Recycling Modification Site Information for Improved Glycopeptide Analysis

Robert J. Chalkley and Peter R. Baker
Mass Spectrometry and Proteomics Resource, University of California San Francisco, USA

Introduction

A selection of different glycan modifications decorate each glycosylation site in a protein. Hence, if a given glycopeptide is identified, there is a high probability that other glycoforms of the same modification site are also present in the sample. If information about known modification sites could be used to direct glycopeptide data analysis tools, then this would lead to increased glycopeptide identification. In this study we evaluate the benefits of altering the search space for glycopeptide identification by filtering at the protein or modified peptide level to improve the ability to identify glycopeptide spectra. This is achieved by adapting Protein Prospector to be able to use a site database to filter sites considered as modified.

A database search engine is given a list of protein and modifications to consider for searching. By reducing the number of proteins and/or modifications considered the confidence measure for modified peptide identifications will increase (as the software calculates a confidence assuming only those things considered are possible), but one has to be careful to consider everything likely and not to restrict search space to the extent that one does not have a good measure of reliability of results.

For N-linked glycosylation the consensus modification motif (N-!P-S/T/C) can be used to filter sites considered, but there is no motif for O-linked glycosylation. This, combined with the higher frequency of Ser and Thr residues, means that many more sites of modification must be considered. Hence, searching for O-glycosylation significantly reduces the confidence of other identifications. However, if a site database of known modification sites could be used, then this should address this problem. The use of such a database is the focus of this poster.

Experiment to Compare Search Filtering Strategies

- Sample: Mouse Liver Lysate
- Glycopeptides enriched using Lectin Weak Affinity Chromatography (LWAC)
- Glycopeptide-enriched fraction separated by high pH RPLC
- High pH fractions analyzed by LC-MS/MS using LTQ-Orbitrap Velos, acquiring sequential HCD and ETD on precursors

A previous analysis of this dataset has been published[1]
All results on this poster are based only on ETD MS data analysis

Search	Parameters
N O	Considered N-linked glycosylation (in consensus motif) and O-linked glycosylation
N	Considered N-linked glycosylation (in consensus motif)
N O Acc	Considered N-linked glycosylation (in consensus motif) and O-linked glycosylation on list of proteins identified in Search #1
SiteDB	Considered N-linked glycosylation and O-linked glycosylation on peptides containing glycosylation sites identified in Search #1
SiteDB Acc	Considered N-linked glycosylation and O-linked glycosylation on peptides containing glycosylation sites identified in Search #1, only considering proteins identified in Search #1

- All searches +/- 12 ppm tolerance on precursor; +/ 0.6 Da on fragments
- Searches 1, 2 and 4 considered 16711 proteins (all SwissProt Mouse entries)
- Searches 3 and 5 considered 849 proteins
- Searches 4 and 5 considered 467 potential modification sites (populated based on results from Searches 1 and 2)

Global Glycopeptide Results

Search	# Unique Glycopeptides
N O	341
N	405
N O Acc	585
SiteDB	704
SiteDB Acc	719

- In total 638 N-linked and 152 O-linked unique glycopeptides were identified among searches, from 199 proteins.
- Calculating **FDR at peptide+glycosylation level**:
  - N-linked FDR = 0/638 = ?
  - O-linked FDR = 4/152 = 2.6 %
- O-linked glycopeptide IDs have higher FDR because many more O-linked glycopeptides are considered due to the much higher frequency of Ser or Thr in comparison to Asn (and in consensus motif)
  - More accurate FDR estimations are produced by measuring random matches for a given modifications type.

- In general, decreasing the search space leads to increased glycopeptide ID
- More glycopeptides are identified when only N-linked are considered (Search N) than when both N- and O-linked are considered (Search N O), despite 39 O-linked glycopeptides being identified in the N O search.
  - Considering O-linked glycosylation dramatically increases search space.
- Restricting based on identified glycosylation sites is more beneficial for improving the number of glycopeptide IDs than filtering at protein accession level. (SiteDB > N O Acc).

Example of Effect of Different Search Space for N-linked Glycosylation

Glucosylceramidase

Search	# Unique Glycopeptides
N O	2
N	4
N O Acc	5
SiteDB	8
SiteDB Acc	9

Number of peptides / peptides +mods within precursor mass tolerance that are considered for matching to a given spectrum

			N O		N		N O Acc		SiteDB		SiteDB Acc	
Site	**DB Peptide**	**Mod**	**Expect**	**# Precursor**	**Expect**	**# Precursor**	**Expect**	**# Precursor**	**Expect**	**# Precursor**	**Expect**	**# Precursor**
N78	RMELSVGAIQANR	HexNAc@12		1951		571	0.026	167		520	0.0056	36
N78	RMELSVGAIQANR	HexNAc2Hex2@12	0.0015	9111	1.20E-04	686	1.0E-04	593	6.2E-05	368	4.4E-06	26
N289	DLGPALANSSHDVK	HexNAc2Hex2@8	0.0046	10977	2.40E-04	587	3.0E-04	717	8.5E-05	206	1.0E-05	24
N289	DLGPALANSSHDVK	HexNAc2Hex3@8		12161	0.024	699	0.027	772	0.0056	160	7.7E-04	22
N289	DLGPALANSSHDVK	HexNAc2Hex3Fuc@8		13530	0.0071	891	0.007	879	0.0012	152	1.7E-04	21
N289	DLGPALANSSHDVK	HexNAc3Hex5SA@8		10306		2171		682	0.0044	45	0.0013	13
N289	DLGPALANSSHDVK	HexNAc4Hex5FucSA2@8		8451		2713		566	0.0032	30	0.0011	10
N289	DLGPALANSSHDVK	HexNAc4Hex5FucSAOx2@8		7061		2429		461	0.01	25	0.004	10
N289	DLGPALANSSHDVK	HexNAc4Hex5SAOx2@8		7026		2297		452	0.012	23	0.005	10

The quality of the match is the same in each search; the only difference is the number of other peptides (with or without modifications) considered. Considering more peptides will always decrease the expectation value, and could lead to a better random match.

- HexNAc modification only found in Acc# searches
  - Many tryptic peptides possible with precursor of mass 1647.8
  - At this mass, restricting # proteins reduces peptides considered more dramatically than restricting modification sites
- Glycopeptides with large sugar structures only identified in searches using modification site database
  - For high mass precursors restricting # modification sites more dramatically reduces peptides considered than restricting proteins, as there are not many unmodified tryptic peptides (with max 1 missed cleavage site) with a mass >3000 Da.
- Combination of accession number and site modification filtering leads to most IDs, but very few precursors considered
  - Is it getting too close to matching purely on basis of mass? How reliable are the lowest confidence results?

Site Database Benefit for O-linked Glycosylation

Extra O-linked glycoforms can also be identified, but fewer glycoforms are typically present per site, so using a database only of sites identified in an initial search is less beneficial than for N-glycosylation.

Laminin subunit alpha-5

			N O	Site DB
Site	**Mod**	**DB Peptide**	**Expect**	**Expect**
S3711	HexNAcHexSA	QAKPSVSPLLWH	8.20E-05	5.30E-06
S3711	HexNAcHexSA2	QAKPSVSPLLWH		0.022

O-linked glycosylation identification would benefit more than N-linked glycosylation from an extensive repository of sites identified in previous studies.

The Uniprot .dat file and UniCarbKB[2] contain information about a few O-linked glycosylation sites, but not enough to be useful for populating a site database repository for searching.

Conclusions

- Adjusting the search space is very beneficial for increasing identification of glycopeptides
- A typical search that allows for O-linked glycosylation has a significant detrimental effect on N-linked glycopeptide ID, as it increases the number of precursor peptides (mostly modified) by up to twenty-fold.
- The reliability of O-linked glycopeptide IDs in combined N- and O-linked glycopeptide searches is typically lower, as so many more O-linked modification sites are considered when searching, so FDRs should be estimated separately.
- Restricting peptides considered based on protein or based on known modification sites both lead to identification of more N-linked glycoforms for sites confidently identified in an initial broader search.
- Filtering based on protein accessions is most beneficial for glycopeptides bearing a small glycan (truncated N-linked or O-linked) structures.
- Filtering based on identified modification sites is significantly more effective for identifying large N-linked glycoforms.

- Modification site databases will soon be associated with databases on the public Protein Prospector website:

prospector2.ucsf.edu

- See also poster TP372 for benefits of a modification site database for other PTM types.

References

[1] *Mol Cell Proteomics* (2015) **14** 8 2103-10
[2] *Nucleic Acids Res* (2014) **42** D215-21

Acknowledgements

We thank Krista Kaasik for acquiring the data used in this analysis. This work was supported by NIGMS 8P41GM103481 and the Howard Hughes Medical Institute.