Glycopeptidomics: Characterizing Global Glycoprotein and Site Heterogeneity

Robert J. Chalkley, Shouling Xu, Peter R. Baker and Katalin F. Medzihradszky

[1]NIGMS Mass Spectrometry Resource, University of California San Francisco, USA

Introduction

Due to the complexity and difficulty of glycopeptide analysis most researchers remove glycans prior to carbohydrate analysis. However, this strategy loses information about the modification status and heterogeneity on individual proteins and sites. Our laboratory employs lectin weak affinity chromatography (LWAC) for enriching glycopeptides[1], and this approach seems to show pan-specificity for all types of glycosylation[1,2]. Using this enrichment approach, combined with ETD and beam-type CID fragmentation, and development of new data analysis approaches using Protein Prospector, we are able to produce global maps of glycosylation patterns along with information about site-specific heterogeneity on individual proteins. In this presentation I will present results from the plant *Arabidopsis*.

Sample Preparation and Data Acquisition

Glycopeptides were enriched from *Arabidopsis* flower using three rounds of POROS-WGA LWAC chromatography. The glycopeptides were then fractionated by high pH reverse phase chromatography before analysis using UPLC (nanoAcquity) interfaced to a LTQ-Orbitrap Velos (Thermo). Spectra were acquired in a data-dependent mode, acquiring either ETD only or sequential HCD and ETD on each precursor.

Data Analysis

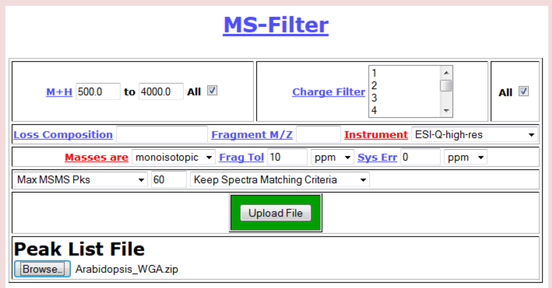
Data was analyzed using the freely available Protein Prospector package (prospector.ucsf.edu). An iterative analysis approach was applied.

1. A search was performed allowing for mass modifications between 140-2000 Da on serine, threonine or asparagine residues of any *Arabidopsis* protein. This produces a list of glycoproteins and glycopeptide compositions (based on the masses of the modifications).
2. A second search is then performed considering only glycoproteins identified in the initial search, allowing mass modifications from 140-3000 Da. This produces a more extensive list of glycan compositions present.
3. Finally, a search is performed against all *Arabidopsis* proteins, only considering defined glycan compositions identified in the second search.

Efficiency of Glycopeptide Enrichment

Glycopeptide CID spectra produce diagnostic oxonium ions that allow identification of glycopeptide spectra[3]. The m/z 204.087 HexNAc oxonium ion is present in almost every glycopeptide spectrum, and at high mass accuracy it cannot be mistaken for a peptide-derived fragment.

The program MS-Filter in Protein Prospector is able to filter peak list files for the presence / absence of particular masses or mass losses.

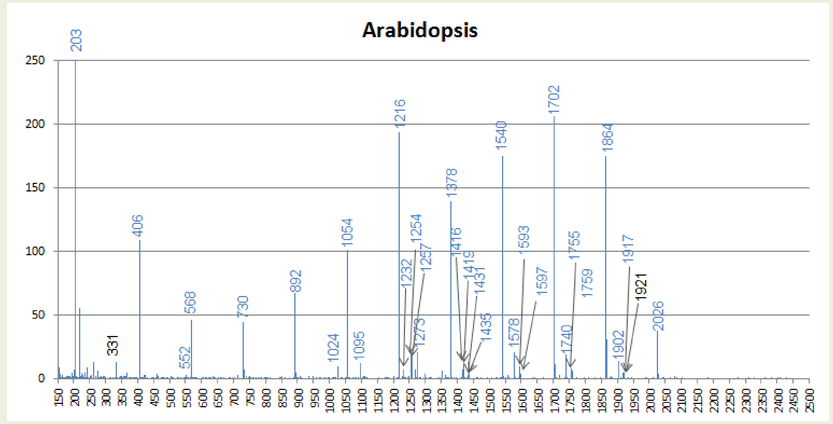


Peak lists from high pH reverse phase fractions of *Arabidopsis* peptides enriched using WGA LWAC chromatography were examined for the presence of the m/z 204.087 ion

V20141028-25_FTMSms2hcd.txt 814/3866 retained
V20141101-01_FTMSms2hcd.txt 2336/3615 retained
V20141028-53_FTMSms2hcd.txt 2649/3595 retained
V20141101-03_FTMSms2hcd.txt 2837/3712 retained
V20141028-43_FTMSms2hcd.txt 110/1589 retained
V20141028-42_FTMSms2hcd.txt 133/1544 retained
V20141028-41_FTMSms2hcd.txt 2509/3341 retained
V20141028-49_FTMSms2hcd.txt 1984/3637 retained
V20141028-40_FTMSms2hcd.txt 2553/3443 retained
V20141028-31_FTMSms2hcd.txt 472/2194 retained
V20141101-04_FTMSms2hcd.txt 2988/3933 retained
V20141028-29_FTMSms2hcd.txt 455/2172 retained
V20141101-02_FTMSms2hcd.txt 2289/3639 retained
V20141028-26_FTMSms2hcd.txt 391/3363 retained
V20141028-47_FTMSms2hcd.txt 1778/3665 retained
V20141028-32_FTMSms2hcd.txt 434/2141 retained
V20141028-51_FTMSms2hcd.txt 2088/3577 retained

After LWAC, 50 % of all acquired spectra featured the m/z 204.087 ion; many fractions contain over 70 % glycopeptides.

Mass Modification Searching for Glycopeptide ID



203 – HexNAc
406 – HexNAc2
552 – HexNAc2Fuc
568 – HexNAc2Hex
714 – HexNAc2HexFuc
730 – HexNAc2Hex2
862 – HexNAc2Hex2Xyl
892 – HexNAc2Hex3
1024 – HexNAc2Hex3Xyl
1054 – HexNAc2Hex4
1095 – HexNAc3Hex3
1216 – HexNAc2Hex5
1227 – HexNAc3Hex3Xyl
1232 – HexNAc2Hex5 + MetOx
1254 – HexNAc2Hex5 + K
1257 – HexNAc3Hex4
1273 – HexNAc3Hex4 + MetOx
1298 – HexNAc4Hex3
1378 – HexNAc2Hex6
1416 – HexNAc2Hex6 + K
1419 – HexNAc3Hex5
1431 – HexNAc2Hex6 + Fe
1435 – HexNAc3Hex5 + MetOx
1540 – HexNAc2Hex7
1578 – HexNAc2Hex7 + K
1593 – HexNAc2Hex7 + Fe
1597 – HexNAc3Hex6 + MetOx
1702 – HexNAc2Hex8
1740 – HexNAc2Hex8 + K
1755 – HexNAc2Hex8 + Fe
1759 – HexNAc3Hex7 + MetOx
1864 – HexNAc2Hex9
1902 – HexNAc2Hex9 + K
1917 – HexNAc2Hex9 + Fe
2026 – HexNAc2Hex10

Histogram of mass modifications observed on peptides from LWAC-enriched *Arabidopsis* peptides. Peaks in blue can be explained as a glycan-related masses.
N.B. m/z 331 = HexNAc + lysine (peptides where there is a missed cleavage); a significant percentage of the m/z 406 peak represents peptides bearing two single O-GlcNAc modifications.

1452 unique glycopeptides (2055 glycopeptide spectra) were identified (at an estimated 0.3 % FDR); about a third of these were O-GlcNAc modifications of nuclear and cytoplasmic proteins; the rest were N-linked glycosylation. Many truncated N-linked structures were observed. No complex O-linked glycans were identified. Several types of O-glycosylation occur in plants, through serines, threonines and hydroxyprolines. One potential reason for not observing them is that most are very large structures, so glycopeptides with these glycans attached would be very difficult to identify.

This dataset represents by far the largest O-GlcNAc dataset produced thus far in a plant, and also the largest plant glycopeptide dataset we are aware of.

Non-Consensus N-glycosylation

Protein Prospector now permits searching for modifications within a consensus motif (see poster W-319 for more details). However, by not specifying the motif we were able to detect non-consensus N-glycosylation on Asn88 of an ER-localized HSP70 family protein.

66 **Acc. #:** AT5G28540.1 **Species:** HEAT SHOCK PROTEIN 70 (HSP 70) FAMILY PROTEIN **Name:** Symbols: BIP1
Acc. #: chr5:10540665-10543274 **Species:** UNREADABLE **Name:** chr5:10540665-10543274 REVERSE LENGTH=669
Protein MW: 73629.9 **Protein pI:** 5.1 **Protein Length:** 669

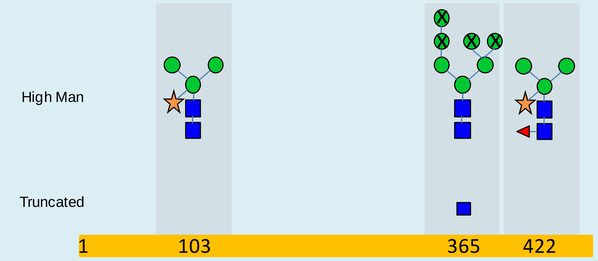
Num Unique	% Cov	Best Disc Score	Best Expect Val
8	11.5	3.51	2.1e-6

m/z	z	ppm	DB Peptide	Variable Mods	Protein Mods	Score	Expect	# in DB
747.8657	2	1.2	ITPSWVGFTDSER			36.4	3.3e-6	1
528.6677	3	3.4	LIGEAAKNQAAVNPER	HexNAc@8=39	HexNAc@88=39	37.5	4.9e-6	1
1074.4838	3	1.7	LIGEAAKNQAAVNPER	HexNAc2Hex7@8=38	HexNAc2Hex7@88=38	23.9	0.0016	1
926.1094	3	5.2	LIGEAAKNQAAVNPER	HexNAc3Hex3@8=38	HexNAc3Hex3@88=38	22.9	0.0020	1
942.4970	2	2.4	LIGEAAKNQAAVNPER	HexNAc@8|13	HexNAc@88|93	22.3	2.1e-4	1
596.6686	3	0.60	IINEPTAAAIAYGLDKK			25.4	0.0032	2
573.3265	2	0.044	NALETYVYNMK			15.0	0.0022	1
833.3649	3	-1.6	EALEWLDENQNSEKEEYDEK			41.1	2.1e-6	1

Site-specific Glycosylation

Glycopeptide analysis allows determination of glycosylation heterogeneity at individual sites in proteins. Different mixtures of glycans were observed at different sites within the same protein. High mannose structures were observed at most sites, but some sites also displayed truncated, complex or hybrid structures. Xylose and fucose residues were only observed on short or truncated structures.

The glycans detected on Purple Acid Phosphatase, a known glycoprotein[4], are presented below, as an example of the types of glycosylation observed.



Glycan compositions observed at specific sites on Purple Acid Phosphatase 26.
Residues containing 'X' were observed both present and absent. Linkage cannot be determined from glycopeptide data, so for high mannose a possible structure is depicted.

Summary and Conclusions

•WGA lectin weak affinity chromatography enriches practically all glycopeptides; not just those containing terminal GlcNAc or sialic acid residues.

•Protein Prospector is able to identify glycoproteins and glycopeptides in an unbiased fashion at high sensitivity from ETD data.

•Many truncated structures are identified, which are not identified in released glycan studies.

•For a similar type of study in mouse, see poster Tu-427

•For more details on advanced modification searching options in Protein Prospector see poster W-319.

Protein Prospector is open-source software freely available for use on the web, or a link to download it locally can be provided upon request.

http://prospector.ucsf.edu

If you have any questions, or want help trying the software please contact us:

ppadmin@cgl.ucsf.edu

Acknowledgements

This work was supported by NIH NIGMS grant 8P41GM103481 and the Howard Hughes Medical Institute.

References

1. Trinidad, J.C., Schoepfer, R., Burlingame, A.L. Medzihradszky, K.F. *Mol Cell Proteomics* (2013) 12 12 3474-3488
2. Medzihradszky, K.F., Kaasik, K., Chalkley, R.J. *Mol Cell Proteomics* (2015) M115.050393
3. Medzihradszky K.F., Kaasik K., Chalkley R.J. *Anal Chem* (2015) 87 5 3064-3071
4. Veljanovski, V., Vanderbeld, B., Knowles, V.L., Snedden, W.A., Plaxton, W.C. *Plant Physiol* (2006) 142 1282-1293