Expanding the glycoforms detected in complex glycopeptide datasets

Katalin F. Medzihradszky[1,2]; Peter R. Baker[1]; Adam Pap[2]; Z. Darula[2], Robert J. Chalkley[1]

[1]Mass Spectrometry Facility, University of California San Francisco, San Francisco, CA

[2]Laboratory of Proteomics Research, Biological Research Centre of the Hungarian Academy of Sciences, Szeged, Hungary

Introduction

Glycopeptides produce characteristic fragment ions under CID/HCD/EThcD fragmentation.

- Low mass glycan oxonium ions allow easy recognition of whether a spectrum is of a glycopeptide.
- For N-linked glycopeptide spectra fragments corresponding to the intact peptide with HexNAc (Y1) and HexNAc2 (Y2) are ubiquitous.
- For O-linked glycopeptide spectra unmodified intact peptide (Y0) and with a HexNAc attached (Y1) are observed for mucin-type O-linked glycans.

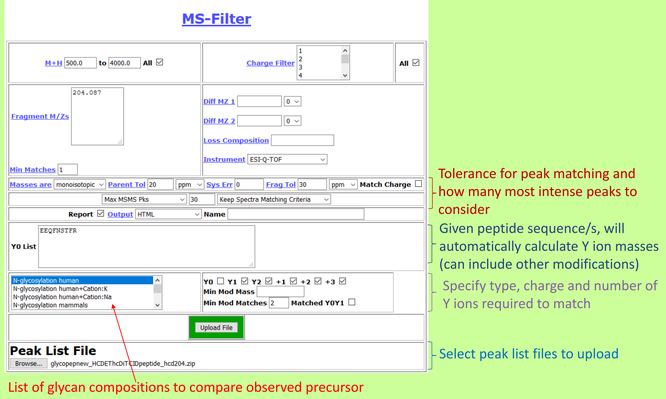
In this study we investigated how many extra glycoforms for peptides could be identified by taking glycopeptides identified by database searching of EThcD data, then using software to search peak lists for additional spectra that contain expected Y ions.

Methods

1. EThcD data of a complex O-glycopeptide dataset were searched using Protein Prospector and/or Byonic to identify glycopeptides.
2. A script was written to take these results, calculate expected Y0 and Y1 ions, then search peak list files for their presence in different charge states. Many candidate glycopeptide spectra were identified, but there were also many false positives.
3. MS-Filter in Protein Prospector was then adapted to be able to perform the same task, but supports both N- and O-glycosylation and has additional parameters

   It allows varying the fragment mass tolerance; the number of most intense peaks in the peak list file to consider; can filter based on the number of Y ions detected; and suggests candidate glycan compositions based on the mass difference observed between the precursor ion and peptide mass.

MS-Filter for Glycosylation

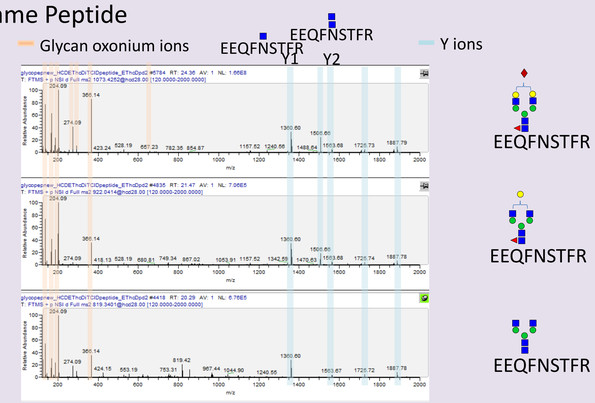


Optimal Parameters

For HCD and EThcD data tried varying number of peaks required to match and among 'n' most intense peaks

- Requiring matching only one peak (Y0 for O-linked; Y1 for N-linked) leads to too many false positives
- Using top 30 peaks compared to top 20 matches about 40% more spectra, but only about 10% more identifiable glycopeptide spectra.

Similarity of N-linked Glycopeptide Spectra of Different Glycoforms of Same Peptide



MS-Filter Example Output

Link to download filtered peak list file

MS-Filter Report

glycopeptenew_HCDEThcDITCIDpeptide_hcd.mgf 37/11496 retained

Archive file (glycopeptenew_HCDEThcDITCIDpeptide_hcd204.zip)

Matched Y ions

M/Z	Charge	Fraction	RT	MSMS Info	Peptide	Mod Mass	Possible Mods	Y1, z=3	Y1, z=2	Y1, z=1	Y2, z=3	Y2, z=2	Y2, z=1
854.0206	3	glycopeptenew_HCDEThcDITCIDpeptide_hcd	427	3036	EEQFNSTFR	1402.5252		0	1	1	0	0	1
976.0596	3	glycopeptenew_HCDEThcDITCIDpeptide_hcd	828	4022	EEQFNSTFR	1768.6422	HexNAc4Hex5Fuc	0	0	1	0	0	1
868.0244	3	glycopeptenew_HCDEThcDITCIDpeptide_hcd	881	4139	EEQFNSTFR	1444.5365	HexNAc4Hex3Fuc	0	1	1	0	0	1
782.5475	4	glycopeptenew_HCDEThcDITCIDpeptide_hcd	892	4170	EEQFNSTFR	1969.6460		0	0	1	0	1	0
782.5475	4	glycopeptenew_HCDEThcDITCIDpeptide_hcd	895	4173	EEQFNSTFR	1969.6460		0	1	1	0	1	1
874.0875	3	glycopeptenew_HCDEThcDITCIDpeptide_hcd	899	4182	EEQFNSTFR	1462.7256		0	0	1	0	0	1
1018.0620	3	glycopeptenew_HCDEThcDITCIDpeptide_hcd	932	4247	EEQFNSTFR	1894.6493		0	0	1	0	0	1
1384.0632	4	glycopeptenew_HCDEThcDITCIDpeptide_hcd	944	4268	EEQFNSTFR	4375.7089		0	1	1	0	0	1
861.6736	3	glycopeptenew_HCDEThcDITCIDpeptide_hcd	946	4274	EEQFNSTFR	1425.4842		0	0	1	0	0	1
872.3223	3	glycopeptenew_HCDEThcDITCIDpeptide_hcd	947	4275	EEQFNSTFR	1457.4301		0	0	1	0	0	1
861.6736	3	glycopeptenew_HCDEThcDITCIDpeptide_hcd	949	4277	EEQFNSTFR	1425.4842		0	1	1	0	1	0
1187.2681	5	glycopeptenew_HCDEThcDITCIDpeptide_hcd	953	4285	EEQFNSTFR	4774.7891		0	1	1	0	1	1
679.5487	4	glycopeptenew_HCDEThcDITCIDpeptide_hcd	959	4295	EEQFNSTFR	1557.6508		1	1	1	0	0	0
1392.0490	4	glycopeptenew_HCDEThcDITCIDpeptide_hcd	963	4302	EEQFNSTFR	4407.6518		0	1	1	0	1	1
799.2953	4	glycopeptenew_HCDEThcDITCIDpeptide_hcd	965	4310	EEQFNSTFR	2036.6374		0	0	1	0	0	1
862.3800	3	glycopeptenew_HCDEThcDITCIDpeptide_hcd	970	4318	EEQFNSTFR	1427.6033		0	0	1	0	0	1
862.3800	3	glycopeptenew_HCDEThcDITCIDpeptide_hcd	973	4323	EEQFNSTFR	1427.6033		0	0	1	0	1	0
873.3572	3	glycopeptenew_HCDEThcDITCIDpeptide_hcd	974	4325	EEQFNSTFR	1460.5348	HexNAc4Hex4	0	1	1	0	1	1
976.0589	3	glycopeptenew_HCDEThcDITCIDpeptide_hcd	999	4372	EEQFNSTFR	1768.6400	HexNAc4Hex5Fuc	0	1	1	0	0	1
989.7346	3	glycopeptenew_HCDEThcDITCIDpeptide_hcd	1008	4388	EEQFNSTFR	1809.6671	HexNAcSHex4Fuc	0	1	1	0	0	1
1389.2213	3	glycopeptenew_HCDEThcDITCIDpeptide_hcd	1014	4397	EEQFNSTFR	3008.1272		0	0	1	0	1	1
908.5300	6	glycopeptenew_HCDEThcDITCIDpeptide_hcd	1016	4403	EEQFNSTFR	4288.6216		0	0	1	0	0	1
819.3401	3	glycopeptenew_HCDEThcDITCIDpeptide_hcd	1025	4418	EEQFNSTFR	1298.4837	HexNAc4Hex3	0	0	1	0	0	1
935.7172	3	glycopeptenew_HCDEThcDITCIDpeptide_hcd	1026	4419	EEQFNSTFR	1647.6149	HexNAcSHex3Fuc	0	0	1	0	0	1
863.3860	5	glycopeptenew_HCDEThcDITCIDpeptide_hcd	1048	4458	EEQFNSTFR	3155.3790		0	0	1	0	1	0
868.0239	3	glycopeptenew_HCDEThcDITCIDpeptide_hcd	1075	4523	EEQFNSTFR	1444.5348	HexNAc4Hex3Fuc	0	1	1	0	1	1
922.0414	3	glycopeptenew_HCDEThcDITCIDpeptide_hcd	1236	4835	EEQFNSTFR	1606.5876	HexNAc4Hex4Fuc	0	0	1	0	0	1
1073.0911	3	glycopeptenew_HCDEThcDITCIDpeptide_hcd	1721	5784	EEQFNSTFR	2059.7364	HexNAc4Hex5FucNeuAc	0	0	1	0	0	1
1017.9345	4	glycopeptenew_HCDEThcDITCIDpeptide_hcd	1749	5838	EEQFNSTFR	2911.1940		0	0	1	0	0	1

Performance

Test Dataset: Human serum sample used for Human Glycoproteomics Initiative Study

- HCD and EThcD acquired on every precursor

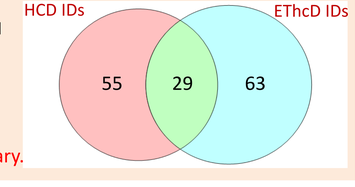
Comparing results for alphabetically first 31 glycopeptides identified:

- 241 spectra extracted; 64 with assigned glycan composition
  - After manual assessment: 52/64 correct
  - Of incorrect, 10 are correct peptide but wrong glycan because wrong precursor monoisotopic mass reported

From extracted spectra I can assign 84 to glycopeptides (by manual analysis/verification)

- 55 of these were not reported in EThcD results

EThcD results identified 92 glycopeptides for these peptides



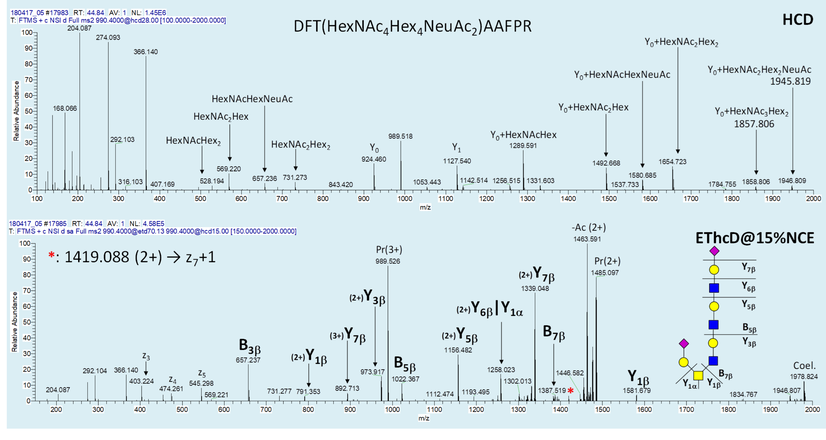
EThcD and MS-Filter driven HCD results are complementary.

Performance on O-Glycopeptide Data

Glycopeptides were enriched from a tryptic digest of human urine samples by LWAC using WGA. The resulting mixtures were analyzed by LC/MS/MS using HCD, and fragment ion-triggered EThcD. This Hepatitis A virus cellular receptor 2 peptide (highlighted in red) was identified from a full database search (SwissProt, human proteins), when only the most common urinary, mucin-type glycans were permitted. We illustrate the power of MS-Filter with data obtained from a single file. For HCD scans to be retained, both Y0 and Y1 had to be present among the 10 most abundant ions, within 10 ppm. MS-Filter allowed the identification of an additional eight glyvoforms of this peptide after manual validation (for example, data representing the glycoform highlighted in green are displayed below).

Peptide	Scan#	RT	Precursor m/z	z	MH+	Delta mass	Potential glycan composition	Δ (ppm)
DFTAAFPR	17477	43.85	770.9904	3	2310.9565	1386.4991	HexNAc3Hex3NeuAc	5.1
DFTAAFPR	17746	44.37	770.9890	3	2310.9525	1386.4950	HexNAc3Hex3NeuAc	2.2
DFTAAFPR	18200	45.30	868.0211	3	2602.0486	1677.5912	HexNAc3Hex3NeuAc2	2.2
DFTAAFPR	18472	45.82	868.0211	3	2602.0486	1677.5912	HexNAc3Hex3NeuAc2	2.2
DFTAAFPR	17651	44.18	778.3441	3	2332.9653	1408.5080	HexNAc3Hex3NeuAcNa	24.0
DFTAAFPR	17106	43.15	892.6960	3	2676.0736	1751.6161	HexNAc4Hex4NeuAc	-4.6
DFTAAFPR	17698	44.28	989.7345	3	2967.1890	2042.7315	HexNAc4Hex4NeuAc2	5.8
DFTAAFPR	17983	44.83	989.7317	3	2967.1805	2042.7231	HexNAc4Hex4NeuAc2	1.7
DFTAAFPR	19642	48.12	624.6002	3	1871.7859	947.3285	HexNAcHexNeuAc2	5.8
DFTAAFPR	17548	44.00	973.6469	5	4864.2033	3939.7459	-	-

Example Glycopeptide Only Identified using MS-Filter



Summary

- MS-Filter has been adapted to be able to look for peak lists containing Y ions for specified peptides
- It is able to suggest glycan compositions based on mass difference between unmodified peptide and observed mass.
- MS-Filter is able to identify many more O-glycoforms of peptides than database searching of HCD (or EThcD) data.
- Analyzing HCD data using MS-Filter provides complementary results to those achieved by EThcD database searching analysis.
- False discovery rate is higher than ideal; still requires manual verification.
- Majority of errors are due to incorrect monoisotopic peak assignment.

Acknowledgements

This work was supported by the Dr Miriam and Sheldon G. Adelson Medical Research Foundation and by the following grants: New Szechenyi Plan GOP-1.1.1-11-2012-0452, the Economic Development and Innovation Operative Programmes GINOP-2.3.2-15-2016-00001, and GINOP-2.3.2-15-2016-00020 from the Ministry for National Economy.