PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2017, 44(10): 821~829
www.pibb.ac.cn

N-连接完整糖肽的质谱分析策略和研究方法 *

朱伯婧　智　渊　孙士生 **
(西北大学生命科学学院，西安 710069)

通讯作者简介

孙士生，西北大学生命科学学院教授，博士生导师，中组部“青年千人”和陕西省“百人计划”获得者．2011 年在西北大学获得博士学位，2011.9~2017.1 在美国约翰•霍普金斯大学病理系 Hui Zhang 教授实验室从事博士后研究．一直从事糖蛋白质组学相关新技术的建立和新软件的开发、新的肿瘤标志物发现和重要糖蛋白的功能研究工作．其建立的 NGAG 方法(*Nature Biotechnology*, 2016)不仅可定量比较不同生物和临床样品间 N- 连接糖链和糖蛋白整体水平的差异，还可进一步分析各蛋白上的糖基化位点及各位点糖链结构和糖基化率．曾作为样本制备主要负责人全面参与了美国临床蛋白质组肿瘤分析(CPTAC)项目，并顺利完成了人类卵巢癌蛋白质组和糖蛋白质组图谱的绘制．其研究成果发表在 *Nature Biotechnology*, *Cell*, *Analytical Chemistry* 和 *Journal of Proteome Research* 等杂志．

摘要　蛋白质糖基化作为最普遍、最重要的蛋白质修饰，一直是组学研究的焦点之一．近十几年来，N- 连接糖蛋白质组学研究普遍采用的方法是将糖链与所修饰的多肽分开进行分析．该策略虽降低了分析难度，却也丢失了糖链与蛋白质糖基化位点间重要的对应关系信息．近年来，完整糖肽的质谱分析策略和方法逐步建立起来．总体而言，要实现对完整糖肽的直接质谱分析，首先需要从复杂样品中富集完整糖肽以消除非糖基化多肽对完整糖肽分析的影响，然后在质谱分析中还需要根据糖肽特性调整相应质谱分析参数，最后在后续数据分析中还需要开发相应的分析软件以完成完整糖肽中多肽序列和糖链组成或结构的鉴定．本文即从以上三个主要方面系统阐述目前 N- 完整糖肽分析中常用的质谱和数据分析策略和方法，并进一步在糖肽谱图识别、母离子单同位素分子质量校正、数据库选择以及假阳性率评估和控制等方面都进行了逐一探讨．完整糖肽的直接质谱分析有助于获取糖链和糖基化位点间的对应关系信息，可为生物标志物发现和疾病致病机理等研究提供更有力的糖蛋白质组学研究工具．

关键词　糖蛋白质组学，完整糖肽，质谱，糖蛋白
学科分类号　Q7
DOI: 10.16476/j.pibb.2017.0254

蛋白质的糖基化修饰是最普遍、最重要的蛋白质翻译后修饰之一．糖基化增加了蛋白质分子的结构复杂性和功能多样性．糖基化不仅对所修饰蛋白的折叠、结构、运输、定位和生物活性的保持等有重要影响[1]，而且可通过糖和蛋白质之间的特异性识别调节生物的各种生理过程，如细胞的生长、分裂、分化、识别、黏附等[2-5]．其中 N- 连接糖蛋白普遍存在于细胞表面和各种体液中，细胞表面糖蛋白是药物分子作用的重要靶点，而体液糖蛋白又是疾病相关生物标志物的主要来源[6]．因此，开展各种生理病理条件下的 N- 糖蛋白质组学定量研究有助于阐明蛋白质糖基化的功能，并对可用于癌症预判和诊断的新生物标志物的发现也具有重要意义．

* 中组部“青年千人”配套经费(361010001)资助．
** 通讯联系人．
Tel: 029-88302142, E-mail: suns@nwu.edu.cn
收稿日期：2017-07-04，接受日期：2017-09-07

随着生物质谱的快速发展及其相关分析方法的进步，糖蛋白质组学有了长足发展．糖蛋白质组学研究的难点在于糖链分子较大且存在微观不均一性，另外糖链自身可以在串联质谱中碎裂且与肽段的碎裂规律不同，因此蛋白质组学的质谱解析方法和软件难以完整地鉴定糖肽上的肽段序列和糖链结构．在之前的研究中，普遍采用的分析策略是将糖蛋白 / 糖肽分离富集后，利用酶解或化学方法将糖链和糖基化的多肽分开，然后对释放的糖链或去糖基化的多肽进行单独质谱分析，从而获得待检测样本中的糖链和 / 或糖蛋白的整体表达和变化情况．其中经典的富集方法包括酰肼化学富集、凝集素富集、亲水亲和富集等[7-11]．这一研究策略目前已经在多种癌症相关研究中得到了很好的应用[6, 12-13]．但是，该策略的局限性在于无法获得糖链和糖基化位点之间的对应关系．近年来，用于完整糖肽分析的方法学也相继建立起来，这些方法可进一步获得糖链和糖基化位点间的对应关系信息[14-20]．

完整糖肽的直接质谱分析需要解决以下几方面的问题：首先，很多糖蛋白在复杂样品中的丰度相对较低，而蛋白糖基化的微不均一性特征(每个糖基化位点上可以连接多达几十个不同糖链)进一步降低了完整糖肽的浓度，因此完整糖肽分析首先需要对糖蛋白 / 糖肽进行有效分离富集，即去除非糖基化蛋白 / 非糖基化多肽．其次，由于 N- 完整糖肽的分子质量普遍较大，且糖链部分比多肽部分更容易碎裂，因此在质谱分析中需要调整相应质谱分析参数以完成完整糖肽鉴定．最后，由于完整糖肽是由糖链和多肽两部分组成的，同一糖基化位点即使在同一样品中也可同时被多种糖链修饰，而目前还没有完善的糖肽数据库，因此完整糖肽质谱数据的后续分析和软件开发成为了完整糖肽分析中的关键环节之一．

本文即从以上三个主要方面(包括完整糖肽富集、质谱解析和数据处理)对近年来发展的 N- 连接完整糖肽的质谱分析策略和研究方法进行系统综述，三部分内容可进一步细化为糖肽富集方法、碎裂模式选择、质谱参数设置、糖肽图谱识别、母离子单同位素分子质量校正、完整糖肽的多肽序列和糖链结构(或组成)的鉴定以及假阳性率评估控制等方面(图 1)．

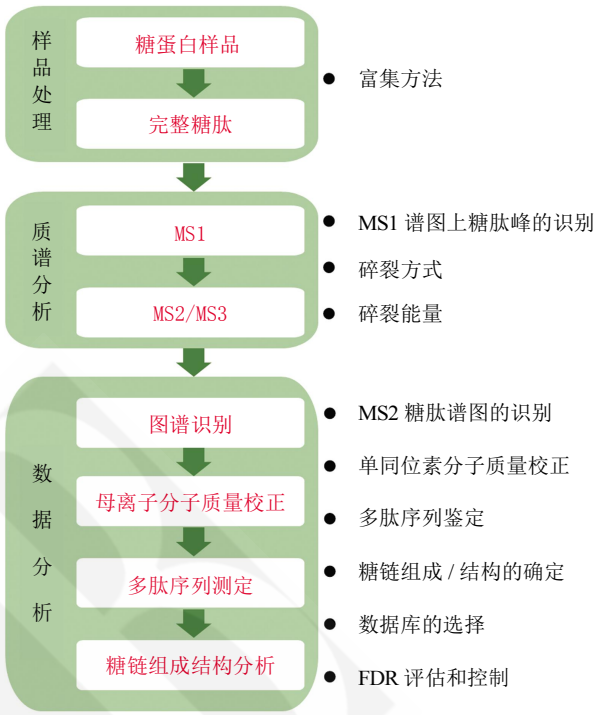

Fig. 1　The workflow for enrichment, mass spectrometry analysis, and data analysis of N-linked intact glycopeptides

图 1　N-连接完整糖肽的富集、质谱分析和后续数据分析的工作流程图

1　完整糖肽的富集策略

在糖蛋白质组学研究中，研究对象往往是一些复杂的生物样品体系，样品中蛋白质的丰度动态变化范围很大，同时很多糖蛋白丰度较低，对所有的或者特定的糖蛋白群进行富集才会有利于鉴定到更多的糖蛋白．此外，由于糖基化肽段往往只占蛋白质酶解肽段的 2%～5%[21]，在质谱分析的过程中，其信号很容易受到非糖肽的抑制，因此糖蛋白质组学研究所面临的首要问题是糖蛋白 / 糖肽的有效分离富集，即去除非糖蛋白 / 非糖肽．

目前可用于完整糖肽富集的方法主要有亲水相互作用色谱(hydrophilic interaction liquid chromatography，HILIC)[22-23]、凝集素(lectin)亲和[24]和硼酸亲和[25]等方法．HILIC 基于富含多羟基的糖链结构的强亲水特性，使糖肽与非糖基化肽段得以

分离，且能保持糖链结构的完整性，既适用于 N-糖肽，也适用于 O- 糖肽[22]. 但是由于糖链和固定相之间的相互作用力较弱，只连接单糖或寡糖的糖肽(如 O- 糖肽)经常无法很好地被 HILIC 富集[23]. 凝集素是一类糖结合蛋白，能专一地识别结构特异的单糖或聚糖中特定的糖基序列并与之结合，它们与糖链的结合是非共价且可逆的. 糖蛋白或糖肽被凝集素捕获之后，通常用特定的单糖通过竞争结合凝集素将糖蛋白或糖肽洗脱下来，从而实现对糖蛋白的富集.

2 完整糖肽的质谱分析

2.1 MS1 水平识别完整糖肽

在质谱分析过程中，由于生物临床蛋白样品的高度复杂性，经糖肽富集后的样品通常依然掺杂很多非糖基化的多肽，这些非糖基化多肽常常会干扰完整糖肽的质谱分析. 而目前已有一些简单有效的方法从一级质谱(first stage of mass spectrometry，MS1)的谱图中筛选出完整糖肽的质谱峰，这些初级筛选可有效减少非糖基化多肽的质谱分析时间，从而增加完整糖肽的鉴定机会. 目前报道的方法有：

2.1.1 分子质量直接识别

其根本依据是分子质量这一概念是通过定义 ^{12}C 的分子质量为 12 来确定的，其他元素的分子质量都通过与 ^{12}C 的分子质量比较来取整确定. 元素可根据质量划分为恰好是正整数(如 ^{12}C，12.00000)，或大于正整数(如 ^{1}H，1.00783；^{14}N，14.00307)或小于正整数(如 ^{16}O，15.99491；^{32}S，31.97207)[26]三类. 通常在多肽中，N 和 H 含量最高，它们的实际质量都比取整后的分子质量稍大一点，而在糖肽中，O 和 S 含量相对较高，它们的实际质量都比取整后的分子质量稍小一点. 因此对于分子质量相等的糖肽和无糖基化的普通多肽，糖肽的分子质量通常小于多肽[27]. 这一特点可以用来区分糖肽与普通多肽. Froehlich 等[28]利用此法通过模拟实验，得到了一个可以区分糖肽母离子与肽段母离子的简单分类器，对糖肽 / 多肽的质量的整数部分和小数部分进行方程式划分归类，来筛选糖肽. 在质谱质量准确度(mass accuracy)为 10ppm 的条件下，该方法识别完整糖肽可以达到 89%以上的灵敏度和 93%以上的特异性. 随着更高分辨率和质量准确度质谱的出现和普及，其识别完整糖肽的特异性将可进一步增强. 不过此方法要求糖肽中的糖链部分占有较大的分子质量比例，即糖肽中氧原子(O)含量较高，与普通多肽可区分性较强.

2.1.2 同位素分布识别

由于多数完整糖肽(尤其是 N- 糖肽)比非糖基化多肽含有更多的氧元素(O)，而自然界中 O 元素的同位素 ^{16}O 与 ^{18}O 之间含有 2u 差异，有别于多肽的其他主要组成元素 C、H、N 等(同位素间均只有 1u 差异). 因此，完整糖肽与普通多肽通常有着不同的同位素丰度分布，其特征性的同位素峰型理论上可以用于识别 MS1 检测到的分子是否为糖肽母离子峰. 然而该方法应用的前提是检测到准确的同位素丰度，目前在实际应用中还很少单独使用. 不过分子质量直接识别和同位素丰度配合使用，结合高分辨率高精准度的质谱分析，在 MS1 母离子水平识别完整糖肽是具有理论可行性的，MS1 层面上直接识别完整糖肽，然后有选择性地进行 MS2 分析可增加糖肽母离子选择的比例，从而达到最大限度分析和鉴定完整糖肽的目的.

2.1.3 母离子带电荷数筛选

由于完整糖肽普遍分子质量相对较大，而多数质谱仪 m/z 检测范围相对较小，多数情况下质谱检测到的完整糖肽母离子峰所带电荷通常在 3+ 或以上. 因此必要时可排除带有 1+ 和 2+ 价位的母离子[17]，以尽可能降低非糖基化多肽的干扰，从而增加完整糖肽被挑选进行 MS2 分析的概率.

2.2 质谱碎裂模式及参数设置

完整糖肽与普通多肽在一些理化特性上存在着较大差异，因此在糖肽的 MS2 分析时需要对一些分析参数重新进行调整和优化.

2.2.1 碎裂方式的选择

在质谱分析中，可用于完整糖肽分析的质谱碎裂方式包括碰撞诱导裂解(collision-induced dissociation，CID)、高能碰撞裂解(higher energy collisional dissociation，HCD)、电子转移裂解(electron transfer dissociation，ETD)，以及组合型的 EThcD 和 CID+HCD 等. CID 碎裂能量相对较低，常用于易碎裂的糖链部分的结构解析. 但由于 CID 普遍存在离子阱的“三分之一效应”[29]，而糖肽的分子质量普遍较大，因此很多情况下 CID-MS2 图谱无法检测到氧鎓离子信号(氧鎓离子分子质量普遍较小)，不利于完整糖肽 MS2 谱图的识别[30]. HCD 模式能够提供关于完整糖肽的广泛信息，包括 b/y 离子(肽段碎片离子)及糖肽(或糖链)特有的氧鎓离子[31]，因此在完整糖肽分析中使

用最广泛[31-35]. ETD 主要产生糖肽的 c/z 离子，碎裂后能够保持糖链完整地连接在肽段碎片上，便于在完整糖肽水平鉴定糖基化位点．但 ETD 需要的碎裂时间一般较长，且对于连有较大糖链的糖基化位点鉴定效果不佳[30]．此外，EThcD 作为 ETD 和 HCD 的结合，采用双分裂产生离子，因此既可产生丰富的肽段序列又可获得糖链定位信息[19]．另外，CID 可提供很好的糖链碎片信息，而 HCD 则可提供氧鎓离子(用于糖肽谱图识别)和高丰度的 Y1 特征离子(提供“肽段 +N- 乙酰葡糖胺”碎片信息)，两者结合(CID+HCD)同样可增加糖肽鉴定的准确性．但多种碎裂模式增加了每种糖肽的质谱分析时间，因此要以降低鉴定数量为代价[36]．关于各种碎裂模式对完整糖肽鉴定的进展已有很好的中文综述文章[30].

2.2.2 碎裂级别的选择

质谱分析中可以选择不同的碎裂级别，目前完整糖肽鉴定用到的质谱碎裂级别多数仅到二级质谱(second stage of mass spectrometry，MS2)，少数方法用到了三级质谱(third stage of mass spectrometry，MS3)．MS1 鉴定的方式即利用质谱仪精确测量酶解片段的分子质量并搜库比较，实现糖肽的鉴定；MS2 是在 MS1 的基础上再选择部分糖肽做进一步的破碎，并对碎片进行深入分析和比较，鉴定出该肽段的序列和糖链的组成，并结合 MS1 的结果从而实现对完整糖肽的鉴定，MS2 能够得到全部或部分肽段的序列，具有更高的可靠性；MS3 是在 MS2 的基础上再选择部分碎片离子当作母离子做进一步的破碎．Jia 等[37]在 MS2 过程中发生岩藻糖的中性丢失时自动收集 MS3 的信息，用于进行核心岩藻糖化糖蛋白的精确大规模鉴定．相较 MS2，MS3 具有更好的图谱质量，其碎片离子的丰度分布更均匀，干扰信号更少．但是 MS3 需要更长的运行时间[20].

2.2.3 MS2/3 中碎裂能量的选择

糖链的碎裂能量与肽段不同，通常来说，糖链在较低的能量下即可碎裂，而肽段则需要较高的能量．因此在糖肽的质谱分析中需要选择适当的碎裂能量来达到其糖链和肽段部分的最佳综合碎裂效果．如之前的研究表明，低能量的 HCD 碎裂能得到丰富的氧鎓离子，而中高能量的 HCD 碎裂则会产生一系列 Y 离子(肽段 + 糖链碎片)[38].

3 完整糖肽的数据分析

3.1 完整糖肽谱图的识别

一些非糖基化多肽的质谱谱图会干扰完整糖肽谱图的数据分析，增加鉴定假阳性率．因此在完整糖肽谱图的数据分析过程中，有必要再通过一些特征峰对来自完整糖肽的 MS2 谱图进行有效的识别．目前绝大多数对完整糖肽的谱图识别均使用氧鎓离子(oxonium ions)筛选法．氧鎓离子是糖肽上糖链碎裂后产生的单糖、二糖或寡糖碎裂后形成的碎片离子，这些碎片离子在非糖基化多肽的 MS2 谱图中并不存在，因此是糖肽的特征碎片离子[39]．有关完整糖肽产生的各种氧鎓离子信息可见综述[30]．依据氧鎓离子对完整糖肽谱图的识别又分为单氧鎓离子识别与多氧鎓离子识别 (图 2a).

3.1.1 单氧鎓离子识别

单氧鎓离子识别一般仅根据有无 N- 乙酰己糖胺(HexNAc)的单电荷谱峰(204.08u)来判断谱图是否为糖肽谱图[14]．该方法操作简单，但强噪音干扰常导致假阳性的出现，因此在分析中最好设置峰相对高度等参数以降低噪音导致的错误匹配．但是，由于 O- 糖链多数也含有 HexNAc 糖基，因此 O- 糖肽 MS2 谱图中多数也都含有 HexNAc 单电荷谱峰(204.08u)，该方法难以区分 N- 和 O- 糖肽.

3.1.2 多氧鎓离子识别

多氧鎓离子识别是根据多个氧鎓离子的谱峰信号来判断谱图是否为糖肽谱图[17, 35, 40]，一般常用的氧鎓离子有 138.055u(HexNAc 的碎片)、168.066u (HexNAc-2H_2O)、186.076u (HexNAc-H_2O)、204.087u (HexNAc)、292.103u(N- 乙酰神经氨酸，NeuAc，唾液酸的一种)和 366.140u(已糖 Hex+HexNAc)等[30]．这些离子联合使用可以有效地降低完整糖肽谱图识别的错误率．通过多氧鎓离子间的相对丰度值还可用于区分 N- 和 O- 糖肽谱图[41]．需要注意的是，由于氧鎓离子的分子质量一般较小(低至 138u)，因此很多氧鎓离子难以在 CID 碎裂模式下检测到，而在 HCD 碎裂模式下，很多质谱仪也需要对 MS2 起始扫描范围进行设置(如设置 MS2 测定的起始检测荷质比为 100).

3.2 完整糖肽单同位素分子质量的测定

完整糖肽鉴定中，通常第一步用到的即是母离子的单同位素峰(monoisotopic peak)的分子质量，

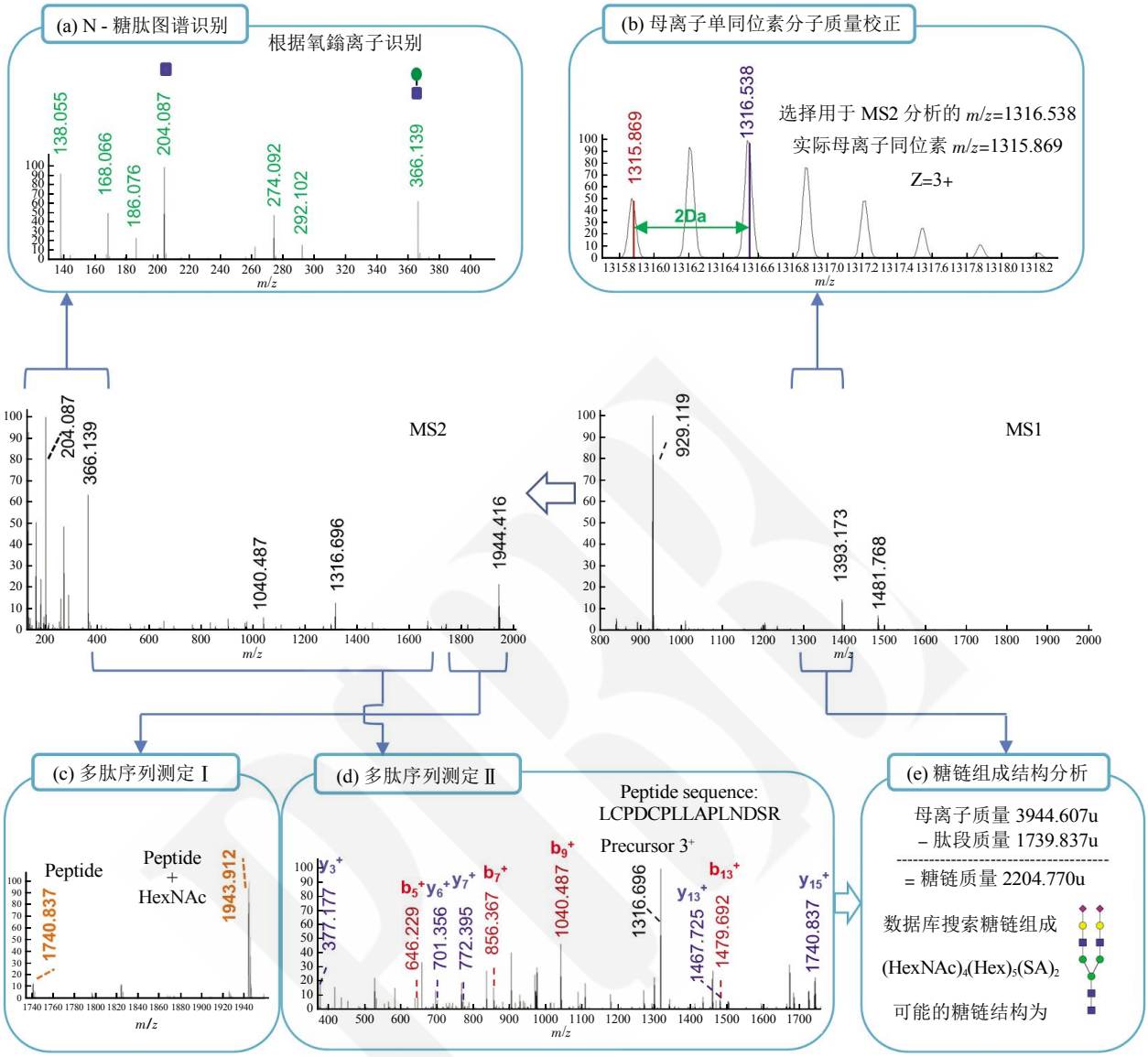

Fig. 2　An example of the intact glycopeptide identification from its MS spectra

图 2　完整糖肽谱图的数据分析实例

完整糖肽的质谱谱图首先通过(a)特征氧鎓离子识别，然后通过(b)人工手动校正母离子的单同位素分子质量，接着由(c)Y0 和 Y1 碎片离子得到多肽分子质量及糖基化位点，初步确定可能的多肽序列，再由(d) b/y 离子匹配多肽片段，最终确定多肽序列，最后通过(e)母离子质量和多肽质量的差异得到糖链质量，搜库匹配出糖链组成，推测出可能结构.

单同位素峰是一个同位素峰簇中质荷比最小的峰(即第一个峰). 由于完整糖肽中糖链含有的 O 元素远大于普通多肽，它的同位素峰型有别于普通多肽. 在完整糖肽的质谱分析中，由于 MS1 检测中的第一个同位素峰通常不是同位素峰簇中最高的，因此质谱自动选择进行 MS2 分析的通常不是单同位素峰，这会导致母离子分子质量偏差. 而该分子质量偏差会导致糖肽图谱无法匹配或匹配错误的情况，对糖肽鉴定造成影响，因此母离子匹配前需要对其单同位素峰的分子质量进行校正(图 2b). 校正方法有：a. 借助已有软件的校正功能. 蛋白质组学中常用数据处理软件 MaxQuant 和 Extract_MSn 来校正单同位素峰，但在某些情况下，软件判断结果有误[42]. b. 平均计算法. 计算在目标糖肽洗脱时间左右的时间段中母离子的平均质谱，平均同位素峰簇对应的母离子质量最小的峰被选为单同位素峰. 这一方法提高了同位素峰簇模式的分布和单同位素峰质量的确定，从而提高了糖肽图谱的匹配[35].

c．人工手动检查．专业技术人员根据经验来校正单同位素峰．该方法效率低，耗时长，且无法复制迁移到其他实验室使用，具有一定的局限性．

3.3 完整糖肽上多肽序列的鉴定

在完整糖肽的谱图鉴定中，多数情况下首先鉴定糖肽的多肽序列部分(图 2c 和 2d)．具体分为以下两种：a．先母离子，后碎片离子匹配．先基于母离子质量匹配糖肽图谱，由于大多数情况下同一质量可能对应许多个具有相同分子质量的糖肽结构，因此还需再通过 MS2 来确定糖肽中多肽的序列和糖链的结构[17]．b．先碎片离子，后母离子匹配．根据 MS2 中是否出现氧鎓离子峰来判断该图谱是否为糖肽的图谱，再根据 Y 离子峰、b/y 离子峰结合 MS1 得到的分子质量信息分别来推导糖链部分的组成以及多肽部分的序列[32, 43-44]．

3.4 完整糖肽上糖链结构 (或组成) 的分析

与 DNA 和蛋白质序列的线性结构不同，糖链通常含有分支结构且连接方式各不相同，糖链结构的复杂性极大地增加其结构解析的难度．目前糖链组成和结构分析主要通过以下几种方式：a．差异分子质量匹配．首先从糖肽母离子与肽段的质量差计算出糖链的质量(图 2e)，然后在糖数据库里精确匹配搜索糖链组成[35]．b．MS2/3 碎片离子匹配．通过 MS2 或 MS3 的碎片离子信息推导出糖链的结构[14]．c．图谱图形匹配．不论糖链结构如何，糖肽和相应的去糖基化多肽会有相似的 MS2 模式，即会有相似的图谱图形[32]．因此可以用氧鎓离子作为糖肽图谱的特征离子来选择完整糖肽的图谱，通过图形比较，将每个完整糖肽的 MS2 分配给一个特定的含糖基化位点的多肽．随后再通过完整糖肽与多肽部分间的质量差与糖数据库相匹配，确定占据每个位点的糖．GPQuest 算法[35]中利用图谱图形匹配的方法，将 HCD 产生的复杂样本碎片用来识别糖基化位点特定的完整糖肽．Yang 等[32]利用此法，进行了艾滋病病毒表面糖蛋白 gp120 和重组 gp120 的糖型分析．结合图 2 我们可以看到完整糖肽谱图的数据分析实例．

3.5 糖链和多肽数据库的选择

数据库搜索过程中，选择不同的数据库也会对搜索时间和鉴定结果造成影响．目前常用的数据库主要有以下几种：

3.5.1 选择全基因组数据库

如 Qin 等[43]在 Uniprot 数据库中搜索 fetuin-A 的糖基化位点，Woo 等[18]在人类蛋白质数据库中鉴定人类癌症细胞系蛋白质的糖肽．Woodin 等[45]提供了现有糖基化数据库的资源列表及详细描述，在此不再赘述．需要指出的是，由于糖肽数据库普遍是由将基因组数据库中所有含 N-X-S/T(X 不能是脯氨酸)序列的多肽与所有可能糖链逐一组合构建而成(即所有可能组合)，因此基于全基因组的完整糖肽数据库普遍较大．由此可见，基于全基因组数据库进行糖肽鉴定虽然可以更全面地鉴定各种糖肽，但是由于数据库较大，在匹配速度较慢的同时还极大地增加了假阳性率．由于 O- 糖基化位点没有固定的保守序列，因此基于全基因组蛋白序列构建的 O- 糖肽数据库将更加巨大，这也是目前 O- 糖肽较难分析的重要原因之一．

3.5.2 通过鉴定结果自行创建样本特有数据库

Parker 等[14]采用串联质谱得到的 N- 糖数据建立了一个 N- 糖基化修饰的数据库，又用 PNGase F 处理富集到的糖肽去除 N- 糖后经过质谱分析与数据库搜索建立了包含每个糖肽碎片的糖肽中心级联数据库．之后在这两个自建数据库中搜索 N- 糖肽的 CID 和 HCD 的 MS2 谱，从 161 个大鼠脑组织的糖蛋白中有效地识别出了 863 个完整的 N- 连接糖肽．Toghi Eshghi 等[35]利用串联质谱分析样品的含糖基化位点多肽，建立了样品的含糖基化位点多肽的图谱库，然后与完整糖肽的谱图进行比较，从而将每个完整糖肽的谱图分配给不同的含糖基化位点多肽，通过差异分子质量确定占据每个位点的糖．自建样本特有数据库具有数据库相对较小、鉴定准确和运行速度快等优势．但此法搜索结果的全面性小于全基因组数据库．

3.6 假阳性率评估及控制

在 MS2 谱图的数据库搜索后，需要对其可能的假阳性匹配进行综合评估和控制．其中建立诱饵数据库(decoy database)是假阳性率(false discovery rate，FDR)评估中最常用的方法．这种方法是将显然不正确的“诱饵(decoy)”序列添加到搜索空间，以相同的参数搜索，会对应显然不正确的搜索结果，但这些不正确的结果在一些鉴定方法中有可能会被视为是正确的．因此这种不正确被视为正确现象的数量可以作为假阳性数量的一个很好的估计，即 FDR 根据匹配到 decoy 糖肽的图谱百分比来评估．Decoy database 方法具体的实现方式也不尽相同，常用以下两种方法：

3.6.1 序列反向

使多肽序列顺序反转，创建反向数据库(reverse

database)是最常见的方法．Reverse database 由把所有已鉴定的含糖基化位点多肽的氨基酸序列相结合后反转序列顺序，分成与目标数据库具有相同长度的 decoy 肽段得到．从而 reverse database 和目标数据库有相似的肽段数量、肽段长度和母离子分子质量，但有完全不同的 MS2 图谱，因此，目标肽段与 reverse database 的匹配类似于目标数据库中的随机匹配．GPQuest[35]中即采用此法进行 FDR 的评估．Parker 等[14]把肽匹配图谱(the peptide-spectrum matches，PSMs)的肽段序列顺序反转后与碎片离子匹配，根据匹配频率评估 FDR，与此法原理类似.

3.6.2 随机匹配

Yang 等[32]用 gp120 去糖基化多肽的 b/y离子与不含任何 gp120 糖肽的糖肽搜索匹配，即采用完全不相干的数据库与目标数据库进行匹配，这种方式即为随机匹配.

3.6.3 假阳性控制

以上仅是估算假阳性率的方法，但在实际分析中，找到影响假阳性率的因素并在实际数据分析中通过设置搜索参数来排除影响因素，尽可能的降低假阳性率才是糖肽分析的关键．例如在完整糖肽分析中，多个实验室均指出 Y1 离子的辅助作用[32]、单同位素峰的校正和 b/y 离子丰度参考[46]都可有效降低假阳性率，从而增强完整糖肽的鉴定.

4 总结和展望

近几年来，随着糖蛋白质组学相关研究技术的发展，通过质谱鉴定完整糖肽的多肽组成和糖链结构的方法相继建立起来．人们在完整糖肽的富集、质谱分析、谱图识别、多肽部分鉴定和糖链结构确定等方面已初步建立了一套系统全面的分析策略和工作流程．这些分析策略和软件工具的建立和开发使我们对糖蛋白各糖基化位点上特征性的糖链结构信息的了解成为可能．但要实现更精确的糖肽的鉴定和分析，还有赖于以上分析流程中很多环节的进一步探索和突破．另外，完整糖肽高通量的定量方法也有待进一步发展．相信随着技术方法的不断革新与改进，完整糖肽的分析研究会不断推进．这些糖蛋白质组学研究工具将作为基因组学、蛋白质组学、糖组学和代谢组学等的有力补充，为蛋白质糖基化结构和功能分析、新的生物标志物发现和致病机理等研究提供强大的技术支持.

参 考 文 献

[1] Hirabayashi J, Arata Y, Kasai K. Glycome project: Concept, strategy and preliminary application to *Caenorhabditis elegans*. Proteomics, 2001, **1**(2): 295−303

[2] Rudd P M, Elliott T, Cresswell P, *et al*. Glycosylation and the immune system. Science, 2001, **291**(5512): 2370−2376

[3] Zhang Y, Zhao J H, Zhang X Y, *et al*. Relations of the type and branch of surface N-glycans to cell adhesion, migration and integrin expressions. Molecular and Cellular Biochemistry, 2004, **260**(1−2): 137−146

[4] Haltiwanger R S. Regulation of signal transduction pathways in development by glycosylation. Current Opinion in Structural Biology, 2002, **12**(5): 593−598

[5] Haltiwanger R S, Lowe J B. Role of glycosylation indevelopment. Annual Review of Biochemistry, 2004, **73**: 491−537

[6] Drake P, Schilling B, Gibson B, *et al*. Elucidation of N-glycosites within human plasma glycoproteins for cancer biomarker discovery. Methods in Molecular Biology (Clifton, NJ), 2013, **951**(951): 307−322

[7] Kaji H, Saito H, Yamauchi Y, *et al*. Lectin affinity capture, isotope-coded tagging and mass spectrometry to identify N-linked glycoproteins. Nature Biotechnology, 2003, **21**(6): 667−672

[8] Zhang H, Li X J, Martin D B, *et al*. Identification and quantification of N-linked glycoproteins using hydrazide chemistry, stable isotope labeling and mass spectrometry. Nature Biotechnology, 2003, **21**(6): 660−666

[9] Yu L, Li X L, Guo Z M, *et al*. Hydrophilic interaction chromatography based enrichment of glycopeptides by using click maltose: a matrix with high selectivity and glycosylation heterogeneity coverage. Chem-Eur J, 2009, **15**(46): 12618−12626

[10] Zhang W, Wang H, Tang H, *et al*. Endoglycosidase-mediated incorporation of 18O into glycans for relative glycan quantitation. Anal Chem, 2011, **83**(12): 4975−4981

[11] Cao Q C, Zhao X Y, Zhao Q, *et al*. Strategy integrating stepped fragmentation and glycan diagnostic ion-based spectrum refinement for the identification of core fucosylated glycoproteome using mass spectrometry. Analytical Chemistry, 2014, **86**(14): 6804−6811

[12] Zhao J, Simeone D M, Heidt D, *et al*. Comparative serum glycoproteomics using lectin selected sialic acid glycoproteins with mass spectrometric analysis: Application to pancreatic cancer serum. Journal of Proteome Research, 2006, **5**(7): 1792−1802

[13] Tarp M A, Clausen H. Mucin-type O-glycosylation and its potential use in drug and vaccine development. Biochimica Et Biophysica Acta-General Subjects, 2008, **1780**(3): 546−563

[14] Parker B L, Thaysen-Andersen M, Solis N, *et al*. Site-specific glycan-peptide analysis for determination of N-glycoproteome heterogeneity. Journal of Proteome Research, 2013, **12**(12): 5791−5800

[15] Wu S W, Pu T H, Viner R, *et al*. Novel LC-MS2 product dependent

parallel data acquisition function and data analysis workflow for sequencing and identification of intact glycopeptides. Analytical Chemistry, 2014, **86**(11): 5478−5486

[16] Bern M, Kil Y J, Becker C. Byonic: advanced peptide and protein identification software. Current protocols in bioinformatics, 2012, Chapter **13**: Unit1320−Unit1320

[17] Sun S, Shah P, Eshghi S T, *et al*. Comprehensive analysis of protein glycosylation by solid-phase extraction of N-linked glycans and glycosite-containing peptides. Nature Biotechnology, 2016, **34**(1): 84−88

[18] Woo C M, Iavarone A T, Spiciarich D R, *et al*. Isotope-targeted glycoproteomics (IsoTaG): a mass-independent platform for intact N- and O-glycopeptide discovery and analysis. Nature Methods, 2015, **12**(6): 561−567

[19] Yu Z, Zhao X, Tian F, *et al*. Sequential fragment ion filtering and endoglycosidase-assisted identification of intact glycopeptides. Analytical and Bioanalytical Chemistry, 2017, **409**(12): 3077−3087

[20] Zeng W F, Liu M Q, Zhang Y, *et al*. pGlyco: a pipeline for the identification of intact N-glycopeptides by using HCD-and CID-MS/MS and MS3. Scientific Reports, 2016, **6**: 25102

[21] Sun B Y, Ranish J A, Utleg A G, *et al*. Shotgun glycopeptide capture approach coupled with mass Spectrometry for comprehensive glycoproteomics. Molecular & Cellular Proteomics, 2007, **6**(1): 141−149

[22] Zauner G, Deelder A M, Wuhrer M. Recent advances in hydrophilic interaction liquid chromatography (HILIC) for structural glycomics. Electrophoresis, 2011, **32**(24): 3456−3466

[23] Jensen P H, Karlsson N G, Kolarich D, *et al*. Structural analysis of N- and O-glycans released from glycoproteins. Nat Protoc, 2012, **7**(7): 1299−1310

[24] Madera M, Mechref Y, Novotny M V. Combining lectin microcolumns with high-resolution separation techniques for enrichment of glycoproteins and glycopeptides. Analytical Chemistry, 2005, **77**(13): 4081−4090

[25] Sparbier K, Wenzel T, Kostrzewa M. Exploring the binding profiles of ConA, boronic acid and WGA by MALDI-TOF/TOF MS and magnetic particles. J Chromatogr B, 2006, **840**(1): 29−36

[26] Lehmann W D, Bohne A, Von Der Lieth C W. The information encrypted in accurate peptide masses - improved protein identification and assistance in glycopeptide identification and characterization. J Mass Spectrom, 2000, **35**(11): 1335−1341

[27] Juhasz P, Martin S A. The utility of nonspecific proteases in the characterization of glycoproteins by high-resolution time-of-flight mass spectrometry. International Journal of Mass Spectrometry & Ion Processes, 1997, **169**/**170**(6): 217−230

[28] Froehlich J W, Dodds E D, Wilhelm M, *et al*. A classifier based on accurate mass measurements to Aid large scale, unbiased glycoproteomics. Molecular & Cellular Proteomics, 2013, **12**(4): 1017−1025

[29] Hoffmann E D, Charette J J, Stroobant V, *et al*. Mass spectrometry : Principles and Applications. UK: John Wiley & Sons, 2007

[30] 曾文锋, 张 扬, 刘铭琪, 等. N-糖肽的规模化质谱解析方法进展. 生物化学与生物物理进展, 2016, **43**(6): 550−562
Zeng W F, Zhang Y, Liu M Q, *et al*. Prog Biochem Biophys, 2016, **43**(6): 550−562

[31] Scott N E, Parker B L, Connolly A M, *et al*. Simultaneous glycan-peptide characterization using hydrophilic interaction chromatography and parallel fragmentation by CID, higher energy collisional dissociation, and electron transfer dissociation MS applied to the N-linked glycoproteome of campylobacter jejuni. Molecular & Cellular Proteomics, 2011, **10** (2): M000031-MCP201-1

[32] Yang W, Shah P, Eshghi S T, *et al*. Glycoform analysis of recombinant and human immunodeficiency virus envelope protein gp120 *via* higher energy collisional dissociation and spectral-aligning strategy. Analytical Chemistry, 2014, **86**(14): 6959−6967

[33] Kolarich D, Jensen P H, Altmann F, *et al*. Determination of site-specific glycan heterogeneity on glycoproteins. Nature Protocols, 2012, **7**(7): 1285−1298

[34] Hart-Smith G, Raftery M J. Detection and characterization of low abundance glycopeptides *via* higher-energy C-Trap dissociation and orbitrap mass analysis. Journal of the American Society for Mass Spectrometry, 2012, **23**(1): 124−140

[35] Toghi Eshghi S, Shah P, Yang W, *et al*. GPQuest: a spectral library matching algorithm for site-specific assignment of tandem mass spectra to intact N-glycopeptides. Anal Chem, 2015, **87** (10): 5181−5188

[36] Segu Z M, Mechref Y. Characterizing protein glycosylation sites through higher-energy C-trap dissociation. Rapid Communications in Mass Spectrometry, 2010, **24**(9): 1217−1225

[37] Jia W, Lu Z, Fu Y, *et al*. A strategy for precise and large scale identification of core fucosylated glycoproteins. Molecular & Cellular Proteomics, 2009, **8**(5): 913−923

[38] 江 静, 应万涛, 钱小红, 等. 亲水相互作用色谱结合串联质谱多重碎裂模式在完整糖肽研究中的应用. 分析化学, 2014, **42**(02): 159−165
Jiang J, Ying W T, Qian X H, *et al*. Chin J Anal Chem, 2014, **42**(02): 159−165

[39] Carr S A, Huddleston M J, Bean M F. Selective identification and differentiation of N- and O-linked oligosaccharides in glycoproteins by liquid chromatography-mass spectrometry. Protein Science: A Publication of the Protein Society, 1993, **2**(2): 183−196

[40] Pompach P, Chandler K B, Lan R, *et al*. Semi-automated identification of N-glycopeptides by hydrophilic interaction chromatography, nano-reverse-phase LC-MS/MS, and glycan database search. Journal of Proteome Research, 2012, **11** (3): 1728−1740

[41] Toghi Eshghi S, Yang W, Hu Y, *et al*. Classification of tandem mass spectra for identification of N- and O-linked glycopeptides. Scientific Reports, 2016, **6**: 37189

[42] 袁作飞, 邬 龙, 刘 超, 等. 规模化蛋白质鉴定中母离子的准确检测技术研究. 生物化学与生物物理进展, 2013, **43**(01): 80−92
Yuan Z F, Wu L, Liu C, *et al*. Prog Biochem Biophys, 2013, **43**(01): 80−92

[43] Qin H, Cheng K, Zhu J, *et al*. Proteomics analysis of O-GalNAc glycosylation in human serum by an integrated strategy. Analytical Chemistry, 2017, **89**(3): 1469−1476

[44] Joenvaara S, Ritamo I, Peltoniemi H, *et al*. N-Glycoproteomics - An automated workflow approach. Glycobiology, 2008, **18** (4): 339−349

[45] Woodin C L, Maxon M, Desaire H. Software for automated interpretation of mass spectrometry data from glycans and glycopeptides. The Analyst, 2013, **138**(10): 2793−2803

[46] Frese C K, Altelaar A F M, Van Den Toorn H, *et al*. Toward full peptide sequence coverage by dual fragmentation combining electron-transfer and higher-energy collision dissociation tandem mass spectrometry. Analytical Chemistry, 2012, **84**(22): 9668−9673

Mass Spectrometry-based Strategies and Methods for N-linked Intact Glycopeptide Analysis*

ZHU Bo-Jing, ZHI Yuan, SUN Shi-Sheng**

(*College of Life Sciences, Northwest University, Xi'an* 710069, China)

Abstract As one of the most common and important protein modifications, glycosylation has been one of the focuses of the proteomic researches. In the last decades, most of N-linked glycoproteomic studies focused mainly on the analysis of either released glycans or de-glycosylated peptides. While this strategy reduced the complexity of glycoprotein analysis, it lost glycosite-specific glycosylation information. Several strategies and methods for intact N-glycopeptide analysis have been established during the last few years. Generally, to achieve the identification and quantification of intact glycopeptides, the first step is to enrich glycopeptides from complex samples to reduce the affects from non-glycosylated peptides, then the mass spectrometry parameter settings need to be adjusted to satisfy the fragment features of glycopeptides, importantly the related software also need to be developed for the precise identification of the peptide sequence and glycan structures or compositions of the intact glycopeptides. These three main aspects of the strategies for mass spectrometry-based intact glycopeptide analysis are discussed in this paper. Some further details, such as the recognition of intact glycopeptide spectra, precursor monoisotopic mass correction, database selection, as well as the false discovery rate (FDR) evaluation and control, are further discussed. Direct intact glycopeptide analyses, with the recovery of the glycosite-specific glycosylation information, will provides a powerful tool for biomarker discovery and the mechanism studies on various diseases.

Key words glycoproteomics, intact glycopeptide, mass spectrometry, glycoprotein

DOI: 10.16476/j.pibb.2017.0254

* This work was supported by a grant from The Thousand Talents Plan (361010001).

**Corresponding author.

Tel: 86-29-88302142, E-mail: suns@nwu.edu.cn

Received: July 4, 2017 Accepted: September 7, 2017