Glycan Structure De Novo Sequencing with Tandem Mass Spectrometry

Kaizhong Zhang

Joint work with Gilles Lajoie, Bin Ma, Baozhen Shan

Department of Computer Science

Department of Biochemistry

University of Western Ontario

University of Waterloo

Outline

- Glycosylation
- Glycan structure determination
- Two similar mathematical models
    - Their algorithms and complexities
- Experiments and Results

Post-translational modifications

- Some amino acids are modified after the protein is synthesized.
- In most cases these PTM are important to the proteins’ functions
- Most PTM can be simply regarded as a new amino acid for bioinformaticians’ point of view.
  - Oxidation M’ = M+16
- Glycosylation is one exception

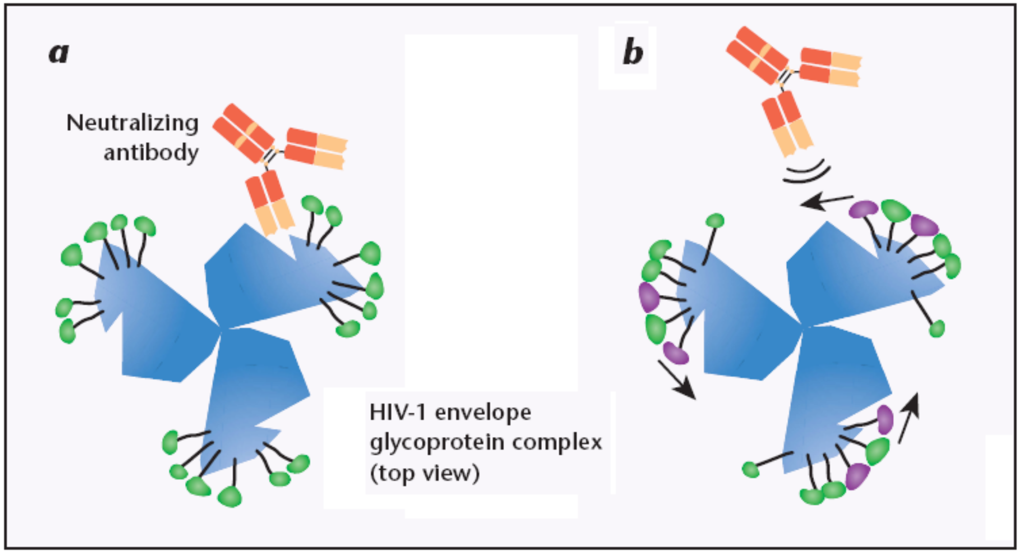
Glycosylation is an important PTM

- In humans more than half of the proteins are believed to be glycosylated.
- The glycan portions have been associated with a wide range of biological functions
  - such as protein folding, solubility, protein localization and trafficking, protection against enzyme degradation, antigenicity and cell-cell recognition.
- Alteration in glycosylation is known to be involved in long list of diseases
  - such as carcinoma of the mammary gland, lung, colon and pancreas, rheumatoid arthritis, Gaucher and Tay Sachs disease

An Example of Protein Glycosylation

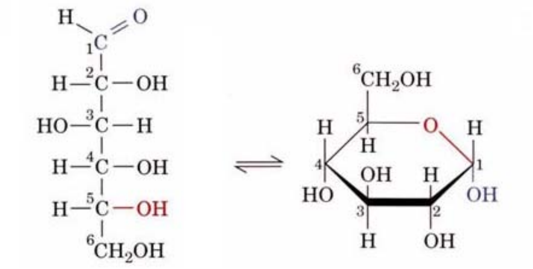
- Structural variation in glycans

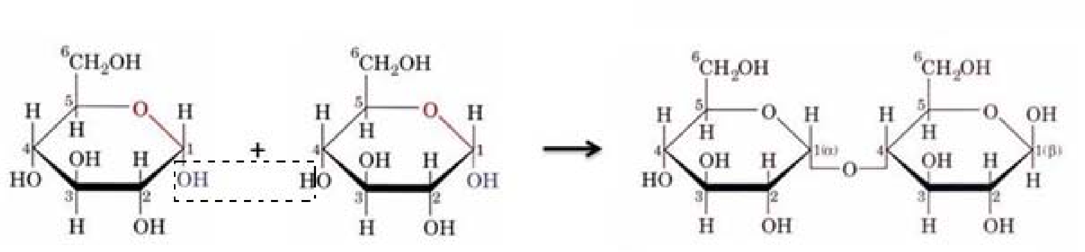
HIV-1: nature's master of disguise *Nature* **422**, 307-312, 2003

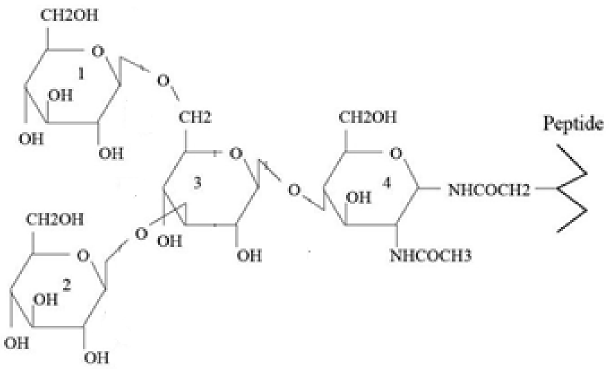


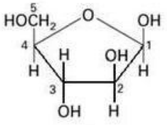
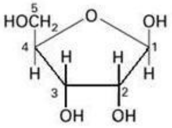
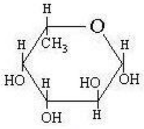
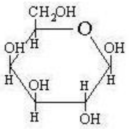
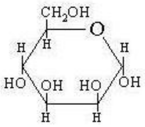
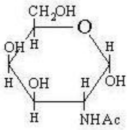
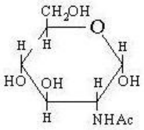
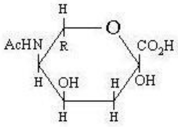
Simple sugars and glycopeptides

$C_n(H_2O)_n$



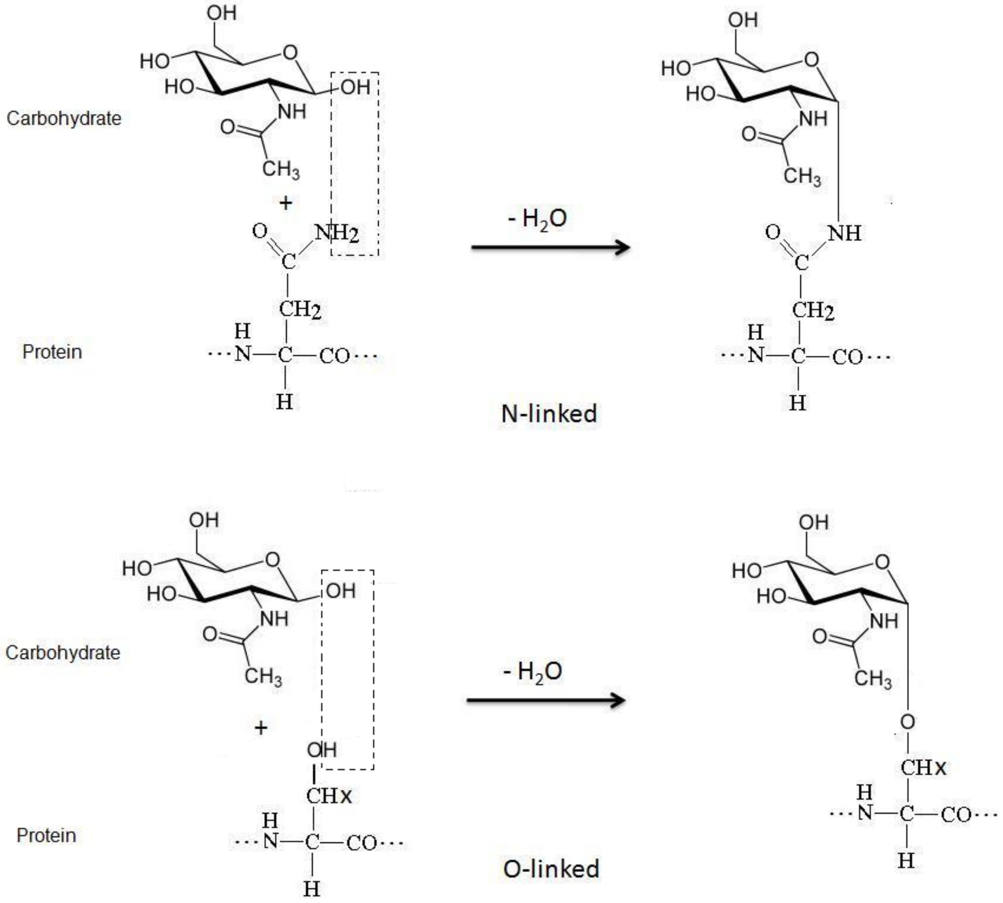




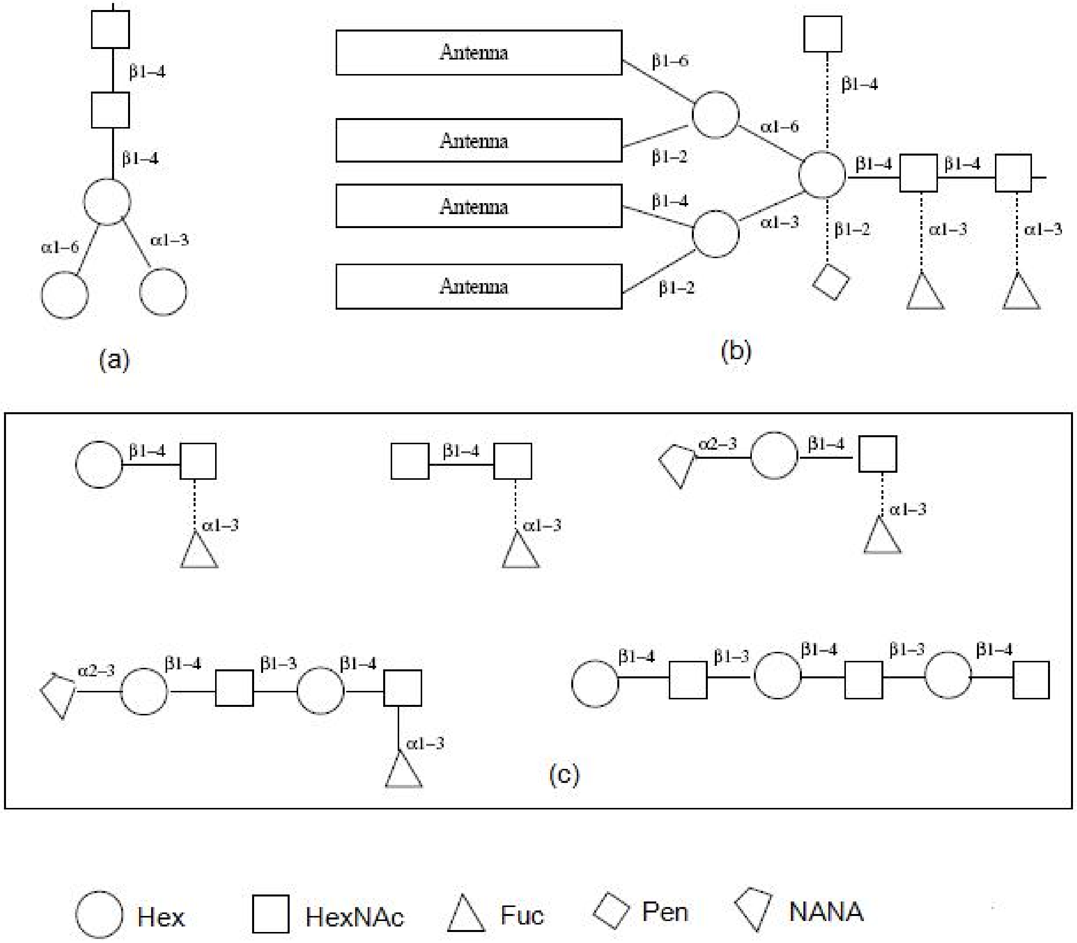
Monosaccharide	Composition	Generic/Abbreviation	Monosaccharide Residue Mass (-$H_2$0)	
			Monoisotopic	Average
 Xylose  Ribose	$C_5O_5H_{10}$	Pentose / Pen	132.0427	132.1
 Fucose	$C_6O_5H_{12}$	Deoxyhexose / Fuc	146.0579	146.1
 Galactose  Mannose	$C_6O_6H_{12}$	Hexose / Hex	162.0528	162.1
 N-Acetylgalactosamine  N-Acetylglucosamine	$C_8O_6NH_{15}$	HexNAc	203.0794	203.2
 Sialic Acid	$C_{11}O_9NH_{19}$	NANA	291.0954	291.3

Glycoproteins

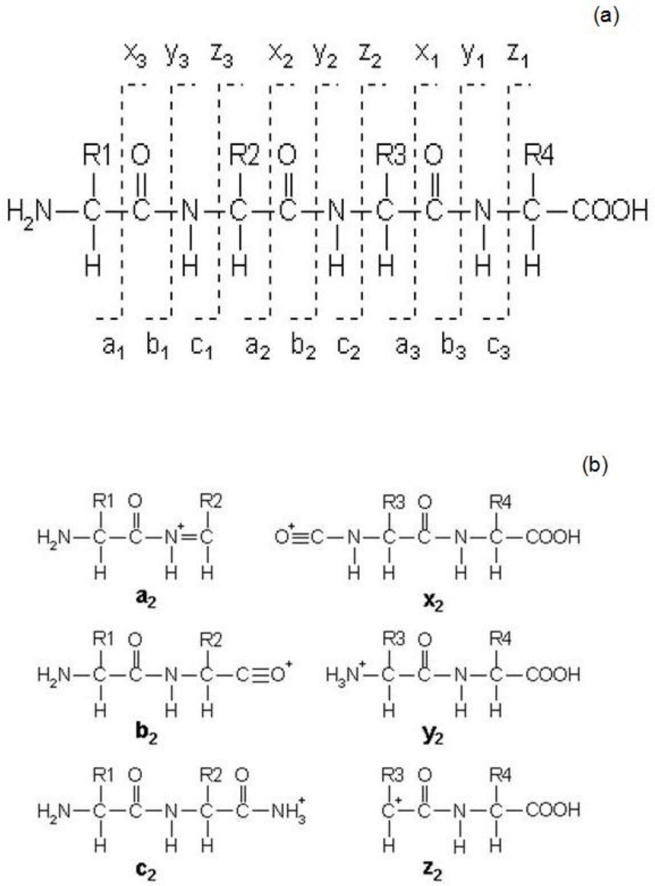
NXT/S X: except P

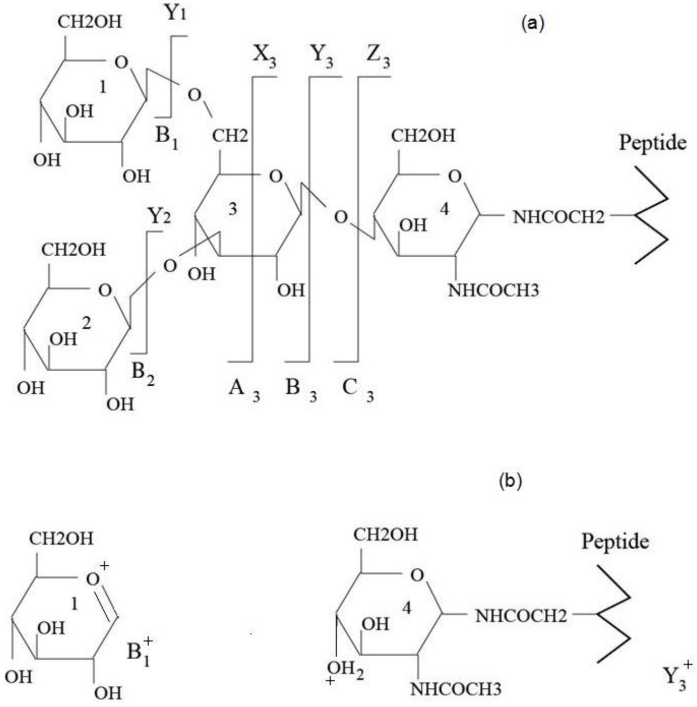


N-linked glycans



Glycopeptide Fragmentation


(a)
x3 y3 z3 x2 y2 z2 x1 y1 z1
R1 O R2 O R3 O R4
H2N—C—C—N—C—C—N—C—C—N—C—COOH
a1 b1 c1 a2 b2 c2 a3 b3 c3
(b)
a2
x2
b2
y2
c2
z2

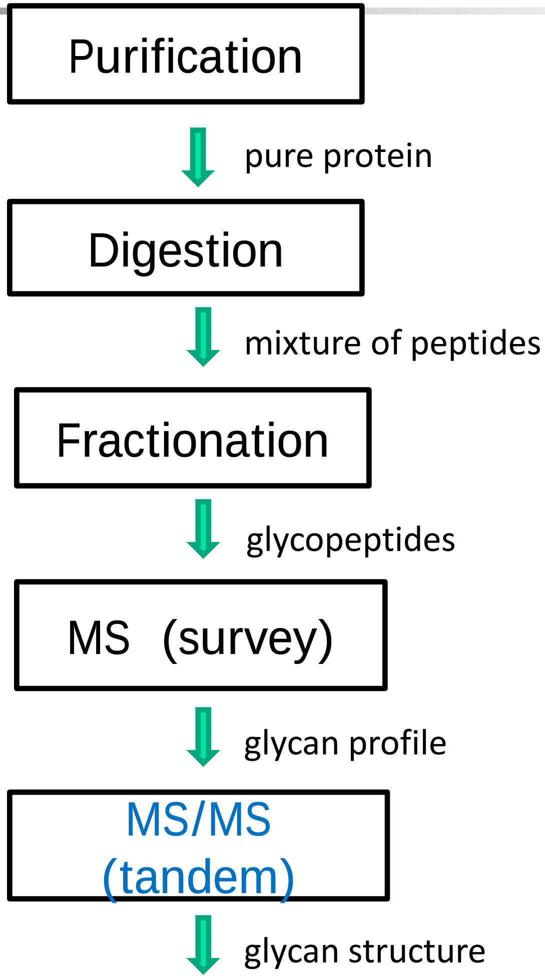


Glycopeptide fragmentation

- ETD tends to fragment peptides.
- CID or HCD tends to fragment glycans.
  - Y-ions are linked to the peptide.
  - One y-ion may associate with several b-ions.
  - Peaks for y-ions and b-ions are separated.

Glycoprotein Mass Spectrometry

- Work flow



Purification

pure protein

Digestion

mixture of peptides

Fractionation

glycopeptides

MS (survey)

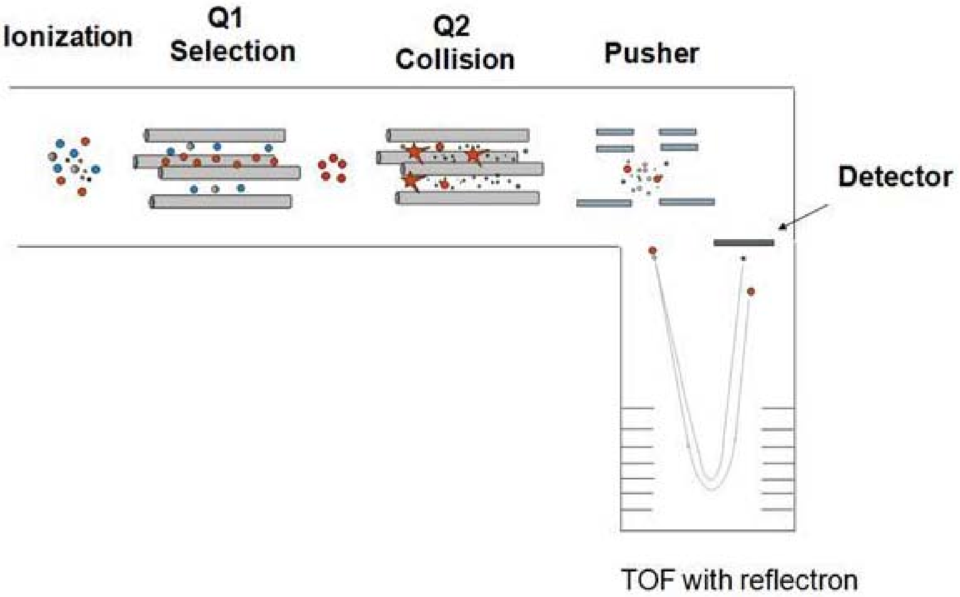
glycan profile

MS/MS (tandem)

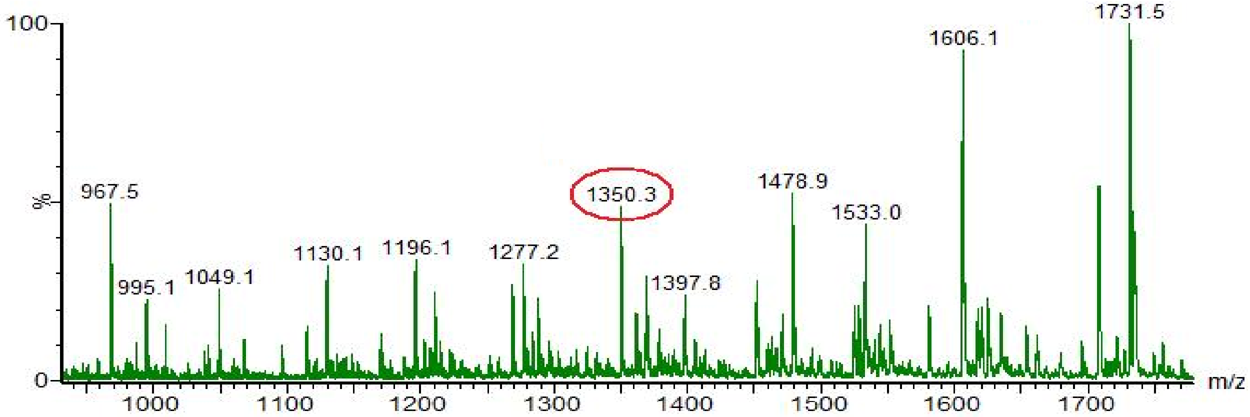
glycan structure

Tandem Mass spectrometry

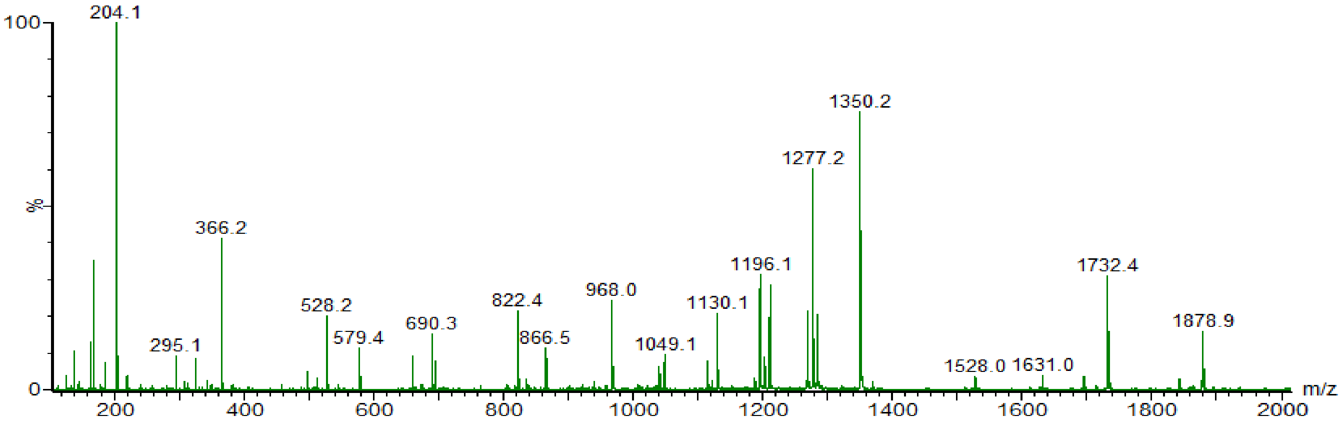
- Two stage of mass analysis Q-TOF2
  - First : select a precursor ion
  - Second: scan the product ions



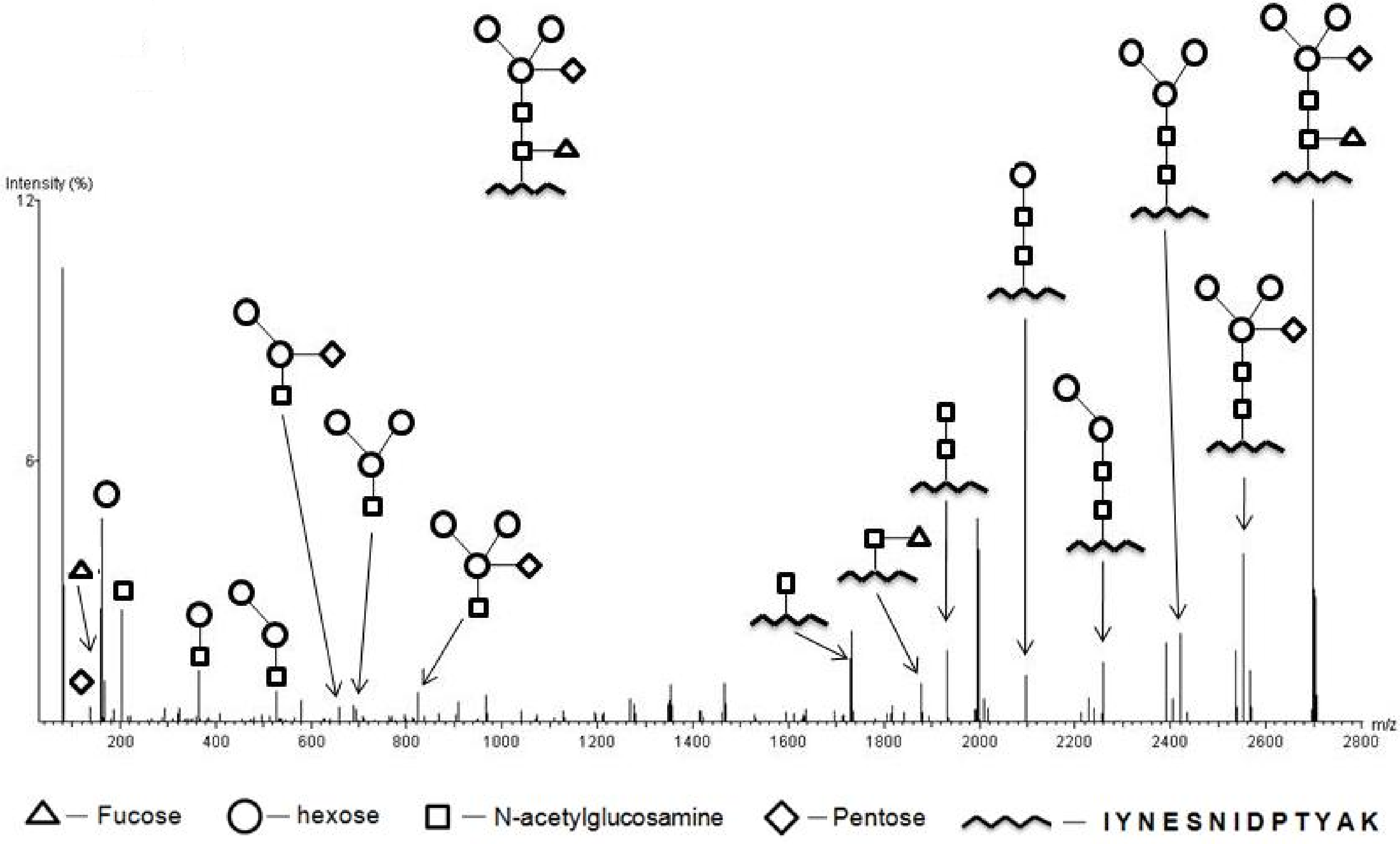
- Fractionation by RP-HPLC
  - GPa IYNESNIDPTYAK
  - GPb LHFHDCFVQGCDASVLLDDTSNFTGEK
  - GPc DSTTASLSSANSDLPAPFFNLSGLISAFSNK
- MS survey scan



- MS/MS tandem scan



Correspondence between the structure and the spectrum



Determine glycan composition

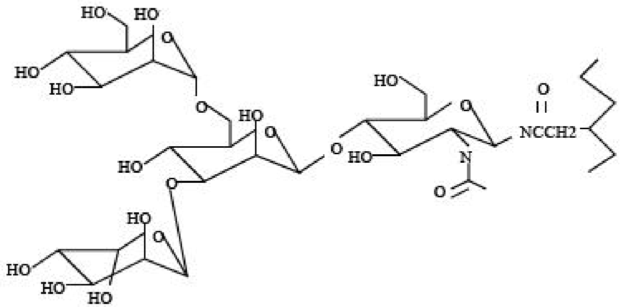
- Use y-ions
- Let m_p be the mass of the peptide
- Let M be the glycopeptide ion mass
- $DP(M)$ will be the score of the optimal path that corresponds to the composition

$$DP(m_p) = 0$$

$$DP(m) = f(m) + \max_{g \in \Sigma} f(m - m(g))$$

Problem Formulation

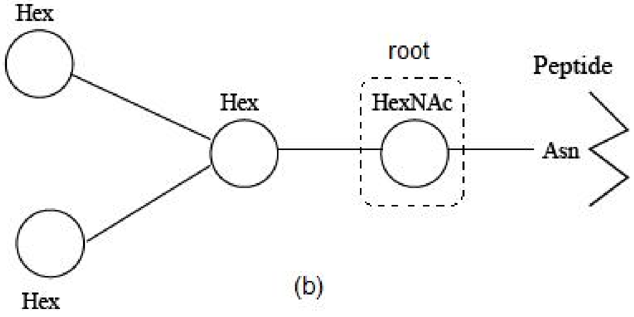
- Glycan tree representation



(a)

$$t = (g; t_1, t_2, t_3, t_4)$$

$$g \in \Sigma$$



(b)

$$T = (HexNAc;\ t_1)$$

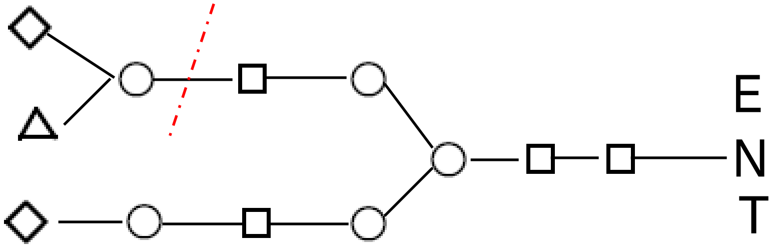
$$t_1 = (Hex; t_1^{'}, t_2^{'})$$

$$t_1^{'} = (Hex)$$

$$t_2^{'} = (Hex)$$

Glycan De novo Sequencing Problem

- Spectrum S defines $f(m)$
  - $f(m)$ high if the peak at/around m is high
  - $f(m)<=0$ if no peak at/around m
- A tree structure T defines mass set $ms(T)$
  - E.g. Any subtree T' defines two mass values $m(T')$ and $M-m(T')$.



Modeling *de novo* sequencing

- The score between a tree T and a spectrum S is defined by

Simple model

$$\boxed{score(S,T) = \sum_{T' \text{ subtree of } T} f(m(T')) + f(M - m(T'))}$$

- Another way is

Mass set model

$$\boxed{score(S,T) = \sum_{m \in ms(T)} f(m)}$$

Problem Statement

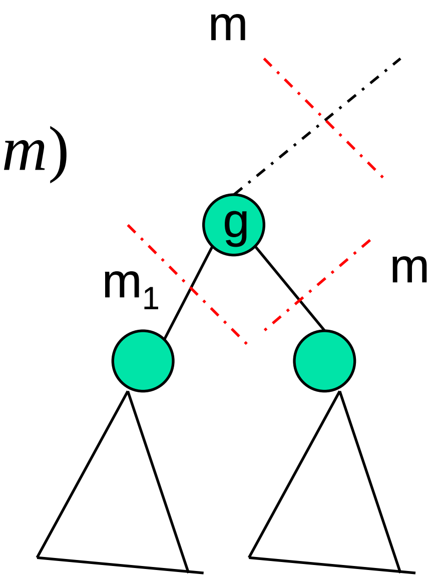
Glycan De Novo sequencing:

Given an MS/MS spectrum *S* with a precursor mass *M* and *f(m)*, find a glycan tree *T* such that $m(T)=M$ and $score(S,T)$ is maximized.

Algorithm under simple model

- *D(m)* be the score of the optimal subtree with mass *m*.

$$D(m) = \max_{\substack{g\in\Sigma \\ m_1+m_2+m(g)=m}} D(m_1)+D(m_2)+f(m)+f(M-m)$$



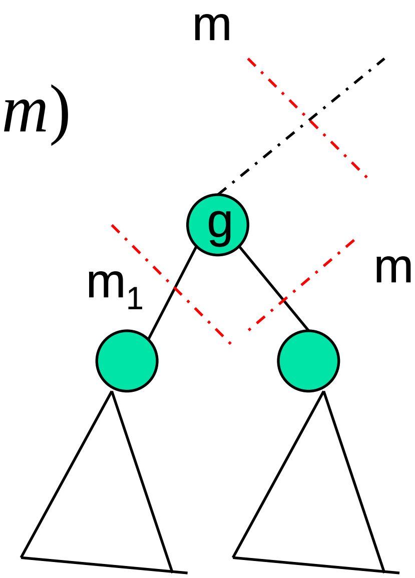
- *D(M)* will be the optimal tree.
- Time complexity $O(M^2)$
- Maximum degree is 2.

Algorithm under simple model

- *$D_2(m)$* be the score of the optimal subforest with at most two subtrees and mass *m*.

$$D(m) = \max_{\substack{g\in\Sigma \\ m_1+m_2+m(g)=m}} D_2(m_1)+D_2(m_2)+f(m)+f(M-m)$$

$$D_2(m) = \max_{0\le m_1\le m-m_1} D(m_1)+D(m-m_1)$$



- *$D(M)$* will be the optimal tree.
- Time complexity $O(M^2)$
- Maximum degree is 4.

Simple model v.s. mass set model

- The simple model “encourages” the algorithm to reuse some of the most intense peaks multiple times. This may cause problems.
- The mass set model can solve such problem.
- Unfortunately, mass set model is NP-hard.

Reduction for NP-hardness Proof

- Exact Cover by 3-Sets

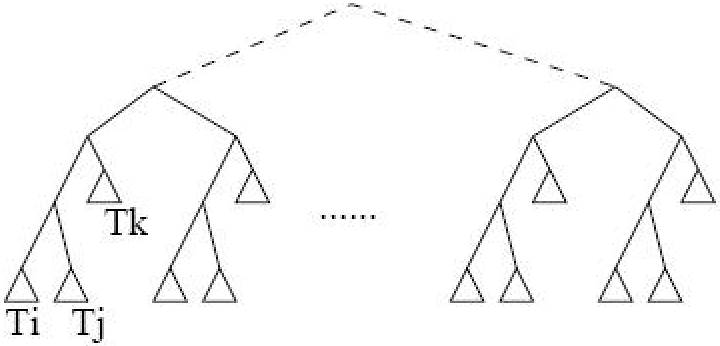
  Given $E = \{e_1, ..., e_n\}$ $S = \{s_1, ..., s_{n'}\}$ $s_l = \{e_i, e_j, e_k\}$

  Find $S^* \subseteq S, \quad s.t. \quad S^* \text{exact cover } E$

- Glycan *De Novo* Sequencing
  - Given $S = \{(m_1, h_1), ..., (m_n, h_n)\} and \quad \mathrm{M} \quad Score(S,T) = \sum_{m \in \Delta(T)} f(m)$
  - Find $T, \quad s.t. \quad Score(S,T) \text{ maxmize and } m(T) = \mathrm{M}$

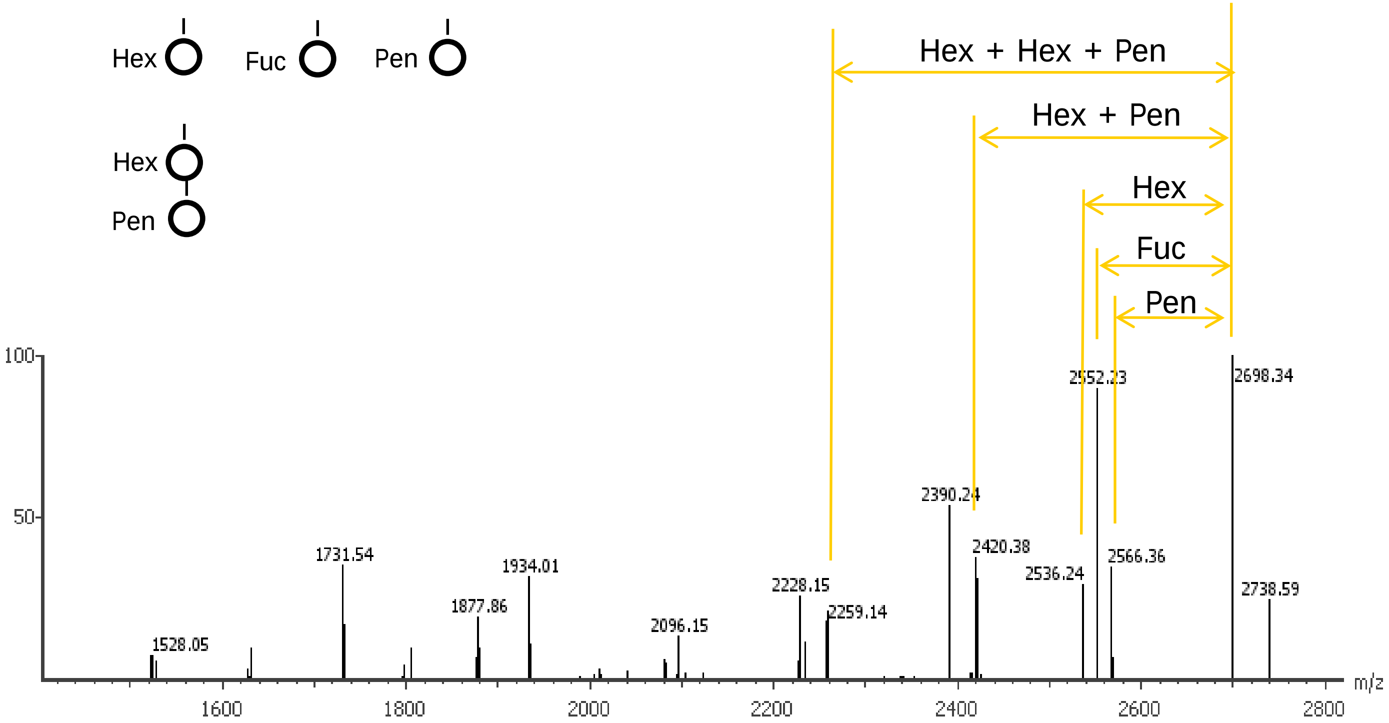
Idea of NP-hardness Proof

- Design a spectrum S and values of $f(m)$.
- With S and $f(m)$, e_i corresponds to T_i, each three-subtree group corresponds to a 3-set.
- If there is an exact cover, an optimal solution tree can be constructed.
- If there is an optimal solution tree, there is an exact cover.



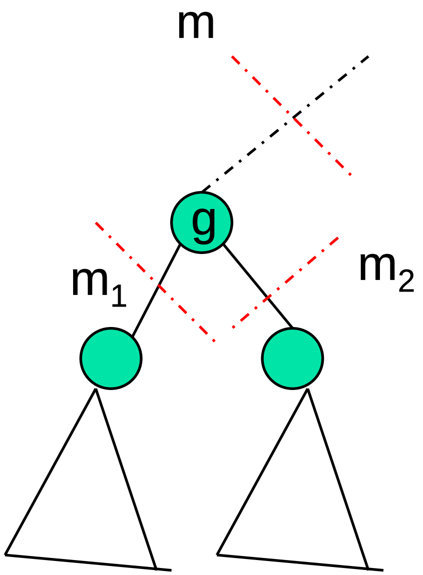
Heuristic algorithm

- Iterative construction, from smaller tree to larger tree
- Use y-ions, b-ions, and “internal” b-ions
- Keep first J trees with highest scores for each size



Heuristic algorithm

- Difficulty: when merging two subtrees together, some peaks may be reused.
- Solution: keep many subtrees (masses used) for m_1 and m_2, when merge, adjust the scores for reused peaks.



Software implementation - GlycoMaster

- Biochemistry considerations
  - Core of N-linked glycans
  - Parent of pentose node is hexose node
  - Fucose and sialic acid are leaf node
  - Parent of fucose node is hexNac node

Parameters

- Scoring function

m – fragment mass

I – peak intensity

Δm – mass error

σ – mass accuracy

d – penalty factor

J – number of trees kept

$$f(m) = \delta(m) \times \xi(m)$$

$$\delta(m) = \begin{cases} \log(I) \times e^{-\Delta m/\sigma} & \text{if } m \text{ matches a peak} \\ -d & \text{otherwise} \end{cases}$$

$$\xi(m) = \begin{cases} b & \text{if } m \text{ is a B} - \text{fragment} \\ a & \text{if } m \text{ is a Y} - \text{fragment} \end{cases}$$

- Software tool - GlycoMaster

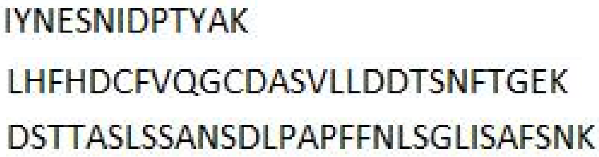
Experiments

- Cationic isozyme peanut peroxidase is a N-linked glycoprotein with three glycosylation sites.
- RP-HPLC separation.
- Mass Spectrometry Instrument： Q-TOF2
  - positive ion, ESI MS/MS mode, with borosilicate nano tips.

Experiments

- Glycopeptide-containing fractions identification
  - Q-TOF instrument operating in precursor ion discovery (PID) mode.
  - Identify typical simple sugar peaks.
- Tandem mass spectrometry for glycopeptides
  - MS/MS was triggered when sugars in the fraction were detected.
  - CID fragmentation breaks glycosidic bonds.
- 16 spectra were obtained, interpreted by human and Glycomaster program separately.
- All compositions agree. 15 structures agree.

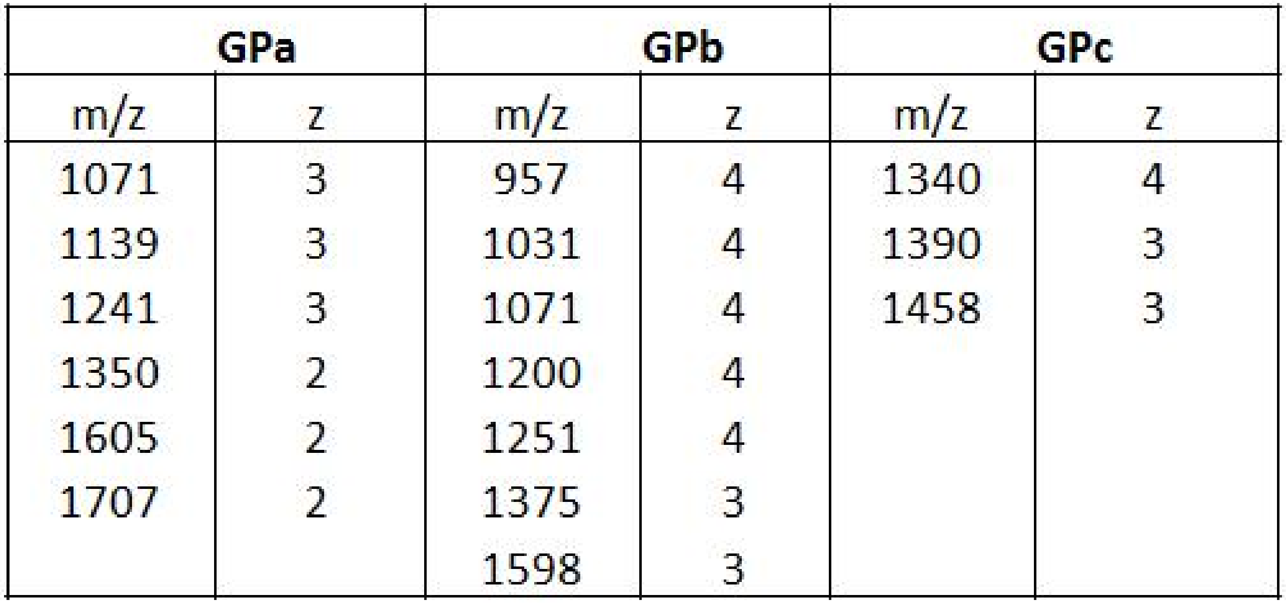
MS/MS tandem scan



IYNESNIDPTYAK

LHFHDCFVQGCDASVLLDDTSNFTGEK

DSTTASLSSANSDLPAPFFNLSGLISAFSNK



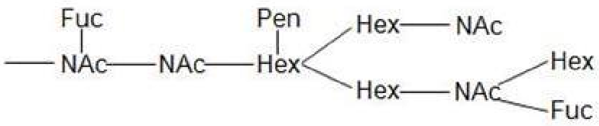
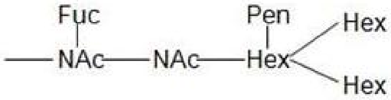
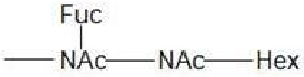
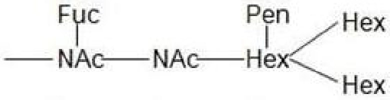
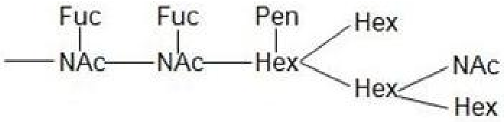
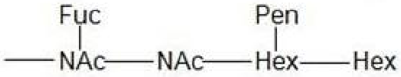
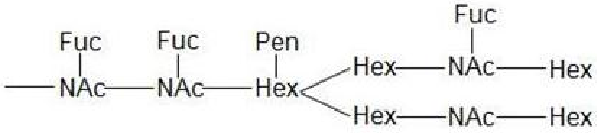
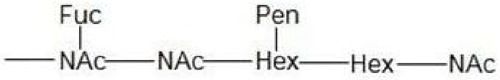
GPa		GPb		GPc	
m/z	z	m/z	z	m/z	z
1071	3	957	4	1340	4
1139	3	1031	4	1390	3
1241	3	1071	4	1458	3
1350	2	1200	4		
1605	2	1251	4		
1707	2	1375	3		
		1598	3		

Training

- 8 MS/MS spectra
  - GPa1071, GPa1241, GPa1605, GPa1707
  - GPb1301, GPb1251, GPb1598
  - GPc1390
- Constants
  - J = 1000
  - a = 5.2
  - b = 2.3
  - d = 2.8

Testing

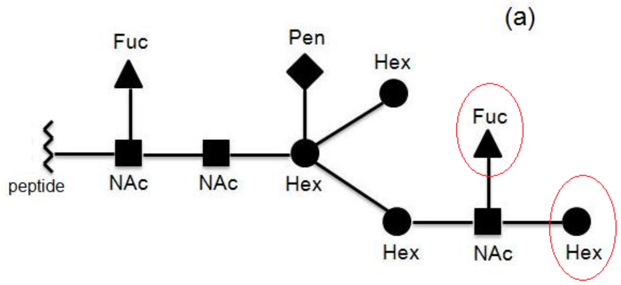
- 8 MS/MS spectra
  - GPa1139, GPa1350
  - GPb957, GPb1071, GPb1200, GPb1375
  - GPc1340, GPc1458
- Use PEAKS for data pre-processing
  - Deisotoping
  - Charge deconvolution

Glycopeptide	Glycan structure	Running time (s)
GPa 1139	 Fuc Pen NAc NAc Hex Hex NAc Hex NAc Hex Fuc	132
GPa 1350		98
GPb 957		54
GPb 1071		101
GPb 1200	 Fuc Fuc Pen NAc NAc Hex Hex Hex NAc Hex	119
GPb 1375		83
GPc 1340	 Fuc Fuc Pen Fuc NAc NAc Hex Hex NAc Hex Hex NAc Hex	201
GPc 1458		84

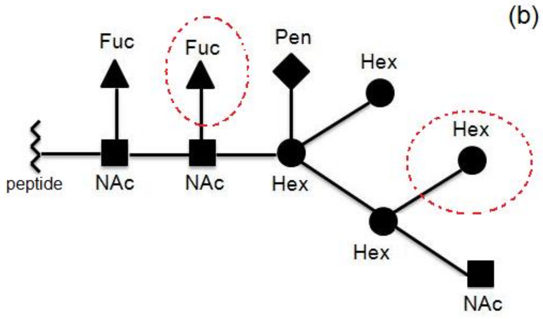
Validation

- All compositions are correct
- 7 of 8 glycan structures are same as manual interpretation
- 1 of 8 glycan structure is slightly different from manual interpretation (GPb1200)

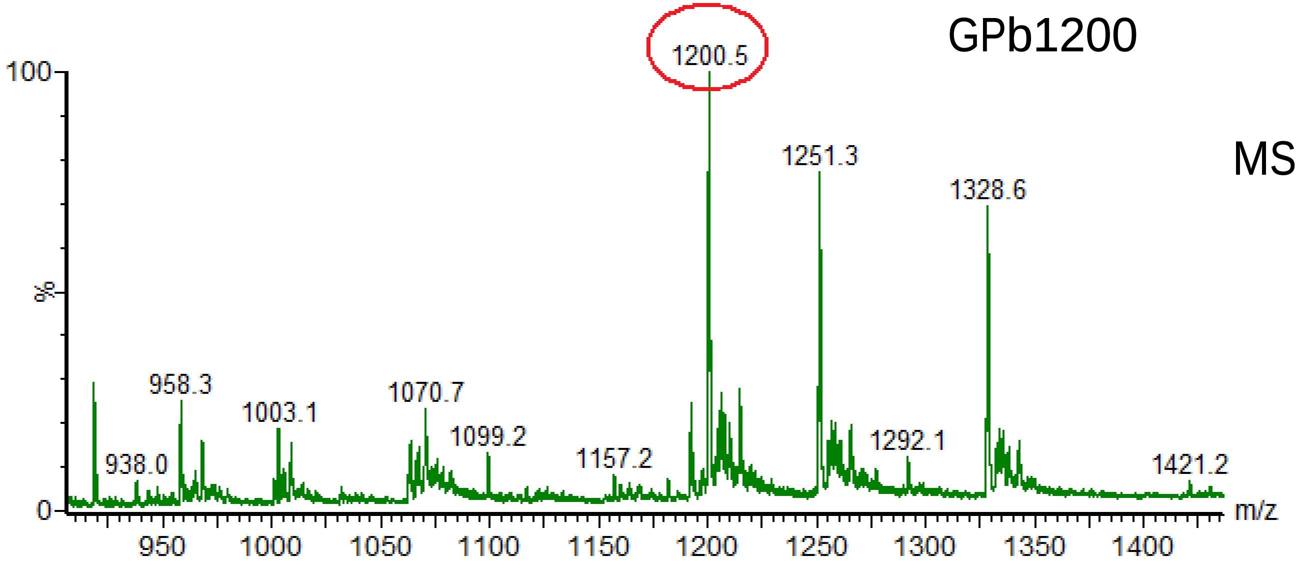
manual

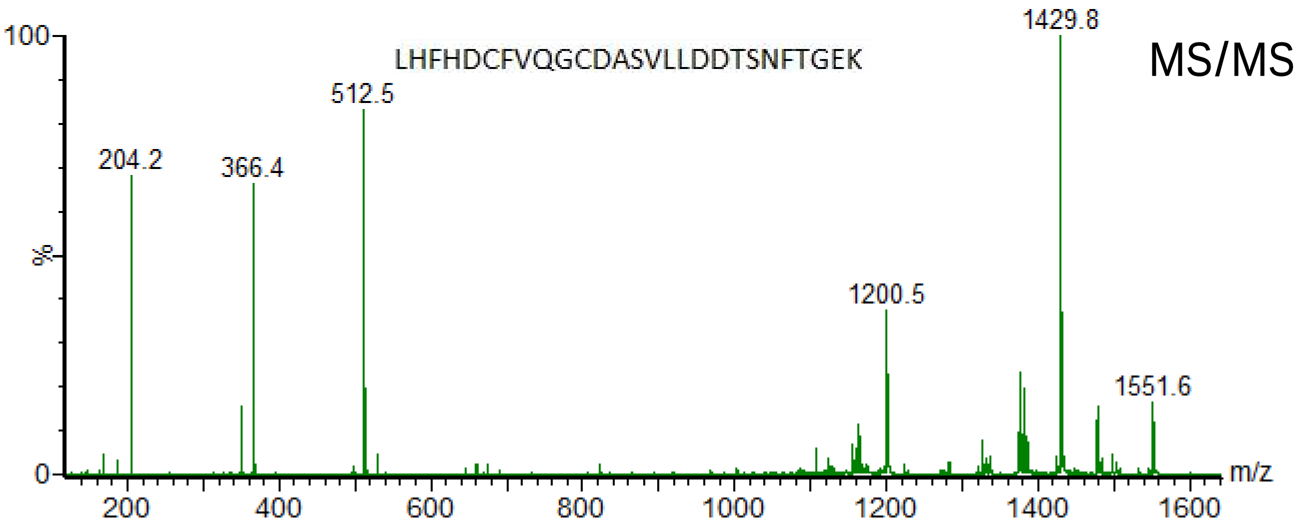


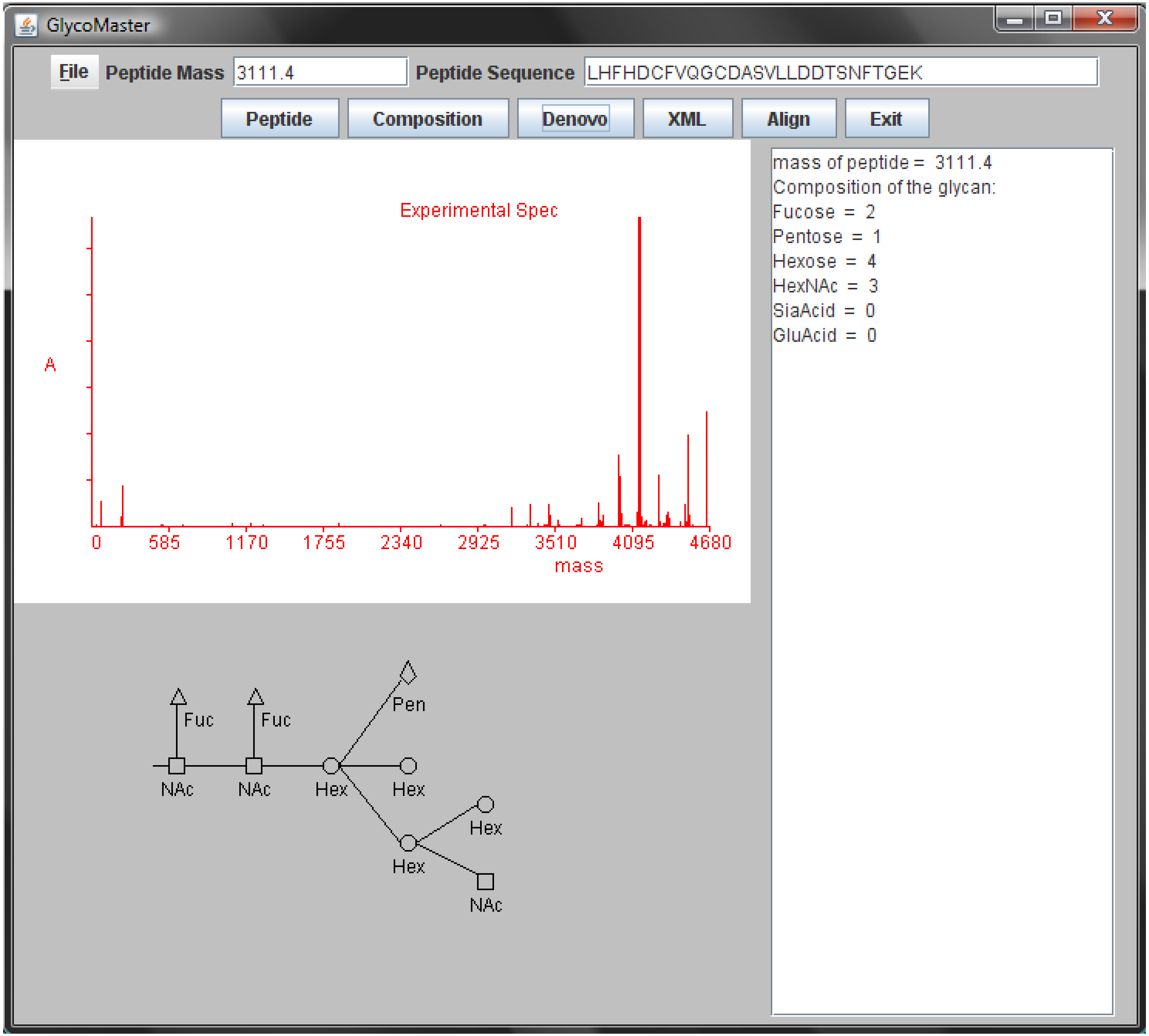
computed



GPb1200


MS

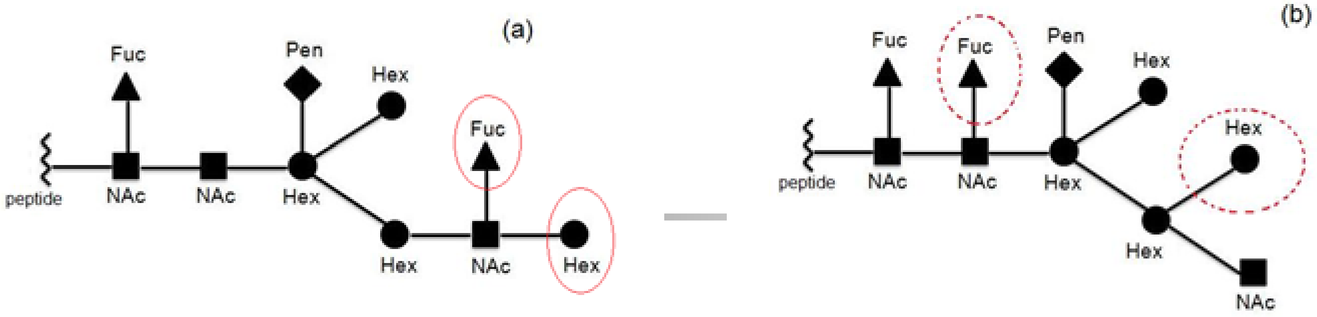

MS/MS

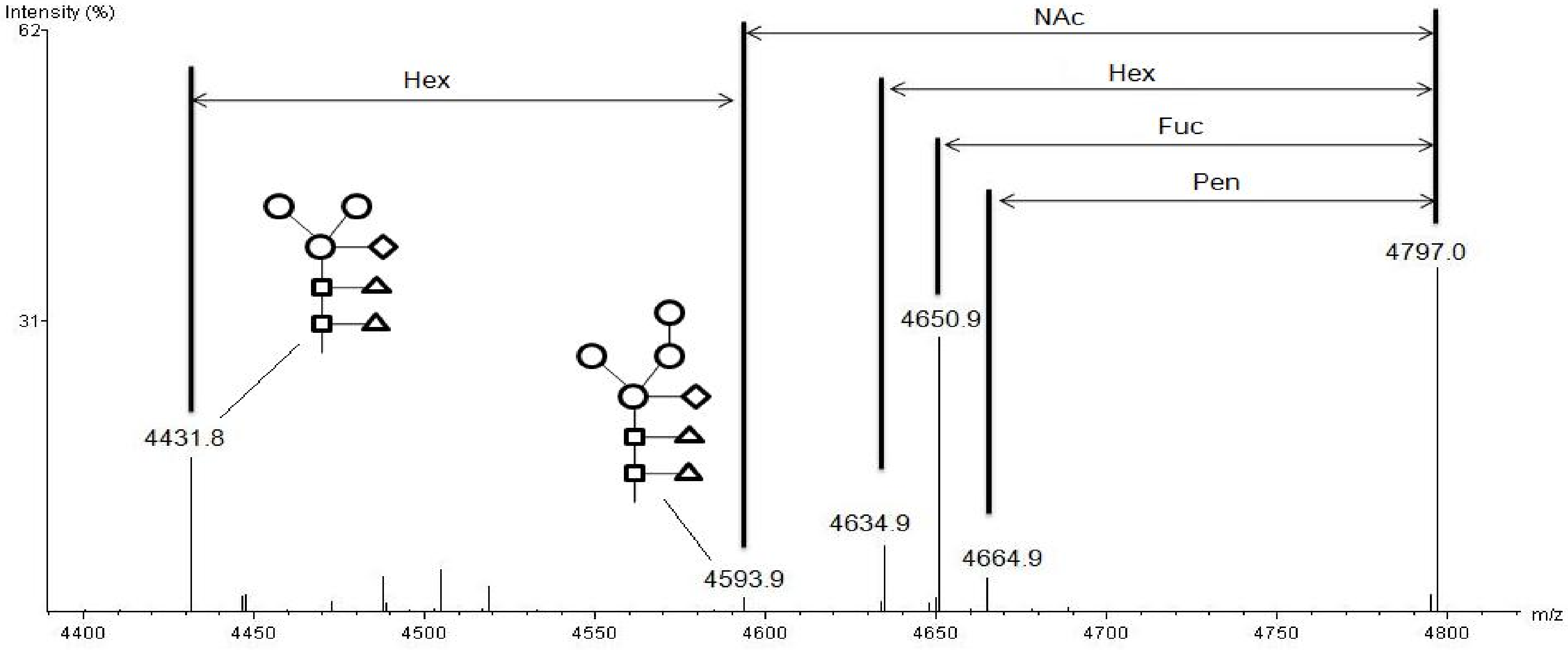


Peptide Mass 3111.4

Peptide Sequence LHFHDCFVQGCDASVLLDDTSNFTGEK

```
mass of peptide =  3111.4
Composition of the glycan:
Fucose  =  2
Pentose  =  1
Hexose  =  4
HexNAc  =  3
SiaAcid  =  0
GluAcid  =  0
```

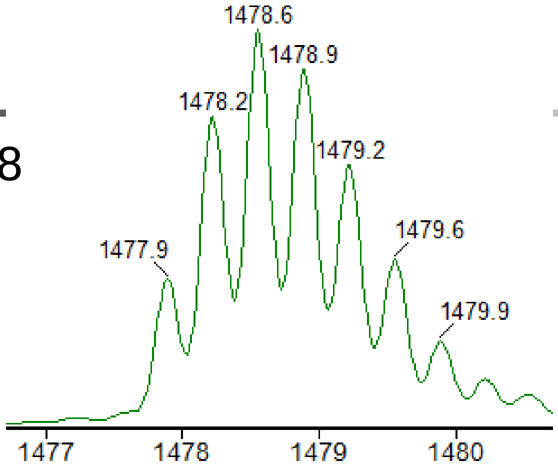




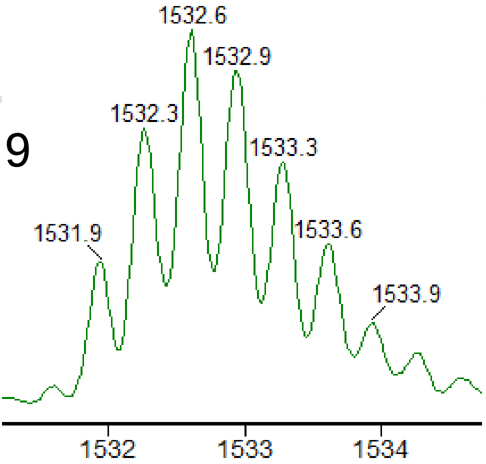
(b) matches two more peaks

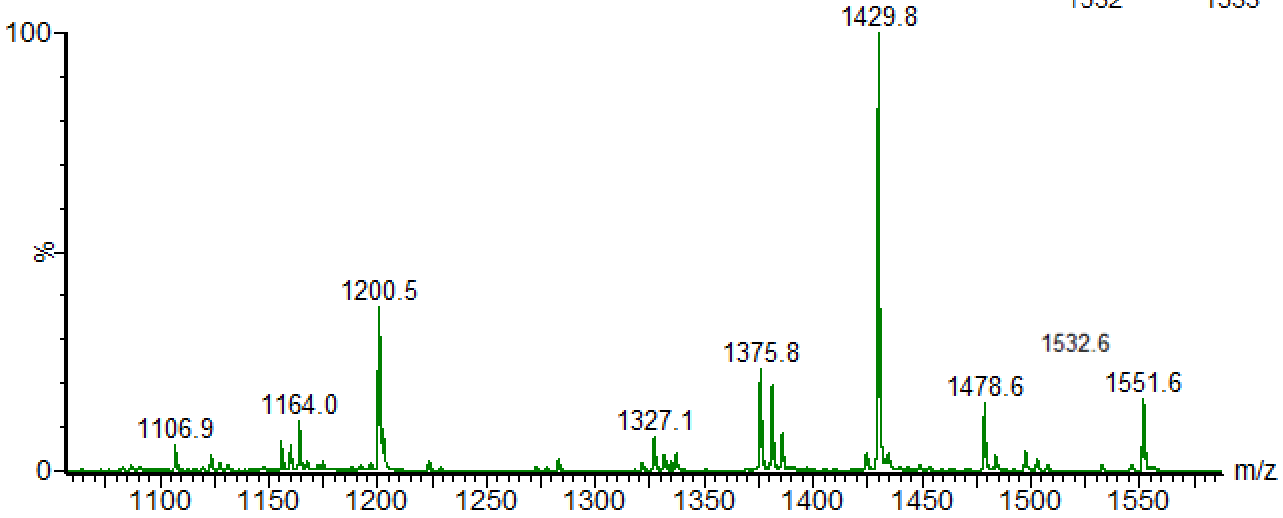
Peaks in raw data

4431.8



4593.9





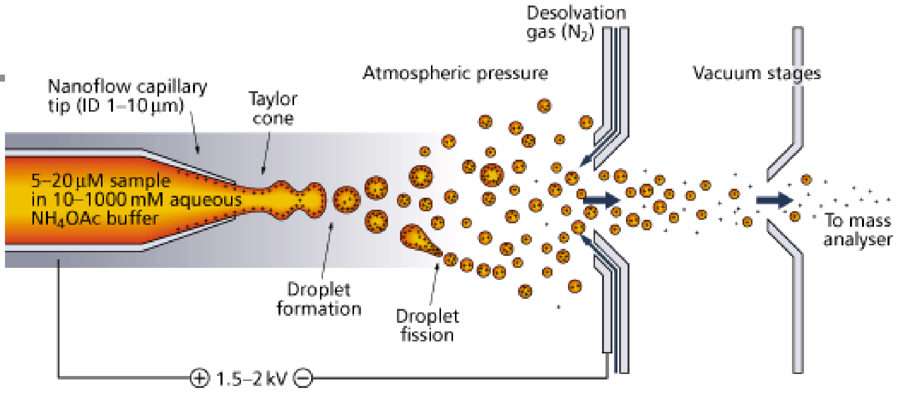
Conclusion

- A polynomial time algorithm is provided under simple model of glycopeptide *De Novo* sequencing
- A more realistic model is proved to be NP-hard
- A new heuristic algorithm is introduced, which works very well in practice.

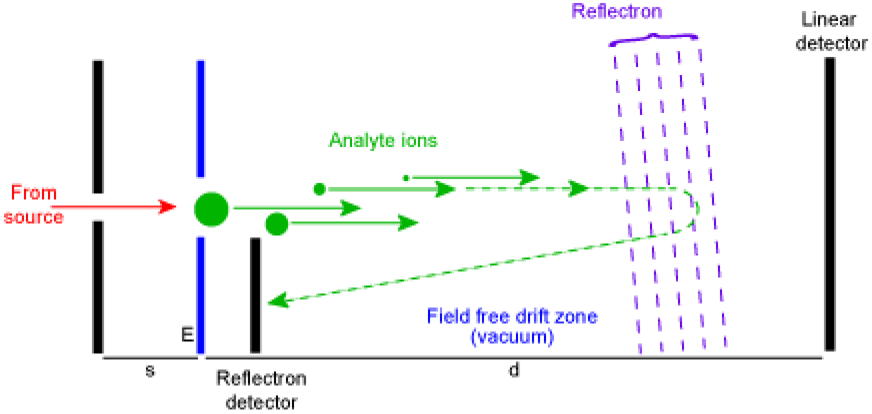
Future work

- Integrate with database search
- Combine ETD-CID or ETD-HCD for full characterization
  - Experimentally, the Thermo^{TM} LTQ Orbitrap is capable of alternating between HCD and ETD for the same precursor ion during an LC/MS/MS analysis.

- ESI


Desolvation gas (N_2)
Atmospheric pressure
Vacuum stages
Nanoflow capillary tip (ID 1–10 µm)
Taylor cone
5–20 µM sample in 10–1000 mM aqueous NH_4OAc buffer
To mass analyser
Droplet formation
Droplet fission
⊕ 1.5–2 kV ⊖

- TOF


Reflectron
Linear detector
Analyte ions
From source
Field free drift zone (vacuum)
E
s
d
Reflectron detector

- 1. if $z_i + z_j = z_{i'} + z_{j'}$, then $\{i,j\} = \{i',j'\}$.
- 2. if $z_i + z_j + z_k = z_{i'} + z_{j'} + z_{k'}$, then $\{i,j,k\} = \{i',j',k'\}$.
- 3. $z_1 < z_2 < \ldots < z_n = \mathrm{O}(\mathrm{n}^{12})$.
- 4. if $i \neq j$, $|z_i - z_j| \geq n^6 + 2$.
- 5. if $\{i,j\} \neq \{i',j'\}$, $|z_i + z_j - z_{i'} + z_{j'}| \geq n^6 + 2$.
- 6. if $\{i,j,k\} \neq \{i',j',k'\}$, $|z_i + z_j + z_k - z_{i'} + z_{j'} + z_{k'}| \geq n^6 + 2$.

- N = $n \times \max z_i$
- M = $4 \times n$
- $x_i = \text{M} + z_i$
- e_i corresponds to x_i
- $\{e_i, e_j, e_k\}$ corresponds to $x_i + x_j + x_k + 2$

Related Work

- Brute-force
  - All topology

    Gaucher et al. *Anal. Chem.* **72**:2231-2236, 2000
  - N-linked glycan

    Goldberg et al. *Proteomics.* **5**:865-875, 2005
- Spectrum graph
  - Glycan composition

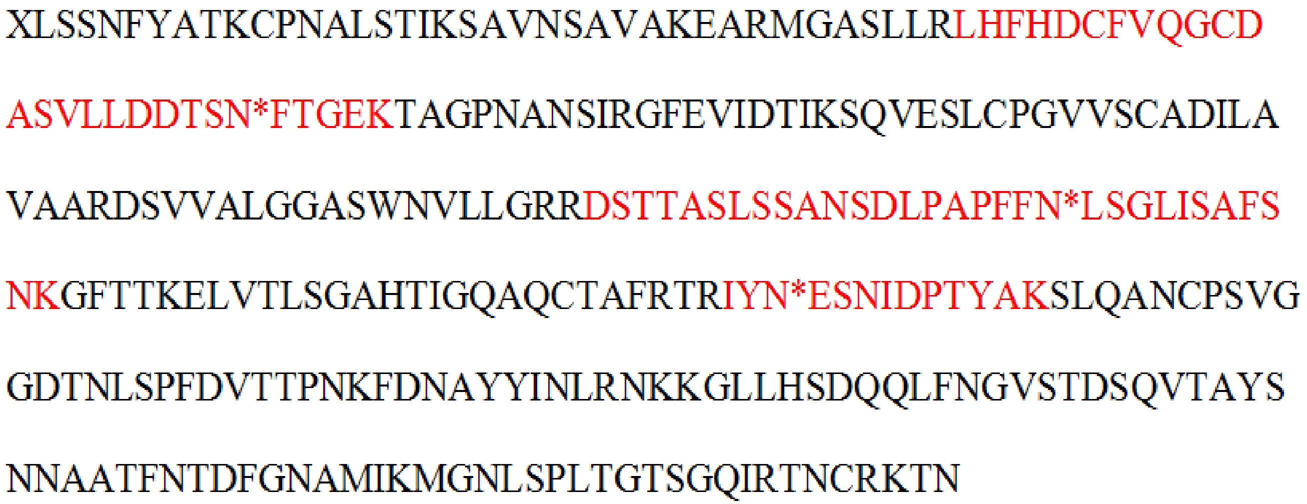
    Mizuno et al. *Anal. Chem.* **71**:4764-4771, 1999
  - Structures of released glycans

    Ethier et al. *Rapid Commun. Mass Spectrom.* **17**:2713-2720, 2003
- Dynamic programming – linear structures

  Tang et al. *Bioinformatics* **21**:i431-i439, 2005

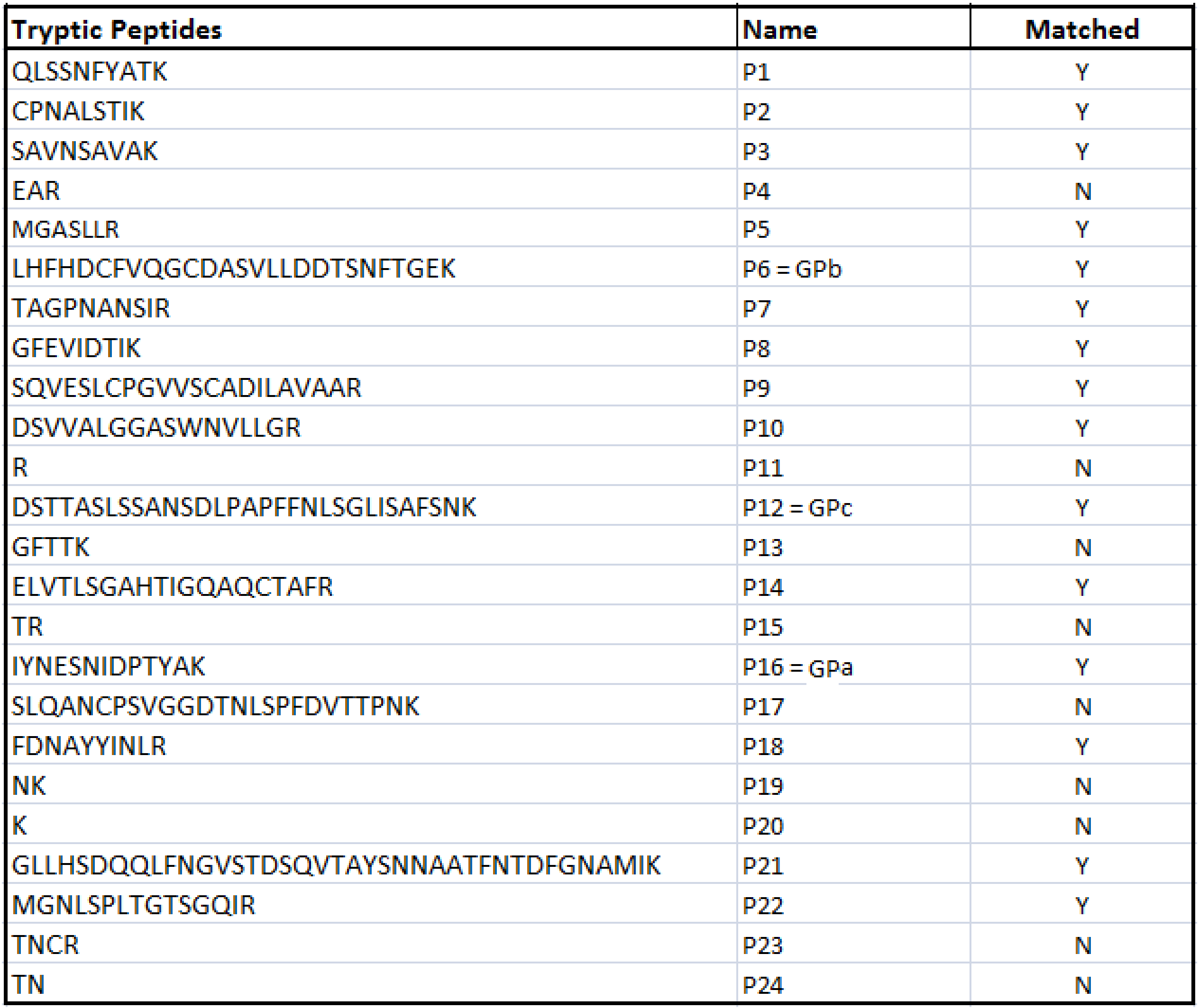
Example: Cationic isozyme peanut peroxidase

- Purification by RP-HPLC

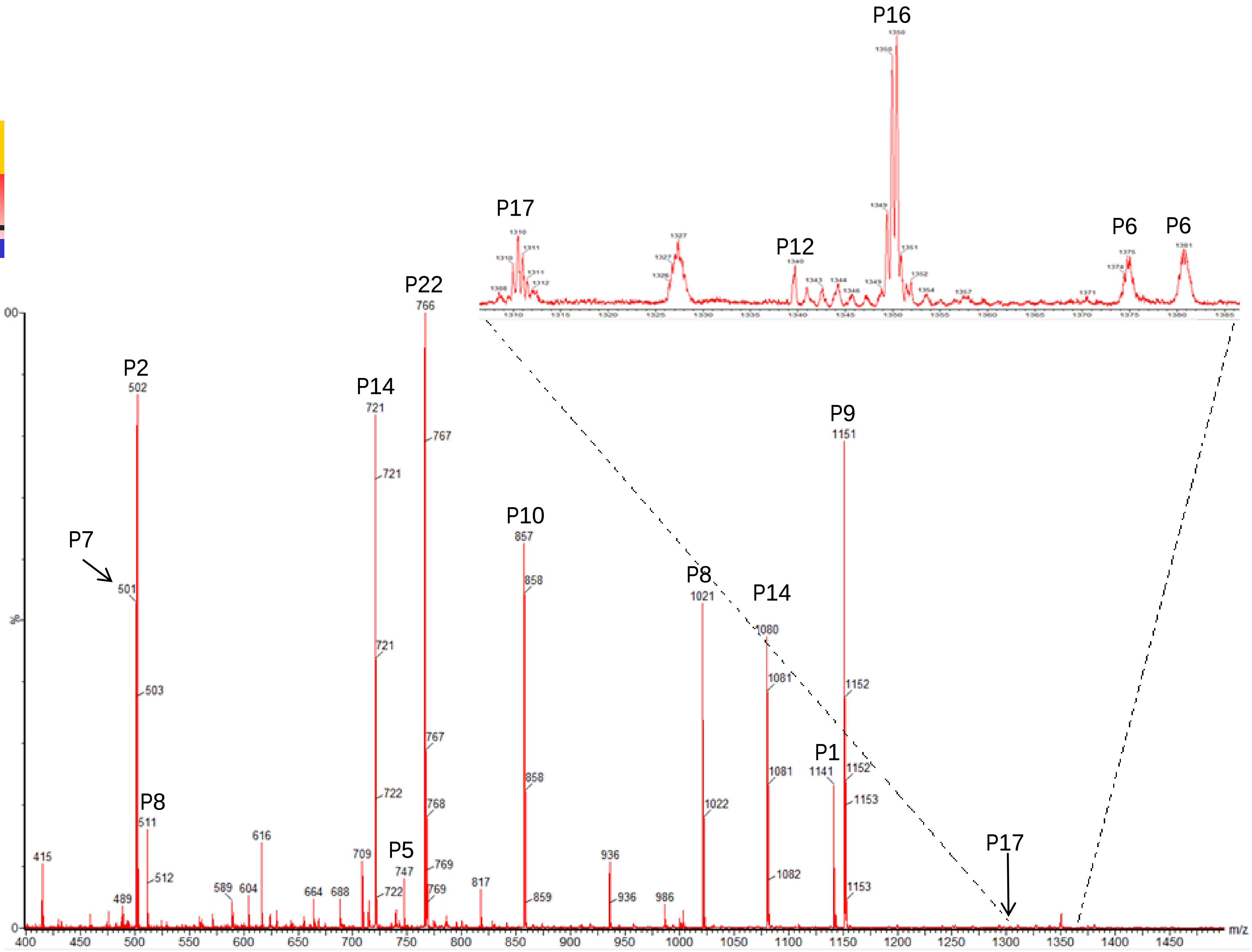


```
XLSSNFYATKCPNALSTIKSAVNSAVAKEARMGASLLRLHFHDCFVQGCD
ASVLLDDTSN*FTGEKTAGPNANSIRGFEVIDTIKSQVESLCPGVVSCADILA
VAARDSVVALGGASWNVLLGRRDSTTASLSSANSDLPAPFFN*LSGLISAFS
NKGFTTKELVTLSGAHTIGQAQCTAFRTRIYN*ESNIDPTYAKSLQANCPSVG
GDTNLSPFDVTTPNKFDNAYYINLRNKKGLLHSDQQLFNGVSTDSQVTAYS
NNAATFNTDFGNAMIKMGNLSPLTGTSGQIRTNCRKTN
```

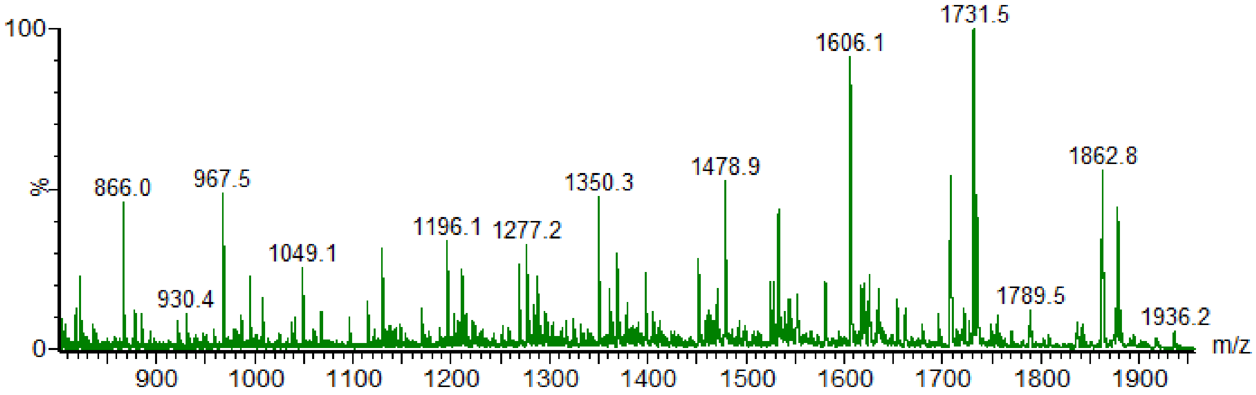
- Digestion with trypsin : cutting at R or K



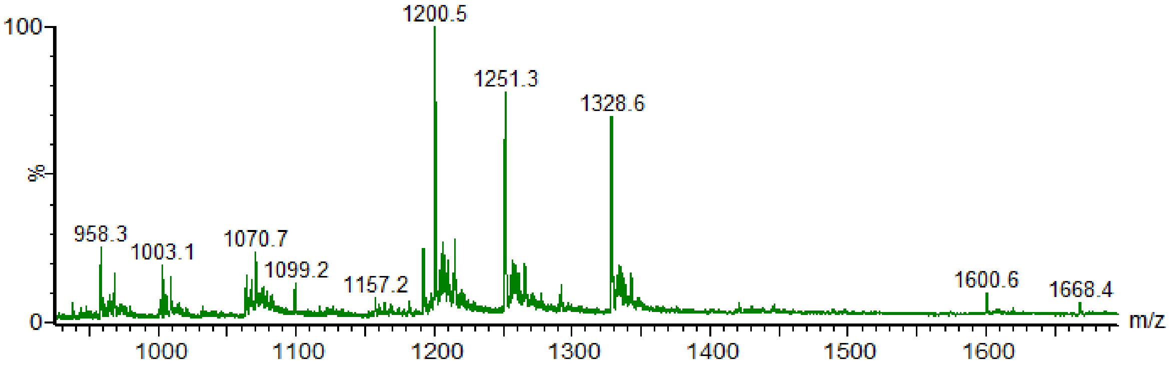
Tryptic Peptides	Name	Matched
QLSSNFYATK	P1	Y
CPNALSTIK	P2	Y
SAVNSAVAK	P3	Y
EAR	P4	N
MGASLLR	P5	Y
LHFHDCFVQGCDASVLLDDTSNFTGEK	P6 = GPb	Y
TAGPNANSIR	P7	Y
GFEVIDTIK	P8	Y
SQVESLCPGVVSCADILAVAAR	P9	Y
DSVVALGGASWNVLLGR	P10	Y
R	P11	N
DSTTASLSSANSDLPAPFFNLSGLISAFSNK	P12 = GPc	Y
GFTTK	P13	N
ELVTLSGAHTIGQAQCTAFR	P14	Y
TR	P15	N
IYNESNIDPTYAK	P16 = GPa	Y
SLQANCPSVGGDTNLSPFDVTTPNK	P17	N
FDNAYYINLR	P18	Y
NK	P19	N
K	P20	N
GLLHSDQQLFNGVSTDSQVTAYSNNAATFNTDFGNAMIK	P21	Y
MGNLSPLTGTSGQIR	P22	Y
TNCR	P23	N
TN	P24	N


P16
P17
P12
P6
P6
P22
766
767
767
768
769
769
P2
502
501
503
P7
P8
511
512
P14
721
721
721
722
722
P9
1151
1152
1152
1153
1153
P10
857
858
858
859
P8
1021
1022
P14
1080
1081
1081
1082
P1
1141
P5
747
P17
415
489
589
604
616
664
688
709
817
936
936
986
100
%
0
400
450
500
550
600
650
700
750
800
850
900
950
1000
1050
1100
1150
1200
1250
1300
1350
1400
1450
m/z

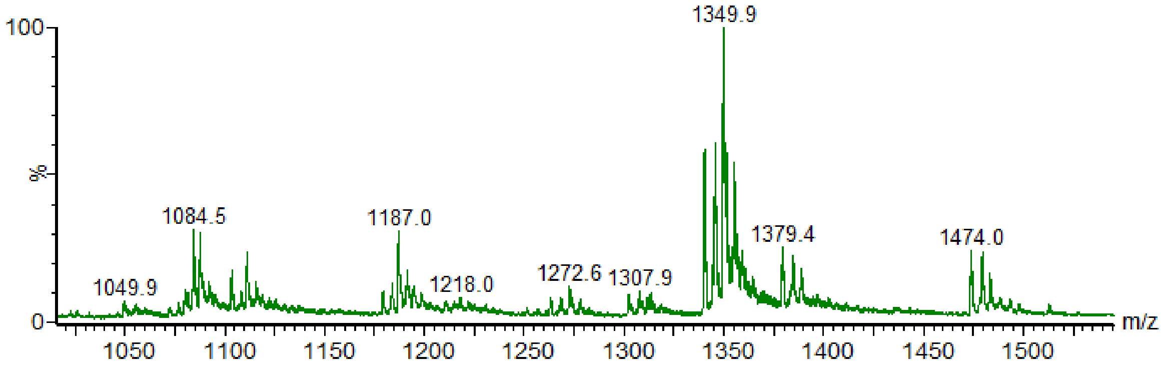
GPa


100
%
0
866.0
967.5
930.4
1049.1
1196.1
1277.2
1350.3
1478.9
1606.1
1731.5
1789.5
1862.8
1936.2
m/z
900
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900

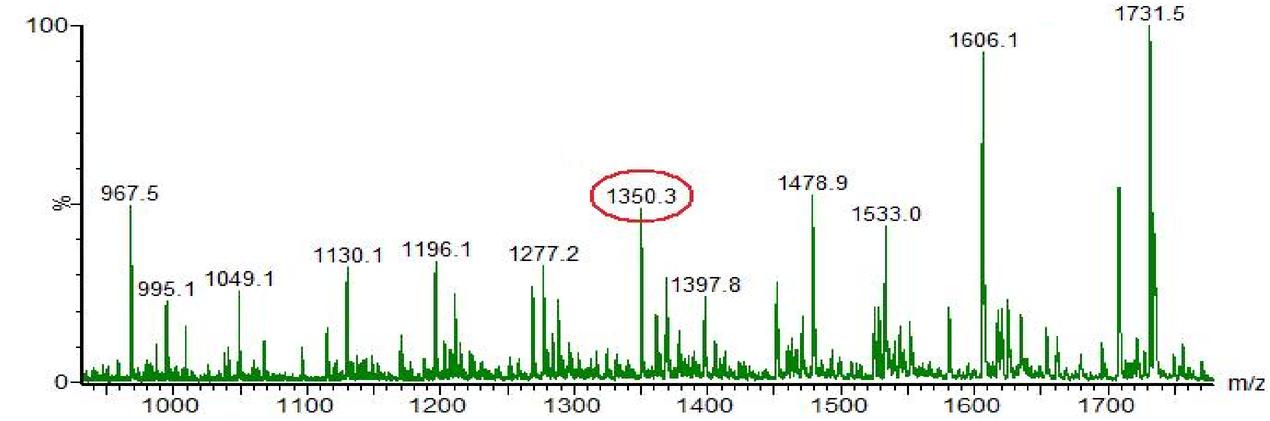
GPb


100
%
0
958.3
1003.1
1070.7
1099.2
1157.2
1200.5
1251.3
1328.6
1600.6
1668.4
m/z
1000
1100
1200
1300
1400
1500
1600

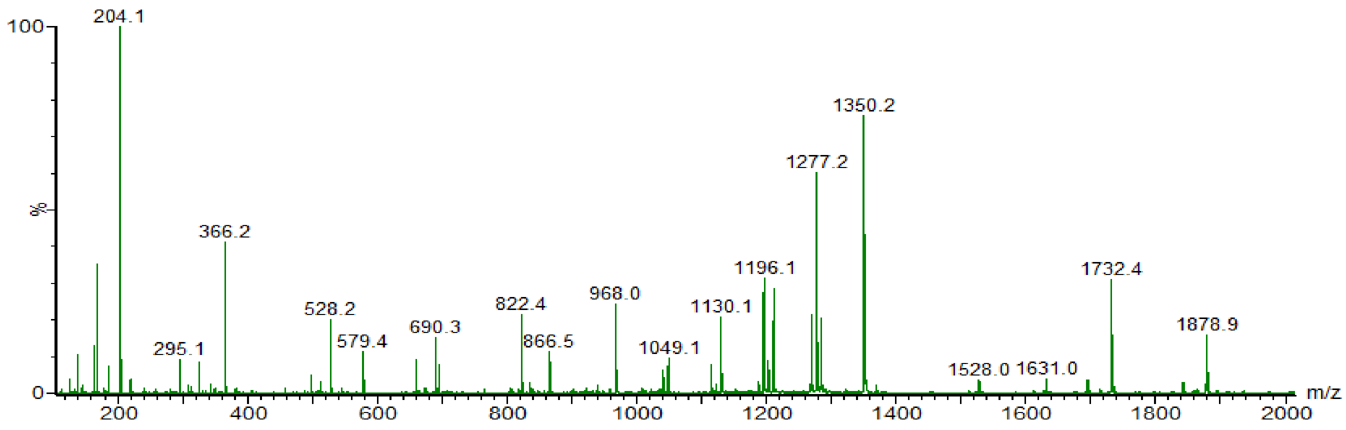
GPc


100
%
0
1049.9
1084.5
1187.0
1218.0
1272.6
1307.9
1349.9
1379.4
1474.0
m/z
1050
1100
1150
1200
1250
1300
1350
1400
1450
1500

- Fractionation by RP-HPLC
  - GPa IYNESNIDPTYAK
  - GPb LHFHDCFVQGCDASVLLDDTSNFTGEK
  - GPc DSTTASLSSANSDLPAPFFNLSGLISAFSNK
- MS survey sca



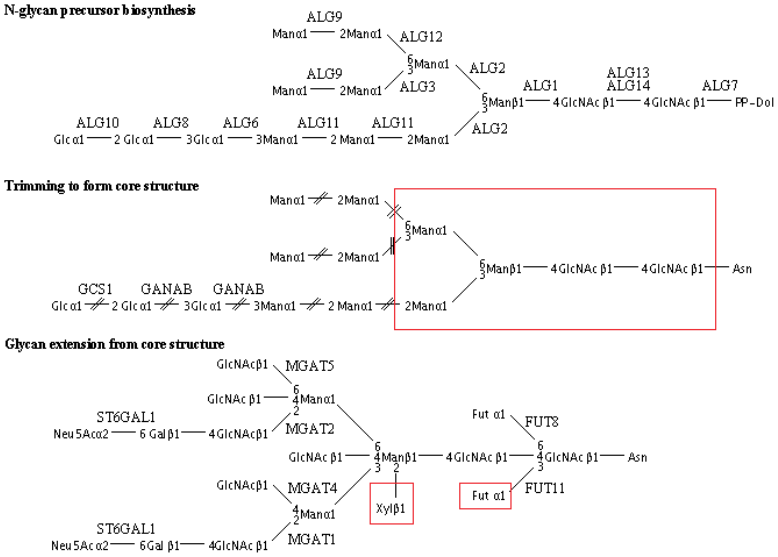
- MS/MS tandem scan



Software implementation - GlycoMaster

Biosynthetic pathways

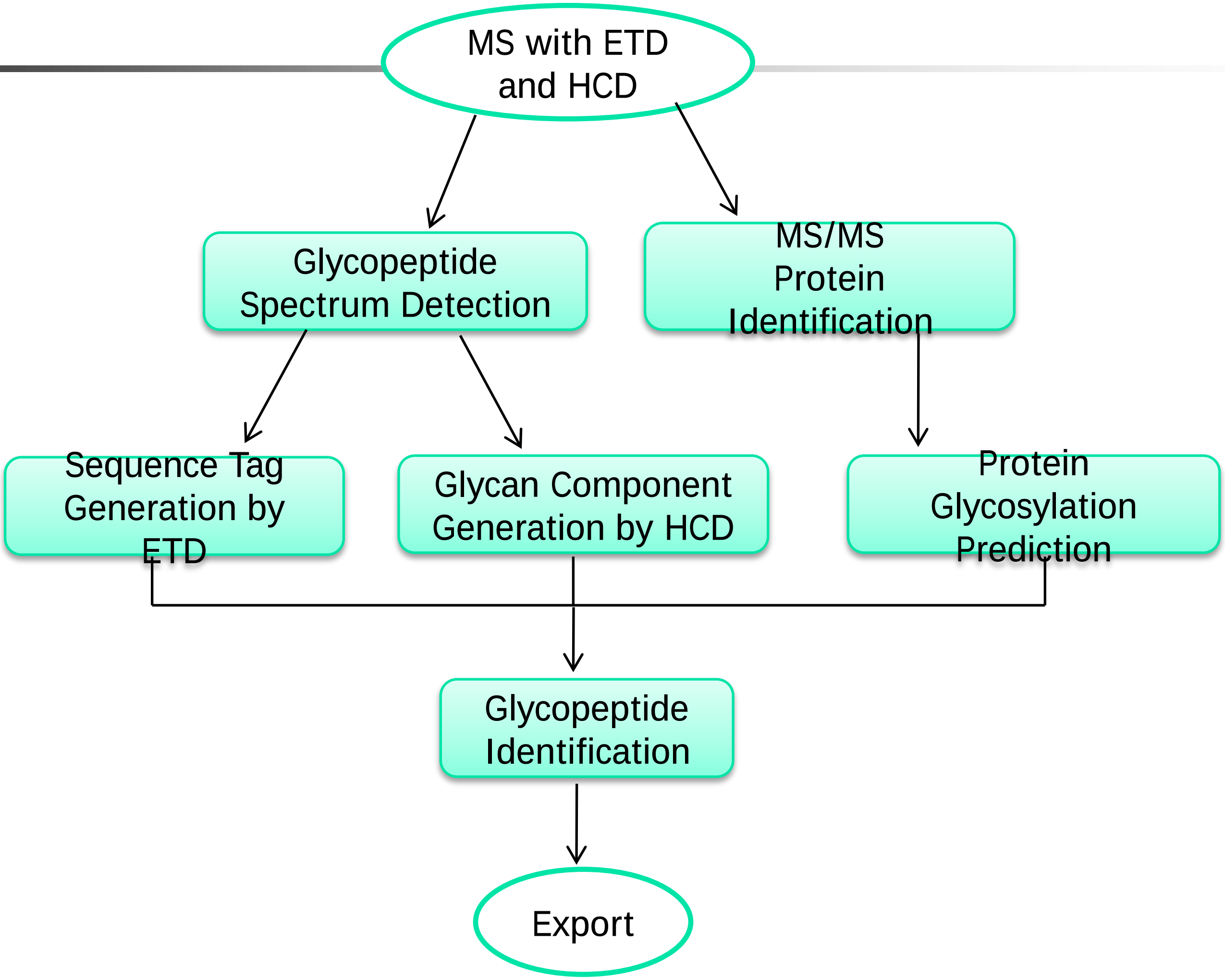
http://www.genome.jp/kegg/pathway


N-glycan precursor biosynthesis
ALG9
Manα1 — 2Manα1
ALG12
6 3 Manα1
ALG9
Manα1 — 2Manα1
ALG3
ALG2
ALG1
ALG13
ALG14
ALG7
6 3 Manβ1 — 4GlcNAc β1 — 4GlcNAc β1 — PP-Dol
ALG10
ALG8
ALG6
ALG11
ALG11
Glc α1 — 2 Glc α1 — 3Glc α1 — 3Manα1 — 2 Manα1 — 2Manα1
ALG2
Trimming to form core structure
Manα1 — 2Manα1
6 3 Manα1
Manα1 — 2Manα1
6 3 Manβ1 — 4GlcNAc β1 — 4GlcNAc β1 — Asn
GCS1
GANAB
GANAB
Glc α1 — 2 Glc α1 — 3Glc α1 — 3Manα1 — 2 Manα1 — 2Manα1
Glycan extension from core structure
GlcNAcβ1
MGAT5
GlcNAc β1 — 6 4 2 Manα1
ST6GAL1
Neu5Acα2 — 6 Gal β1 — 4GlcNAcβ1
MGAT2
Fut α1
FUT8
GlcNAc β1 — 6 4 3 Manβ1 2 — 4GlcNAc β1 — 6 4 3 GlcNAc β1 — Asn
GlcNAcβ1
MGAT4
Xylβ1
Fut α1
FUT11
4 2 Manα1
ST6GAL1
Neu5Acα2 — 6Gal β1 — 4GlcNAc β1
MGAT1

- Core of N-linked glycans
- Parent of pentose node is hexose node
- Fucose and sialic acid are leaf node

Science **291**:2351-2356, 2001

Data Analysis



MS with ETD and HCD

Glycopeptide Spectrum Detection

MS/MS Protein Identification

Sequence Tag Generation by ETD

Glycan Component Generation by HCD

Protein Glycosylation Prediction

Glycopeptide Identification

Export