Toward Automated Identification of Glycan Branching Patterns Using Multistage Mass Spectrometry with Intelligent Precursor Selection

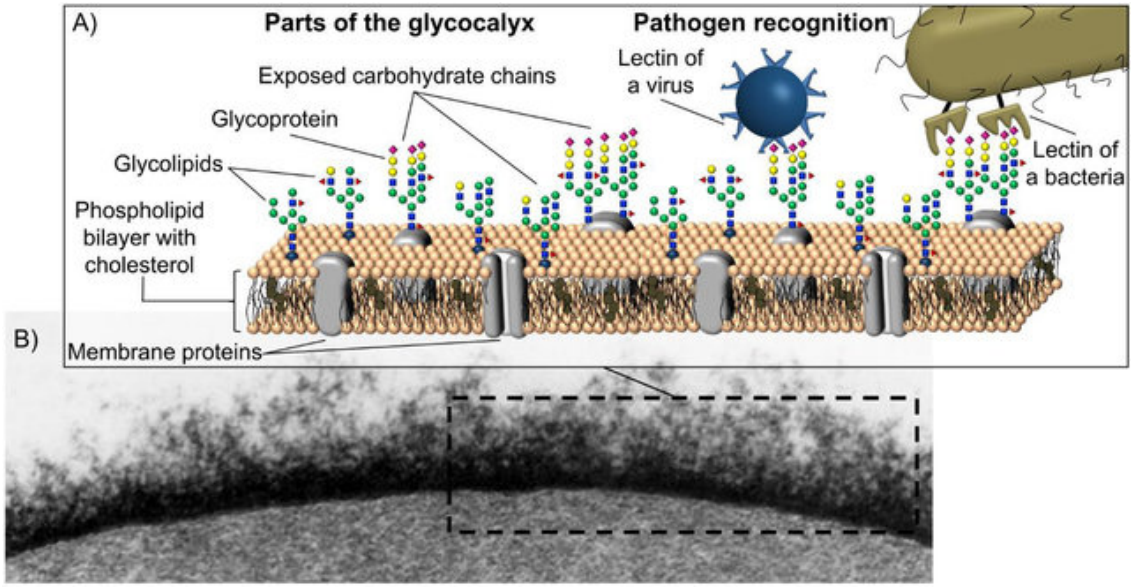
孙世伟

中国科学院计算技术研究所

Aug.25 2021

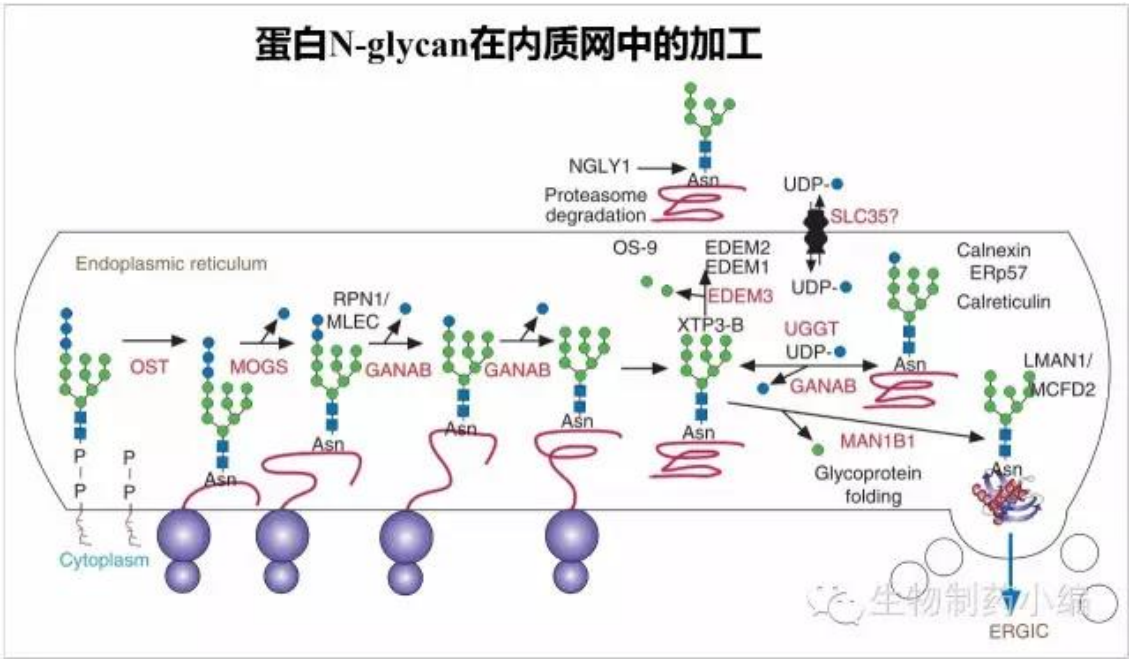
糖组学：全面了解生命基础的重要一环

糖基化：最复杂的翻译后修饰



- 细胞表面覆盖一层糖被，超过半数的蛋白质存在糖基化修饰现象。
- 可溶性糖蛋白：血浆蛋白除白蛋白外都是糖蛋白，酶、肽类激素、抗体、补体、生长因子、干扰素等。
- 膜结合糖蛋白：酶、受体、凝集素及运载蛋白
- 结构糖蛋白：纤粘连蛋白、层粘连蛋白等

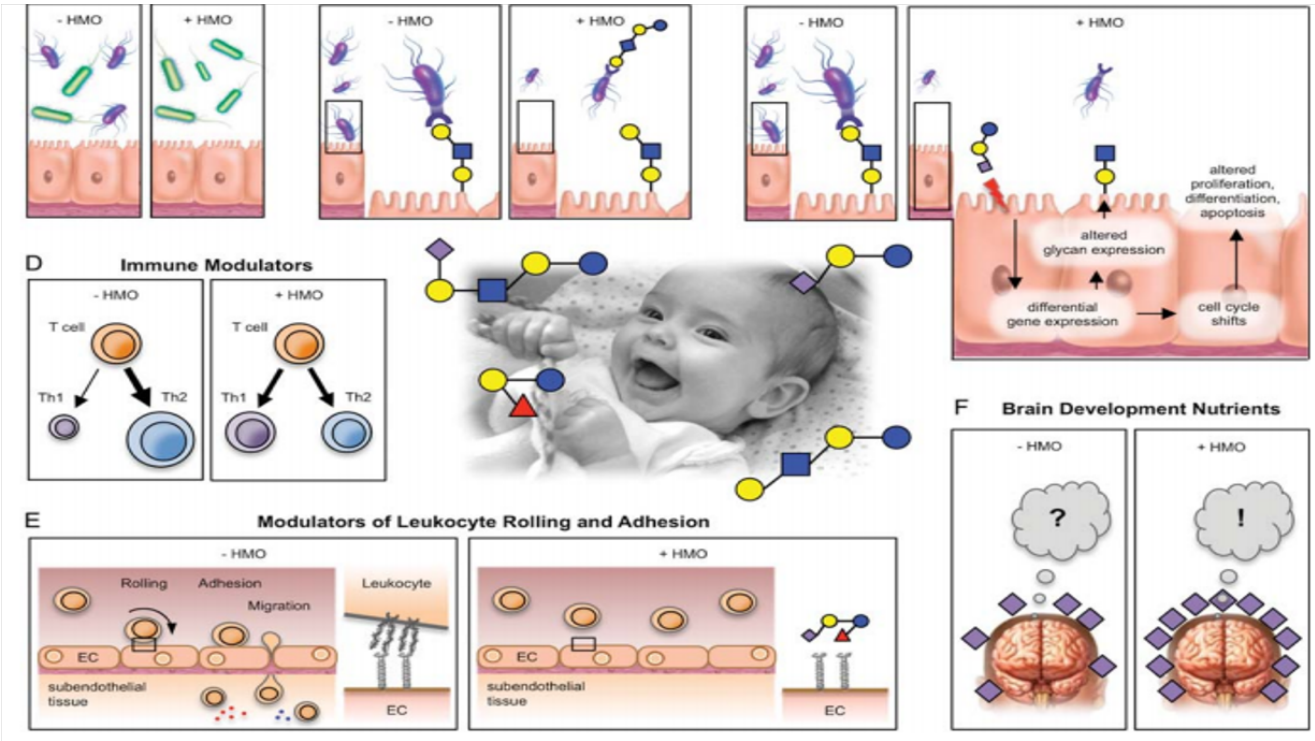
糖的作用



- 真核生物中，一半以上的蛋白质发生糖基化；
- 蛋白质在糖基化后才能正确折叠并具有生物学活性

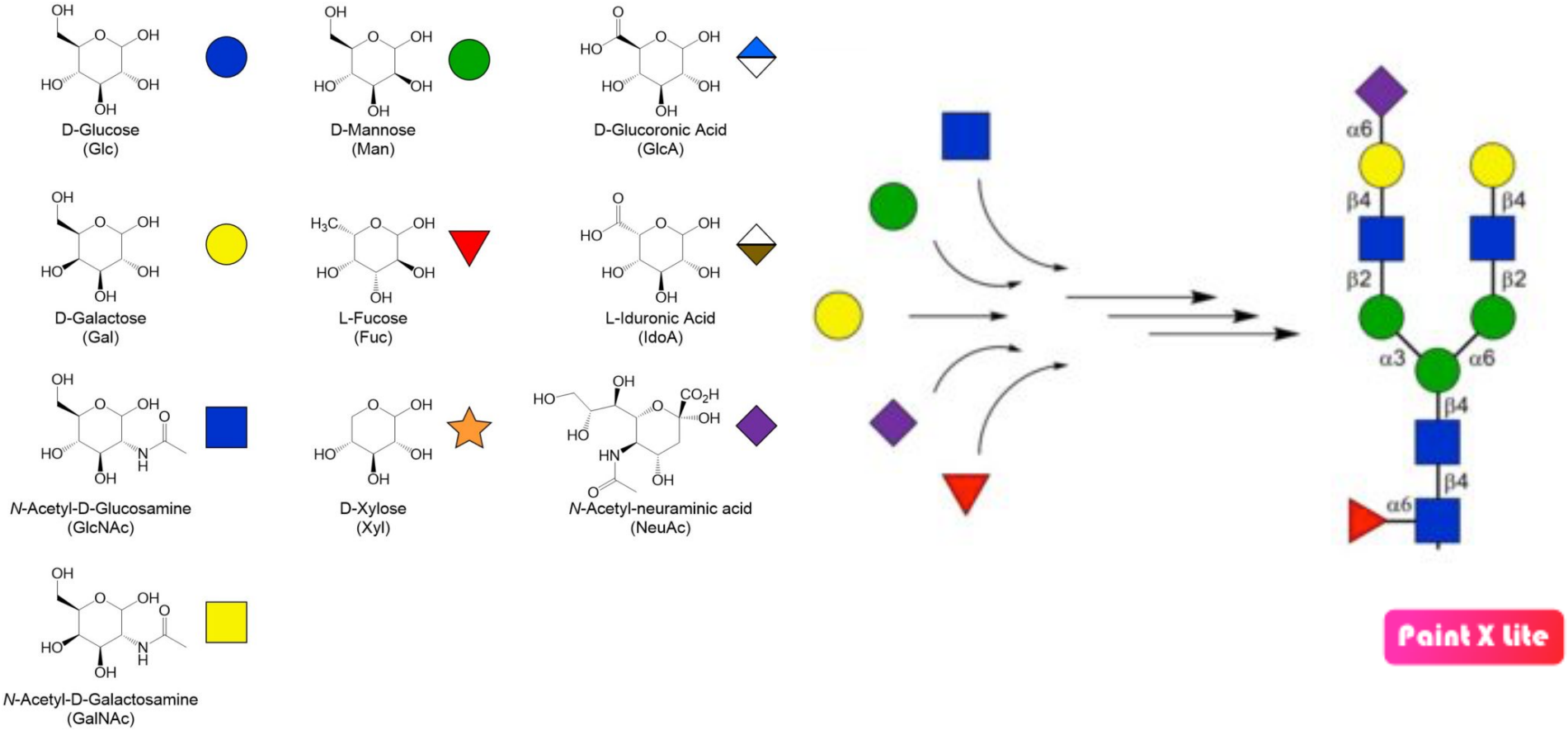
1. Moremen KW, Tiemeyer M, Nairn AV. Nat Rev Mol Cell Biol. 2012 Jun 22 ;13(7) :448-62. doi : 10.1038/nrm3383.

母乳寡糖的分离分析和功能研究



- 对感染的防御功能
- 对肠道微生态平衡的维护功能
- 对免疫系统的调节功能

糖的基本组成及常见构型



结构更复杂：树形结构，糖链具有多样性和复杂性

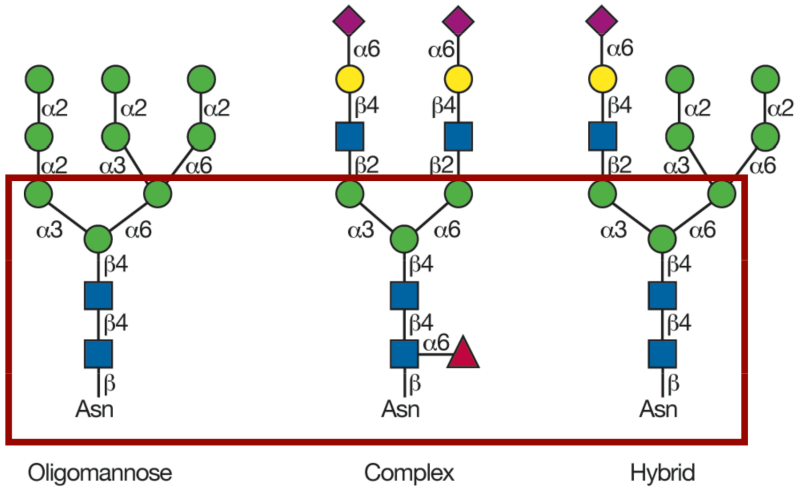
基因组学、蛋白组学和糖组学

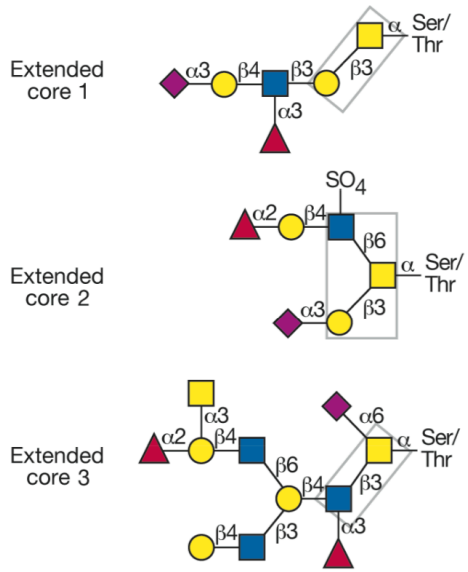
Nucleic Acid: 5′ T-C-A-G-G-A-T-C-C-A-A-C-G-C 3′

Protein: Ser-Asn-Pro-Leu-Cys-Lys-Gln-Thr-Gly-Lys-Arg

N-terminal C-terminal

Glycan

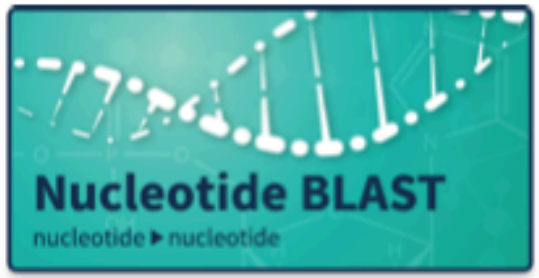




现有计算平台不足以支撑糖组学研究

基因组学

ATTTCTTAACTTGATG



蛋白组学

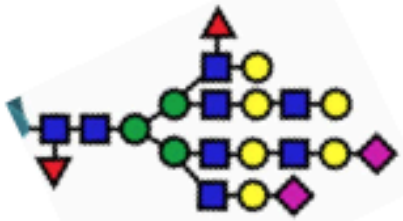
SEQUEST

MASCOT

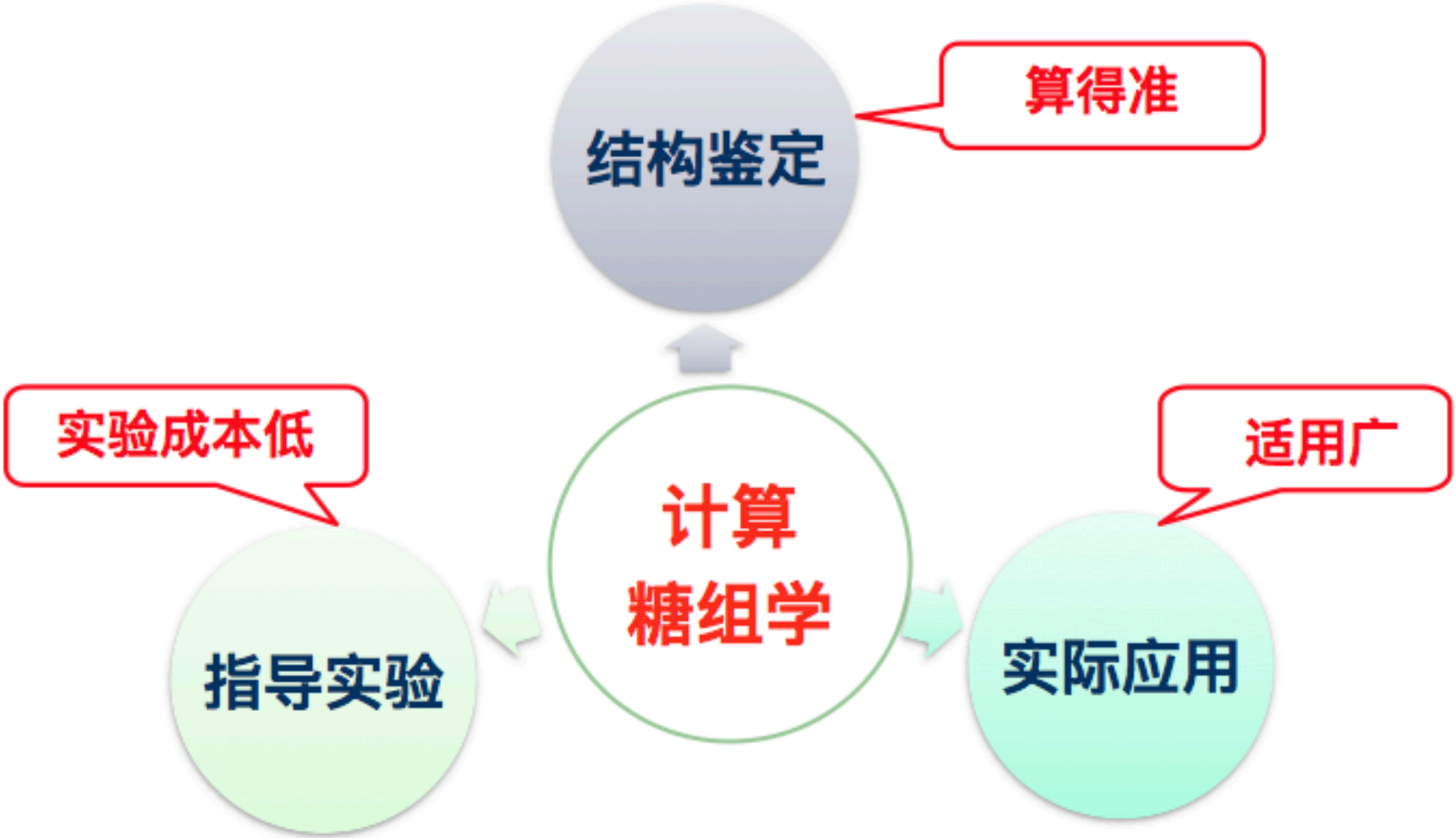
PEAKS

pFind

糖组学

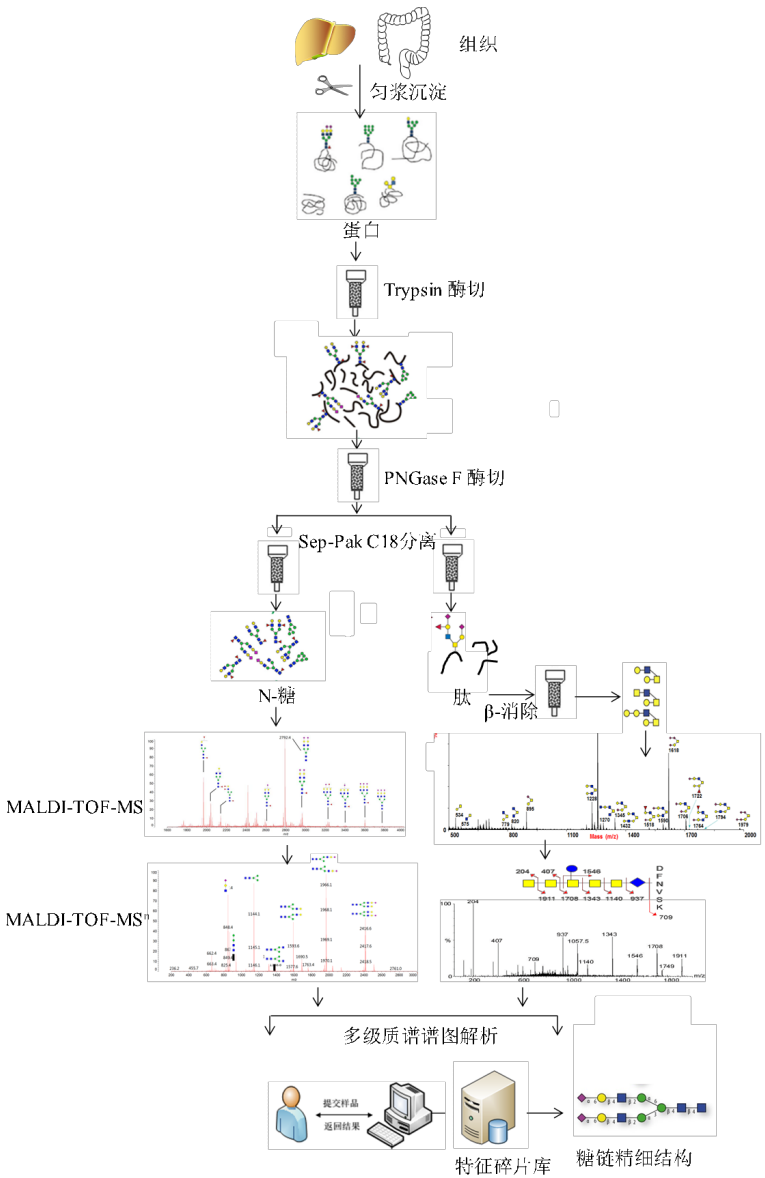


糖组学平台的宗旨

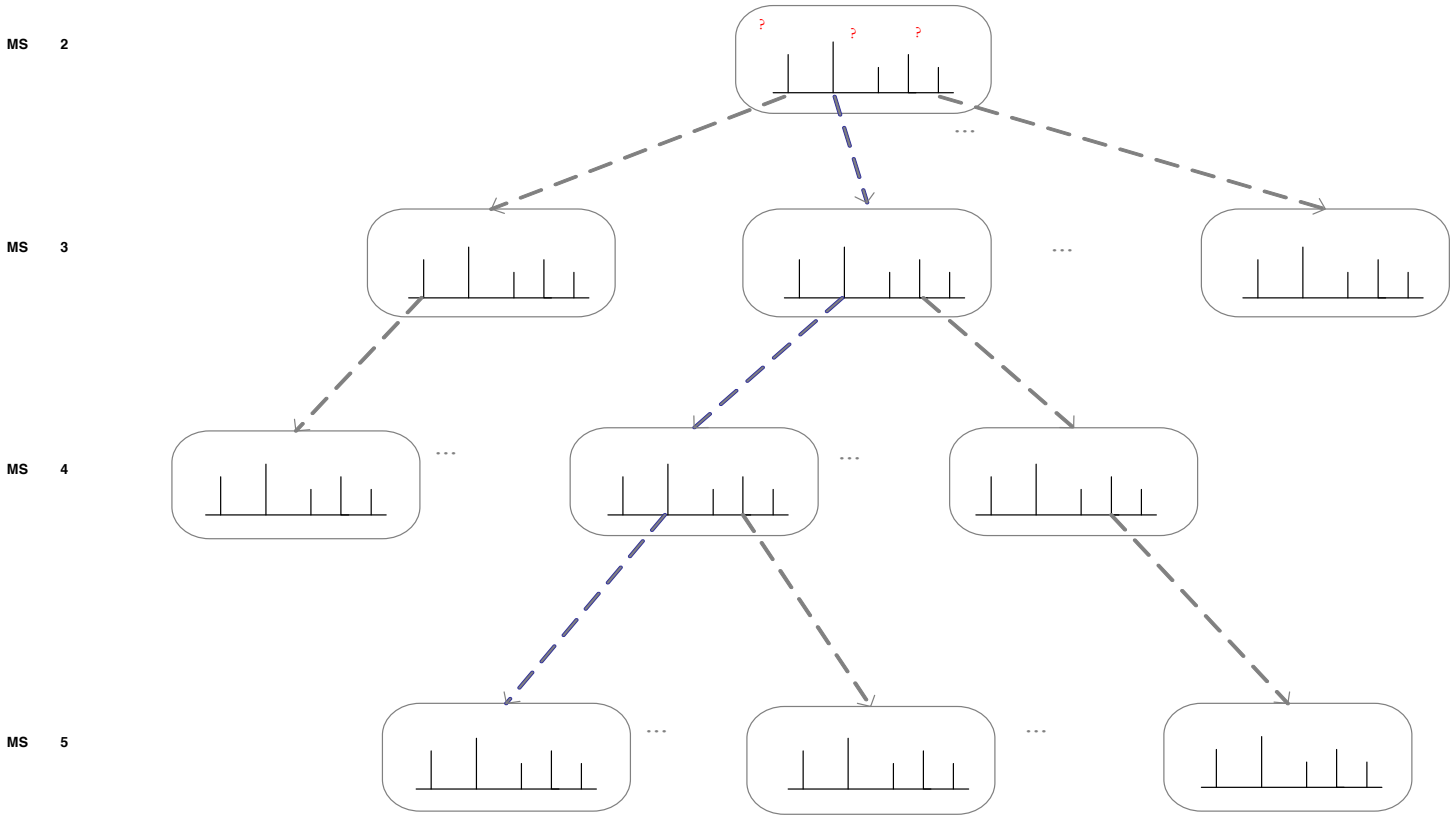


数据来源：质谱技术

多级质谱技术用于糖链分析

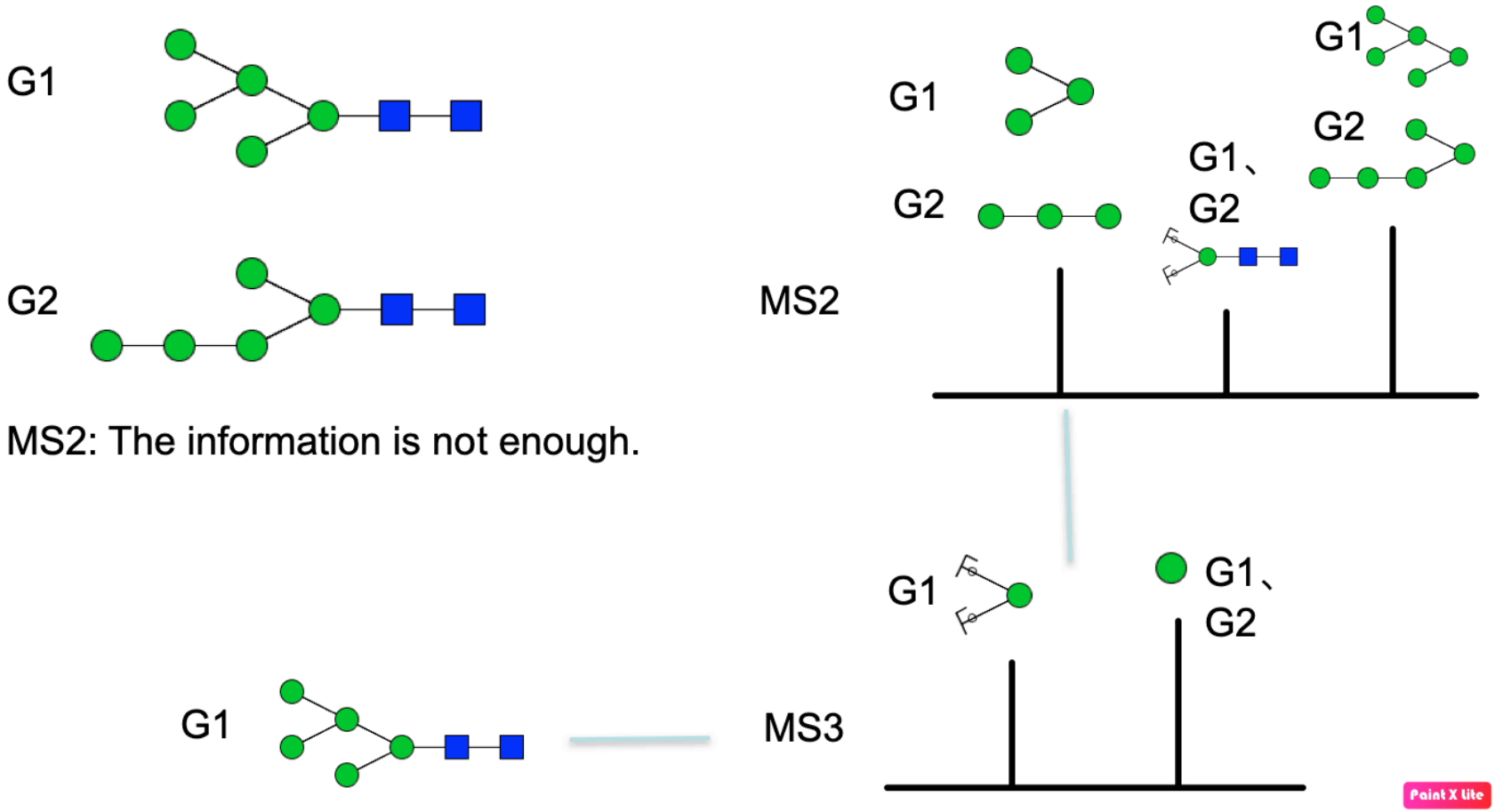


传统打谱方式



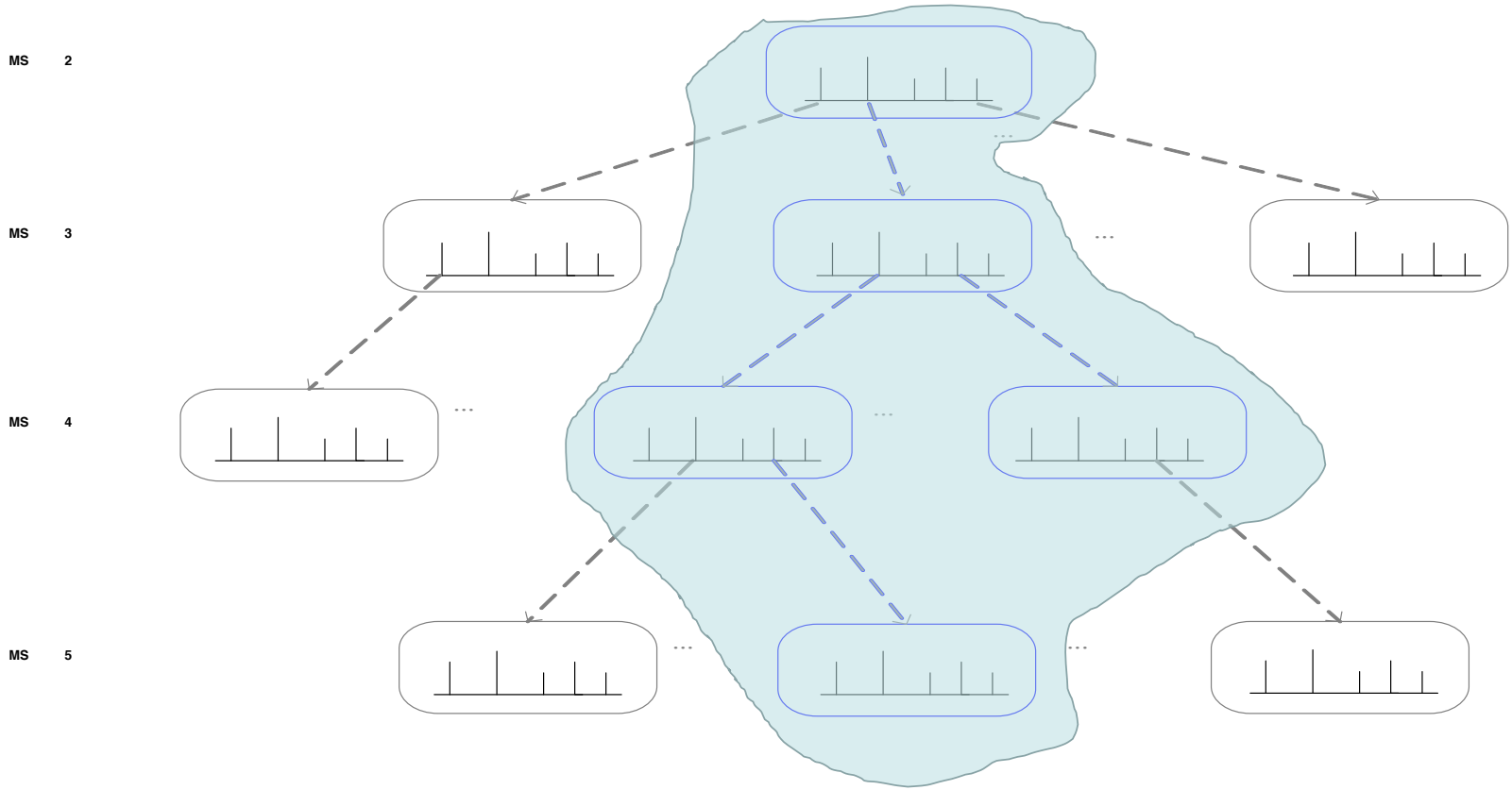
除一级质谱外，其他质谱都选择强度前五名的峰作为母离子产生下一级质谱

质谱实例



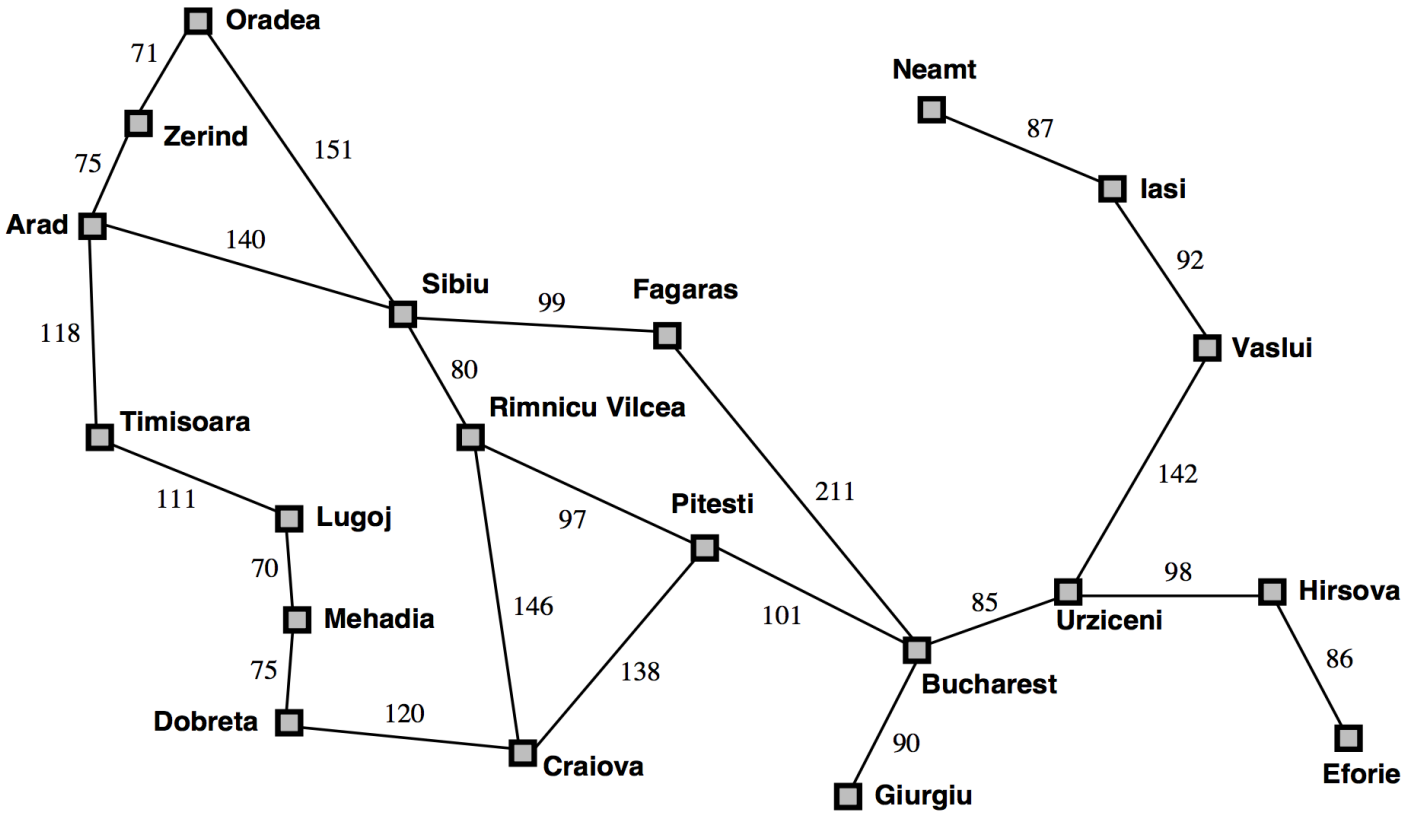
- 质谱图的横坐标是质荷比(m/z)= 糖碎片离子质量;
- 纵坐标是离子强度 = 糖碎片离子丰度;

我们的目标



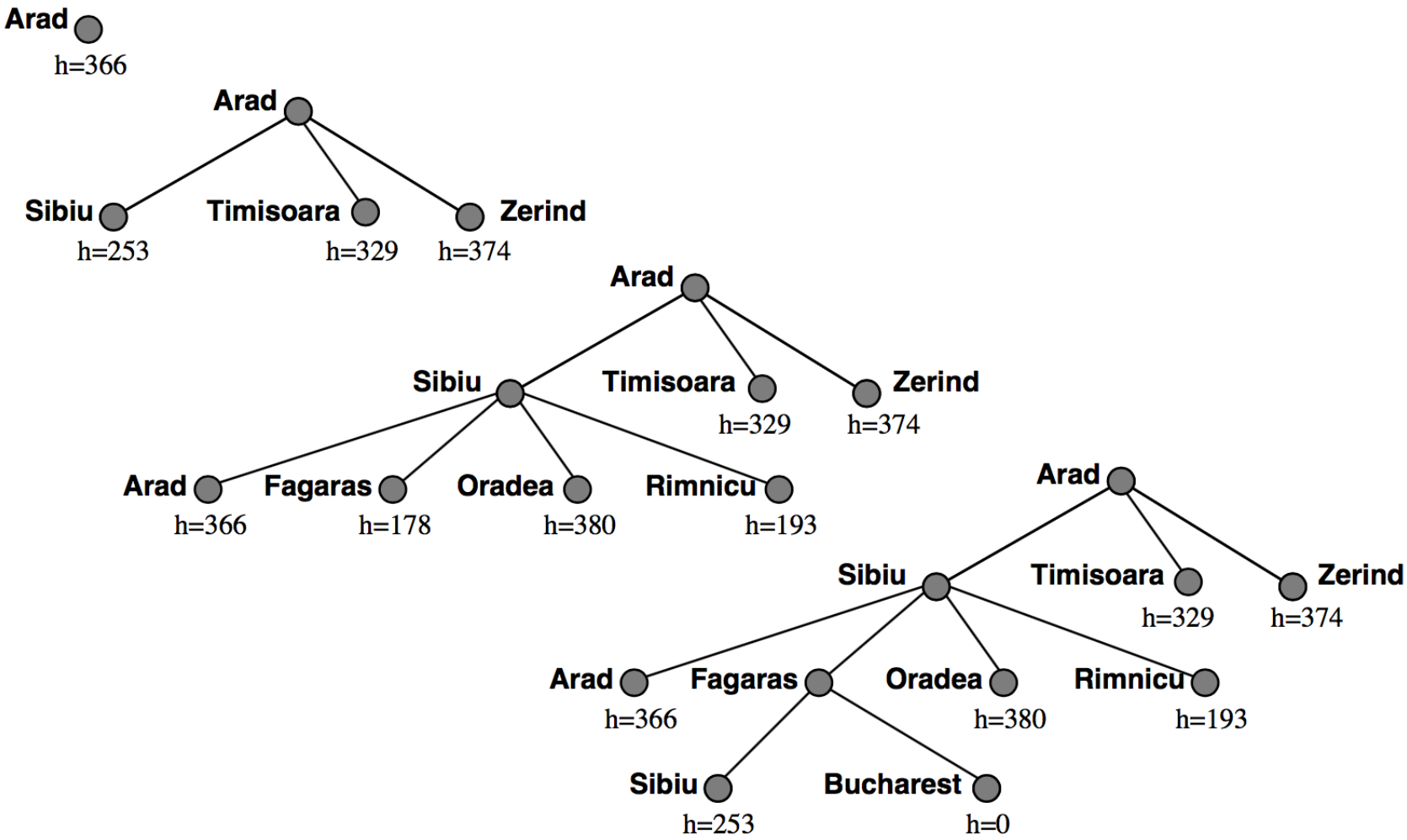
问题：如何指导实验，用最少量的打谱次数完成鉴定任务

Best First Search



Straight–line distance to Bucharest	
Arad	366
Bucharest	0
Craiova	160
Dobreta	242
Eforie	161
Fagaras	178
Giurgiu	77
Hirsova	151
Iasi	226
Lugoj	244
Mehadia	241
Neamt	234
Oradea	380
Pitesti	98
Rimnicu Vilcea	193
Sibiu	253
Timisoara	329
Urziceni	80
Vaslui	199
Zerind	374

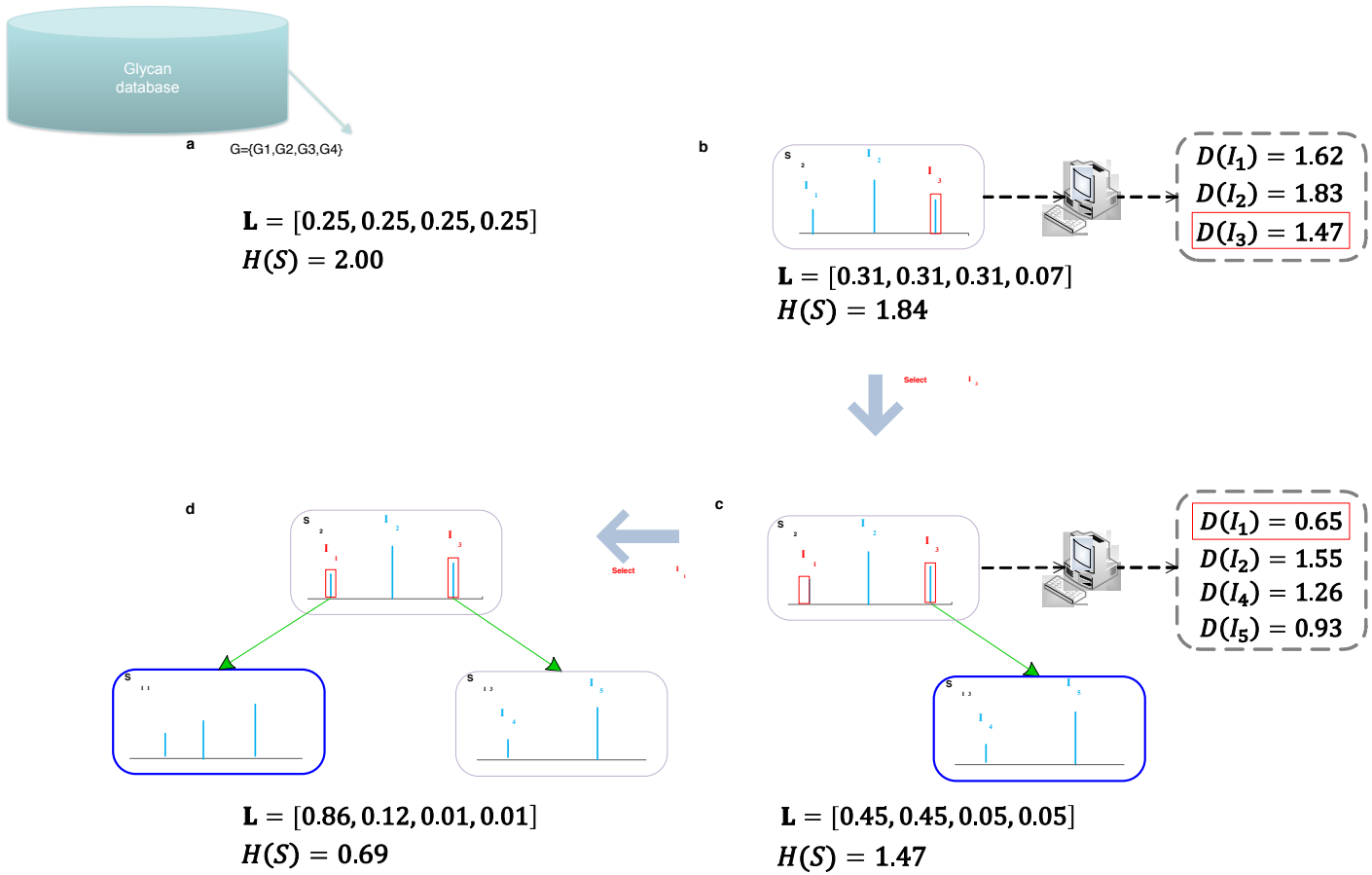
Best First Search



两个关键定义

- 目标点是什么:
  一个候选糖的概率是 1，其余都是 0，即熵是 0;
- 如何衡量每个峰到目标点的距离:
  信息增益（期望增益）

峰到达终点状态的距离：系统的熵



- 峰产生谱，更新每个候选糖的概率，系统的熵值即为该峰距离最终状态的距离；
- 峰的生成谱未知，求熵值的期望值

多糖鉴定流程

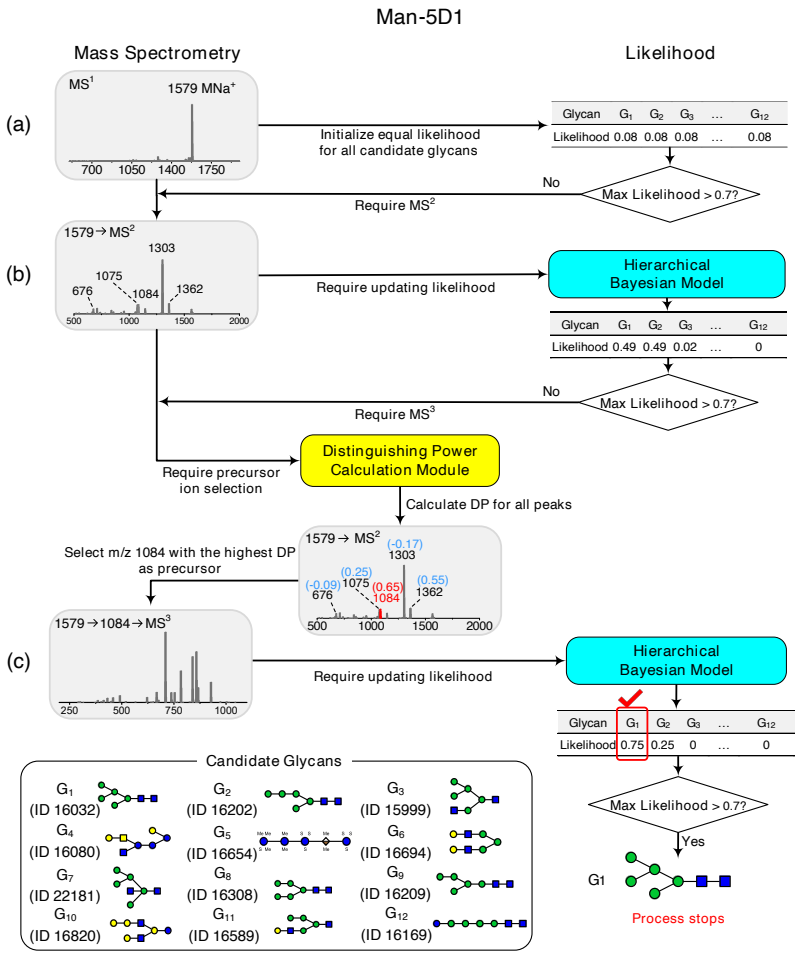
打分公式

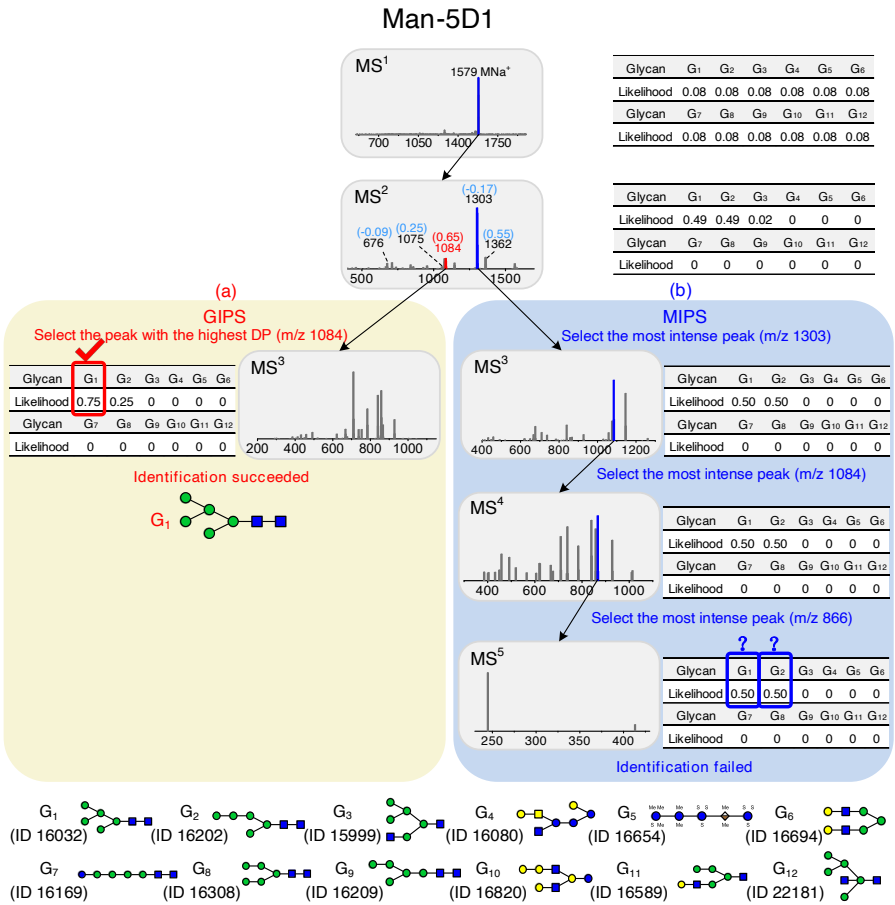
$$f(G_i, S_1, \cdots, S_m) = \sum_{k=1}^{m} \sum_{l=1}^{|S_k|} I(G_i, S_k^{(l)}).$$

正则化方法

$$p(G_i|S_1, \cdots, S_m) = \frac{e^{f(G_i, S_1, \cdots, S_m)}}{\sum_{j=1}^{n} e^{f(G_j, S_1, \cdots, S_m)}}.$$

GIPS 系统鉴定流程





1. Shiwei Sun, Chuncui Huang, etc. Analytical Chemistry, 2018
2. Yaojun Wagn, Dongbo Bu, etc. Bioinformatics, 2019

GIPS 方法的性能测试

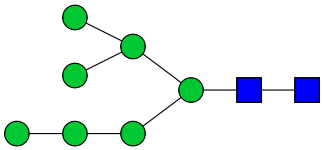
Glycans	MNa^+ (m/z)	Number of Candidates	Probability of actual glycans			
			MS^2	MS^3	MS^4	MS^5
A2	2792	2	0.90			
Hybrid-Octa	1824	9	0.43	0.90		
NGA3	1906	7	0.33	0.92		
NGA4	2152	5	0.23	0.73		
Man-5D1	1579	12	0.49	0.75		
Man-6	1783	12	0.33	0.69	0.75	
Man-7D3	1987	9	0.33	0.59	0.93	
Man-7D1	1987	9	0.23	0.96		
LNFPI	1100	6	0.79			
LNFPII	1100	6	0.59	0.99		
B-Hexa-T2	1304	2	0.99			
Globo-H-Hexa	1304	2	0.99			
pLNH	1375	8	0.64	0.96		
LNH	1375	8	0.27	0.90		

GIPS 与最高峰选择法（MIPS）结果比较：

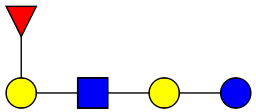
- GIPS 打谱次数少，样品需求量低一个数量级；
- MIPS 成功率：50%；
- GIPS 成功率：100%；

高相似度同分异构体区分

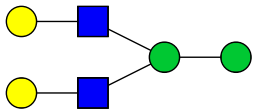
Other isomeric pairs



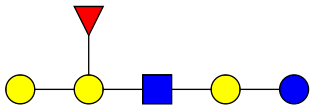
Man7D-1



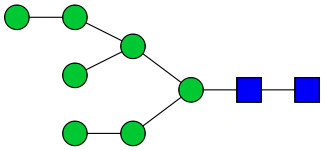
LNFP-I



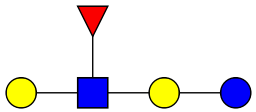
H-Hexa- T2



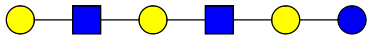
pLNH



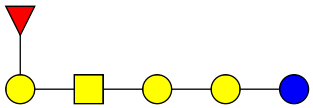
Man7D-3



LNFP-II



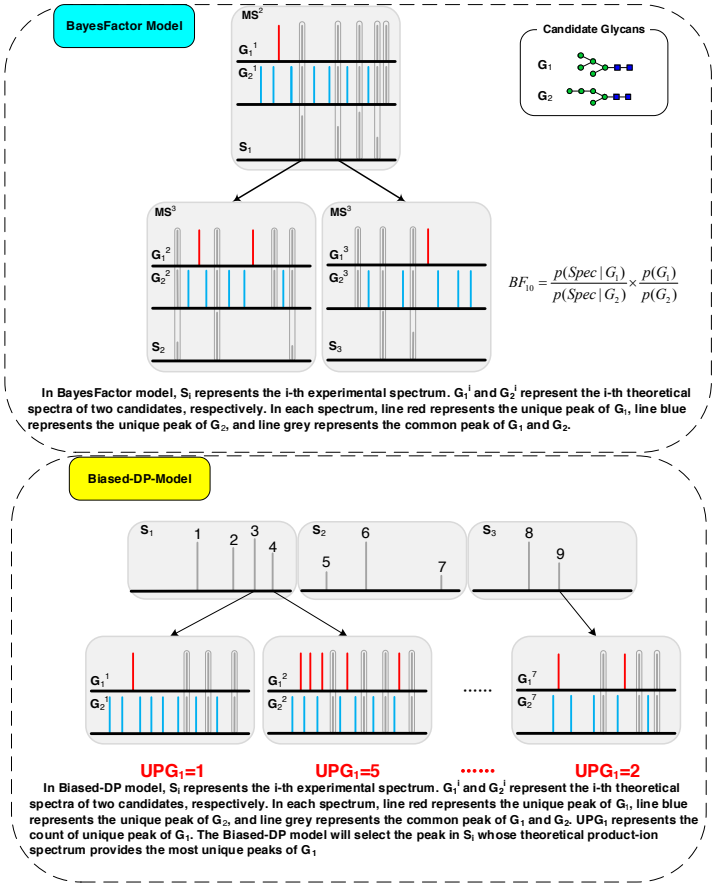
Globo-H-Hexa



LNH

were also assigned using this approach.

GIPS II：增强版的 GIPS



- 问题：候选糖中存在高度覆盖的问题；
- 解决思路：宏观之后转微观；

---

a. Chuncui Huang, etc.
Carbohydrate Polymers, 2019

GIPS II：增强版的 GIPS

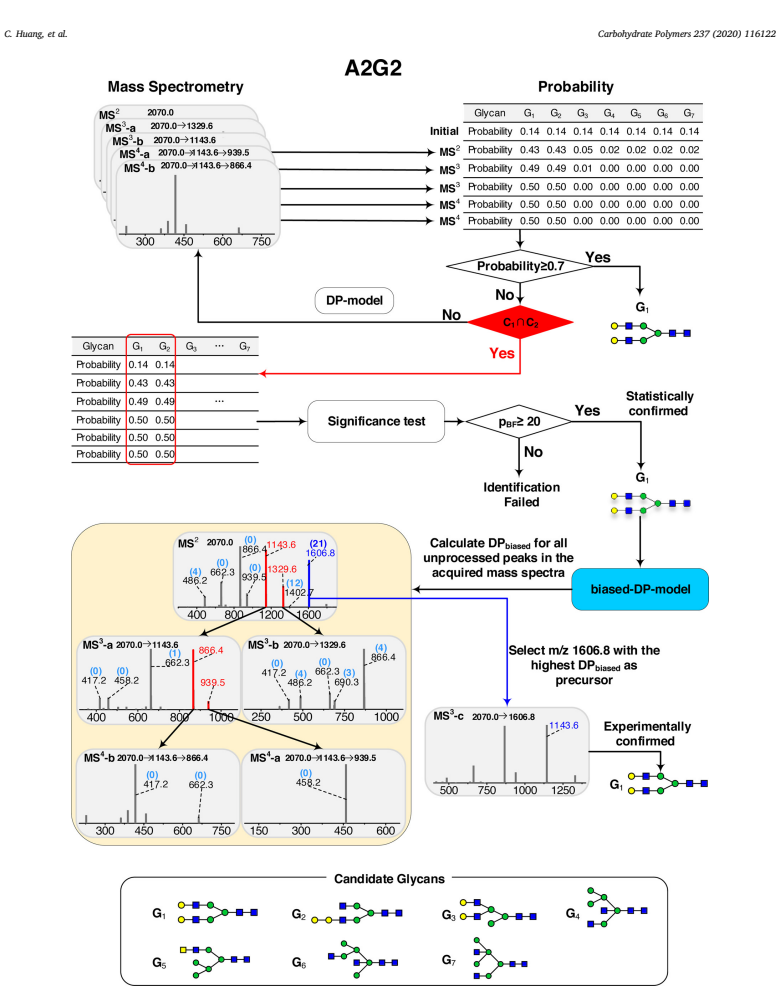
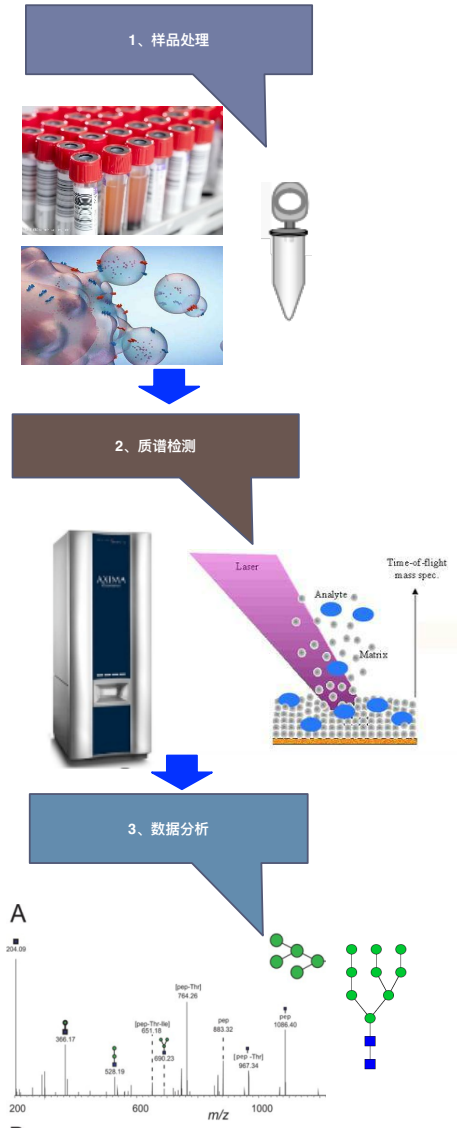
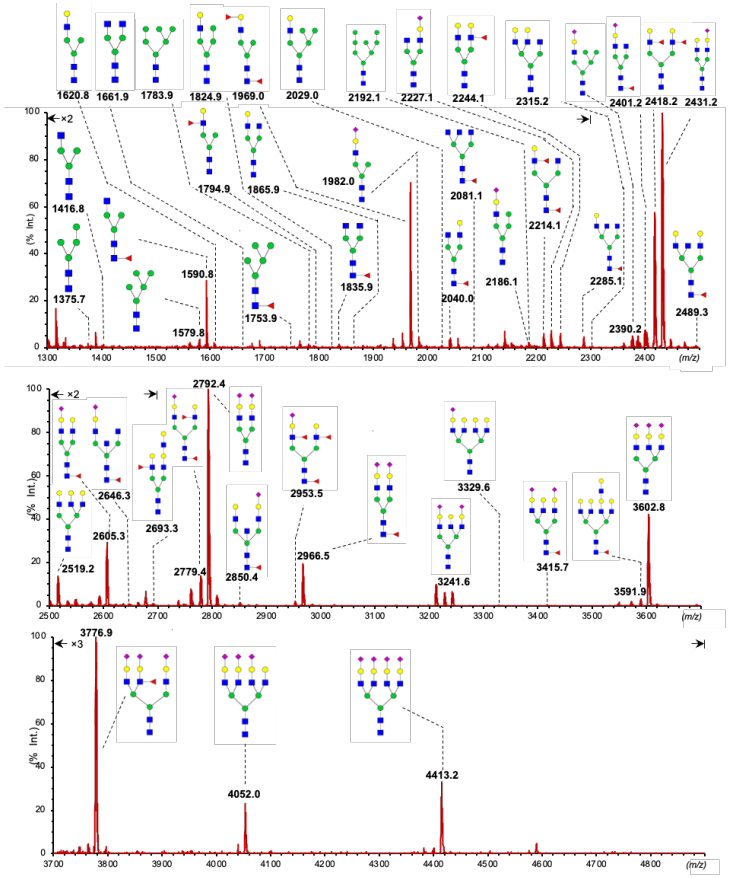


Fig. 2. Identification process of A2G2 using GIPS-II.
Seven candidate glycans G_1 to G_7 were extracted from the database, and candidate glycans G_1 and G_2 were calculated with the identical highest probability using DP-model. MS^3 was required using GIPS-II with the significant test and biased-DP-model to identify A2G2. The fragment ions in red were selected as precursors using DP-model, and fragment ions in blue were selected using biased-DP-model as the precursors for the next round of product-ion scanning. The calculated DP_{biased} values are in brackets above the m/z values in the spectrum. C_1 and C_2 represent two different conditions; C_1: the number of used spectra ≥ 5; C_2: the used spectra have no more information to distinguish between G_1 and G_2.

6



实战：血清样本



- 鉴定了 45 个高甘露糖型、复杂型和杂交型 N 糖(42 例有文献支持)
- DP 模型鉴定出 43 个；
- 2 个 N- 聚糖(m/z 2315.2 和 3602.8) 即使经过 6 轮 MS 扫描也无法确定(表 S2)。启动 bias-DP 模型进行统计和实验验证结构；

1. Hui Wang, etc. Journal of Proteomics, 2020
2. Chuncui Huang, Hui Wang etc. Carbohydrate Polymers,2020

混合物分析方法简介

假设

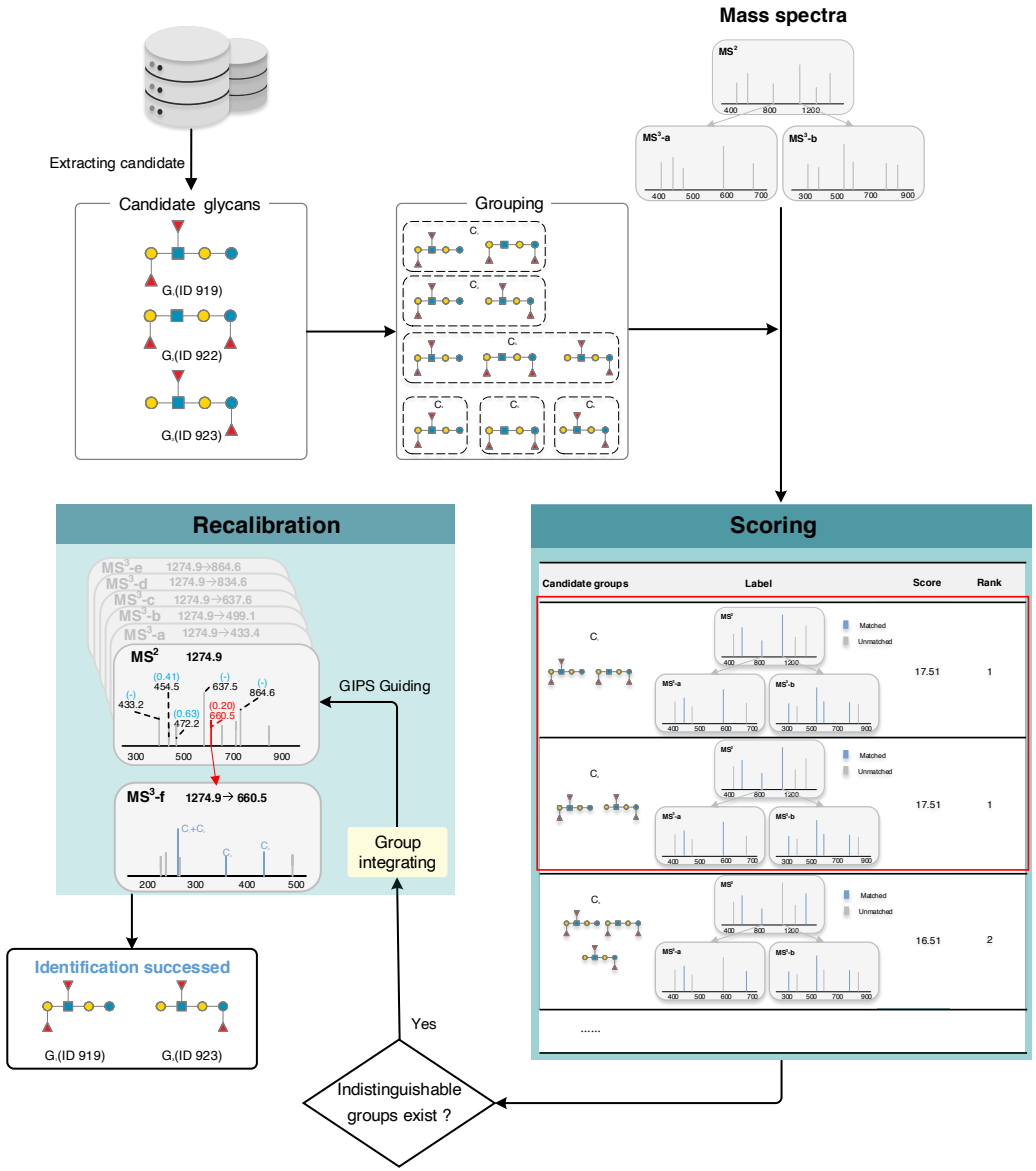
剃刀原理：如无必要，勿增实体

问题形式化描述

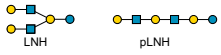
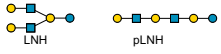
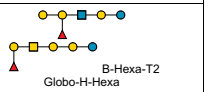
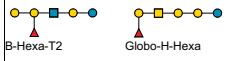
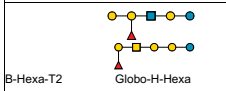
给定集合 U(实验谱）和由候选糖(峰的集合）组成的集合 S，寻找一个 S 的最小子集，使得能够覆盖集合 U。

- 输入：多级谱，候选糖的理论谱；
- 输出：候选糖集合的子集
- 约束：子集尽可能小，选出的子集对谱的解释尽可能好
- GIPS-II

混合物分析模型 :GIPS-mix

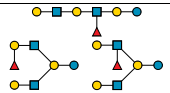
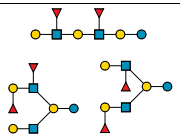
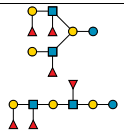


准确性验证：手工混合标准糖验证

Samples	Mixed Components	MNa^+	Ratio[1]	NoG[2]	NoB[3]	NoM[4]	Assignment
1	LNFP-I & LNFP-II	*m/z* 1100	1:1	13	3	7	
			1:3				
2	LNnFP-I & LNFP-II	*m/z* 1100	1:1	13	3	7	
3	LNH & pLNH	*m/z* 1375	1:1	8	2	3	
			1:3				
			3:1				
4	B-Hexa-T2 & Globo-H-Hexa	*m/z* 1304	1:1	11	4	15	
			1:3				
			3:1				
5	LNDFH-I & LNDFH-II	*m/z* 1274	1:1	5	3	7	
6	LNDFH-I & LNnDFH-II	*m/z* 1274	1:1	5	3	7	

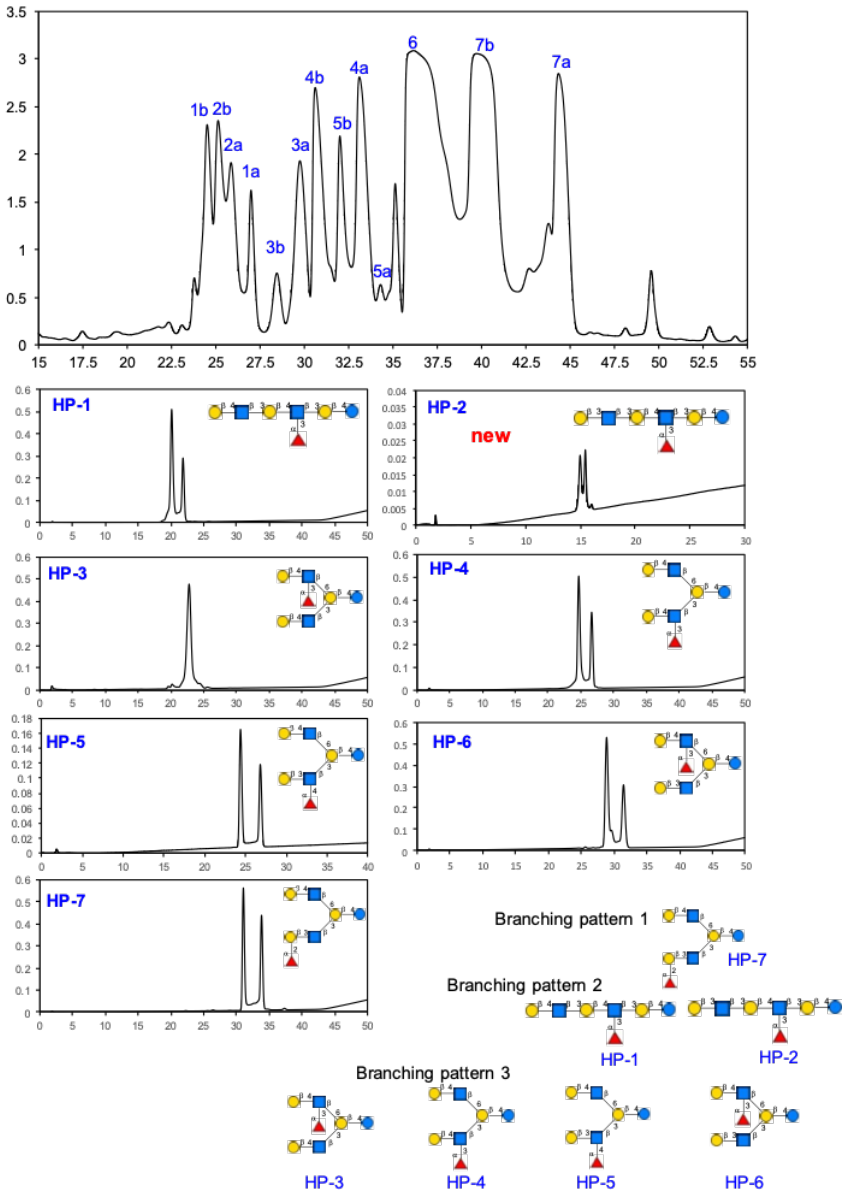
- 手工混合的不同比例的标准糖，成功鉴定；

准确性验证：母乳寡糖(DP4 ~ DP9)

Fractions	Selected MNa$^+$	NoG^1	NoB^2	NoM^3	Assignment by GIPS-Mix	References
DP4	*m/z* 926	3	1	1		30
DP5	*m/z* 1100	17	3	7		19, 30,31
DP6	*m/z* 1274	9	3	7		30, 31
	m/z 1375	10	2	3		30-33
DP7	*m/z* 1549	26	5	31		30,32-34
DP8	*m/z* 1723	23	9	511		30, 35, 36
DP9	*m/z* 1898	19	6	63		30, 37, 38

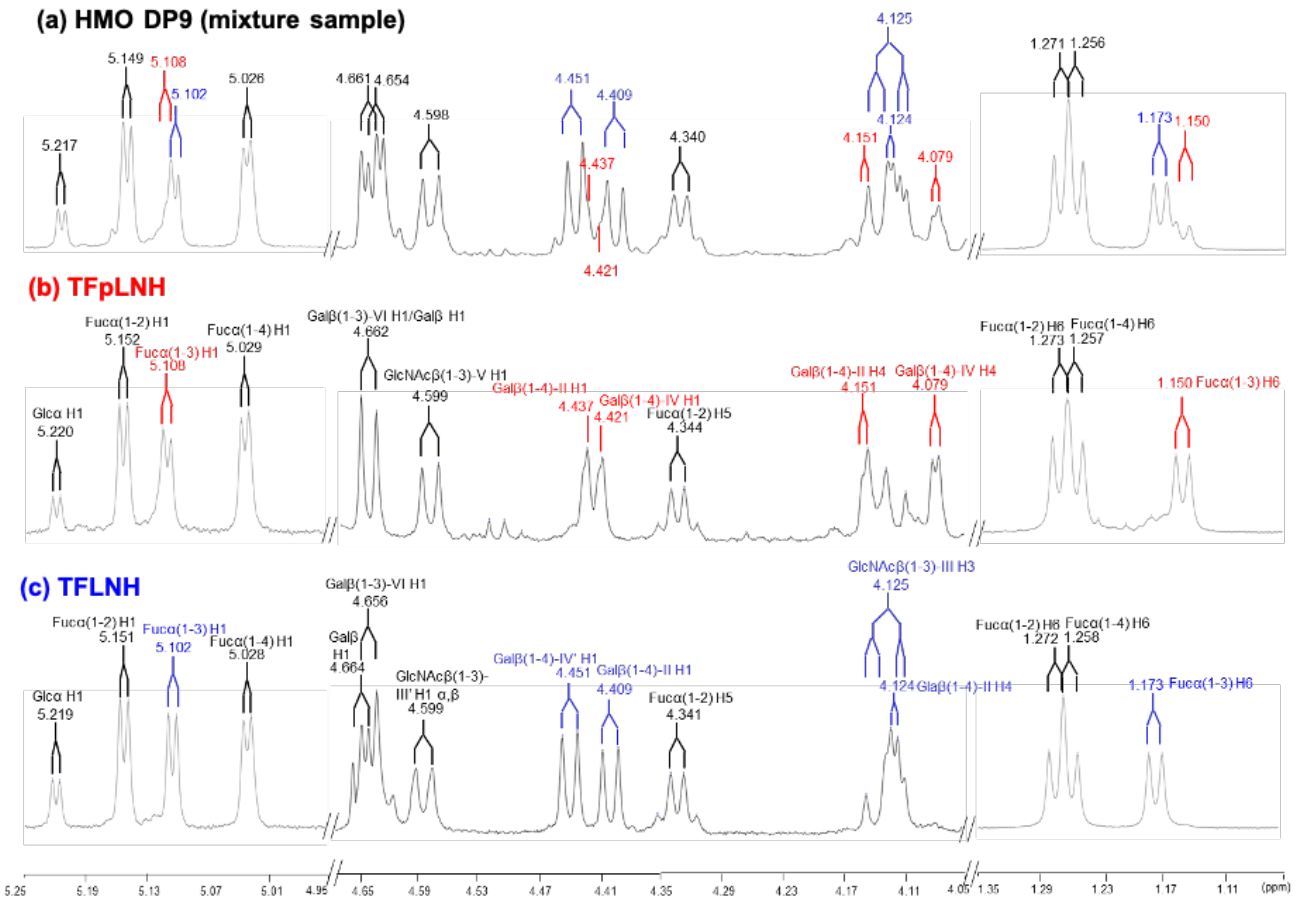
- HPLC 粗分的组分 DP4 ~ DP9, 鉴定的出的 HMO 结构，大部分都有文献报道；
- DP7 和 DP9 我们进行了实验验证：
- 对于组分 DP7, 我们通过 HPLC 分离将组分接出，用 ESI-CID-MS/MS 负离子模式进行验证；
- 对于组分 DP9，我们用 NMR 进行了实验验证；

DP7 结果准确性验证：ESI-CID-MS/MS



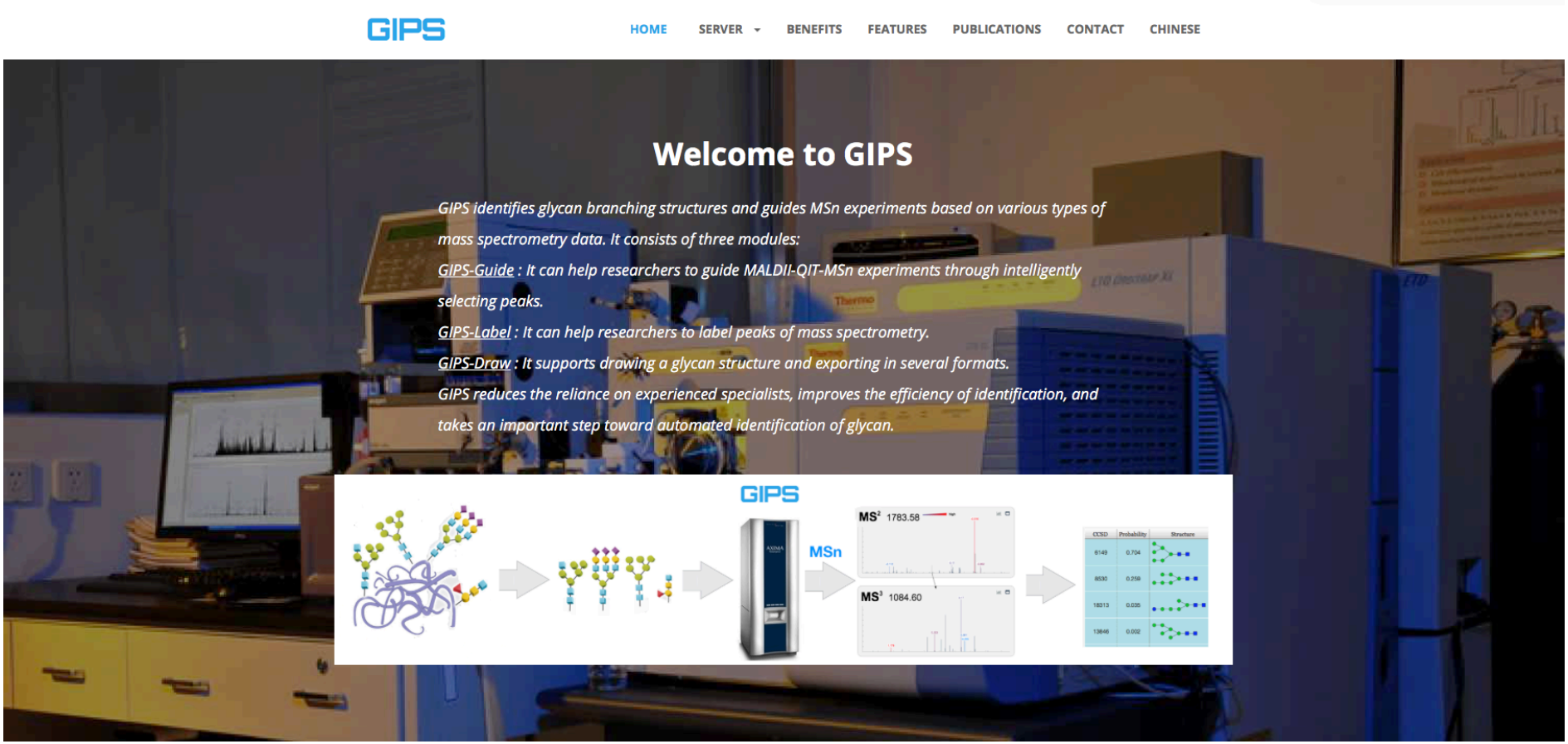
- 将 DP7 组分进一步细分，将各个组分接出
- 用标准化的 ESI-CID-MS/MS 的方法进行分析；
- 得出 7 个糖结构，这 7 个结构分别归属于 3 种分支结构，与我们鉴定的分支结构一致。

DP9 结果准确性验证：NMR



- DP9 的 NMR 谱图与标准品的 NMR 图谱
- 比较发现：DP9 混合组分中都能找到标准品的特征峰，而且为主要组分；

我们的网站：http ://glycan.ict.ac.cn



致谢

谢谢大家

孙世伟 dwsun@ict.ac.cn