

糖蛋白质组学的信息学资源和方法*

刘 杭 姚 臻 杨芃原 张 扬**

(复旦大学医学院生物医学研究院, 上海 200032)

摘要 糖基化修饰是生物体内最常见、最重要的蛋白质翻译后修饰之一。哺乳动物体内超过 50% 的蛋白质都会发生糖基化修饰。糖蛋白广泛分布于各种组织的细胞膜表面, 执行着重要的生物学功能。随着高通量、高灵敏度和高分辨率的蛋白质组学时代的来临, 许多基于串联质谱技术解析糖链结构的生物数据库和分析软件也应运而生。本文综述了目前文献中最常用的糖类生物信息学资源, 包括各种糖蛋白的数据库以及质谱解析糖类的相关工具和新技术、新方法。

关键词 糖蛋白质组学, 质谱, 糖信息学, 数据库, 软件

学科分类号 Q53

DOI: 10.16476/j.pibb.2016.0115

糖蛋白广泛存在于真核、原核以及微生物中, 是生物体内重要的生理活性物质。带有分支结构的糖链分子通过羟基以共价方式连接到蛋白质的羟基或羧基上, 形成糖肽键。依据糖肽键的不同, 糖蛋白可分为 O-糖蛋白和 N-糖蛋白^[1]。目前研究主要以 N-糖蛋白为主。糖链的构象影响蛋白质的正确折叠、细胞内定位、分子功能以及抗原性等^[2]。据统计, 细胞内超过 50% 的蛋白质都会发生糖基化修饰, 这些糖蛋白广泛分布于各种组织的细胞膜表面, 是接收细胞外部信号刺激的主要途径, 而且与细胞之间信息交流、信号转导、细胞识别、免疫分子互作有着密切关联。此外, 在临床研究发现, 糖蛋白不仅参与肿瘤的发生、发展和转移, 蛋白质上的聚糖和其糖链结构的变化也参与了代谢性疾病、心脑血管疾病、先天性遗传病等重大疾病的发生发展^[3]。糖蛋白与胚胎干细胞的发育过程也有关系^[4]。由此可见, 对于糖蛋白的研究具有十分重要的生物学意义和应用前景。

糖组学是一门新兴的学科, 目标是表征和研究生物体内合成的糖分子, 包括对细胞内所有的糖基转移酶进行分类以及发现糖的真实结构。而糖蛋白质组学是研究被糖链修饰的蛋白, 不仅包含糖组学的研究部分, 还要揭示蛋白质上发生修饰位点和蛋白质本身的功能。

单糖与单糖之间存在多种连接方式, 经常形成复杂的树状分支结构。生物体内糖链不是由基因组直接编码形成, 因此不能从 DNA 序列中直接预测出糖链结构。基因组和蛋白质组信息学算法基本上都不能直接移植到糖组学中, 因此糖信息学方法相对落后。目前, 预测糖基化修饰位点相对简单, 可以从蛋白质序列直接预测或通过常规蛋白质组学技术得到。而鉴定糖的拓扑结构信息学方法不多, 相对可信的都是通过糖谱图库检索的思路。也有人通过从头预测方法从谱图中直接推断糖链结构, 但由于计算复杂度呈指数增长, 不适合长糖链的鉴定。发展一种通过质谱数据自动解析糖链结构的算法是目前糖信息学领域中最热门的研究。

1 糖蛋白相关的数据库

1.1 糖蛋白库

SWISS-PROT 是一个被生物学家广泛使用的高可信、高注释的蛋白质序列库。据统计, 在 SWISS-PROT 中大约有 70% 的蛋白质含有潜在的

* 国家自然科学基金资助项目(31300680)。

** 通讯联系人。

Tel: 021-54237482, E-mail: zhangyang@fudan.edu.cn

收稿日期: 2016-07-01, 接受日期: 2016-08-29

N-糖基化位点 Asn-X-Ser/Thr(X 为除了脯氨酸以外的任一种氨基酸)的序列模式, 而基于数据库中注释的信息证明生物体内大约有 50% 以上的蛋白质会发生糖基化^[5-7].

PDB (Protein Data Bank, <http://www.rcsb.org/pdb/home/home.do>) 数据库是目前世界上最大的蛋白质空间结构数据库, 拥有大约 11 万多种蛋白质结构. 其中仅有 9% 左右的蛋白注释含有 N-、O-糖信息, 原因是 PDB 库中大部分蛋白质一方面来源于细菌, 另一方面是在大肠杆菌中重组表达得到的. 糖链的存在会影响蛋白质晶体的生长, 因此蛋白质在结晶前一般都会将糖链切除. 在这几种情况下, 即使蛋白来源于哺乳动物或者植物, 它们都不会被糖基化. 尽管如此, 从 PDB 中还是可以查询到 4 375 个 N-糖结构和 4 106 个 O-糖结构.

O-GLYCBASE (<http://www.cbs.dtu.dk/databases/OGLYCBASE/>) 是一个经过修正的 O-和 C-糖基化蛋白质的数据库 (版本 6 包含了 242 种非冗余记录). 该数据库收录的标准是蛋白质至少含有 1 个经过实验验证的 O-或 C-糖基化位点^[8].

1.2 糖结构库

糖结构库是糖组学研究的首要资源, 是糖肽质谱鉴定的数据基础. KEGG GLYCAN 包含 1 万多个糖结构信息, 已经被业界评为糖组学最重要的数据资源. KEGG GLYCAN 中大部分记录来自 CarbBank 数据库. 1999 年 CarbBank 计划终止后, 在糖生物学家的不懈努力下保留了这些数据且在其基础上又添加了 750 多种糖结构^[9]. 此外, KEGG 提供的 Pathway 数据库也收集了 41 条糖生物合成的通路和 100 多个糖生物合成酶.

欧洲的 GLYCOSCIENCES. de^[10-11] 和美国的 CCSD (The Complex Carbohydrate Structure Database)^[12-13] 也是糖组学的巨大免费资源. CCSD 涵盖来自 14 000 多篇文献中收集的 23 500 个左右的碳水化合物结构, 是迄今为止最大的糖类相关数据来源, 但由于资金问题已停止更新, 其大部分数据已被 GLYCOSCIENCES. de (近 40 000)、KEGG Glycan database (10 989)、CFG Glycan Database (8626)^[14]、Bacterial Carbohydrate Structure Database (BCSDB, 11 605)^[15], 以及 GlycoBase (377)^[16] 收录.

UniCarbKB 是在 GlycoSuiteDB 的基础上, 结合 EUROCarbDB 搜索标准, 整合 MonosaccharideDB、GlycoWorkBench、GlycanBuilder 和 GlycoBase 等资源的糖蛋白结构数据库. 该数据

库包含糖类型、核心结构类型、连接方式、异构体、质量数、化学式和检测方法, 并应用于疾病诊断、预防(疫苗)和治疗等方面. UniCarbKB 目前包含 890 篇参考文献, 3 740 个糖结构、400 个糖蛋白和 598 个糖基化位点^[17].

随着每年数据累积, PDB 库中越来越多的糖结构信息出错, 因此基于 PDB 产生了 2 个二级数据库. GlycoconjugateDataBank 通过 “getCARBO” 软件将 PDB 中的 N-糖与 KEGGGLYCAN 库进行比对, 以完成对糖三维结构的检测和注释^[18]. GlycoMapsDB 采用了构象分析软件 CAT (Conformational Analysis Tools) 通过检测糖苷键扭转角 φ/ψ 的合理性来进行质控^[19].

GlycomeDB 是目前最大的糖链结构数据库. GlycomeDB 从 7 个糖结构数据库, 包括 BCSDB、CarbBank、CFG、GlycoBase (Dublin)、GlycoBase (Lille)、GlyAffinity、GLYCOSCIENCES. de 和 KEGG 中获取所有糖链结构, 经过汇总、纠正和去冗余, 最后将所有记录转译成 GlycoCT 和 Glyde 两种通用格式. 用户可以通过一个 JAVA 程序 “GlycoUpdateDB” 实现数据库的本地化^[20].

GlycoBase 数据库及其搜索工具 autoGU, 可以解析高效液相色谱 (HPLC) 获得的糖谱数据. GlycoBase 包含超过 350 个高效液相色谱洗脱的 2-AB 标记的 N-糖链结构与外切糖苷酶酶解产物. AutoGU 可以对外切糖苷酶酶解得到的 HPLC 指纹图谱进行谱峰注释. 这个工具对低浓度的聚糖分析最有帮助, 不仅能提供定量信息, 而且可以与质谱解析得到糖链结构相互验证. GlycoExtractor 网站^[21] 可以提取糖链的液相色谱信息, 包括峰数据、峰面积和 GU (GlycanUnit) 值.

也有人尝试将糖链结构用类似氨基酸的线性方式表达, 创建了 GlyDB 数据库. 利用蛋白质组学最常用的搜索引擎 Sequest 完成实验谱图和数据库之间的检索. 此方法只能处理低能 CID 下获得的糖肽谱图. 糖肽在低能 CID 下大部分碎片都是由糖链部分碎裂产生的^[22].

目前不同数据库对糖链结构的存储编码方式不同, 因此不同数据库之间进行数据分析时往往会产生重复记录. 而 GlyTouCan 数据库则对所有的糖结构资源进行整合, 设立唯一检索编码, 以简化分析流程^[23].

表 1 列出了当前常用的糖链结构数据库资源.

Table 1 Glycan structure database

表 1 糖链结构数据库

数据库	编码方式	注释信息	网址	网站状态
GlycomeDB	GlycoCT	糖结构、物种分类	http://www.glycome-db.org/	M
EUROCarbDB	GlycoCT	糖结构、物种分类、质谱数据	http://www.ebi.ac.uk/eurocarb/	S
CarbBank	IUPAC extended	糖结构、物种分类	http://www.boc.chem.uu.nl/sugabase/carbbank.html	S
KEGG GLYCAN	KCF	糖结构、与 KEGG 其他库的交叉注释信息	http://www.genome.jp/kegg/glycan/	S
GLYCOSCIENCES.de	LINUCS	糖结构、物种分类、蛋白三维结构、核磁数据、PDB 库索引号	http://www.glycosciences.de/	M
CFG Glycan Structure DB	Glycominds linear code	糖结构、物种分类、糖芯片数据	http://www.functionalglycomics.org/	S
BCSDB	BCSDB linear code	糖结构、物种分类、核磁数据	http://www.glyco.ac.ru/bcsdb3/	M
UniCarbKB	Figure	糖结构、质谱数据、糖蛋白信息	http://unicarbkb.org/	M
GlycoBase (Dublin)	Motif based	糖结构、物种分类、质谱数据、色谱数据	http://glycobase.ucd.ie/	S
GlycoBase (Lille)	Linkage path	糖结构、物种分类、核磁数据	http://glycobase.univ-lille1.fr/base/	M
JCGGDB	CabosML	糖结构、物种分类、糖基转移酶、糖芯片、核磁数据、质谱数据	http://jcgddb.jp/	M
GlyTouCan	GlycoCT	糖结构	http://glytoucan.org/	M

M: 数据库可用; S: 数据库不可用。

2 糖蛋白鉴定相关的软件

2.1 基于肽段序列的预测软件

从功能基因组的时代起,就有很多研究通过蛋白质序列来预测蛋白质上可能发生的修饰及位点,其中以支持向量机(support vector machine, SVM)算法最为纯熟,广泛运用于磷酸化肽段预测.预测修饰是基于修饰位点附近氨基酸组成的全局分析. GlySeqDB 是一个分析糖基化修饰位点规律性的数据库,数据集主要来源于 PDB 和 SwissProt. 用 GlySeqDB 能够统计出已知 N-、O-糖基化位点周边氨基酸序列规律. GlySeqDB 每周利用“pdb2linucs”工具采集 PDB 的数据,过去冗余,统计出糖基化位点周围的氨基酸序列信息.

丹麦技术大学生物序列分析中心(CBS, <http://www.cbs.dtu.dk/>)开发的 NetNGlyc 软件,是通过人工神经网络算法(artificial neural network, ANN)分析包含 Asn-X-Ser/Thr 序列子的蛋白来预测 N-糖基化位点. NetNGlyc 会先利用 SignalP 程序来过滤掉无信号肽的蛋白质序列. 基于 ANN 算法, CBS 还提供其他多种蛋白质翻译后修饰的预测,包括 O-GalNAc 型、O-(alpha)-GlcNAc 型、O-(beta)-GlcNAc 型,分别由 NetOGlyc、DictyOGlyc、YinOYang 软件完成.

除了 ANN 算法,有人通过 SVM^[24]和随机森林(random forest, RF)^[25]算法提高以氨基酸序列来预测糖基化位点的效率. 但仅从序列来预测蛋白质修饰,假阳性太高,若结合相关质谱数据可以大大降低错误率.

2.2 基于质谱数据的分析软件

目前国际上糖结构相关的质谱分析软件较多,总体上分为两大类:一类是数据库依赖型,另一类是数据库非依赖型(*de novo*),前者大都为业内著名的糖结构数据库提供.

SWISSPROT 在基于质谱数据的糖修饰分析方面提供了 FindMod、GlycanMass 和 GlycoMod 三大免费工具. FindMod 是通过搜索肽指纹图谱(peptide mass fingerprint, PMF)确定修饰种类的软件. 先比对 PMF 数据与理论肽之间的所有质量差值,在 SWISSPROT 数据库(22 种修饰,5 153 个肽段)中搜索这些差值来确定修饰类型. GlycanMass 可用来计算寡糖质量数的简易小工具,用户只需设定 7 类单糖和一些无机物的上限个数,便可得到化合物的总分子质量. GlycoMod 与 FindMod 基本一致,只是在 FindMod 的基础上添加了计算糖链组成的模块. GlycoMod 能用来预测从糖蛋白上解离下来的糖链结构及其修饰状态和衍生还原端,也能计算已知肽序列上的糖组成. Gusakov 等^[26]研究了

丝状真菌 *Chrysosporium lucknowense* 不同突变株的糖蛋白表达情况, 利用 GlycoMod 成功找到了 28 个酶的 N-糖链结构。

德国癌症研究中心提供的 GlycoFragment 和 GlycoSearchMS 是一套计算和搜索糖链质谱数据的简易工具。用户以 IUPAC 格式输入结构式, GlycoFragment 便可得到糖链的 A-、B-、C-、X-、Y- 理论碎片离子及质量数。GlycoSearchMS 将 GlycoFragment 产生的所有理论碎片和 SweetDB 数据库中所有糖链结构质量数进行比对, 给出候选结果。GlycoFragment 和 GlycoSearchMS 可以在复杂样品中自动鉴定糖组成。目前版本的 GlycoSearchMS 提取了 SweetDB 数据库中大约 5 000 个 N-糖和 1 200 个 O-糖的理论质谱图数据^[27]。

GlycoX 是用来计算糖基化位点与糖蛋白中寡糖结构的 MATLAB 开发的程序。用非特异性酶酶切的糖蛋白在糖基化位点上会留下一个很小的标签用以确定糖基化位置。在输入糖肽精确质量数以及糖的串联质谱数据后, 程序能够计算每个糖基化位点上糖的组分及其对应位点。GlycoX 纯粹计算谱图的碎片信息, 不与数据库的理论碎片进行比较^[28]。

PARC(Palo Alto Research Center)机构同时开发了 Cartoonist(对一级质谱数据进行峰注释)、Cartoonist2(对二级质谱数据进行峰注释)以及 Peptonist(同时采集一级和二级质谱数据确定糖肽组成)这三个糖组学软件。Peptonist 先对二级谱图进行糖基化肽段的分析, 再配合对应的一级谱图推断每条肽段上是否发生 N-糖基化。Cartoonist 和 Peptonist 采用了专家系统来精确归属 N-糖蛋白, 而 Cartoonist2 仅适用于分析 O-糖蛋白, 较少采用专家系统的帮助^[29-30]。之后, 他们又在新的实际样品中进一步测试算法的可信度, 准确度超过 50%^[31]。

GlycoPepDB^[32]是糖肽组成注释的网站, 算法主要利用实验数据与 N-糖库中的糖肽质量数比对来预测结果。GlycoPepDB 能搜索生物相关的碳水化合物组成以及预测糖型模式, 引用了“智能搜索”概念, 因此比目前 N-糖工具更加先进。GlycoPepID^[33]是其网站提供的搜索工具, 用户在提供蛋白质序列、酶切种类、质量数容差范围、电荷状态等信息后就可对提交的质量数列表进行糖肽组成的扫描。

用多级质谱产生的糖链碎片理论上可以以自下而上的方式(n 级 $\rightarrow$ 1 级质谱数据)全面的解析糖链

结构。FragLib^[34-35]运用这个思想产生了全甲基化糖链碎片库, 配套软件可以存储、组织、比较谱图和评估结果, 包括残基连接方式、单糖种类、糖链构象和分支结构等。该方法已成功运用到未知样品。

GlycoMiner^[36]是自动解析 N-糖肽的软件。软件主要分两步, 第一步是通过诊断离子(marker ion)判别谱图是否属于糖肽, 即对低质量范围的离子进行扫描, 并给出对应的得分 S_{marker} (marker score), 第二步是确定组成并给出质量打分 S_{quality} (spectrumquality score)。GlycoMiner 的核心思想是在谱图中搜索预定义的糖相关碎片组合(lossvector)来确定裸肽的质量数, 然后在理论肽段库中找到裸肽质量数。最好的情况是糖肽上的糖链中性丢失到糖肽键的位置, 这样匹配度最高。最后对匹配峰进行糖组成注释, 并通过 S_{quality} 给出候选结果。 S_{quality} 同时考虑了罚分、质量数差异、噪音数目等因素。在 AGP 和转铁蛋白两个标准糖蛋白测试中, 控制假阴性率和假阳性率都在 0.1% 的条件下, GlycoMiner 能搜索到的 338 个糖肽, 其中 196 个进手工验证为正确, 其余谱图质量较低。

Gao 等开发了一种大规模产生 N-糖结构库的方法, 利用此库可以高通量地鉴定低能 CID 来源的糖肽谱图, 并给出糖链初级结构和组成。预测过程主要以检索引擎 Glyquest 实现, 该软件目前还有些缺陷: 不能获得肽段的未知修饰、不能区分同分异构体以及只能处理简单样品。

2.3 其他软件

目前, 一些综合性强的免费软件运用较广泛, 如 KegDraw 和 GlycoWorkbench。KegDraw 是用来绘制糖结构的 JAVA 软件。KegDraw 采用两种模式来绘制糖结构: 化合物模式和糖模式, 绘制的结构可以在 KEGG 的各个数据库中进行查询。而 KCaM 软件能够分析 KegDraw 绘制的糖结构, 将绘制的结构和数据库中的已知糖结构进行相似度匹配, 获取最接近的查询结果^[37]。

GlycanBuilder 是集成在 GlycoWorkBench 中的一个模块, 两者结合不但可以完成糖结构的图形化标注, GlycoWorkBench 可以计算特定参数下, 如衍生物修饰类型和碎片离子类型、产生的所有理论碎片, 并与 CFG 和 KEGG 等糖库进行一一比对^[38-40]。

SysBioWare 是一个半自动谱图解析软件。SysBioWare 读取原始谱图后, 经过降噪和小波分析后自动检测同位素峰。检测到的单同位素质量数自动关联到原始谱图中, 并做糖组成预测。

SysBioWare 在人的尿液样品中得到很好地运用^[41].

表 2 列举了当前通过质谱数据来鉴定糖蛋白以

及解析糖蛋白结构的相关软件.

Table 2 Glycan analysis software for Mass Spectrometry

表 2 糖相关的质谱解析软件

名称	描述	网址
GlySeq	计算糖基化位点近邻氨基酸的组成概率	http://www.glycosciences.de/tools/glyseq/
Glycofragment	计算寡糖理论碎片	http://www.glycosciences.de/tools/GlycoFragments/
GlycoSearchMS	实现实际谱图与理论碎片匹配	http://www.glycosciences.de/sweetdb/start.php?action=form_ms_search
GlycoPeakfinder	搜索 GLYCOSCIENCES.de 来注释糖峰	http://www.eurocarbdb.org/applications/ms-tools/
GlycoWorkBench	糖型拓扑结构构建以及用实际谱图数据搜索公共库的综合工具	http://www.eurocarbdb.org/applications/ms-tools/
GlycoMod	糖或糖肽组成预测工具	http://www.expasy.org/tools/glycomod/
GlycanMass	单糖组成计算器	http://www.expasy.org/tools/glycomod/glycanmass.html
GlycoMiner	糖肽解析的软件平台	http://w3.chemres.hu/ms/glycominer
GlycoX ^[28]	通过精确质量数预测糖肽组成	-
GlycoExtractor	辅助 HPLC 中抽提出糖谱数据软件	http://glycobase.nibrt.ie:8080/DemoGlycoExtractor/MainPage
SysBioWare ^[41]	糖链结构质谱数据注释平台	-
Glyquest	N-糖结构库的搜索引擎	http://BioPharmaSoft.com
GlycoPepDB	糖肽结构库及搜索引擎	http://hexose.chem.ku.edu/sugar.php

3 从头预测算法

在蛋白质组学时代, 从头预测技术是在无背景数据库作检索情况下, 直接在谱图中推测肽段序列. 而在糖组学中, 此方法仍被沿用. STAT 和 StrOligo 是最早提出的两大算法, 可以根据质谱数据直接预测糖链拓扑结构.

糖拓扑分析工具(Saccharide Topology Analysis Tool, STAT)是一个基于网页开发的计算程序, 能够迅速地从最多包含 10 个残基的糖链谱图中提取结构信息. 当输入母离子质量数、可能的单糖组成单元、电荷信息以及碎片离子质量数后, 程序从多级谱图数据中分析产生所有可能的单糖连接信息. 将 STAT 的结果与糖结构可能的分支结构结合起来, 可以进一步筛选出最合理的结构. 由于计算复杂度的限制, STAT 不能解析太长的糖链, 也不适合解析糖链环内断裂数据^[42].

StrOligo 算法主要处理 MAIDI 产生的质谱数据, 可鉴定复杂 N-糖. 经实验测试, StrOligo 准确率在 80%左右. StrOligo 算法包含两部分内容: a. 利用最小二乘法进行同位素峰的过滤; b. 基于糖肽母离子和其串联数据, 通过构建质量数的“关系树”来确定糖结构. 由于存在随机因素, “关系树”内可能融入错误的糖肽质量数, 但并不影响鉴

定结果, 因为真实结果还是占主导地位^[43].

在缺乏糖结构数据库作为检索源情况下, 从头预测方法是从质谱数据推断糖结构的一种不错的选择. 但由于此方法依赖于穷举计算, 迫于计算压力无法完成包含 10 个以上单糖的糖链预测. 另外, 此方法不能鉴定糖链的分支结构和出现位置, 不能区别同分异构体.

OSCAR (oligosaccharide subtree constraint algorithm)^[44]方法已被多篇文献引用. OSCAR 算法是利用自上而下的思想利用多级谱图的碎片通路推断糖链拓扑结构, 包括分支结构和连接方式. 因为糖链合成有逻辑上的限制, 这种从头预测的方法可以在不用预先生物合成或外部比较的条件下获取分子的拓扑结构. OSCAR 目前在全甲基化糖链拓扑结构确立中运用较多.

Peltoniemi 等^[45]用常规蛋白质组学技术路线(分子筛→反相色谱→质谱)得到经胰酶酶解的 N-糖肽谱图. 在无已知糖链结构库支持的情况下, 运用分枝界定算法可以自动从串联质谱数据中同时得到糖链结构和肽段序列.

4 糖结构的命名和数据交换

各种糖结构数据库都以自定义的命名方式来存储数据, 这样就很难在各个数据库之间比较和共享

糖结构数据. 为此, 一系列公认的命名方案诞生了, 但糖序列的复杂性使得很难用一种简洁的语言来描述. 早在 1969 年 IUPAC 就初步对糖类进行了系统性的命名. IUPAC-IUBMB(International Union of Pure and Applied Chemistry and International Union of Biochemistry and Molecular Biology)在 1996 年定义了糖类的正式命名, 其中包括了各类单糖、寡糖、聚糖等规范命名. 不过将其应用于日常使用的糖命名中过于繁琐, 科学家们尝试将糖命名更简洁, 并且更容易让计算机识别. SWEETDB 采用了 LINUCS (linear notation for unique description of carbohydrate sequences)方式来描述糖结构, 它可以将二维结构转换成唯一的序列描述, 但 LINUCS 的描述也很难让人轻松读懂.

KEGG 的 KCF(KEGG chemical function)格式尝试将糖结构的描述图形化, 并更方便地与其他糖数据库相接. 而计算机技术的发展使得图形化表示糖更加方便: KEGDraw 和 GlycoWorkBench 采用的图形表示方法则更加简明、通俗易懂.

不过计算机并不擅长处理图形数据, 随着可扩展标记语言 XML(extensible markup language)技术的成熟, 我们可以轻松地将图形化、符号化的糖转换成计算机更易识别并处理的语言. CabosML(carbohydrate sequence markup language)和 GLYDE(GLYcan data exchange)分别是日本与美国科学家开发出的用 XML 格式来描述的糖规则. 两者语法上略有出入, 而其实质则基本一致, 这样的命名方法大大降低了计算机处理糖结构信息的难度, 使得建立一个糖结构数据库变得更加简单. 同时对于糖类生物学家来说, 这样的表示方法也相对易读^[46].

5 糖链结构鉴定的最新研究方法

5.1 人工合成糖数据集的比对

最近的研究表明糖链的串级谱图强度和其本身的组成有着内在的联系. 将糖链的串级谱图和人工合成的已知结构的糖串级谱图进行比对已成为一个非常有效的鉴定手段. 而此方法的关键在于需要足够多的人工合成的糖质谱数据来对未知的糖进行匹配. 现有的人工合成糖数据集中包含了每个糖的串级和三级谱图, 每个糖结构都重复 3 次采集. 实验谱图通过欧氏距离与糖谱图库进行比对, 并从分析物中挑出最优的串级谱图做三级质谱.

整个方法的优缺点: 即使不熟悉糖链结构分析, 也可通过质谱轻松鉴定糖链, 而且只需要质谱

无需其他分析手段, 只需少量的样品即可. 对于每个二级碎片峰, 只有计算出可能为糖链才会对其进行三级质谱鉴定, 大大节约了时间并节省了样品. 此方法能鉴别出只有构象差异的糖, 如甘露糖和葡萄糖, 还能通过多级谱图区分同分异构体. 此方法的缺点在于糖链的鉴定非常依赖于人工合成库的大小, 无法补充从头预测方法带来的优势^[47-48].

5.2 人工修饰的数据库

人工修饰的蛋白质序列库可用来降低糖肽鉴定中肽段的错误匹配, 且减少糖肽谱图手工验证的工作量. 其工作原理是用 PNGase(肽段 N-糖苷酶)F 酶将糖链从糖蛋白上游离下来, 肽段修饰位点上的天冬酰胺变成天冬氨酸, 产生了 0.9840 u 的质量位移. 在蛋白质序列搜索时在天冬酰胺上加入了去酰胺化的可变修饰即可完成对肽段的搜索. 在糖基化位点确认中将 Asn-X-Ser/Thr 替换为 J-X-Ser/Thr 用以代表去糖基化的 Asp-X-Ser/Thr. 这样数据库检索时就不必加入可变修饰, 而直接搜索修改数据库中的 J, 即能确认 N-糖基化位点. 由于天冬酰胺本身的去酰胺化修饰会引起 N-糖基化肽段的检索错误, 因此最好的方法是在天冬酰胺去糖基化的过程中加入 ^{18}O 的同位素标记, 这样观测到的糖基化位点的质量数偏移就变成了 3 u 了. 此方法只能鉴定 N-糖修饰位点, 无法追踪位点特异性的糖链修饰^[49]. 而后 Thaysen-Andersen 等^[50]研究了 this 质量数偏移和糖肽的定量关系.

5.3 糖基转移酶预测糖链结构

糖链是由一系列糖基转移酶以不同方式组合而成的. 若事先通过表达谱实验了解到特定细胞或特定阶段的糖基转移酶表达情况, 就可能预测出样本中所有可能的糖型结构. Kawano 等^[51]提出了一种利用糖基转移酶转录组数据来预测糖型结构. 该方法首先通过创建一个糖基转移酶参与键形成的化学反应库, 计算每个化学反应在不同糖型结构中共同出现的频率. 可以计算出化学反应共同出现得分, 并考虑适当的罚分. 作者利用 KEGG GLYCAN database 作为理论糖库, 准确率可达 81%. 而在实际运用中, 利用人类癌症细胞的基因表达谱, 可成功预测出带有唾液酸和 SialylLewisX 的糖型结构.

5.4 结合质谱技术分析同一生理过程的全部糖蛋白以及糖结构信息

目前糖肽的完整鉴定主要是基于两步法策略. 两步法策略是从完整糖肽上用糖肽酶或化学方法, 包括广泛使用的 N-聚糖酶(如 PNGase F 和 PNGase

A)来释放 N- 糖链, 或替代化学裂解(如 β - 消除反应或肼)来释放 O- 或 N- 连接的糖链, 分别用质谱来鉴定去糖基化肽段和糖链^[52], 即第一步为目标样品构建一个去糖基化肽段库, 为第二步的糖肽谱图检索提供数据库. Zhang 等^[53]利用固相提取 N- 糖链和糖肽的化学酶解离方法检测到近百个糖蛋白的糖链结构及其位点信息. 数据处理中利用 GlycoMod 搜索 CFG 数据库确定糖链, 利用 GPQuest 进行完成糖肽鉴定. Liu 等^[54]基于两步法策略开发的 GRIP 方法可以规模化地鉴定到人类血清样品中 745 个 N- 糖肽及其糖链组成, 不仅观测到免疫球蛋白在全局水平上的微观不均一性, 而且近一半的糖型结构含唾液酸(唾液酸极难离子化, 成为糖组学的技术瓶颈), 是迄今为止最高通量的糖肽组学数据. 两步法策略实验步骤复杂, 很难保证实验重复性, 必然会被一步法策略更替. 目前的研究表明, 基于同一个母离子在不同质谱碎裂条件形成的谱图(一母多谱)构建的一步法搜索引擎必将成为糖肽质谱解析的唯一途径.

6 总结与展望

近年来, 糖相关数据的收集和信息学工具的开发以及各个数据库中相关数据的交叉引用的工作进展迅速, 但相对于基因组和蛋白质组领域的信息学方法来说, 糖信息学的研究依然属于起步阶段. 而糖谱图的自动解析算法依然处于瓶颈阶段, 这大大阻碍了高通量蛋白质组学研究中快速可靠地鉴定糖型结构.

糖链结构最流行的鉴定方法还是依赖于可重现性高的质谱数据去匹配已知糖链结构的理论碎片. 这种方法主要由两个步骤完成: a. 需要事先定义一批已知结构的糖链, 并在其基础之上产生对应的理论碎片离子作为检索数据库; b. 在测试数据集中, 根据实验数据和检索数据库的最优匹配程度自定义数学公式作为打分和排序原则. 此方法虽然简单可行, 但也忽略了两个重要的问题: a. 糖链中不是每个糖苷键都会在质谱中发生断裂; b. 由于理论谱中不存在峰强度的信息, 因此不能利用峰强度去匹配实际样本.

日本糖学会在这方面作了改进. 为了模拟实验谱图, 他们利用人工合成糖形成的碎片信息作为检索数据库, 虽然可以同时解决以上两个问题, 但花费昂贵. 因此, 发展一套既经济又有效的质谱数据解析技术是目前生物信息学在糖组学研究方面的首

要任务. 此外, 糖结构的表述标准和格式, 以及糖组学和基因组学 / 蛋白质组学的整合等都是现在研究的重点.

参 考 文 献

- [1] Jensen P H, Karlsson N G, Kolarich D, *et al.* Structure analysis of N- and O-Glycans released from glycoproteins. *Nature Protocol*, 2012, **7**(7): 1299–1310
- [2] Angelyn Larkin, Barbara Imperiali. The expanding horizons of asparagine-linked glycosylation. *Biochemistry*, 2011, **50**: 4411–4426
- [3] Bieberich E. Synthesis, processing, and function of N-glycans in N-glycoproteins. *Adv Neurobiol*, 2014, **9**: 47–70
- [4] Fujitani N, Furukawa J, Araki K, *et al.* Total cellular glycomics allows characterizing cells and streamlining the discovery process for cellular biomarkers. *Proc Natl Acad Sci USA*, 2013, **110**(6): 2105–2110
- [5] Lima T, Auchincloss A H, Coudert E, *et al.* HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. *Nucleic Acids Res*, 2009, **37**(Database issue): D471–D478
- [6] Schneider M, Lane L, Boutet E, *et al.* The UniProtKB/Swiss-Prot knowledgebase and its Plant Proteome Annotation Program. *J Proteomics*, 2009, **72**(3): 567–573
- [7] Lam P V, Goldman R, Karagiannis K, *et al.* Structure-based comparative analysis and prediction of N-linked glycosylation sites in evolutionarily distant eukaryotes. *Genomics Proteomics Bioinformatics*, 2013, **11**(2): 96–104
- [8] Cornelia Caragea, Jivko Sinapov, Adrian Silvescu, *et al.* Glycosylation site prediction using ensembles of Support Vector Machine classifiers. *BMC Bioinformatics*, 2007, **8**(438): 1–13
- [9] Kanehisa M, Goto S, Furumichi M, *et al.* KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res*, 2010, **38**(Database issue): D355–D360
- [10] Lutteke T, Böhne-Lang A, Loss A, *et al.* GLYCOSCIENCES.de: an Internet portal to support glycomics and glycobiology research. *Glycobiology*, 2006, **16**(5): 71R–81R
- [11] Toukach P, Joshi H J, Ranzinger R, *et al.* Sharing of worldwide distributed carbohydrate-related digital resources: online connection of the bacterial carbohydrate structure database and GLYCOSCIENCES.de. *Nucleic Acids Res*, 2007, **35** (Database issue): D280–D286
- [12] Döbete S, Bock K, Smith D, *et al.* The complex carbohydrate structure Database. *Trends Biochem Sci*, 1989, **14**(12): 475–477
- [13] Döbete S, Albersheim P. CarbBank. *Glycobiology*, 1992, **2**(6): 505
- [14] Raman R, Raguram S, Venkataraman G, *et al.* Glycomics: an integrated systems approach to structure-function relationships of glycans. *Nat Methods*, 2005, **2**(11): 817–824
- [15] Herget S, Toukach P V, Ranzinger R, *et al.* Statistical analysis of the Bacterial Carbohydrate Structure Data Base (BCSDB): characteristics and diversity of bacterial carbohydrates in comparison with mammalian glycans. *BMC Struct Biol*, 2008: 35

- [16] Campbell M P, Royle L, Radcliffe C M, *et al.* GlycoBase and autoGU: tools for HPLC-based glycan analysis. *Bioinformatics*, 2008, **24**(9): 1214–1216
- [17] Campbell Matthew P, PetersonRobyn, MariethozJulien, *et al.* UniCarbKB: building a knowledge platform for glycoproteomics. *Nucleic Acids Research*, 2014, **42**(D1): D215–D221
- [18] Nakahara T, Hashimoto R, Nakagawa H, *et al.* Glycoconjugate Data Bank: Structures—an annotated glycan structure database and N-glycan primary structure verification service. *Nucleic Acids Res*, 2008, **36**(Database issue): D368–D371
- [19] Frank M, Lutteke T, von der Lieth C W. GlycoMapsDB: a database of the accessible conformational space of glycosidic linkages. *Nucleic Acids Res*, 2007, **35**(Database issue): 287–290
- [20] René Ranzinger, Stephan Herget, Claus-Wilhelm von der Lieth, *et al.* GlycomeDB—a unified database for carbohydrate structures. *Nucleic Acids Res*, 2011, **39**(Suppl 1): D373–D376
- [21] Artemenko N V, Campbell M P, Rudd P M. GlycoExtractor: a web-based interface for high throughput processing of HPLC-glycan data. *J Proteome Res*, 2010, **9**(4): 2037–2041
- [22] Ren JM, Rejtar T, Li L, *et al.* N-Glycan structure annotation of glycopeptides using a linearized glycan structure database (GlyDB). *J Proteome Res*, 2007, **6**(8): 3162–3173
- [23] Aoki-Kinoshita Kiyoko, Agravat Sanjay, Aoki Nobuyuki P, *et al.* GlyTouCan 1.0-The international glycan structure Repository. *Nucleic Acids Research*, 2016, **44**(D1):D1237–D1242
- [24] Caragea C, Sinapov J, Silvescu A, *et al.* Glycosylation site prediction using ensembles of Support Vector Machine classifiers. *BMC Bioinformatics*, 2007, **8**: 438
- [25] Hamby S E, Hirst J D. Prediction of glycosylation sites using random forests. *BMC Bioinformatics*, 2008, **9**: 500
- [26] Gusakov A V, Antonov A I, Ustinov B B. N-Glycosylation in *Chrysosporium lucknowense* enzymes. *Carbohydr Res*, 2008, **343**(1): 48–55
- [27] 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 Res*, 2004, **32**(Web Server issue): W261–W266
- [28] An H J, Tillinghast J S, Woodruff D L, *et al.* A new computer program (GlycoX) to determine simultaneously the glycosylation sites and oligosaccharide heterogeneity of glycoproteins. *J Proteome Res*, 2006, **5**(10): 2800–2808
- [29] 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
- [30] Goldberg D, Bern M, Parry S, *et al.* Automated N-glycopeptide identification using a combination of single- and tandem-MS. *J Proteome Res*, 2007, **6**(10): 3995–4005
- [31] 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
- [32] Go E P, Rebecchi K R, Dalpathado D S, *et al.* GlycoPep DB: a tool for glycopeptide analysis using a "Smart Search". *Anal Chem*, 2007, **79**(4):1708–1713
- [33] Irungu J, Go E P, Dalpathado D S, *et al.* Simplification of mass spectral analysis of acidic glycopeptides using GlycoPep ID. *Anal Chem*, 2007, **79**(8): 3065–3074
- [34] Ashline D, Singh S, Hanneman A, *et al.* Congruent strategies for carbohydrate sequencing. 1. Mining structural details by MSn. *Anal Chem*, 2005, **77**(19): 6250–6262
- [35] Zhang H, Singh S, Reinhold V N. Congruent strategies for carbohydrate sequencing. 2. FragLib: an MSn spectral library. *Anal Chem*, 2005, **77**(19): 6263–6270
- [36] Ozohanics O, Krenyacz J, Ludanyi K, *et al.* GlycoMiner: a new software tool to elucidate glycopeptide composition. *Rapid Commun Mass Spectrom*, 2008, **22**(20): 3245–3254
- [37] Hashimoto K, Goto S, Kawano S, *et al.* KEGG as a glycome informatics resource. *Glycobiology*, 2006, **16**(5): 63R–70R
- [38] Tissot B, Ceroni A, Powell A K, *et al.* Software tool for the structural determination of glycosaminoglycans by mass spectrometry. *Anal Chem*, 2008, **80**(23): 9204–9212
- [39] Ceroni A, Maass K, Geyer H, *et al.* GlycoWorkbench: a tool for the computer-assisted annotation of mass spectra of glycans. *J Proteome Res*, 2008, **7**(4): 1650–1659
- [40] Ceroni A, Dell A, Haslam S M. The GlycanBuilder: a fast, intuitive and flexible software tool for building and displaying glycan structures. *Source Code Biol Med*, 2007, **2**(1): 3
- [41] Vakhrushev S Y, Dadimov D, Peter-Katalinic J. Software platform for high-throughput glycomics. *Anal Chem*, 2009, **81** (9):3252–3260
- [42] Gaucher S P, Morrow J, Leary J A. STAT: a saccharide topology analysis tool used in combination with tandem mass spectrometry. *Anal Chem*, 2000, **72**(11): 2331–2336
- [43] Ethier M, Figeys D, Perreault H. N-glycosylation analysis using the StrOligo algorithm. *Methods Mol Biol*, 2006, **328**: 187–197
- [44] 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 MSn data. *Anal Chem*, 2005, **77**(19): 6271–6279
- [45] 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
- [46] Varki A, Cummings R D, Esko J D, *et al.* Symbol nomenclature for glycan representation. *Proteomics*, 2009, **9**(24): 5398–5399
- [47] Kameyama A, Nakaya S, Ito H, *et al.* Strategy for simulation of CID spectra of N-linked oligosaccharides toward glycomics. *J Proteome Res*, 2006, **5**(4): 808–814
- [48] Ito H, Kameyama A, Sato T, *et al.* Preparation of a glycan library using a variety of glycosyltransferases. *Methods Mol Biol*, 2009, **534**: 283–291
- [49] Atwood J A, Sahoo S S, Alvarez-Manilla G, *et al.* Simple modification of a protein database for mass spectral identification of N-linked glycopeptides. *Rapid Commun Mass Spectrom*, 2005, **19**(21): 3002–3006
- [50] Thaysen-Andersen M, Mysling S, Hojrup P. Site-specific

- glycoprofiling of N-linked glycopeptides using MALDI-TOF MS: strong correlation between signal strength and glycoform quantities. *Anal Chem*, 2009, **81**(10): 3933–3943
- [51] Kawano S, Hashimoto K, Miyama T, *et al.* Prediction of glycan structures from gene expression data based on glycosyltransferase reactions. *Bioinformatics*, 2005, **21**(21): 3976–3982
- [52] Goetz J A, Novotny M V, Mechref Y. Enzymatic/chemical release of O-glycans allowing MS analysis at high sensitivity. *Anal. Chem*, 2009, **81**(23): 9546–9552
- [53] Sun S S, Punit Shah, Shadi Toghi Eshghi, *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
- [54] Liu M Q, Zhang Y, Chen Y H, *et al.* Efficient and accurate glycopeptide identification pipeline for high-throughput site-specific N-Glycosylation analysis. *J Proteome Research*, 2014, **13**(6): 3121–3129

Bioinformatic Resources and Methods for Glycoproteomics*

LIU Hang, YAO Jun, YANG Peng-Yuan, ZHANG Yang**

(The Institutes of Biomedical Sciences of Shanghai Medical School, Fudan University, Shanghai 200032, China)

Abstract Glycosylation is one of the most common and important post-translational modifications (PTMs) of proteins in living organisms. More than 50% proteins are glycosylated among mammals. Glycoproteins are widely distributed in cell membrane surface and body fluids, and play critical roles in biological processes. With the rapid development of biological mass spectrometry technique, as well as biological database and analysis software for glycan structure elucidation, glycoproteomics research field is booming soon. This article summarized the most commonly used bioinformatic resources in literatures, including new developed databases, MS technique and methods.

Key words glycoproteomics, mass spectrometry, glycoinformatics, database, software

DOI: 10.16476/j.pibb.2016.0115

*This work was supported by a grant from The National Natural Science Foundation of China (31300680).

**Corresponding author.

Tel: 86-21-54237482, E-mail: zhangyang@fudan.edu.cn

Received: July 1, 2016 Accepted: August 29, 2016