PIBB 生物化学与生物物理进展
*Progress in Biochemistry and Biophysics*
2017, 44(10): 830~847
www.pibb.ac.cn

# 串联质谱寡糖结构解析方法评介 *

王耀君[1,2] 黄纯翠[3] 高 枫[2] 张敬玮[2] 李 岩[3] 卜东波[2] 孙世伟[2]**
([1] 北京大学光华管理学院，北京 100871；[2] 中国科学院计算技术研究所，北京 100190；[3] 中国科学院生物物理研究所，北京 100101)

**通讯作者简介**

孙世伟，南开大学物理化学理学硕士学位，中国科学院计算技术研究所获得博士学位，现任中国科学院计算技术研究所前瞻研究实验室副研究员，研究工作集中在生物信息学和机器学习．研究方向是蛋白质组学，糖组学中算法及统计问题研究．主要工作获得国家自然基金(No. 31671369，No. 31671369，No. 30800189)的资助．

**摘要** 糖组学的研究与发展对生命科学及生物医药的发展具有重要的推动作用．寡糖结构的解析是糖组学中重要的研究课题之一．串联质谱分析技术以其具有高特异性及高灵敏度的特点成为了广为使用的寡糖结构解析方法．本文首先概述了串联质谱寡糖结构解析的研究背景；然后介绍了现有的寡糖结构解析策略及基于每种策略的经典解析方法，并对所列方法的原理和算法进行逐一分析讨论；最后，总结现有方法的优缺点，对串联质谱寡糖结构研究领域进行了研究展望．

**关键词** 质谱解析，寡糖，算法，寡糖结构解析
**学科分类号** TP311
**DOI**: 10.16476/j.pibb.2016.0393

糖组学是继基因组学和蛋白质组学后生命科学的新兴研究领域．其主要研究内容为：寡糖的空间结构及其在生物机体中所发挥的功能．寡糖作为一种重要的碳水化合物，参与蛋白质折叠和信号转导等多种细胞生命过程，同时，它经常与蛋白质和脂类化合物结合在一起形成糖蛋白和糖脂．糖蛋白常出现于细胞表面，参与细菌和病毒识别以及凝集素等其他蛋白质的识别过程，以此来行使重要的生物学功能．有研究者对蛋白质数据库 SWISS-PROT 中的数据进行了分析，发现超过 50%的蛋白质糖基化修饰[1-3].

寡糖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，寡糖结构的解析是一个包括对糖类分子组成、单糖的连接顺序和连接位点依次逐层解析的过程．图 1 示出了 N 糖 GlcNAc2Man9 结构的多种经典表示方式，虽然表示方式不同，但它们都是以树形结构呈现，每一树形结构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 2 个单糖的糖苷键，数字代表寡糖结构中单糖之间的连接位点，α、β 是指通过半缩醛形成六元环的结构时羟基的相对位置．

与基因组学和蛋白质组学相比，糖组学的研究进展较为缓慢，其主要发展瓶颈在于实现寡糖结构的快速准确解析很难．相对于直链型的 DNA 和氨基酸序列，寡糖的树形结构增加了寡糖结构的解析难度[4-5]．图 2 示出了寡糖结构中 2 个单糖之间通过糖苷键连接的表示方式，每一个单糖通过糖苷键可以跟另外一个或多个单糖相连接形成如同树形的分枝结构，这种分枝结构的复杂性导致了寡糖结构的多样性．

* 国家自然科学基金(31270834, 31671369, 61272318, 30870572, 61303161)资助项目.
** 通讯联系人.
Tel: 010-62600887, E-mail: dwsun@ict.ac.cn
收稿日期：2016-12-31，接受日期：2017-03-21

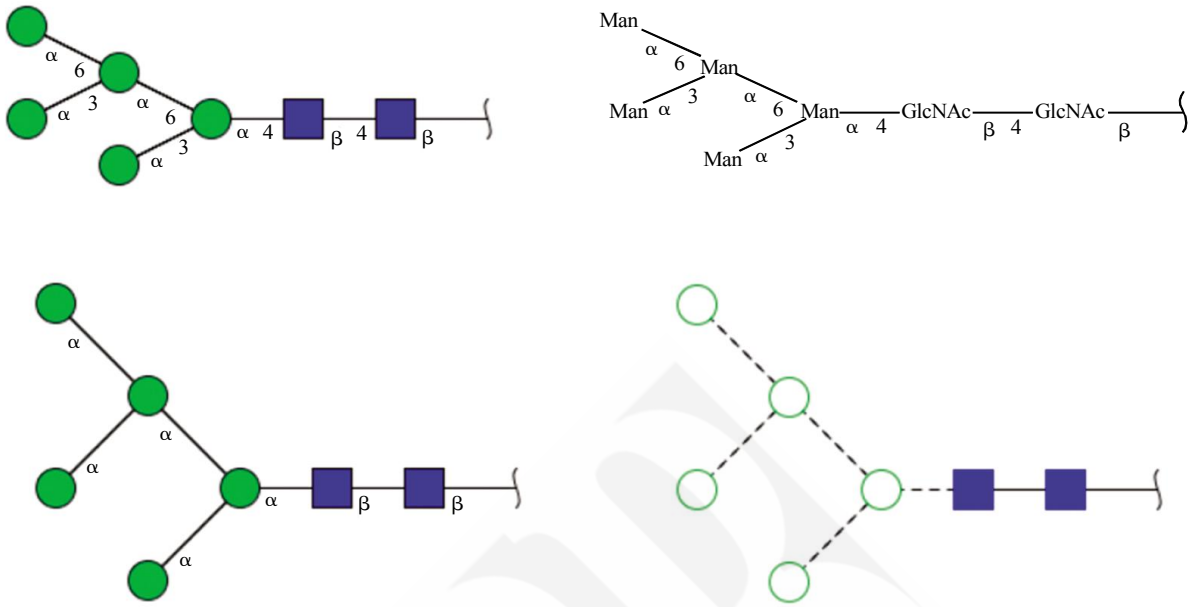


**Fig. 1　Oligosaccharide graphic representation types**

**图 1　寡糖结构的常用表示方式**

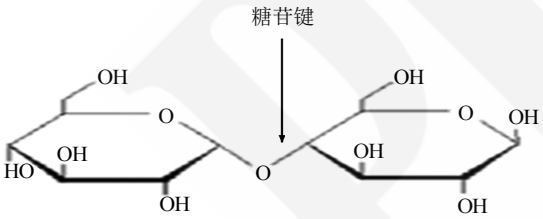


**Fig. 2　Glycosidic bond**

**图 2　糖苷键示意图**

质谱技术的出现和发展极大地促进了蛋白质组学和糖组学的发展，串联质谱技术为蛋白质序列分析及寡糖结构的解析提供了可行性支持．串联质谱解析方法由于具有高通量及高灵敏度等优势，逐渐成为蛋白质鉴定和寡糖结构解析的一种重要的普适性策略[6-8].

串联质谱解析方法的大致流程：首先进行寡糖样品制备，随后进行 HPLC 分离，然后通过 MALDI、ESI 等离子化方式注入质谱仪中，选择某种离子碎裂方式(如 CID、HCD、ETD 等碎裂方式)将寡糖结构碎裂成碎片离子，收集并统计碎片离子对应的质荷比信息($m/z$)和数量信息(Intensity)，得到二级质谱图(MS/MS Spectrum)．最后通过对质谱数据的分析实现寡糖结构解析．在上述解析流程中，样品制备和 HPLC 分离技术的研究已经相对成熟，质谱仪技术随着仪器工业的发展精度越来越高，目前已基本满足质谱解析所需的精度需求，而质谱数据分析是通过分析质谱数据来推断样品中寡糖的组成及结构，是质谱解析流程中最具有挑战的环节．所以，质谱数据分析技术是串联质谱解析方法的发展瓶颈，也是目前寡糖结构解析的主要研究方面[9-13].

早期的质谱数据解析完全是人工分析，生物学家们根据积累的知识和经验对质谱图谱进行研究分析得出样品对应的组成、序列和结构信息．但是对于目前高通量寡糖样品质谱数据解析应用场景下，传统的基于人工手动图谱分析的低效工作方式难以满足当前应用需求．因此，借助于计算机技术的发展，一系列可实现高通量自动或半自动质谱数据解析的软件和方法应运而生．若按解析策略的异同对现有的寡糖串联质谱解析方法进行分类，主要可以归类为以下三种策略：

a．寡糖结构库搜索策略．收集已解析出的寡糖结构，并以对应分子质量建立索引形成寡糖结构库．待解析的质谱解析时，首先用待解析质谱的母离子质量以一定的容差值搜索寡糖结构库，得到候选寡糖结构集，然后对每一候选结构使用质谱仪模

拟碎裂方式产生理论质谱，模拟碎裂是指使用计算机模型模拟寡糖结构在质谱仪中的碎裂方式产生理论质谱的过程，最后，所有候选寡糖结构的理论质谱与待解析质谱进行比较，返回相似度最高的寡糖结构作为解析结果[14-21].

b．*De novo* 策略．根据寡糖结构的生化规则和构成寡糖所有单糖的分子质量作为先验知识，通过质谱图谱中谱峰间的 *m*/*z* 差值，推断寡糖结构的组成糖单元或组成子结构，然后使用算法组装糖单元或寡糖子结构从而形成寡糖结构．如果组装结果为包含多个寡糖候选结构的结构集，则使用某种打分方法从候选集中选择打分最高的候选寡糖结构作为解析结果[22-32].

c．谱库搜索策略．收集已有解析结果的质谱数据，从中选择可信度高的结果以“质谱 - 寡糖结构对”的映射形式收录于数据库中形成谱库．待解析质谱解析时，将待解析质谱与质谱数据库中的质谱进行比较，返回相似度最高的质谱对应的寡糖结构作为解析结果[33-42].

## 1 寡糖结构库搜索策略主要方法

结构库搜索策略对应的方法主要由四部分组成：搜库、模拟糖结构在质谱仪中碎裂的理论谱预测、用于待解析质谱理论谱相似度评价的打分方法设计、结果分析．下面主要根据这四个方面对现有结构库搜索策略涉及的解析方法进行分析研究.

### 1.1 GlycosidIQ[43]

#### 1.1.1 理论谱构造

GlycosidIQ 方法构造的理论谱中包括的离子类型有 A、B、C、X、Y、Z 离子．理论谱构造模型限定产生上述类型离子的碎裂方式共有 4 种，分别是：1 次糖苷键断裂、1 次环内断裂、2 次糖苷键断裂和 1 次环内断裂加 1 次糖苷键断裂．对 4 种碎裂方式产生的离子依据生化规则 过滤，去掉理论上不可能产生的离子，然后形成理论谱.

#### 1.1.2 搜库

GlycosidIQ 使用的结构库为包含有 1 674 个独立糖结构的 GlycoSuiteDB 数据库的衍生库．对衍生库中的每一糖结构构建理论谱形成包含有 1 674 张理论谱的谱库．搜库时，对于未知结构的寡糖质谱按照物种类别和质谱母离子质量进行搜索，得到候选结构集.

#### 1.1.3 打分方法

包含有两种打分方法，Segmentation 打分方法和 Correspondence 打分方法.

Segmentation 打分方法：此打分用来衡量对于候选结构中每一处糖苷键的断裂信息在待解析质谱中的支持度，即对应未知的 B、Y、C、Z 离子，如果每个位置的离子都出现则打分为 1，如果结构中没有找到对应于某个糖苷键处碎裂形成的离子，则打分为 2，如果有 4 个单糖构成的糖结构没有碎裂离子支持则打分为 $4^3$.

Correspondence 打分：待解析质谱中对应匹配离子的强度进行加权．对于同样数目的匹配离子，匹配离子峰 Intensity 值加和，加和值越大的结构打分越高.

GlycosidIQ 对每一候选结构的两种打分结果进行线性相加来得出候选结构的打分.

#### 1.1.4 结果分析

测试数据为 130 张负离子模式的二级质谱，其中有 115 张质谱来自于 O 糖样品，另外 15 张质谱来自于 N 糖样品．候选结构打分第一名对应正确解析结果的占 78%，打分第二名对应正确解析结果的占 17%．对解析结果分析，多数情况下相同打分对应一个或多个候选结构．对所有测试数据结果进行统计分析，打分第一名对应的候选结构数平均为 1.5 个.

#### 1.1.5 方法评价

GlycosidIQ 方法只支持二级质谱数据搜索，方法中用来进行结构筛选的初打分的打分评估对象为糖苷键断裂形成的离子，所以使用的先决条件是待解析质谱中必须含有丰富的糖苷键断裂对应的离子．此方法可以解析负离子模式的质谱，同时也支持 MALDI 和 ESI 离子源产生的正离子模式产生的质谱．方法缺陷是，GlycosidIQ 不能确保得出唯一解析结果，大部分情形下可以把解析结果缩小在包含几个候选结构的候选集中．GlysosidIQ 方法之所以不能够给出唯一解析结果的原因是只使用了二级质谱数据，而由于糖结构的复杂性，在多数情形下只使用二级质谱的信息是难以得出唯一解析结果的.

### 1.2 GlycoFragment/GlycoSearchMS[44]

GlycoFragment/GlycoSearchMS 方法分为两个模块，用于理论谱构造的 GlycoFragment 模块和负责搜库及谱谱比对打分的 GlycoSearchMS 模块.

#### 1.2.1 理论谱库构造

使用 SweetDB 数据库中的寡糖结构作为寡糖结构数据集，运用 GlycoFragment 模块对结构集中的每一个寡糖结构构建理论谱，形成理论谱库.

GlyoFragment 构造的理论谱包括有 6 种离子，分别是：A、B、C、X、Y、Z 离子．形成的理论谱库共包含有 5 000 个 N 糖和 1 200 个 O 糖结构对应的理论谱．

**1.2.2**　搜库

首先进行搜索参数设定，需设定的参数包括 ESI 离子源型号、衍生方式等．搜库时，对于未知结构的寡糖质谱按照质谱母离子质量搜索得到结构候选集，对候选集中的理论谱使用 MSscore 打分方法和待解析的质谱进行谱谱相似度打分，返回打分结果高的结构作为解析结果．

**1.2.3**　打分方法

使用 MScore 打分，公式为：

$$MS_{\text{score}}=\frac{\sum_{1}^{n}[1-(|1-(|P_s-P_t|/Err)]}{n_{\text{input}}}\times 100 \qquad (1)$$

$n$：待解析质谱中谱峰个数；$P_s$：待解析质谱谱峰质荷比 $m/z$ 值；$P_t$：理论谱谱峰质荷比 $m/z$ 值；$Err$：容差值．

**1.2.4**　结果分析

根据 MScore 打分给出打分由高到低排序的候选寡糖结构结果汇总表，点击每一候选结果可以得到对应候选结构和待解析质谱比较的详细信息，包括匹配离子数及匹配离子对应的寡糖子结构标注．在使用过程中，研究人员可根据每一候选结构的打分，辅以谱谱比较的详细信息进行人工筛选，最终得出解析结果．

**1.2.5**　方法评价

GlycoFragment/GlycoSearchMS 方法解析流程非常简单，只使用了一个打分函数对库中糖结构筛选及评估，操作简单、解析速度快．其缺点是：在构造理论谱环节没有使用生化规则过滤，产生的理论谱含有大量噪音数据；打分函数只考虑谱峰的 $m/z$ 值，而未使用谱峰强度 Intensity 信息，致使解析灵敏度低，从而增加了解析结果的假阳性率．与 GlycosidIQ 一样，此方法同样存在质谱数据不足的问题，解析过程中只使用了二级质谱数据，这样势必会有多数解析难以给出唯一结果．

## 2　De novo 策略

*De novo* 策略侧重于算法设计，通过构建算法模型推断质谱数据对应的寡糖结构．由于寡糖结构的复杂性，若只依赖于设计精良算法模型难以给出准确解析结果，于是大部分 *De novo* 寡糖结构解析方法通过引入生化规则对解析结果进行过滤．下面对目前基于 *De novo* 解析策略的方法从三个方面进行分析研究：算法流程设计、生化规则设定以及打分函数设计．

### 2.1　GlycoMod[45]

**2.1.1**　算法流程

a．输入寡糖的质量和设定可选参数．可选参数包括：分子质量容差、离子模式(正离子或负离子模式)、加和物类型(Na 离子、钾离子或三氟乙酸等)、寡糖类型(N 糖或 O 糖)．

b．若解析样品是寡糖化合物，则枚举寡糖的所有可能单糖组合，并引入生化规则进行过滤，然后得出寡糖的所有可能单糖组成．

c．若解析样品是糖肽，根据糖肽对应的蛋白质所属物种，GlycoMod 按糖肽质量搜索 SWISS-PROT 数据库，得出可能被糖基化的肽段及数据库中对应糖基化蛋白质的其他信息．

**2.1.2**　生化规则

a．寡糖组成中单糖个数有限定(表 1)；b．寡糖组成中不可以同时出现 Sulfate 和 Phosphate 这两种单糖；c．Hexose 及 HexNAx 的总数需大于 0；d．Fucose 的个数要小于 Hexose 及 HexNAx 的总数；e．如果 HexNAc 的个数小于等于 2 并且 Hexose 的个数大于 2 时，NeuAc 和 NeuGc 都不可以出现；f．N 糖，非衍生化、甲基化、乙酰化的分子质量上限分别为：8 000 u、10 000 u、13 000 u；g．O 糖，非衍生化、甲基化、乙酰化的分子质量上限分别为：5 000 u、7 000 u、9 500 u．使用穷举法计算寡糖组成，对于 N 糖使用生化规则 a～f；对于 O 糖使用生化规则 a 和 g．

**Table 1　Biochemical rules included in GlycoMod method**

**表 1　GlycoMod 寡糖结构生化规则**

| 单糖名称 | 构成 O 糖的单糖数 | 构成 N 糖的单糖数 |
|---|---|---|
| Hexose | 0～14 | 0～20 |
| HexNAc | 0～14 | 0～20 |
| Deoxyhexose | 0～6 | 0～6 |
| NeuAc | 0～7 | 0～5 |
| NeuGc | 0～7 | 0～5 |
| Pentose | 0～3 | 0～3 |
| Sulphate | 0～6 | 0～3 |
| Phosphate | 0～6 | 0～2 |
| KDN | 0～2 | 0～0 |
| HexA | 0～2 | 0～0 |

**2.1.3** 打分方法

GlycoMod 没有打分方法的设计，使用生化规则筛选后，返回尽可能多的解析结果，候选结构没有打分排名.

**2.1.4** 方法评价

GlycoMod 可以实现对包括非衍生化、甲基化、乙酰化等寡糖及寡糖衍生物寡糖组成的解析. 其主要优势是与 SWISS-PROT/TrEMBL 数据库的对接使得糖基化肽段容易被解析. 方法中使用了普适的寡糖结构单糖组成限定规则，使得给定分子质量可以返回所有可能的寡糖组成候选，这样可以发现很多未被发现的糖组成，有助于进一步发现新的寡糖. 方法的不足在于，只实现了寡糖组成的解析，没有实现寡糖结构的解析.

**2.2 STAT**[46]

**2.2.1** 算法流程

a. 设定初始化参数，参数包括：单糖种类、母离子质量、带电荷数、质量容差、寡糖结构是否包含还原端等.

b. 利用背包算法及根据初始化参数得到的限定规则，枚举给定母离子质量的寡糖结构所有可能组成.

c. 利用背包算法计算质谱数据中每一个离子对应的单糖组成来确定寡糖的子结构.

d. 根据步骤 b 得到的寡糖结构组成枚举寡糖结构所有可能的树形结构，同时以步骤 c 的结果作为限定规则进行过滤，最后得出一个候选结构集.

e. 对候选结构集中所有候选结构逐一计算 *S* 打分，打分由低到高排序.

f. 如果是 N 糖，则过滤掉不含有 N 糖核心的候选结构.

g. 选择打分最高的候选结构进行人工解析.

**2.2.2** 打分方法

**2.2.1** 步骤 e 中使用的 *S* 打分的公式为

$$S = \sum_{i=1}^{n} m_i \tag{2}$$

$m_i$：对于待解析质谱中第 $i$ 个谱峰离子如果由某个候选寡糖结构碎裂形成，则对应的候选结构需断裂的糖苷键的个数；$i$：待解析质谱中第 $i$ 个谱峰离子；$n$：待解析质谱中谱峰离子的个数.

**2.2.3** 方法评价

STAT 方法使用背包算法枚举，使用生化规则过滤得到候选寡糖结构，根据对质谱中的峰离子使用类似于计算寡糖子结构的方法对候选寡糖结构过滤；STAG 方法可以把寡糖候选结构限定在一个较小的范围内，方法简单直观易操作. 其缺点是：计算速度慢、枚举的范围太广；引入的生化规则太少，只有一条针对 N 糖的核心结构过滤规则. 根据方法对应引文所述，在初始限定规则的前提下，$m/z$ 为 1 467 的寡糖可以枚举出 2 290 种候选糖结构；STAT 只可解析由低于 10 个单糖组成的寡糖结构，解析结果需要通过人工解析来最终确定. STAT 中每一步都需要手工选择参数. 需要选择的参数有：单糖种类、母离子质量、带电荷数、是否带还原端；母离子组成；产物离子的筛选和可能组成. STAT 并不算是自动化的方法，每一步都需要依赖于实验人员的经验进行参数设定和结构筛选. 作为最早的寡糖结构 *De novo* 方法，其解析流程和打分方法的设计对后续的解析方法的发展打下了基础.

**2.3 StrOligo**[47-48]

**2.3.1** 算法流程

a. 加载待解析质谱数据，设定可选参数.

b. 质谱去同位素：对待解析质谱进行数据预处理，只保留单一同位素峰，去掉其他同位素峰.

c. 建立关系树：在经过步骤 b 处理过的质谱数据中，如果两个谱峰 $m/z$ 相差正好为 1 个或 2 个单糖的质量，则在关系树种建立 2 个节点，节点值分别对应 2 个谱峰的 $m/z$ 值，节点权重值根据 2 个谱峰的强度分别设定. 对 2 个节点用一条边连线，对所有节点通过节点连线构造关系树，关系树中所有节点对应为质谱中的 B 离子或 Y 离子.

d. 计算寡糖结构的单糖组成. 根据关系树和候选结构打分函数 CompScore 计算寡糖可能的单糖组成. 对于每一个关系树中的节点，枚举所有的可能组合，组合对应的分子质量值接近但不超过关系树中的节点对应的分子质量值. 然后根据质谱母离子 $m/z$ 值用同样方法枚举所有可能单糖组合，并对每一可能组合进行打分. 打分的标准为：母离子对应的候选组合分子质量与母离子质量的差值在给定容差内；候选组合匹配母离子组合关系树中节点的数目越多，打分越高.

e. 计算寡糖结构的树形结构. StrOligo 目前只可用来解析 N 糖. 在流程 d 得到的寡糖组成组合集合中，对每一个组成组合枚举所有可能的糖结构并使用 N 糖的生化规则进行过滤，筛出不可能的结构. 对每一枚举结构通过 StructScore 打分方

法计算其可信度，返回打分最高的候选结构供用户进一步人工解析．根据之前的组合打分以及人工经验，用户选择一些可能的组合计算结构，然后枚举所有可能的结构．对每一候选结构进行模拟碎裂，通过理论碎片离子与关系树的匹配程度对候选结构进行打分．匹配计算采用“向量点积法”．

**2.3.2** 生化规则

复杂型糖(图 3)在 N 糖中普遍存在，所以在 **2.3.1** 步骤(e)中增加了复杂型糖的生化规则：a．Fuc 只能与还原端的 GlcNAc 连接；b．对于哺乳动物 N 糖，三甘露糖核心的中心甘露糖 Mannose 可以再连接一个 GlcNAc；c．天线型寡糖至多可以有 4 分枝；d．与三甘露糖核心的 2 分枝向左连的单糖为 GlcNAc，紧接着可以连接 1 个 Gal；e．复杂型寡糖的树形分枝结构的末端可以是 Neu5Ac 或 Neu5Gc；f．PMP 单元可以连接在与三甘露糖核心还原端的 GlcNAc 上．

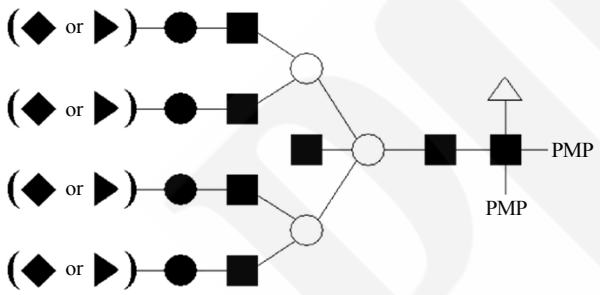


**Fig. 3 Schematic of complex oligosaccharide structures**

**图 3 复杂型寡糖结构构型示意图**

**2.3.3** 打分方法

StrOligo 在解析流程 d 中使用 CompScore 对候选寡糖组成打分，在流程 e 中使用 StructScore 对候选寡糖结构打分．

CompScore：根据关系树跟候选结构组成相吻合的节点数及候选结构分子质量与待解析质谱母离子质量的差值设计线性打分函数．

StructScore：对候选寡糖结构使用模拟碎裂的方式构建理论谱，理论谱跟待解析质谱进行向量点积的运算，结果作为候选结构的打分．

**2.3.4** 方法评价

StrOligo 方法通过去除同位素峰步骤，去除了质谱解析过程中的部分干扰因素，然后分别使用了组成打分和结构打分，实现了从组成到结构的逐步缩小候选结果范围的方式进行寡糖结构解析，提高了解析的灵敏度和解析速度．不足：方法只可以解析 N 糖，且只支持正离子模式的质谱解析；同时，StrOligo 增加的生化规则太多，这么处理对于常见的寡糖可以实现快速准确地解析，但是限制了新结构的发现．

**2.4 GLYCH[8]**

GLYCH 对二级质谱中每个谱峰对应的分子质量，枚举可能的寡糖子结构，然后利用动态规划的思想从小结构逐渐生成大的寡糖树形结构．

**2.4.1** 算法流程

a．待解析质谱预处理．在质谱 *m/z* 维度上，设置一个滑动窗口，其宽度包含 20 个谱峰．将滑动窗口按 *m/z* 值从低到高滑动．在每个窗口中，计算 *m/z* 值在窗口范围内的谱峰强度的平均值以及标准差，然后过滤掉谱峰强度低于均值 3 个标准差的谱峰．

b．枚举谱峰对应的 *PRF* 集．定义 *PRF*={*PRM*, *r*, *b*'}．其中 *PRM* 为前缀质量，即糖结构产生的相应 B 离子质量减去非还原端附加物质量．*R* 为子结构根节点的单糖类型，论文中一共有 6 种类型，见表 2；*b*' 表示单糖之间连接位点，一共有 4 种连接方式(1-2、1-3、1-4、1-6)．待解析质谱中每一个峰对应 A、B、C、X、Y、Z 中的一种离子类型或噪音，假设谱峰 *p* 对应的离子类型为 *i*，子结构根节点对应的单糖类型为 *r*，那么我们可以根据下面公式计算该峰对应的 *PRM*，计算容差设为 0.2 u.

$$\begin{aligned} M_{ir}&=M_p+\delta M_{ir} \quad i=\mathrm{B, C, {}^{0,2}A,\cdots, {}^{3,5}A}\\ M_{ir}&=M_{\mathrm{parent}}-(M_p+\delta M_{ir}) \quad i=\mathrm{Y, Z, {}^{0,2}X,\cdots, {}^{3,5}X}\end{aligned} \tag{3}$$

其中，$M_p$ 表示谱峰的 *m/z*，$\delta M_{ir}$ 表示谱峰选定离子类型 *i* 后与 B 离子的质量差，如果 *i* 为 B 离子，则 $\delta M_{ir}$ 为糖结构附加物质量，如果 *i* 为 C 离子，则为糖结构附加物质量加上 18，如果 *i* 为 A 离子，则 $\delta M_{ir}$ 对应表 3．同理，对应 X、Y 和 Z 离子，可以计算出相应的 $\delta M_{ir}$．所以每个谱峰最多能够计算出 22×6×4=528 种不同的 *PRF*.

c．*PRF* 打分．使用一个简单策略计算每个 *PRF* 的得分，即统计待解析质谱中能产生该 *PRF* 的谱峰个数，记为 $S(M_i, r_i, b_i)$.

d．获取寡糖结构候选集．首先，根据 *PRM* 对所有 *PRF* 进行排序．其次，对于任何 *PRF*(*m*, *r*, *b*)，将 *V*(*m*, *r*, *b*)作为一系列$(M_1, r_1, b_1)$, $(M_2, r_2, b_2)$, …$(M_j=m, r_j, b_j)$对应的寡糖结构的最大得分，其中该寡糖结构的根节点为 $r_j$．那么最大的 *V* 对应的候选结构就是得到的最优候选结构．最后，利用动态规

划算法根据所有的 *PRF* 求解对应的 $V$，从而得到候选结构.

e．候选结构重评估：通过动态规划算法获取候选寡糖结构集合，将候选寡糖结构模拟碎裂产生理论谱，理论谱不仅考虑一次碎裂产生的 B/Y，C/Z，A/X 离子，还考虑两次碎裂产生的离子．然后将理论谱与待解析质谱匹配的峰的个数作为重评估的分值，对候选结构进行重排序.

**2.4.2**　生化规则

a．单链寡糖不同连接位点形成 A/X 离子类型，见表 2.

**Table 2　Biochemical rules included in GLYCH method(a)**

**表 2　GLYCH 包含的寡糖生化规则(a)**

| X/A 离子 | 1-2 | 1-3 | 1-4 | 1-6 |
|---|---|---|---|---|
| $^{0,2}$X/A | − | + | + | + |
| $^{0,3}$X/A | − | − | + | + |
| $^{0,4}$X/A | − | − | − | + |
| $^{1,3}$X/A | + | + | − | − |
| $^{1,4}$X/A | + | + | − | − |
| $^{1,5}$X/A | + | + | − | − |
| $^{2,4}$X/A | + | + | − | − |
| $^{2,5}$X/A | + | + | − | − |
| $^{3,5}$X/A | + | + | − | − |

“+”为在相应连接类型上断裂可以形成 A/X 离子，“−”为对应连接类型上断裂不可形成 A/X 离子；行属性为 2 个单糖的连接的位点；列属性为单糖类型.

b．二分枝寡糖不同连接位点形成 A/X 离子类型，见表 3.

**Table 3　Biochemical rules included in GLYCH method(b)**

**表 3　GLYCH 包含的寡糖生化规则(b)**

| X/A 离子 | 1-2, 1-3 | 1-2, 1-4 | 1-2, 1-6 | 1-3, 1-4 | 1-3, 1-6 | 1-4, 1-6 |
|---|---|---|---|---|---|---|
| $^{0,2}$X/A | + | + | + | − | − | − |
| $^{0,3}$X/A | − | + | + | + | − | − |
| $^{0,4}$X/A | − | − | + | − | + | + |
| $^{1,3}$X/A | − | + | + | + | − | − |
| $^{1,4}$X/A | − | − | + | − | + | + |
| $^{1,5}$X/A | − | − | − | − | + | − |
| $^{2,4}$X/A | + | + | + | − | − | + |
| $^{2,5}$X/A | + | + | + | − | + | − |
| $^{3,5}$X/A | − | + | + | + | + | − |

“+”为在相应连接类型上断裂可以形成 A/X 离子，“−”为对应连接类型上断裂不可形成 A/X 离子；行属性为 2 个单糖的连接的位点；列属性为单糖类型.

c．单糖不同连接位点是否可以形成 A/X 离子类型，见表 4.

**Table 4　Biochemical rules included in GLYCH method(c)**

**表 4　GLYCH 包含的寡糖生化规则(c)**

| X/A 离子 | 1-2 | 1-3 | 1-4 | 1-6 |
|---|---|---|---|---|
| Hex | + | + | + | + |
| HexA | + | + | + | − |
| HexNAc | − | + | + | + |
| Fuc | + | + | + | − |
| Xyl | + | + | + | − |
| NeuAc | − | + | − | − |
| NeuGc | − | + | − | − |

“+”为在相应连接类型上断裂可以形成 A/X 离子，“−”为对应连接类型上断裂不可形成 A/X 离子；行属性为 2 个单糖的连接的位点；列属性为单糖类型.

**2.4.3**　打分方法

GLYCH 的打分主要分为 3 个步骤：首先，对每个 *PRF* 进行打分；其次，利用动态规划计算候选结构打分；最后，对候选结构重 评估.

a．对每个 *PRF* 打分．打分函数设计比较简单，主要统计了待解析质谱中能产生该 *PRF* 峰的个数，个数越多打分越高.

b．动态规划产生候选结构．在整个动态规划过程中，只考虑寡糖结构两分枝情况．首先根据 *PRM* 从小到大对 *PRF* 进行排序，然后根据当前 *PRF* 对应的候选结构与前面 *PRF* 的关联程度计算得分，计算公式如下：

$$V(m, r, b)=S(m, r, b)+\min_{m_1\le m_2\le m}\begin{cases}0 & \text{if } m=\text{mass}(r)\\ V(m_1, r_1, b_1) & \text{if } m=\text{mass}(r)+m_1\\ V(m_1, r_1, b_1)+V(m_2, r_2, b_2) & \text{if } m=\text{mass}(r)+m_1+m_2, b_1\neq b_2\end{cases} \tag{4}$$

其中 $S(m, r, b)$表示 *PRF* 对应的分值，mass($r$)类型为 $r$ 的单糖质量，迭代式分别表示：当前 *PRF* 刚好由 1 个单糖 $r$ 构成；当前 *PRF* 由根节点为 $r$ 的单糖连接 1 个质量为 $m_1$ 的子结构构成的；当前 *PRF* 由根节点为 $r$ 的单糖连接结构质量分别为 $m_1$ 和 $m_2$ 的 2 个分枝结构构成.

c．重评估．将候选结构模拟碎裂构建理论谱，然后将理论谱与待解析质谱匹配的离子峰个数作为候选结构的重评估分值.

**2.4.4** 方法评价

GLYCH 基于二级质谱能够实现对于寡糖结构的解析．通过测试数据集的验证表明，GLYCH 对于直链型和二分枝型寡糖解析效果很好(表 5)，其优点主要包含以下几点：a．只需要二级质谱就能完成寡糖结构的解析；b．在解析过程中不仅考虑 B/Y、C/Z 离子，还同时考虑了 A/X 离子，从而能够在解析寡糖结构的同时，能够解析寡糖结构单糖之间的连接位点信息；c．利用动态规划算法，大大提升了寡糖结构的解析效率，能够解析分子质量更大的糖结构；d．该方法能够很好地扩展到多分枝结构以及 O 糖解析问题．

**Table 5 Results for GLYCH test data set**

**表 5 测试数据验证 GLYCH 结果**

| 类型 | 糖样品 | 正确结果排名 | 打分 | 相同打分候选结构数目 |
|---|---|---|---|---|
| 直链型寡糖 | Hexaose | 1 | 26 | 26 |
| | 3-Sialyllactose | 1 | 19 | 2 |
| | 6-Sialyllactose | 1 | 21 | 2 |
| | Tetraose-a | 1 | 20 | 3 |
| | Tetraose-c | 1 | 20 | 2 |
| 二分枝型寡糖 | Oligomannose | 1 | 25 | 17 |

但是，该策略同样存在不足，主要体现在两个方面：a．在计算质谱谱峰的 *PRF* 时，对于每个谱峰需要进行的枚举次数最多 528 次，在一定程度上影响了寡糖结构解析的效率，特别是进行高通量寡糖结构时解析效率很低；b．该策略的打分函数设计不是很精细，没有考虑质谱的谱峰强度信息等，势必会影响打分的普适性和准确度．

**2.5 CartoonistTwo[49-50]**

**2.5.1** 算法流程

a．待解析质谱数据进行谱峰去噪处理；b．如果有多级质谱的数据，把多级质谱数据合并为一张质谱；c．根据母离子质量计算寡糖的可能单糖组成；d．根据寡糖单糖组成枚举对应的候选寡糖结构，并使用生化规则排除不可能的糖结构，同时限定最大枚举候选结构数不超过 10 000．

**2.5.2** 生化规则

O 糖结构的还原端向左可以连一棵或两棵子树结构，子树结构上如果有分枝结构，则只能是链接一个 Fuc 的情形．即，CartoonistTwo 的生化规则只是对 O 糖分枝结构做了简单的限制．

**2.5.3** 打分方法

CartoonistTwo 方法设计了多个打分函数对解析结果进行评估，分别如下：

a．Basic Scorer：计算候选寡糖结构对应理论谱和待解析质谱匹配峰的个数．

b．Basic+Shedding：在 Basic 打分的基础上加 1 个奖励项，如果根据对应结构的理论谱跟待解析质谱匹配的谱峰可以推断出一条通向根节点的路径则产生 1 个奖励值．

c．Basic+Barking：在 Basic 打分的基础上加 1 个惩罚项，惩罚值的计算方式为理论谱中未在待解析质谱中出现的离子个数乘以 1 个惩罚常量得出惩罚值．

d．Shedding+Barking：在 Basic 打分的基础上同时增加奖励项和惩罚项．

e．Shedding+Barking(Multiple)：在 Basic 打分的基础上同时增加奖励项和惩罚项；对于缺失峰的惩罚值设置权重，权重值的计算方式为：缺失峰的个数除以理论谱谱峰的个数．

**2.5.4** 结果分析

使用 39 组已有确定寡糖结构标注的 O 糖多级质谱数据对 CartoonistTwo 方法进行了验证．结果如表 6，90%的结果打分排在在第一名和第二名的候选结构集中，超过 50%的结果给出正确结果，且解析结果唯一．

**Table 6 Results for CartoonistTwo test data set**

**表 6 测试数据验证 CartoonistTwo 结果**

| | Correct | Tie | Second | Miss | Performance |
|---|---|---|---|---|---|
| Basic Scorer | 7(7) | 27(27) | 2(0) | 3(0) | 0.449(0.502) |
| Basic+Shedding | 9(9) | 25(24) | 2(1) | 3(0) | 0.460(0.514) |
| Basic+Barking | 19(19) | 3(3) | 9(8) | 8(4) | 0.643(0.716) |
| Shedding+Barking | 20(20) | 3(3) | 9(8) | 7(3) | 0.657(0.730) |
| Shedding+Barking(Multiple) | 20(20) | 3(3) | 9(8) | 7(3) | 0.658(0.732) |

**2.5.5** 方法评价

CartoonistTwo 方法的优点体现在两方面：a. 待解析质谱的谱峰预处理和候选结构的打分. 通过统计学方法把待解析质谱中的低谱峰强度的真实谱峰跟噪音谱峰进行了区分. b. 使用了精细的打分方法，具有很高的解析灵敏度.

### 2.6 OSCAR[51]

OSCAR 是 Oligosaccharide Subtree Constraint Algorithm 的缩写，是一种基于多级质谱进行寡糖结构解析的 *De novo* 算法.

**2.6.1** 算法流程

a. 根据待解析质谱母离子质量，在寡糖结构组成库中搜索该质量对应的寡糖结构组成.

b. 枚举寡糖结构组成对应的所有的可能候选结构，然后利用一些生物规则对枚举结构进行过滤.

c. 利用寡糖结构对应的信息，不断过滤掉不合理的候选结构.

**2.6.2** 生化规则

该算法一共使用了近 50 种生物规则，下面主要介绍几个：

a. 寡糖结构规模最多含有 20 个单糖.

b. 对每一个候选糖结构中的各种单糖个数设置个数上限：Hex12 个，HexNAc12 个，Fucose5 个，NeuAc5 个.

c. 根据多级质谱谱峰推断得出的寡糖子结构组成，必须是候选结构中的一个连通子集，如果候选结构中，不存在相应连通子集，则该候选结构可以被过滤掉.

d. 下一级质谱中的谱峰对应的寡糖子结构必须能够从上一级产生的候选结构中碎裂生成.

下面利用一个实例来说明，使用 OSCAR 进行寡糖结构解析的流程，具体过程见图 4. 图 4 中，第 1 列表示待解析样品的多级质谱谱图信息，第 2 列表示质谱对应的可能候选结构，第 3 列对应样品的实际寡糖结构的碎裂情况，用来为解析过程进行注释和对照. 首先，根据母离子质量在寡糖组成库中按照分子质量 1187.6 进行检索，得到对应的寡糖组成 $H_3NR$；然后根据得到的寡糖组成信息枚举所有可能的结构，一共有 26 个候选结构，见图 4 中第 2 列第 1 个方框. 然后根据 894.4($H_3N$-(ene)) 将不满足条件的候选结构进行过滤，得到 18 个候选结构，同理，通过 259.0(H-(oh))将直链结构过滤掉，最后只剩下 1 个分枝结构如图 4 中第 2 列最后一个方框，此时便完成了相应结构的解析.

**2.6.3** 打分方法

OSCAR 没有使用具体的打分方法，只是使用更多的谱信息对候选结构逐步筛选过滤.

**2.6.4** 方法评价

OSCAR 是一种 *De novo* 解析方法，其优点主要有以下几点：a. 该方法不仅能够解析 N 糖和 O 糖的分枝结构，同时能够解析寡糖结构中各单糖之间的连接位点；b. 该方法能够通过选择不同碎片离子构建多级质谱来解析寡糖混合物；c. 该方法解析寡糖结构不依赖寡糖结构库，因此能够实现对于未知结构的解析. 然而，该算法通过多级质谱进行结构解析，仍存在一些不足：a. 该算法的应用必须配备有寡糖的单糖组成库，根据组成库判断相应寡糖结构组成信息，如果组成库中缺失相应谱峰质量对应的寡糖结构或子结构的组成信息，则无法进行后续操作；b. 该算法的输入是多级质谱产生路径，但是没有明确如何选择合理的路径进行解析，不同的实验操作可能会得到不一样的解析结果；c. 如果只通过多级质谱产生路径来作为寡糖结构解析的输入，往往不能得到唯一解析结果.

## 3 谱库搜索策略

谱库搜索策略对应的方法主要由四部分组成：谱库建立、搜库比对、打分、结果分析. 下面的描述和分析从这四部分进行展开.

### 3.1 FragLib[38]

**3.1.1** 谱库建立

FragLib 通过从生化制剂公司购买的有结构标注的寡糖标准品作为谱库的数据源，得到的寡糖标准样品经过甲基化和添加金属离子加和物处理后，使用线性离子阱质谱仪 LTQ，ThermFinnigan 在正离子模式下产生多级谱数据，然后以“质谱 - 寡糖结构”对的形式存放在一个关系数据库中. 谱库数据可以导出为 NIST-MSP 和 XML 等数据格式，便于数据共享.

**3.1.2** 搜库比对

把谱库数据导出为 NIST-MSP 文件格式，利用 NIST 的质谱搜索工具实现谱库搜索和结果打分.

**3.1.3** 打分

使用“$R$ score”打分函数计算待解析质谱和谱库中质谱的相似度打分：

$$R_{\text{score}}=\sum_{i=1}^{N}\left(1-\frac{|\text{IntensityUnknown}_i-\text{IntensityStd}_i|}{\text{IntensityMax}}\right) \quad (5)$$

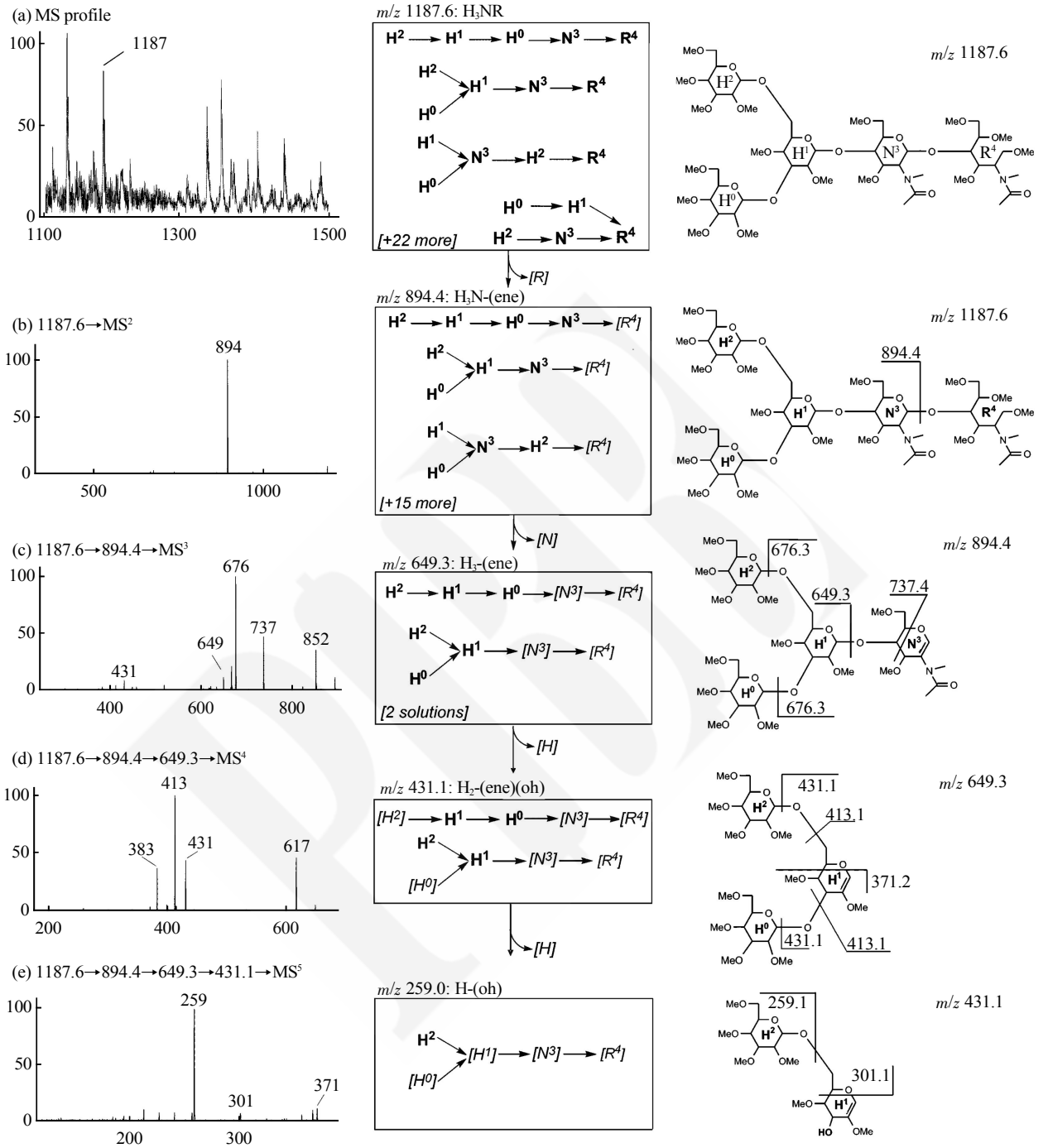


**Fig. 4　OSCAR N Glycan $HexNAc_2Hex_3$ collision path ways for identification**

**图 4　OSCAR N 糖 $HexNAc_2Hex_3$ 基于多级质谱碎裂路径 m/z: 1187.6_894.4_649.3_431.1_259.0 解析流程**

H：Hex，N：HexNAc，R：糖结构还原端单糖 HexNAc，(oh)：碎裂产生 1 个羟基，(ene)：碎裂产生的离子碎片中含有 1 个双键.

IntensityUnKown：待解析质谱的相对谱峰强度；IntensityStd：谱库中谱的相对谱峰强度；IntensityMax：最大相对峰强强度；*N*：谱峰个数. 当 *R* 打分为 0 时代表两张质谱完全不相似，打分为 1000 代表两张参与比较的质谱完全相同.

### 3.1.4　结果分析

待解析寡糖样品产生多级质谱数据，使用 FragLib 方法提供的搜索工具在谱库中搜索对应的结构，然后综合多级谱对应的子结构信息得出寡糖结构.

**3.1.5** 方法评价

FragLib 方法的优点是：a．使用的寡糖样品是经过甲基化后并且添加了金属离子作为加和物，样品如此处理后产生的质谱可重复性高，而且寡糖结构碎裂产生的离子类型多，从而使得质谱包含的信息量丰富．b．支持多级质谱数据的搜索，寡糖结构解析的同时辅以子结构的解析有助于区分同分异构体，另外，通过多个子结构的解析信息的汇总可以发现新的寡糖结构．不足：a．解析过程是半自动化的，搜库比对过程只使用了一个打分函数来计算谱谱相似度，搜索得到候选寡糖结构后需要人工综合多级质谱搜索得到的信息给出判断．这样，解析结果的准确度会依赖于操作者的解谱经验．b．多级谱库的构建过程多级质谱打谱路径没有明确的指导准则，如果包括所有路径的多级谱库是不现实的．因此，FragLib 方法只是提出了一个多级质谱谱库解析方法的框架，实现一个完整的全自动化解析方法还需要进一步改进．

**3.2 Kameyama's Method**[39]

**3.2.1** 谱库建立

从生化制剂公司购买有结构标注的寡糖样品作为质谱谱库数据源．使用基质辅助离子源，三重四级杆离子阱质谱仪对样品处理产生二级质谱和三级质谱；所有寡糖样品都产生二级质谱数据，在二级谱中选取谱峰强度较高的离子峰产生三级质谱．对所有质谱进行数据预处理，使用宽度为 0.7 u 的窗口，$m/z$ 从低到高进行窗口滑动，窗口滑动过程中对窗口内的谱峰合并为单一谱峰，取窗口内最高峰的谱峰强度作为合并谱峰的强度．

**3.2.2** 搜库比对

a．输入寡糖样品的二级质谱数据，与库中所有二级质谱计算相似度，并使用差异性打分函数作为相似度打分函数，返回打分结果．

b．如果排名第一的候选结构差异性打分小于 60，且只有唯一候选则返回第一名的结果作为解析结果．

c．否则，选择使用最优选峰法在二级谱中选择一个离子峰产生三级质谱．

d．如果三级质谱的最小差异性打分小于 70，且只有唯一候选结构则返回差异性打分最小的候选结构作为解析结果．

e．否则，使用最优选峰法从二级谱中选择下一个离子峰产生三级谱，直至最小差异性打分小于 70，返回解析结果．

最优选峰法谱峰选择过程：

a．计算谱库中候选结构中相同 $m/z$ 对应质谱中峰强较大的谱峰作为主要离子峰，计算差异性打分 S1+S2．

b．选择打分最大的 $m/z$ 作为下一级质谱的母离子；如果候选结构大于等于 3 时，对于每一个主要离子峰，计算两两之间的打分，选择最小的打分作为此离子峰的差异性打分．

c．根据步骤 a 和 b 计算得出二级质谱中最优离子峰对应 $m/z$．

**3.2.3** 打分

使用了差异性打分函数，具体公式为：

$$S_1=\sum_{i=1}^{m}(x_i-y_i)^2$$
$$S_2=\sum_{i=1}^{n}(y_i-x_i)^2 \quad (6)$$

定义 $S_1+S_2$ 为谱谱比较的差异性打分，分值越高说明两张质谱的差异性越大，返回分值最小的库结构做为解析结果．

**3.2.4** 结果分析

使用测试数据集验证方法的解析效果，见表 7．对于结构 A、B、C、D、E、H、I 使用二级质谱可以得出解析结果，对应 F、G 使用了三级谱的数据得出解析结果．

**3.2.5** 方法评价

Kameyama's Method 是一种基于谱库搜索策略的寡糖结构解析方法，方法设计了最优选峰法，理论上可以实现以较少的打谱次数实现寡糖结构的解析，在方法引文文献中提供的测试数据上解析效果表现良好．方法的优点是：a．对于没有寡糖结构解析经验的研究者来说，使用 Kameyama's Method 可以实现寡糖结构的解析；b．由于在产生三级质谱时使用最优选峰法挑选有较高区分能力的离子，这样可以在解析过程中减少盲目打谱的次数；c．由于使用了三级质谱数据，所以跟只用二级质谱数据的方法相比，可以得出较为准确的解析结果．方法的不足：此方法的解析准确性和普适性很大程度上依赖于谱库的大小和谱库中质谱数据质量．构建一个包含丰富寡糖质谱数据的大容量谱库需要长时间的积累，需要收集大量的质谱数据，尤其是多级质谱数据；由于寡糖结构的复杂性，导致搜集已有准确解析结果的寡糖质谱数据不易，如果同时收集齐三级质谱、四级质谱，收集难度更大，而且数据

**Table 7 Results for Kameyama's Method test data set**

**表 7 测试数据验证 Kameyama's Method 结果**

| 序号 | 母离子质量 | 候选结构 | | 打分 | | |
|---|---|---|---|---|---|---|
| | | 解析结果 ID | 寡糖标号 | $S_1$ | $S_2$ | $S_1$+$S_2$ |
| 1 | 1417 | ONA-51 | A | 5 | 6 | 11 |
| 2 | 1579 | ONG-a5 | B | 3 | 3 | 6 |
| 3 | 1579 | ONG-cf | C | 5 | 5 | 10 |
| 4 | 1742 | ONG-47 | D | 7 | 7 | 14 |
| 5 | 1563 | ONA-69 | E | 9 | 9 | 18 |
| 6 | 1725 | ONG-cd | F | 4 | 4 | 8 |
| | | ONG-a6 | | 25 | 26 | 51 |
| | 1280/1725 | ONG-cd | | 8 | 8 | 16 |
| | | ONG-a6 | | 129 | 129 | 258 |
| 7 | 1725 | ONG-cd | G | 26 | 25 | 51 |
| | | ONG-a6 | | 5 | 5 | 10 |
| | 1280/1725 | ONG-cd | | 131 | 132 | 263 |
| | | ONG-a6 | | 5 | 6 | 11 |
| 8 | 1888 | ONG-48 | H | 2 | 2 | 4 |
| 9 | 1929 | ONG-a8 | I | 4 | 4 | 8 |
| 10 | 1767 | ONA-a7 | J | 6 | 6 | 12 |
| | | ONA-ad | | 12 | 11 | 23 |
| | 1321/1767 | ONA-a7 | | 13 | 14 | 27 |
| | | ONA-ad | | 62 | 61 | 123 |
| 11 | 2091 | ONG-df | K | 4 | 4 | 8 |
| | | ONG-ac | | 4 | 4 | 8 |
| | 1443/2091 | ONG-df | | 9 | 9 | 18 |
| | | ONG-ac | | 65 | 49 | 114 |

库占用的物理空间会呈现指数级的增长．所以此方法目前不具有普适性．

### 3.3 B2 离子库[33]

B2 离子定义：依照 Domon 和 Costello 提出的寡糖质谱标记法，寡糖结构在非还原端从左向右数第二个糖苷键上氧原子右端位置断裂形成的靠近非还原端的离子称作 B2 离子，如图 5．B2 离子的单糖有 5 种组成，分别是：Hex-Fuc、 Hex-Hex、Hex-HexNAc、 HexNAc-Hex、HexNAc- HexNAc．由于两个单糖之间相互连接的糖苷键类型不同(α、β)以及单糖之间的连接位点不同，所以每种组成对应多种连接方式．此方法作者认为寡糖质谱 B2 离子在直链型、分枝型等各种糖的多级质谱中往往是最容易出现且用它产生下一级质谱，质谱包含有丰富的糖苷键连接信息．

**3.3.1** 谱库建立．正交阻塞法可合成包含有不同单糖类型及多种连接位点的寡糖样品．使用正交阻塞合成法合成寡糖样品，然后对样品进行甲基化及加钠处理，使用型号为 Esquire-LC 和 Bruker-HP ion trap 的质谱仪产生 B2 离子．B2 离子库中存放的是寡糖结构的 B2 离子对应的寡糖子结构及 B2 离子对应的下一级质谱(图 6)．

**3.3.2** 搜库比对

a．待解析寡糖样品使用其他解析方法得到对应寡糖构型．

b．在寡糖样品二级质谱(MS2)中选择 B2 离子产生下一级质谱．

c．搜索 B2 离子库，根据谱谱相似度比较，返回 B2 离子对应寡糖子结构的连接位点信息．

d．对于谱中的非 B2 离子的其他离子，直接

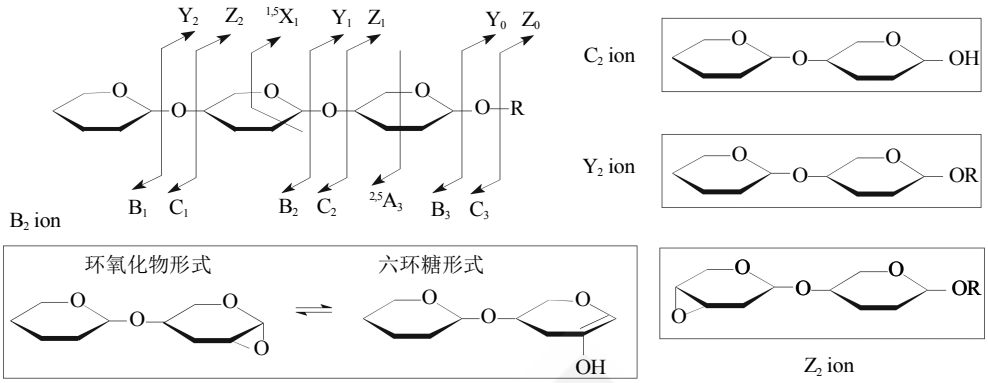


**Fig. 5　Glycosidic bond dissociation manners and ion types**

**图 5　寡糖结构糖苷键碎裂方式及离子类型**

(a) Hex→Fuc (*m*/*z*331)

| | Hexα1→3Fuc | Hexβ1→3Fuc | Hexβ1→4Fuc |
|---|---|---|---|
| *m*/*z*169 | $Y_1$ | $Y_1$ | $Y_1$ |
| *m*/*z*185 | $B_1$ | $B_1$ | $B_1$ |
| *m*/*z*257 | | $^{0,3}X_0$ | |

(b) Hex→HexNAc (*m*/*z*388)

| | Hexβ1→HexNAc | Hexβ1→4HexNAc | Hexβ1→6HexNAc |
|---|---|---|---|
| *m*/*z*185 | $B_1$ | $B_1$ | |
| *m*/*z*203 | $C_1$ | $C_1$ | $C_1$ |
| *m*/*z*208 | $Z_1$ | | |
| *m*/*z*226 | $Y_1$ | $Y_1$ | |
| *m*/*z*245 | | | $^{0,4}A_2$ |
| *m*/*z*259 | | $^{3,5}A_2$ | |
| *m*/*z*275 | | | $^{0,3}A_2$ |
| *m*/*z*287 | | $^{0,2}A_2$-$H_2O$ | |
| *m*/*z*328 | | $^{0,4}X_0$ | $^{0,4}X_0$ |
| *m*/*z*370 | | $B_2$-$H_2O$ | $B_2$-$H_2O$ |

(c) HexNAc→Hex (*m*/*z*388)

| | HexNAcβ1→Hex | HexNAcβ1→4Hex | HexNAcβ1→6Hex |
|---|---|---|---|
| *m*/*z*185 | $Y_1$ | $Y_1$ | $Y_1$ |
| *m*/*z*226 | $B_1$ | $B_1$ | $B_1$ |
| *m*/*z*244 | | $C_1$ | $C_1$ |
| *m*/*z*286 | | | $^{0,4}A_2$ |
| *m*/*z*316 | | | $^{0,3}A_2$ |
| *m*/*z*370 | | $B_2$-$H_2O$ | $B_2$-$H_2O$ |

(d) HexNAc→HexNAc (*m*/*z*429)

| | HexNAcβ1→3HexNAc | HexNAcβ1→4HexNAc | HexNAcβ1→6HexNAc |
|---|---|---|---|
| *m*/*z*226 | $B_1$ or $Y_1$ | $B_1$ or $Y_1$ | $B_1$ or $Y_1$ |
| *m*/*z*244 | $C_1$ | | |
| *m*/*z*286 | | | $^{0,4}A_2$ |
| *m*/*z*316 | | | $^{0,3}A_2$ |
| *m*/*z*370 | | | $^{2,4}X_0$ |

(e) Hex→Hex (*m*/*z*347)

| | Hexβ1→2Hex | Hexβ1→3Hex | Hexβ1→4Hex | Hexβ1→6Hex |
|---|---|---|---|---|
| *m*/*z*185 | $B_1$ / $Y_1$ | $B_1$ or $Y_1$ | $B_1$ or $Y_1$ | $B_1$ or $Y_1$ |
| *m*/*z*203 | $C_1$ | $C_1$ | | $C_1$ |
| *m*/*z*245 | | | | $^{0,4}A_2$ |
| *m*/*z*275 | | | | $^{0,3}A_2$ |
| *m*/*z*329 | $B_2$-$H_2O$ | | $B_2$-$H_2O$ | $B_2$-$H_2O$ |

**Fig. 6　Different B2 ion fragmentation patterns of B2 ion library**

**图 6　B2 离子库中不同类型 B2 离子的离子模式**

产生下一级质谱，在下一级质谱中选择 B2 离子再产生下一级质谱，然后转到步骤 c；以此迭代直到解析出结构中所有连接位点信息.

e. 返回包含有连接位点信息的寡糖解析结果.

### 3.3.3 打分

方法的介绍文献中没有提及使用何种打分方法.

### 3.3.4 结果分析

图 7 作为一个使用 B2 离子库方法进行解析的样例来对解析结果进行分析，样例中的待解析样品的主要成分为母离子质量为 862Th 的寡糖 Galβ1-3(Galβ1-4GlcNAcβ1-6)GalNAcα-OBn,通过加钠和非甲基化处理后产生多级质谱，使用 B2 离子库方法来确定糖结构的连接位点. 一级谱中的离子主要是

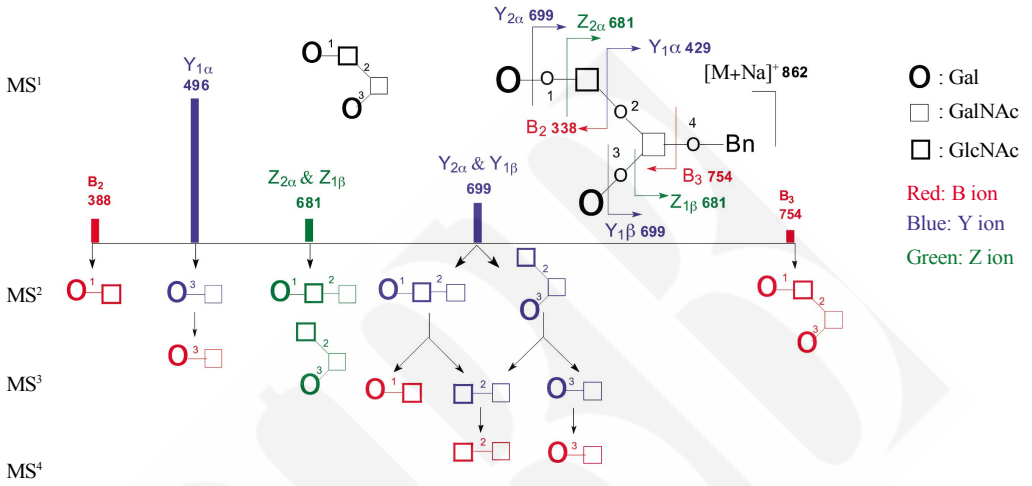


**Fig.7　An example of B2 ion library method**

**图 7　B2 离子库方法解析实例**

以糖苷键碎裂形成的 B、Z、Y 离子. 具体分析步骤为：

a. 在 MS2 质谱中找到 B2 离子，通过质谱仪产生对应的 MS3，然后通过搜索 B2 离子库确定寡糖分枝结构的上分枝非还原端的 2 个单糖间的连接位点.

b. 对于属于非 B2 离子的 Y1α 离子，直接产生 MS3，从 MS3 中选择 B2 离子通过质谱仪产生 MS4，再次搜库确定分枝结构下分枝非还原端的 2 个单糖间的连接位点.

c. 同样，对于属于非 B2 离子的 Y2α 和 Y1β 对应谱峰 699Th，直接产生 MS3；Y2α 和 Y1β 分别对应分枝二糖子结构 Hex-(HexNAc-)HexNAc-O-Bn 和直链二糖子结构 Hex-HexNAc-HexNAc-O-Bn；谱峰 699Th 产生的 MS3 中包含 3 个主峰离子，其中 B2 离子对应的子结构在步骤 a 中已解析出.

d. 对另外 2 个离子产生 MS4，通过 MS4 中的 B2 离子可以确定子结构 HexNAc-HexNAc 的连接位点.

### 3.3.5 方法评价

根据解谱经验得出由不同连接位点构成的 B2 离子产生的下一级质特异性很高，把不同的单糖及位点组合的单糖对应的 B2 离子谱存于数据库中用于寡糖连接位点的解析. 对于待解析的寡糖样品，在确定了分枝结构的前提下，通过 B2 离子库搜索，可以进一步确定连接位点.

B2 离子库方法对仪器类型及样品制备比较敏感，方法只验证了使用 ESI 离子阱质谱仪并且对寡糖样品非甲基化加钠处理的实验环境下解析的有效性，对于其他实验环境下的解析效果有待进一步验证. 方法中没有包含明确的打分函数，任何计算机辅助的解析方法在实际应用时是需要设计打分函数的，且不同的打分函数设计对解析结果有一定影响. 同时，B2 离子库方法只是用来解析寡糖的连接位点，实现糖结构的解析需要先使用其他解析方法来确定寡糖的分枝结构，然后进一步实现位点的解析. 总之，B2 离子库方法只是提出了一种寡糖糖结构连接位点的解析思路，无法直接用来解析寡糖结构.

## 4 现有策略的总结及研究展望

结构库搜索策有 GlycosidIQ 和 GlycoFragment/GlycoSearchMS 方法，结构库搜索的的优势是速度快、操作简单．缺陷是：寡糖结构库搜索策略依赖于理论质谱的预测．目前，由于寡糖结构的复杂性，已有的研究工作对于寡糖质谱的形成机制认识仍然有局限，导致现有的涉及到质谱理论谱预测的解析方法，理论质谱预测的精度都不高，从而影响了解析结果的准确性．同时，这两种结构库搜索方法只用到了二级质谱数据，二级质谱仅提供了有限的信息，从而导致大部分解析无法得出唯一解析结果．

在蛋白质组学中，蛋白质序列库是通过基因翻译得到蛋白质序列来实现的，翻译得到的序列跟实际序列相比准确性高，而且可以覆盖相应物种的所有可能的蛋白质序列．然而对于寡糖来说没有可以生成寡糖结构的模板，构建寡糖结构库只能是依赖于收集先前解析得出的寡糖结构作为结构库数据源．在先前的解析结果中难免存在解析错误的情形，所以导致收集在库中的寡糖结构准确性无法保证．同时，由于没有模板，可能的寡糖结构是没法预估，从而导致库中结构数目的上界无法确定．

*De novo* 策略的方法有：GlycoMod、STAT、CartoonistTwo、StrOligo、GLYCH 和 OSCAR．*De novo* 策略方法解析准确度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，部分离子的缺失会导致 *De novo* 策略无法得出准确的解析结果．此外，由于 *De novo* 策略要枚举所有可能的寡糖结构，通常解析速度较慢．目前的这几种基于 *De novo* 策略的解析方法在多数情况不能给出唯一解析结果，得到唯一解析结果需要进一步人工解析．

谱库搜索策略包括有：FragLib、Kameyama’s Method 及 B2 离子库．策略的优点是：与结构库搜索策略相比，谱库搜索策略使用已知的真实质谱而不是预测出的理论谱与待解析质谱进行比较，从而使得数据比较更加准确，解析结果更可信．由于使用的谱库依赖于已解析过的质谱，所以基于谱库搜索策略的解析方法对于谱库中未收录的质谱则无法给出准确解析．为了实现寡糖结构的准确解析，谱库搜索策略寡糖结构解析方法除了需要收集二级质谱数据还需要收集大量的多级质谱数据，这种策略的瓶颈在于质谱数据的收集．由于目前没有成规模的寡糖谱库，所以基于谱库策略的解析方法在实际的寡糖解析中目前难以应用．

串联质谱寡糖结构解析方法大部分产生于2000～2005 年期间，近几年从计算的角度来解决寡糖解析的方法几乎没有，研究者们试图从其他非计算的角度来提高寡糖解析的准确度．如，牛津大学的 Harvey 等[52]尝试用离子淌度质谱，配合正、负离子 MALDI 来解析 N 糖．此方法不仅能够得到传统多级谱图，还可以根据离子迁移率区分同分异构，以及对样品中含量较少的糖通过改变加合物进行信号放大再解析．这种方法从生物实验的角度提高了质谱所反映的寡糖结构信息量．

实现寡糖结构准确解析目前还有很大的研究空间．笔者认为未来关于串联质谱寡糖结构解析有如下几个研究趋势：

a．多质谱仪联用解析

由于不同类型质谱仪的质谱产生机理不同，某些类型质谱仪易于产生糖苷键碎裂的离子，有些质谱仪易于产生糖结构环内碎裂离子．如果对同一样品分别使用不同类型的质谱仪产生多种质谱，这样可得到包含有丰富寡糖结构信息的质谱，从而利用更多的结构碎裂信息提高解析准确度．例如，对同一寡糖样品分别使用 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寡糖结构解析，与独立使用两张质谱进行解析相比，合并谱可以提高解析准确度．多质谱仪联用解析方案的实施难点在于：需要借助多台不同类型的质谱仪，通过牺牲时间和成本来提高解析准确度．

多仪器联用解析的具体可研究内容有：不同类型仪器产生的质谱数据中离子类型分布规律的研究；不同类型寡糖结构解析所需要的质谱仪类型研究；质谱数据去噪及多同样品多质谱数据合并方法的研究；适用于多质谱仪寡糖解析方法的研究．

b．多级质谱解析

多级质谱解析成为了区分寡糖同分异构的重要方式．串联多级质谱可以有助于解析同分异构寡糖结构的分枝、连接位点和异头物等．

多级质谱技术可通过串联质谱仪，以层次递进的方式产生多级质谱，使用适用于多级质谱解析的打分算法集合多级质谱数据进行寡糖结构解析．多级质谱通过提供更多的糖结构信息，实现同分异构体的逐步区分．通常情况下，使用二级质谱可以对

直链型的蛋白质肽段实现准确鉴定，但是对于寡糖结构，二级质谱往往难以给出准确唯一的解析结果．其原因在于：第一，超过 80%的寡糖由少于 7 种单糖组成，且不同的单糖对应相同的分子质量，例如 Gal 和 Glu 这两种单糖具有相同的分子质量 162.05 u，这样导致同一分子质量的寡糖具有多个同分异构体．第二，多数寡糖结构呈现树形分枝结构．这样导致同一分子质量的寡糖具有多个同分异构体，寡糖结构碎裂后得到的多数离子具有相同的质量．这样，二级质谱中的谱峰难以提供足够的信息实现寡糖结构的准确解析，需要借助多级质谱技术，来提高解析准确度．

多级质谱解析的具体研究内容有：适用于多级质谱解析的结构库搜索方法研究；适用于多几周解析的 *De novo* 解析方法研究．

c．现有解析方法联用

现有方法由于方法设计原理不同，具有不同的侧重点．如 GlycoMod 侧重于解析寡糖结构组成的解析，GLYCH 对于直链型的糖可以实现准确解析，B2 离子库方法在已知寡糖分枝结构的前提下，可以实现寡糖位点的解析．对现有方法的有效联用可以提高寡糖结构的解析准确度．

现有解析方法联用的研究内容有：同一样品多种方法结果的分析研究；多方法联用体系的设计研究．

## 5 总　　结

串联质谱技术以其高通量、高灵敏度的特点逐步成为寡糖结构解析的主要方法．现有的解析方法基本实现了寡糖组成的准确解析以及部分分枝结构的准确解析．对于多数寡糖结构，现有的解析方法未能给出准确的分枝结构和位点信息，而这些信息对于糖组学的研究发展及医疗诊断中的实际应用至关重要．要实现从组成、分枝结构、位点三方面实现寡糖结构的准确解析，串联质谱寡糖解析过程中所涉及到的诸多实验问题及算法问题亟待改进．希望本文的研究可为对本领域的研究者在糖组学中开展下一步研究工作提供参考，为准备进入本领域的研究者提供帮助．

**致谢**　感谢编辑与审稿人在文章改进过程中提出的宝贵意见和建议．

## 参 考 文 献

[1] Goldberg D, Bern M, North S J, *et al*. Glycan family analysis for deducing N-glycan topology from single MS. Bioinformatics, 2009, **25**(3): 365−371

[2] Leymarie N, Zaia J. Effective use of mass spectrometry for glycan and glycopeptide structural analysis. Analytical Chemistry, 2012, **84**(7): 3040−3048

[3] Aoki-Kinoshita K F. An introduction to bioinformatics for glycomics research. PLoS Computational Biology, 2008, **4** (5): e1000075

[4] Domon B, Costello C E. A systematic nomenclature for carbohydrate fragmentations in FAB-MS/MS spectra of glycoconjugates. Glycoconjugate Journal, 1988, **5**(4): 397−409

[5] Stephens E, Maslen S L, Green L G, *et al*. Fragmentation characteristics of neutral N-linked glycans using a MALDI-TOF/TOF tandem mass spectrometer. Analytical Chemistry, 2004, **76**(8): 2343−2354

[6] Liu W T, Ng J, Meluzzi D, *et al*. Interpretation of tandem mass spectra obtained from cyclic nonribosomal peptides. Analytical Chemistry, 2009, **81**(11): 4200−4209

[7] Bunkenborg J, Matthiesen R. Interpretation of collision-induced fragmentation tandem mass spectra of posttranslationally modified peptides. Methods in Molecular Biology, 2007, **367** (1007): 169−194

[8] Tang H, Mechref Y, Novotny M V. Automated interpretation of MS/MS spectra of oligosaccharides. Bioinformatics, 2005, **21**(suppl 1): i431−i439

[9] Von Der Lieth C W, Lutteke T, Frank M. The role of informatics in glycobiology research with special emphasis on automatic interpretation of MS spectra. Biochimica et Biophysica Acta, 2006, **1760**(4): 568−577

[10] Rakus J F, Mahal L K. New technologies for glycomic analysis: toward a systematic understanding of the glycome. Annual Review of Analytical Chemistry, 2011, **4**(4): 367−392

[11] El-Aneed A, Cohen A, Banoub J. Mass spectrometry, review of the basics: electrospray, MALDI, and commonly used mass analyzers. Appl Spectrosc Rev, 2009, **44**(3): 210−230

[12] Hart G W, Copeland R J. Glycomics hits the big time. Cell, 2010, **143**(5): 672−676

[13] Frank M, Schloissnig S. Bioinformatics and molecular modeling in glycobiology. Cellular and Molecular Life Sciences: CMLS, 2010, **67**(16): 2749−2772

[14] Sadygov R G, Cociorva D, Yates J R, 3rd. Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. Nature Methods, 2004, **1**(3): 195−202

[15] Xu H, Wang L, Sallans L, *et al*. A hierarchical MS2/MS3 database search algorithm for automated analysis of phosphopeptide tandem mass spectra. Proteomics, 2009, **9**(7): 1763−1770

[16] Kim S, Mischerikow N, Bandeira N, *et al*. The generating function

of CID, ETD, and CID/ETD pairs of tandem mass spectra: applications to database search. Molecular & Cellular Proteomics: MCP, 2010, **9**(12): 2840-2852

[17] Edwards N J. Protein identification from tandem mass spectra by database searching. Methods in Molecular Biology, 2011, **694**(694): 119-138

[18] Dimaggio P A, Jr., Floudas C A, Lu B, *et al*. A hybrid method for peptide identification using integer linear optimization, local database search, and quadrupole time-of-flight or OrbiTrap tandem mass spectrometry. Journal of Proteome Research, 2008, **7** (4): 1584-1593

[19] Schwilkowski B, Zhang N, Aebersold R. ProbID: A probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. Proteomics, 2002, **2** (10): 1406-1412

[20] Zhang J, Li J, Liu X, *et al*. A nonparametric model for quality control of database search results in shotgun proteomics. BMC Bioinformatics, 2008, **9**(1): 29

[21] Victor K G, Murgai M, Lyons C E, *et al*. MAZIE: a mass and charge inference engine to enhance database searching of tandem mass spectra. Journal of the American Society for Mass Spectrometry, 2010, **21**(1): 80-87

[22] Zhang J, Xin L, Shan B, *et al*. PEAKS DB: *de novo* sequencing assisted database search for sensitive and accurate peptide identification. Molecular & Cellular Proteomics: MCP, 2012, **11**(4): M111 010587

[23] Ma B, Johnson R. *De novo* sequencing and homology searching. Molecular & Cellular Proteomics: MCP, 2012, **11**(2): O111 014902

[24] Iossifov I, Ronemus M, Levy D, *et al*. *De novo* gene disruptions in children on the autistic spectrum. Neuron, 2012, **74**(2): 285-299

[25] Zhang D, Liu H, Zhang S, *et al*. An effective method for *de novo* peptide sequencing based on phosphorylation strategy and mass spectrometry. Talanta, 2011, **84**(3): 614-622

[26] Hughes C, Ma B, Lajoie G A. *De novo* sequencing methods in proteomics. Methods in Molecular Biology, 2010, **604** (604): 105-121

[27] Chi H, Sun R X, Yang B, *et al*. pNovo: *de novo* peptide sequencing and identification using HCD spectra. Journal of Proteome Research, 2010, **9**(5): 2713-2724

[28] Peltoniemi H, Joenvaara S, Renkonen R. *De novo* glycan structure search with the CID MS/MS spectra of native N-glycopeptides. Glycobiology, 2009, 19(7): 707-714

[29] Ma B, Lajoie G. *De novo* interpretation of tandem mass spectra. Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [*et al*], 2009, Chapter 13(Unit 13.10)

[30] Bocker S, Rasche F. Towards *de novo* identification of metabolites by analyzing tandem mass spectra. Bioinformatics, 2008, **24** (16): i49-i55

[31] Frank A M, Savitski M M, Nielsen M L, *et al*. *De novo* peptide sequencing and identification with precision mass spectrometry. Journal of Proteome Research, 2007, **6**(1): 114-123

[32] Frank A, Pevzner P. PepNovo: *De novo* peptide sequencing *via* probabilistic network modeling. Analytical Chemistry, 2005, **77**(4): 964-973

[33] Xue J, Laine R A, Matta K L. Enhancing MS n mass spectrometry strategy for carbohydrate analysis: A b 2 ion spectral library. Journal of Proteomics, 2015, **112**(112): 224-249

[34] Lam H, Zhang X, Li Y Z, *et al*. Understanding the improved sensitivity of spectral library searching over sequence database searching in proteomics data analysis. Proteomics, 2011, **11** (6): 1075-1085

[35] Barsnes H, Vaudel M, Colaert N, *et al*. compomics-utilities: an open-source Java library for computational proteomics. BMC Bioinformatics, 2011, **12**(1): 70

[36] Ye D, Fu Y, Sun R X, *et al*. Open MS/MS spectral library search to identify unanticipated post-translational modifications and increase spectral identification rate. Bioinformatics, 2010, **26**(12): i399-406

[37] Yen C Y, Meyer-Arendt K, Eichelberger B, *et al*. A simulated MS/MS library for spectrum-to-spectrum searching in large scale identification of proteins. Molecular & Cellular Proteomics: MCP, 2009, **8**(4): 857-869

[38] Zhang H, Singh S, Reinhold V N. Congruent strategies for carbohydrate sequencing. 2. FragLib: An MS n spectral library. Analytical Chemistry, 2005, **77**(19): 6263-6270

[39] Kameyama A, Kikuchi N, Nakaya S, *et al*. A strategy for identification of oligosaccharide structures using observational multistage mass spectral library. Analytical Chemistry, 2005, **77**(15): 4719-4725

[40] Elias J E, Gibbons F D, King O D, *et al*. Intensity-based protein identification by machine learning from a library of tandem mass spectra. Nature Biotechnology, 2004, **22**(2): 214-219

[41] Yates J R, Morgan S F, Gatlin C L, *et al*. Method to compare collision-induced dissociation spectra of peptides: Potential for library searching and subtractive analysis. Analytical Chemistry, 1998, **70**(17): 3557-3565

[42] Stein S E, Scott D R. Optimization and testing of mass-spectral library search algorithms for compound identification. Journal of the American Society for Mass Spectrometry, 1994, **5**(9): 859-866

[43] Joshi H J, Harrison M J, Schulz B L, *et al*. Development of a mass fingerprinting tool for automated interpretation of oligosaccharide fragmentation data. Proteomics, 2004, **4**(6): 1650-1664

[44] Lohmann K K, Von Der Lieth C-W. GlycoFragment and GlycoSearchMS: web tools to support the interpretation of mass spectra of complex carbohydrates. Nucleic Acids Research, 2004, **32**(suppl 2): W261-W266

[45] Cooper C A, Gasteiger E, Packer N H. GlycoMod - a software tool for determining glycosylation compositions from mass spectrometric data. Proteomics, 2001, **1**(2): 340-349

[46] Gaucher S P, Morrow J, Leary J A. STAT: a saccharide topology analysis tool used in combination with tandem mass spectrometry. Analytical Chemistry, 2000, **72**(11): 2331-2336

[47] Ethier M, Saba J A, Ens W, *et al*. Automated structural assignment of derivatized complex N - linked oligosaccharides from tandem mass spectra. Rapid Communications in Mass Spectrometry, 2002,

**16**(18): 1743−1754

[48] Ethier M, Saba J A, Spearman M, *et al*. Application of the StrOligo algorithm for the automated structure assignment of complex N‐linked glycans from glycoproteins using tandem mass spectrometry. Rapid communications in Mass Spectrometry, 2003, **17** (24): 2713−2720

[49] Goldberg D, Sutton-Smith M, Paulson J, *et al*. Automatic annotation of matrix-assisted laser desorption/ionization N-glycan spectra. Proteomics, 2005, **5**(4): 865−875

[50] Goldberg D, Bern M, Li B, *et al*. Automatic determination of O-glycan structure from fragmentation spectra. Journal of Proteome Research, 2006, **5**(6): 1429−1434

[51] Lapadula A J, Hatcher P J, Hanneman A J, *et al*. Congruent strategies for carbohydrate sequencing. 3. OSCAR: An algorithm for assigning oligosaccharide topology from MS n data. Analytical Chemistry, 2005, **77**(19): 6271−6279

[52] Harvey D J, Scarff C A, Crispin M, *et al*. MALDI-MS/MS with traveling wave ion mobility for the structural analysis of N-linked glycans. Journal of the American Society for Mass Spectrometry, 2012, **23**(11): 1955−1966

# Survey on Algorithms for Tandem Mass Spectrometry Oligosaccharide Identification*

WANG Yao-Jun[1,2], HUANG Chun-Cui[3], GAO Feng[2], ZHANG Jing-Wei[2], LI Yan[3], BU Dong-Bo[2], SUN Shi-Wei [2]**

([1] *Guanghua School of Management, Peking University, Beijing* 100871, China;
[2] *Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190, China;
[3] *Institute of Biophysics, Chinese Academy of Sciences, Beijing* 100101, China)

**Abstract** Glycomics research plays an important role in life sciences and biomedicine industry. Oligosaccharide structure identification is one of the important research topics of glycomics. The development of high-throughput mass spectrometry technology in recent decade contributes significantly to oligosaccharide structure identification. In this review, we introduced the research background of tandem mass spectrometry assisted glycan structure identification. Then, we reviewed the current strategies used for glycan identification and analyzed the key technologies of existing methods. Finally, we summarized the advantages and disadvantages of the existing methods. The outlook on utilization of tandem mass spectrometry to assist oligosaccharide structure identification is also discussed.

**Key words** mass spectrometry, oligosaccharide, algorithm, oligosaccharide identification
**DOI**: 10.16476/j.pibb.2016.0393

* This work was supported by grants from The National Natural Science Foundation of China (31270834, 31671369, 61272318, 30870572, 61303161).
**Corresponding author.
Tel: 86-10-62600887, E-mail: dwsun@ict.ac.cn
Received: December 31, 2016 Accepted: March 21, 2017