谱，进而对每个候选结构对应的理论质谱进行谱谱比对打分，其中，所述候选结构根据当前质谱的母离子质量进行结构库搜索得出；

[0016] 4) 如果无法根据当前的谱谱比对打分结果得出唯一的匹配结构，则将所述下一级质谱作为当前质谱，重复步骤2) 进行下一级的质谱分析，直至得出唯一的匹配结构。。

[0017] 其中，所述步骤2) 中，所述产生下一级质谱的离子根据各候选离子对应的信息熵选出。

[0018] 其中，所述步骤2) 中，所述产生下一级质谱的离子的选择方法如下：

[0019] 21) 根据当前的质谱的谱峰，选择产生下一级质谱的候选离子；

[0020] 22) 计算各候选离子对应的信息熵；其中，对于每一个候选离子，模拟生成其所有可能形成的下一级可能质谱；然后每个可能质谱存在的条件下，更新每个候选结构的后验

概率，计算得到后验概率的信息熵作为该可能质谱的信息熵；对于每一个候选离子，计算该候选离子所有可能质谱的平均信息熵，将这个平均信息熵作为该候选离子对应的信息熵；

[0021] 23）根据所述各候选离子对应的信息熵选择产生下一级质谱的离子。

[0022] 其中，所述步骤 23）还包括：选择各候选离子中对应信息熵最小的作为产生下一级质谱的离子。

[0023] 其中，所述步骤 23）还包括：结合各候选离子的丰度和的对应信息熵选择产生下一级质谱的离子；其中，根据丰度筛除部分候选离子，然后在剩余的候选离子中，选择对应信息熵最小的作为产生下一级质谱的离子。

[0024] 其中，所述步骤 21）还包括：从候选谱峰离子中，排除在所有的候选结构中对应于相同的子结构的离子。

[0025] 其中，所述步骤 22）中，候选离子对应的信息熵的计算方法包括下列步骤：

[0026] 221）对于当前实验质谱中的候选离子 i，分析其再次碎裂后可能产生的各个碎片离子，得到每个碎片离子所对应的理论峰，每个理论峰在谱中出现或者不出现构成所有理论谱的集合 $\{s_{i,1}, s_{i,2}, \cdots, s_{i,j}, \cdots, s_{i,n}\}$，其中 $s_{i,j}$ 表示第 i 个离子产生的第 j 种可能的质谱；

[0027] 222）然后计算每一个候选结构在下一级质谱实验中产生质谱 $s_{i,j}$ 的概率 $P(s_{i,j}|G_k, S^1, \cdots S^M)$，其中 M 表示当前已得到的实验谱的个数；进一步地，计算各个可能质谱 $s_{i,j}$ 分别存在的情况下，各个候选结构是正确结构的概率，并计算该概率集合的信息熵 $H(s_{i,j})$；进而计算第 i 个离子产生的所有可能质谱 $s_{i,j}$ 得到的信息熵的均值 $H(S_i)$。

[0028] 其中，所述步骤 3）还包括：根据概率 $P(G_i|S^1)$ 和理论谱中的每个峰在实验谱中出现的概率，计算获得二级实验质谱的条件下，候选结构 G_i 对应理论谱的谱谱比对打分；其中，$P(G_i|S^1)$ 表示在一级质谱 S^1 的条件下，当候选结构数目为 m 时，各候选结构 G_i 是正确结构的概率，$P(G_i|S^1) = \frac{1}{m}$。

[0029] 其中，所述步骤 3）还包括：基于获得二级实验质谱的条件下，候选结构 G_i 对应理论谱的谱谱比对打分，利用层次贝叶斯模型，通过将上一级质谱的后验概率以先验概率的方式代入当前所获得的下一级质谱的方式，依次递推地计算出获得更高级别的实验质谱的条件下，候选结构 G_i 对应理论谱的谱谱比对打分。

[0030] 其中，所述生物大分子结构为糖结构，所述生物大分子样品为糖样品。

[0031] 与现有技术相比，本发明具有下列技术效果：

[0032] （1）本发明能够提升糖结构同分异构体的鉴定准确度。

[0033] （2）本发明能够以最少的打谱次数提供尽可能多的糖结构碎裂信息，在实现准确鉴定的同时，可以显著降低多级质谱鉴定的开销，例如花费较少的样品量及鉴定时间。

附图说明

[0034] 以下，结合附图来详细说明本发明的实施例，其中：

[0035] 图 1 示出了 N 糖 G1cNAc2Man9 糖结构的多种经典表示方式；

[0036] 图 2 示出了糖结构中典型的糖苷键的表示方式；

[0037] 图 3 示出了本发明一个实施例中的多级质谱糖鉴定方法的流程示意图；

[0038] 图 4 示出了本发明一个实施例中基于层次贝叶斯模型的质谱比较打分方法示意图；

[0039] 图 5 示出了对 Man5N 糖样品，分别使用两种打谱方法进行多级质谱鉴定的结果比较；

[0040] 图 6 示出了对 Man6N 糖样品，分别使用两种打谱方法进行多级质谱鉴定的结果比较。

具体实施方式

[0041] 如前文所述，糖结构的复杂性导致了明显的同分异构现象，即同一分子质量会对应多个不同的糖结构。在结构库搜索中，若简单按照质谱母离子质量搜索，则会返回多个难以区分的同分异构体。例如，糖分子量为 1579.454 在 Carbbank 糖结构库中会对应有 175 个同分异构的糖结构。发明人针对糖结构的这种复杂性以及现有糖结构鉴定策略存在的缺陷，通过建立有效融合多级质谱数据的糖鉴定打分模型，以及在打谱过程中选择鉴别能力最强的离子进行打下一级谱，实现了综合利用多级质谱信息的开销较低的糖结构鉴定方案，提高了鉴定准确度。

[0042] 下面结合附图和实施例对本发明做进一步地描述。

[0043] 图 4 示出了本发明一个实施例的多级质谱糖鉴定方法的流程示意图，包括下列步骤：

[0044] 步骤 1：对糖类样品进行二级质谱鉴定，得到糖结构候选集。在一个例子中，首先对糖类样品纯化后进行质谱分析，获得一级质谱；然后选取目标离子碎裂获得二级质谱，再以二级质谱的母离子质量进行糖结构库搜索，搜索结果作为鉴定结果的候选集。

[0045] 步骤 2：计算在实验二级谱存在的情况下，候选集中各候选结构是正确结构的概率并排除可能性很低的结构，得到新的候选集。此时，如果候选集中仅剩唯一的候选结构如果候选集中某个结构是正确结果的概率具有绝对优势，则直接给出鉴定结果，本次鉴定结束，如果候选集中的各个结构的概率平均，不足以判断哪个是真正的正确结果有至少两个候选结构，则执行步骤 3。

[0046] 步骤 3：在已有的实验质谱的所有峰中，选择潜在区分能力最大的峰进行更进一级的质谱分析（即进行下一级的质谱分析）。需强调的是，本实施例中，选择潜在区分能力最大的峰进行打谱，这与传统的产生多级质谱的方法不同，传统的产生多级质谱的方法是：从一级谱开始，在质谱中选取丰度最大的离子打下一级质谱，以此方式逐级产生多级质谱。

[0047] 步骤 4：得到新的实验质谱后，重新计算各个候选结构是正确结果的概率，具体方法是使用层次贝叶斯模型，将信息（即后验概率）将步骤 2 得到的各个候选结构的概率以先验概率的方式代入下一级的新质谱的鉴别计算中，得出下一级质谱实验后的新谱谱比对打分后的新的各个候选结构是正确结果的概率值，根据新的谱谱比对打分进一步排除匹配度低的候选结构。此时，如果候选集中仅剩唯一的候选结构，则直接给出鉴定结果，本次鉴定结束，如果候选集中有至少两个候选结构，则回到步骤 3。

[0048] 步骤 4 中，上述判定鉴定是否结束的具体方法并不是唯一的。例如，在另一个实施例中，所述步骤 4 中完成当前一级的质谱试验后，如果候选集中仅剩唯一的候选结构，或者存在某种结构的概率值相对其它结构具有绝对优势，则直接给出鉴定结果，本次鉴定结束；

否则，回到步骤 3。

[0049] 在一个实施例中，上述步骤 4 中使用层次贝叶斯模型，先前实验质谱的鉴别信息（后验概率）以先验概率的方式，有机融合入新实验质谱加入后各个候选结构的概率，进而得到在新的实验谱存在的情况下，各个候选结构是正确结构的概率。

[0050] 下面以一个简单的五级质谱为例进形说明，为了描述简单，在此例中，每级质谱只有一个峰被选择进行再次断裂，在方法的实际应用中，每个质谱中的进一步断裂的峰的个数是不固定的。为便于描述，以 G_i 表示为第 i 个候选糖结构；S^1 表示样品对应的一级质谱；S^2 表示样品对应的二级质谱；S_j^3 表示由二级质谱中的谱峰 j 产生的下一级质谱，即三级质谱；S_k^4 表示由三级质谱中的谱峰 k 产生的下一级质谱，即四级质谱；S_l^5 表示由四级质谱中的谱峰 l 产生的下一级质谱，即五级质谱；$P(G_i|S^1,S^2,S_j^3,S_k^4,S_l^5)$ 表示在多级质谱 S^1、S^2、S_j^3、S_k^4、S_l^5 存在的条件下，糖结构 G_i 是正确结构的概率，该概率（也可称为概率打分或者谱谱比对打分）Score 的计算按下述方案得出：

[0051] a、基于层次贝叶斯模型，首先计算得到二级质谱（S^1、S^2）后，糖结构 G_i 是正确结构的概率 $P(G_i|S^1,S^2)$：

[0052] $$P(G_i|S^1,S^2)=\frac{P(G_i|S^1)\cdot P(S^2|G_i,S^1)}{P(S^2|S^1)}$$

[0053] 其中，在一级谱的条件下各个候选结构的概率是等概率的，因此当候选结构数目为 m 时，各候选结构 G_i 是正确结构的概率 $P(G_i|S^1)=\frac{1}{m}$；

[0054] $$P(S^2|G_i,S^1)=p_{exist}^{u}\cdot(1-p_{exist})^{v}$$

[0055] 其中 p_{exist} 表示理论谱中的每个峰在实验谱中出现的概率；

[0056] $(1-p_{exist})$ 表示理论谱中的每个峰在实验谱中没有出现的概率；u 表示在 S^2 中有 u 个理论谱峰出现；v 表示在 S^2 中有 v 个理论谱峰没有出现；u+v ＝理论峰的总个数，此处理论峰是指假定糖结构为 G_i 且一级质谱为 S^1，继续进行二级质谱实验理论上可获得的离子所对应的谱峰，这些理论峰在实际的实验谱中可能存在，也可能不存在，但单个谱峰在理论谱中存在并且实验谱中也出现的概率 p_{exist} 大致是稳定的，它可以根据统计及先验知识得出，本实施例中 p_{exist} 取 0.7。

[0057] 另外，$P(S^2|S^1)=\sum_{i=1}^{m}P(G_i|S^1)\cdot P(S^2|G_i,S^1)$

[0058] 前文中已描述了 $P(G_i|S^1)$ 和 $P(S^2|G_i,S^1)$ 的计算方法，根据 $P(G_i|S^1)$ 和 $P(S^2|G_i,S^1)$ 即可得到 $P(S^2|S^1)$，进而计算出 $P(G_i|S^1,S^2)$。

[0059] b、基于层次贝叶斯模型，计算得到三级质谱（S^1,S^2,S_j^3）后，糖结构 G_i 是正确结构的概率 $P(G_i|S^1,S^2,S_j^3)$:

[0060] $$P(G_i|S^1,S^2,S_j^3)=\frac{P(G_i|S^1,S^2)\cdot P(S_j^3|G_i,S^1,S^2)}{P(S_j^3|S^1,S^2)}$$

[0061] 其中，$P(G_i|S^1,S^2)$ 为上一级质谱的谱谱比对打分；

[0062] $P(S_j^3|G_i,S^1,S^2)$表示假设糖结构为 G_i且一级质谱为 S^1，二级质谱为 S^2的前提下，获得三级质谱S_j^3的概率。根据假定糖结构 G_i以及已产生的一级质谱 S^1和二级质谱 S^2可获得下一级质谱实验中所有可能出现的理论谱峰，然后再根据S_j^3中实际出现的谱峰，基于单个理论谱峰的出现概率 p_{exist}计算出 $P(S_j^3|G_i,S^1,S^2)$。

[0063] $$P(S_j^3|G_i,S^1,S^2)=p_{exist}^{u3}\cdot(1-p_{exist})^{v3}$$

[0064] 其中，u3 表示在S_j^3中有 u3 个理论谱峰出现；v3 表示在S_j^3 中有 v3 个理论谱峰没有出现。

[0065] $P(S_j^3|S^1,S^2)$根据下述公式得出：

[0066] $$P(S_j^3|S^1,S^2)=\sum_{i=1}^{m}P(G_i|S^1,S^2)\cdot P(S_j^3|G_i,S^1,S^2)$$

[0067] c、基于层次贝叶斯模型，计算得到四级质谱（S^1，S^2，S_j^3，S_k^4）后，糖结构 G_i是正确结构的概率$P(G_i|S^1,S^2,S_j^3,S_k^4)$:

[0068] $$P(G_i|S^1,S^2,S_j^3,S_k^4)=\frac{P(G_i|S^1,S^2,S_j^3)\cdot P(S_k^4|G_i,S^1,S^2,S_j^3)}{P(S_k^4|S^1,S^2,S_j^3)}$$

[0069] 其中，$P(G_i|S^1,S^2,S_j^3)$为上一级质谱的谱谱比对打分；

[0070] $P(S_k^4|G_i,S^1,S^2,S_j^3)$表示表示假设糖结构为 G_i且一级质谱为 S^1二级质谱为 S^2三级质谱为 S_j^3的前提下，获得四级质谱S_k^4的概率。根据假定糖结构 G_i以及已产生的一级质谱 S^1和二级质谱 S^2以及三级质谱 S_j^3可获得下一级质谱实验中所有可能出现的理论谱峰，然后再根据S_k^4中实际出现的谱峰，基于单个理论谱峰的出现概率 p_{exist}计算出 $P(S_k^4|G_i,S^1,S^2,S_j^3)$。

[0071] $$P(S_k^4|G_i,S^1,S^2,S_j^3)=p_{exist}^{u4}\cdot(1-p_{exist})^{v4}$$

[0072] 其中，u4 表示在S_k^4中有 u4 个理论谱峰出现；v4 表示在S_k^4中有 v4 个理论谱峰没有出现。

[0073] $P(S_k^4|S^1,S^2,S_j^3)$根据下述公式得出：

[0074] $$P\left(S_k^4|S^1,S^2,S_j^3\right)=\sum_{i=1}^{m}P\left(G_i|S^1,S^2,S_j^3\right)\cdot P(S_k^4|G_i,S^1,S^2,S_j^3)$$

[0075] d、基于层次贝叶斯模型，计算得到五级质谱（S^1，S^2，S_j^3，S_k^4，S_l^5）后，糖结构G_i是正确结果的概率$\mathrm{P}\left(G_i\middle|S^1,S^2,S_j^3,S_k^4,S_l^5\right)$:

[0076] $$\mathrm{P}\left(G_i\middle|S^1,S^2,S_j^3,S_k^4,S_l^5\right)=\frac{P\left(G_i|S^1,S^2,S_j^3,S_k^4\right)\cdot P(S_l^5|G_i,S^1,S^2,S_j^3,S_k^4)}{P(S_l^5|S^1,S^2,S_j^3,S_k^4)}$$

[0077] $$Score=\mathrm{P}(G_i|S^1,S^2,S_j^3,S_k^4,S_l^5)$$

[0078] 其中，$P\left(G_i|S^1,S^2,S_j^3,S_k^4\right)$为上一级质谱的谱谱比对打分；

[0079] $P\left(G_i|S^1,S^2,S_j^3,S_k^4\right)$表示假定糖结构为$G_i$且一级质谱为$S^1$二级质谱为$S^2$三级质谱为$S_j^3$四级质谱为$S_k^4$的前提下，获得五级质谱$S_l^5$的概率。根据假定糖结构$G_i$以及已产生的一级质谱$S^1$、二级质谱$S^2$、三级质谱$S_j^3$以及四级质谱$S_k^4$可获得下一级质谱实验中所有可能出现的理论谱峰，然后再根据S_l^5中实际出现的谱峰，基于单个理论谱峰的出现概率p_{exist}计算出$P(S_l^5|G_i,S^1,S^2,S_j^3,S_k^4)$。

[0080] $$P(S_l^5|G_i,S^1,S^2,S_j^3,S_k^4)=p_{exist}^{u5}\cdot(1-p_{exist})^{v5}$$

[0081] 其中，u5 表示在S_l^5中有 u5 个理论谱峰出现；v5 表示在S_l^5中有 v5 个理论谱峰没有出现。

[0082] $P(S_l^5|S^1,S^2,S_j^3,S_k^4)$根据下述公式得出

[0083] $$P\left(S_l^5|S^1,S^2,S_j^3,S_k^4\right)=\sum_{i=1}^{m}P\left(G_i|S^1,S^2,S_j^3,S_k^4\right)\cdot P(S_l^5|G_i,S^1,S^2,S_j^3,S_k^4)$$

[0084] 上述多级质谱实验中，可以预先设定谱谱打分阈值，在每一级计算出谱谱打分后，比较该级谱谱打分是否超过阈值，如果未超过，则认为相应的候选结构不匹配，从候选集中剔除该候选结构。当候选集中只剩唯一的候选结构时，即可停止多级质谱实验，直接得出鉴定结果。

[0085] 该实施例中，采用层次贝叶斯模型将上一级质谱信息以先验概率的方式，有机融合入下一级质谱实验中，最终通过糖类样品的多级质谱的谱谱比较，逐步缩小鉴定结果范围，直到筛选出唯一鉴定结果。这种方案克服了现有技术的多级质谱策略的缺陷，提升了糖结构同分异构体的鉴定准确度。同时该实施例也避免了使用多台不同型号的质谱仪，能够在一定程度上减少糖结构鉴定的开销。

[0086] 进一步地，如前文所述，所述步骤 3 中，传统的产生多级质谱的方法是：从一级谱

开始，在质谱中选取丰度最大的离子打下一级质谱，以此方式逐级产生多级质谱。然而，在糖鉴定中，直接选择丰度最大的离子，对于下一级质谱鉴定并不是提升鉴别能力的较优选择。在本发明的一个优选本实施例中，针对传统的“丰度优先离子选择”打谱方式的不足，设计出基于信息熵技术的最优打谱路径选择算法，以使产生的下一级质谱尽可能地提升鉴别能力，以此更快实现糖结构鉴定，同时减少打谱次数，从而降低开销。

[0087] 所述最优打谱路径算法具体包含下列步骤：

[0088] 步骤 31：在当前的质谱中选择产生下一级质谱的候选谱峰离子。为了描述简单起见，在此处假设样品是糖类纯样品、且一个谱峰离子只对应于一个糖结构或糖的子结构。本步骤中，只选择可能产生区分信息的离子作为产生下一级质谱的候选离子。其原因在于：有些离子在所有的候选糖结构中对应于相同的子结构，这些离子产生的下一级质谱对候选结构的区分显然是无效的，因此这类离子可以排除，在当前的质谱的所有谱峰所对应的离子中排除这类离子后就得到了候选谱峰离子。

[0089] 步骤 32：计算各候选离子对应的信息熵。本步骤中采用信息熵来衡量各候选离子对于候选糖结构的区分度。简要地说，信息熵是信息量的度量；信息熵越小，表明各个候选离子对于候选糖结构的区分度越大；反之，信息熵越大，表明区分度越小。假设在质谱形成过程中，糖结构的每个糖苷键的碎裂是等概率的，对于每一个候选离子，模拟生成其所有可能形成的下一级可能质谱；然后每个可能质谱存在的条件下，更新每个候选糖结构的后验概率，计算得到后验概率的信息熵；最后每一个候选离子的区分能力使用该离子所有可能质谱的平均信息熵来衡量。

[0090] 具体计算方法包括：

[0091] 步骤 321：对于当前实验质谱中的候选离子 i，分析其再次碎裂后可能产生的各个碎片离子（即碎裂后的片段的带电离子），得到每个碎片离子所对应的理论峰，每个理论峰在谱中出现或者不出现构成所有理论谱的集合 $\{s_{i,1}, s_{i,2}, \cdots, s_{i,j}, \cdots, s_{i,n}\}$，其中 $s_{i,j}$ 表示第 i 个离子产生的第 j 种可能的质谱。

[0092] 步骤 322：然后计算每一个候选结构在下一级质谱实验中产生质谱 $s_{i,j}$ 的概率 $P(s_{i,j}|G_k, s^1, \cdots s^M)$，其中 M 表示当前已得到的实验谱的个数。

[0093] 进一步地，计算各个可能质谱 $s_{i,j}$ 分别存在的情况下，各个候选结构是正确结构的概率，并计算该概率集合的信息熵 $H(s_{i,j})$；第 i 个离子产生的所有可能质谱 $s_{i,j}$ 得到的信息熵的均值 $H(S_i)$，将 $H(S_i)$ 作为第 i 个候选峰的预计鉴别能力的度量，$H(S_i)$ 越小，表示该峰的预计鉴别能力越强。使用上述方法计算所有候选离子 i 的 $H(S_i)$，选择使得 $H(S_i)$ 最小的候选离子 i 用于产生下一级质谱。

[0094] 步骤 322 可用公式表示如下：计算

[0095] $$\underset{0<i<t}{\mathrm{Min}}\, H(S_i)$$

[0096] $$H(S_i) = \frac{1}{n}\sum_{j=1}^{n} H(s_{i,j})$$

[0097] $$H(s_{i,j}) = -\sum_{k=1}^{m} P(G_k|s_{i,j}, s^1, \cdots, s^M) \cdot \log_2 P(G_k|s_{i,j}, s^1, \cdots, s^M)$$

[0098] $$P(G_k|s_{i,j},s^1,\cdots,s^M)=\frac{P(s_{i,j}|G_k,s^1,\cdots,s^M)\cdot P(G_k,s^1,\cdots,s^M)}{\sum_{k=1}^{m}P(s_{i,j}|G_k,s^1,\cdots,s^M)\cdot P(G_k,s^1,\cdots,s^M)}$$

[0099] $$=\frac{P(s_{i,j}|G_k)\cdot P(G_k|s^1,\cdots,s^M)}{\sum_{k=1}^{m}P(s_{i,j}|G_k)\cdot P(G_k|s^1,\cdots,s^M)}$$

[0100] 其中 t 表示在当前鉴定结果中有 t 个具有分类能力的候选离子峰；

[0101] m 表示在当前鉴定结果中有 m 个未能区分的候选结构；

[0102] n 表示使用离子 i 产生下一级质谱，最多可以产生 n 种可能形态的质谱；

[0103] S_i表示由离子 i 产生的质谱；

[0104] $s_{i,j}$表示第 i 个离子产生的第 j 种可能可能质谱；

[0105] $H(s_{i,j})$ 表示使用离子 i 产生的下一级质谱的第 j 种可能谱 $s_{i,j}$用于糖结构鉴定得到的信息熵；

[0106] s^1，…，s^M表示已产生的质谱数据；如前文所述，M 是已经得到的实验质谱的数目。

[0107] G_k表示第 k 个候选糖结构；

[0108] $P(G_k|s^1,\cdots,s^M)$ 表示利用本次鉴定已产生的质谱数据鉴定为糖结构 G_k的先验概率；

[0109] $P(G_k|s_{i,j},s^1,\cdots,s^M)$ 表示利用质谱 $s_{i,j}$结合本次鉴定已产生的质谱数据可鉴定出样品对应的糖结构为 G_k的概率。

[0110] 在另一个优选实施例中，所述步骤 322 中在计算出各个候选离子的信息熵后，综合使用信息熵和离子丰度，选择用于产生下一级谱的离子。熵最小的离子产生的质谱，理论上对于区分候选结构提供的信息量是最大的。但是，在实际的实验谱中，有时候理论上最优的离子对应的离子丰度很低，这样可能会导致用它来产生下一级实验谱的效果不好。所以，本实施例中选择熵较小的离子的同时还保证离子丰度足够高。为此，设定一个丰度阈值，在超过该丰度阈值的离子中选择熵较小的离子作为产生下一级谱的离子。

[0111] 本发明采用层次贝叶斯模型将已有的实验质谱信息以先验概率的方式，有机融合入下一次新的质谱实验中；采用信息熵来获得最优打谱路径；最终通过糖类样品的多级质谱的谱谱比较，逐步缩小鉴定结果范围，直到筛选出唯一鉴定结果。这种方案克服了现有技术的多级质谱策略的缺陷，理论上能够以最少的打谱次数提供尽可能多的糖结构碎裂信息，提升了糖结构同分异构体的鉴定准确度，同时也降低了多级质谱的鉴定开销。

[0112] 进一步地，以下给出了基于本发明的综合使用信息熵和离子丰度进行离子选择的实施例的初步实验结果。

[0113] 对已知糖结构的 N-Linked 纯糖样品（简称 N 糖）进行多级质谱糖结构鉴定实验，测试本发明优选实施例多级质谱鉴定体系的有效性；测试本发明的最优打谱路径算法的有效性。

[0114] 实验中所使用的质谱仪来自于岛津公司生产的 MALDI-IT-TOF 质谱仪 AXIMA Resonance。N 糖样品来自于 Ludger 公司。实验中使用的糖结构库为 Carbbank 糖结构数据库。Carbbank 糖结构数据库由复杂碳水化合物研究中心 (complex carbohydrate research center，CCRC) 于 1986 年开始创建，数据库称作 CCSD(complex carbohydrate structure

database)，同时开发了一款名为 Carbbank 的糖结构数据库编辑软件，后来对 CCSD 糖结构数据库也被称为 Carbbank 糖结构数据库。

[0115] 实验 1 :Man5N 糖样品糖结构鉴定

[0116] 实验过程 :

[0117] (1) 样品经过处理后注入质谱仪中，产生一级质谱，对一级质谱中按分子质量选择对应离子打二级质谱。二级质谱使用结构库搜索策略搜索结构库。得出候选糖结构。

[0118] (2) 使用本发明优选实施例的多级质谱糖结构鉴定体系对候选结构进行区分。

[0119] (3) 在产生下级质谱的打谱环节分别使用本发明优选实施例的最优打谱路径选择算法和传统打谱路径选择方法进行试验，比较两种方法的鉴定效率。

[0120] 实验结果 :图 5 示出了对 Man5N 糖样品，分别使用两种打谱方法进行多级质谱鉴定的结果比较。横坐标为多级谱，从左向右 2、3、4、5 分别代表二级、三级、四级、五级质谱 ;纵坐标为正确结构对应的概率打分，黑线为使用本发明优选实施例打谱方法的结果，灰线为使用传统“丰度优先离子选择”打谱方法的结果。

[0121] 使用本发明优选实施例最优打谱路径算法打到四级质谱，正确结构对应的概率打分为 0.936 ;使用传统打谱路径方法即使打到五级质谱，正确结构对应的概率打分仅为 0.621。

[0122] 实验 2 :Man6N 糖样品糖结构鉴定

[0123] 实验过程 :如实验 1

[0124] 实验结果 :图 6 示出了对 Man6N 糖样品，分别使用两种打谱方法进行多级质谱鉴定的结果比较。使用本发明优选实施的最优打谱路径算法只需打到三级质谱，正确结构对应的概率打分为 0.937 ;使用传统打谱路径方法即使打到五级质谱，正确结构对应的概率打分仅为 0.731。

[0125] 结果分析 :

[0126] (1) 使用本发明优选实施的多级质谱糖结构鉴定体系可以实现同分异构体的区分 ;

[0127] (2) 本发明优选实施例打谱路径算法比传统方法更有效。实验 1 使用传统方法共打了 5 次谱，本发明优选实施例算法共打了 4 次谱且对应的鉴定结果区分度明显高于传统方法。实验 2 使用传统方法共打了 5 次谱，本发明优选实施例算法共打了 3 次谱且对应的鉴定结果区分度明显高于传统方法。

[0128] 上述实施例中，以糖结构为例对多级质谱鉴定方法做了详细描述，本领域技术人员易于理解，上述多级质谱鉴定方法也可以推广到其它生物大分子结构的鉴定中。

[0129] 最后应说明的是，以上实施例仅用以描述本发明的技术方案而不是对本技术方法进行限制，本发明在应用上可以延伸为其它的修改、变化、应用和实施例，并且认为所有这样的修改、变化、应用、实施例都在本发明的精神和教导范围内。

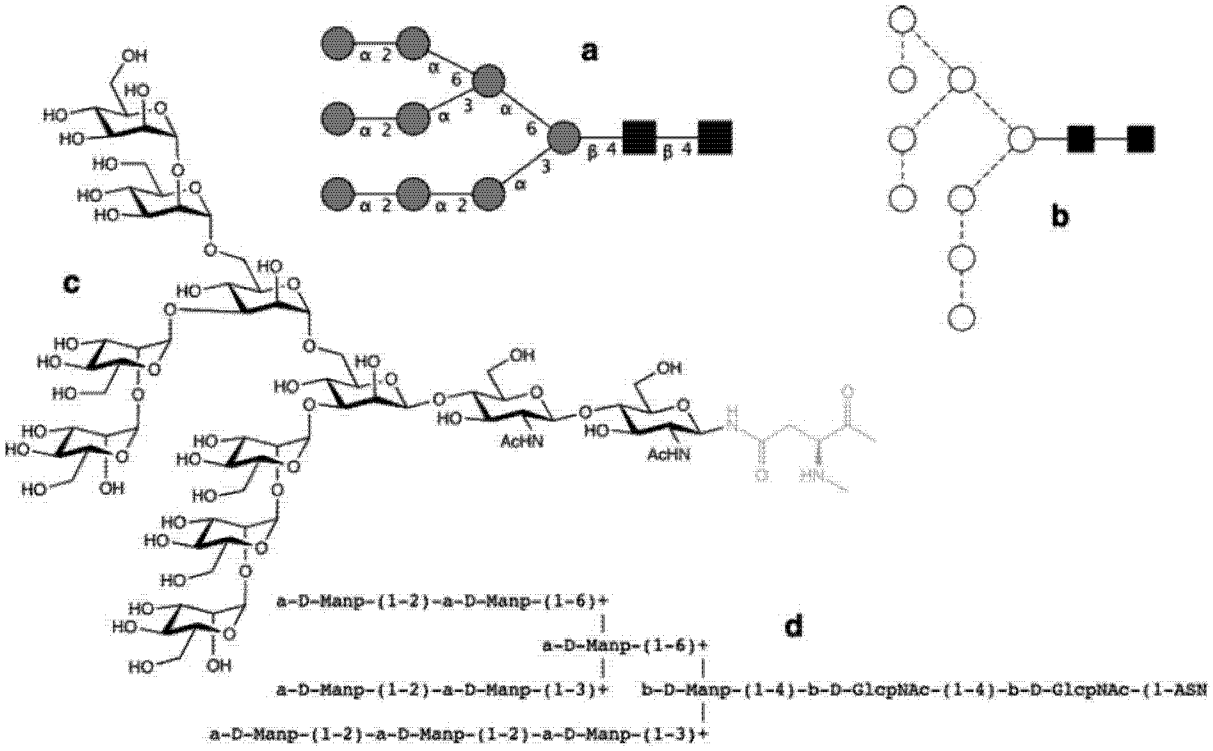

a
b
c
d
a-D-Manp-(1-2)-a-D-Manp-(1-6)+
a-D-Manp-(1-6)+
a-D-Manp-(1-2)-a-D-Manp-(1-3)+
b-D-Manp-(1-4)-b-D-GlcpNAc-(1-4)-b-D-GlcpNAc-(1-ASN
a-D-Manp-(1-2)-a-D-Manp-(1-2)-a-D-Manp-(1-3)+

图 1

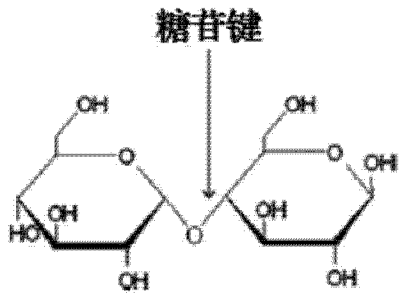

糖苷键

图 2

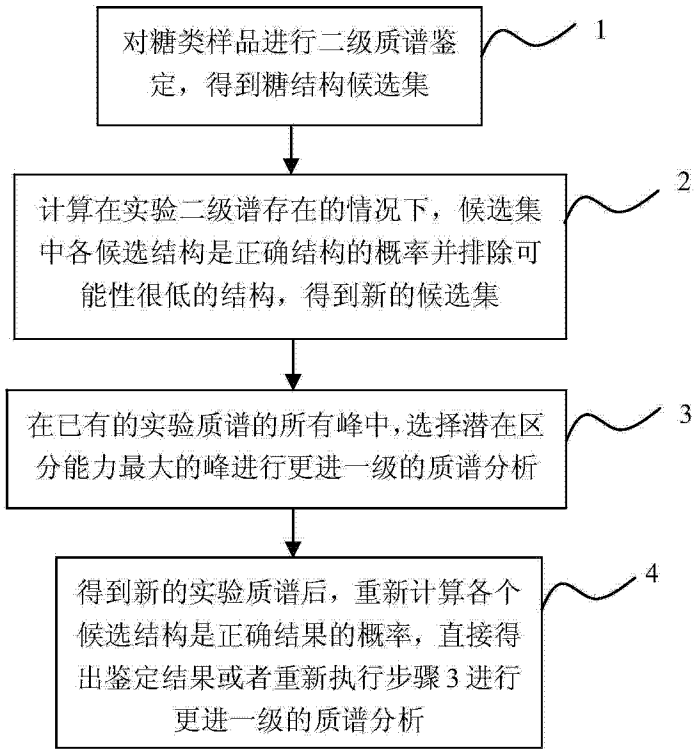

对糖类样品进行二级质谱鉴定，得到糖结构候选集
1
计算在实验二级谱存在的情况下，候选集中各候选结构是正确结构的概率并排除可能性很低的结构，得到新的候选集
2
在已有的实验质谱的所有峰中，选择潜在区分能力最大的峰进行更进一级的质谱分析
3
得到新的实验质谱后，重新计算各个候选结构是正确结果的概率，直接得出鉴定结果或者重新执行步骤 3 进行更进一级的质谱分析
4

图 3

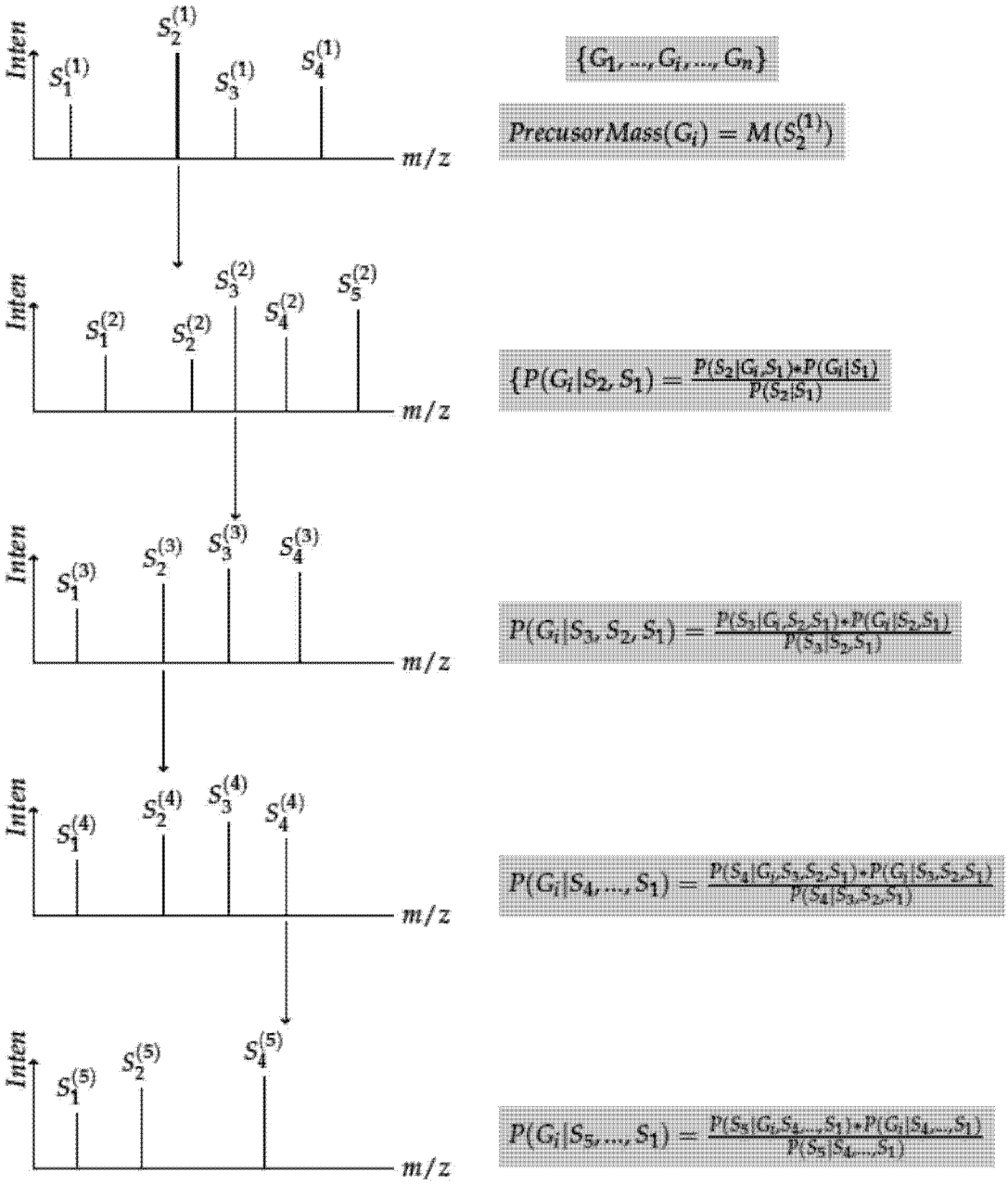

Inten
m/z
$S_1^{(1)}$ $S_2^{(1)}$ $S_3^{(1)}$ $S_4^{(1)}$
$\{G_1, \ldots, G_i, \ldots, G_n\}$
$PrecusorMass(G_i) = M(S_2^{(1)})$
$S_1^{(2)}$ $S_2^{(2)}$ $S_3^{(2)}$ $S_4^{(2)}$ $S_5^{(2)}$
$\{P(G_i|S_2, S_1) = \frac{P(S_2|G_i, S_1) * P(G_i|S_1)}{P(S_2|S_1)}$
$S_1^{(3)}$ $S_2^{(3)}$ $S_3^{(3)}$ $S_4^{(3)}$
$P(G_i|S_3, S_2, S_1) = \frac{P(S_3|G_i, S_2, S_1) * P(G_i|S_2, S_1)}{P(S_3|S_2, S_1)}$
$S_1^{(4)}$ $S_2^{(4)}$ $S_3^{(4)}$ $S_4^{(4)}$
$P(G_i|S_4, \ldots, S_1) = \frac{P(S_4|G_i, S_3, S_2, S_1) * P(G_i|S_3, S_2, S_1)}{P(S_4|S_3, S_2, S_1)}$
$S_1^{(5)}$ $S_2^{(5)}$ $S_4^{(5)}$
$P(G_i|S_5, \ldots, S_1) = \frac{P(S_5|G_i, S_4 \ldots, S_1) * P(G_i|S_4 \ldots, S_1)}{P(S_5|S_4 \ldots, S_1)}$

图 4

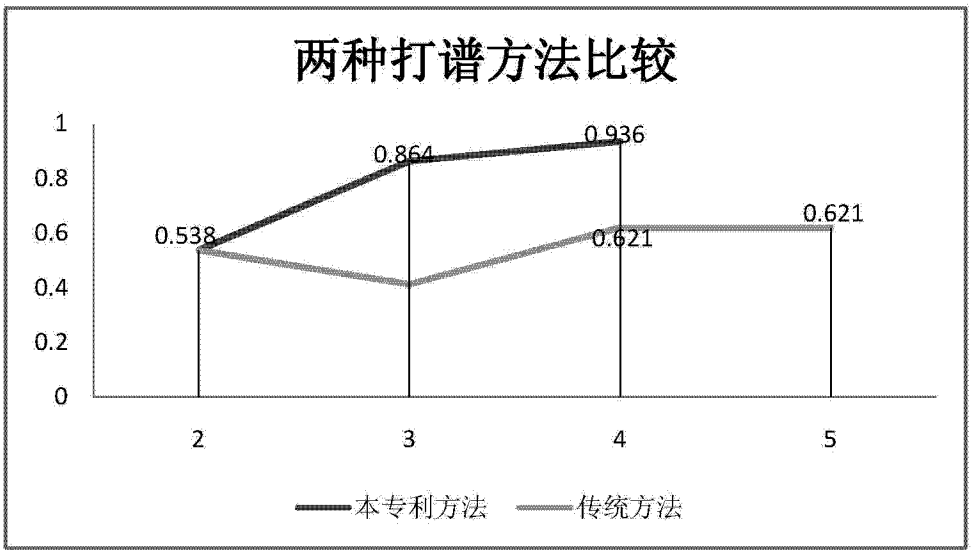

两种打谱方法比较
1
0.8
0.6
0.4
0.2
0
0.538
0.864
0.936
0.621
0.621
2
3
4
5
本专利方法
传统方法

图 5

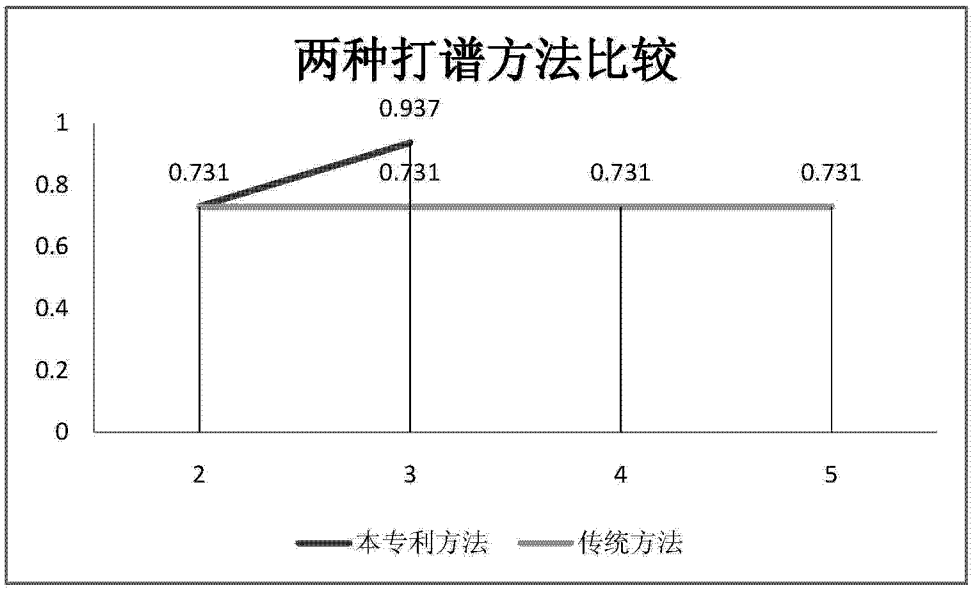

两种打谱方法比较
1
0.8
0.6
0.4
0.2
0
0.731
0.937
0.731
0.731
0.731
2
3
4
5
本专利方法
传统方法

图 6