Detecting Glycosylations in Complex Samples

Thorsten Johl, Manfred Nimtz, Lothar Jänsch, Frank Klawonn

Detecting Glycosylations in Complex Samples

Thorsten Johl, Manfred Nimtz, Lothar Jänsch and Frank Klawonn

Helmholtz Centre for Infection Research, Inhoffenstraße 7, 38124 Braunschweig
Thorsten.Johl@helmholtz-hzi.de

Glycoproteins are the highly diverse key element in the process of cell – cell recognition and host – pathogen interaction. It is this diversity that makes it a challenge to identify the glyco-peptides together with their modification from trypsin-digested complex samples in mass spectrometry studies. The biological approach is to isolate the peptides and separate them from their glycosylation to analyse both separately. Here we present an in-silico approach that analyses the combined spectra by using highly accurate data and turns previously established knowledge into algorithms to refine the identification process. It complements the established method, needs no separation, and works on the most readily available clinical sample of them all: Urine.

Keywords: proteomics, glycosylation, mass spectrometry

1 Introduction

Glycosylations are complex post-translational modifications of proteins and the resulting glycoproteins are often found as part of the cell membrane where the glycosylations extend from the plasma membrane into the intracellular space and form the glycocalyx. Due to a high variability in molecular structure, they play a vital role in cell – cell interactions and immune responses [1] comparable to a key that fits only a specific lock, and several pathogens exploit this system to gain entry into specific cell types. Other roles include helping in protein folding, as transport molecule, lubricant or providing frost resistance. Current estimates state that about half of the eukaryotic proteome is glycosylated, making it the most abundant of modifications and one of the most versatile. Glycosylations most commonly attach either to the asparagine (N) amino acid via nitrogen, or to serine (S) or threonine (T) via oxygen and are hence called N- or O-linked glycosylations. These two types are made up of the same basic mono-saccharides, which can be broken down into four groups of identical masses (**Fig. 1** A). These basic components are arranged into link type-dependent core structures of highly variable overall complexity (**Fig. 1** B). This complexity hampers mass spectrometric (MS) analyses of glycosylated samples, because MS relies on a known total peptide mass for identification before it compares the most intense signals of a fragmentation pattern (MS^2) of this peptide to a number of theoretical signals produced by known peptides of similar mass from a reference database [2]. The best

AIAI, p. xx, 2012.

fitting spectrum is then assumed to be the correct interpretation. But these theoretical spectra do not contain any modification unless each is specified, and to predict all possible glycosylation patterns would soon reach incalculable complexity.

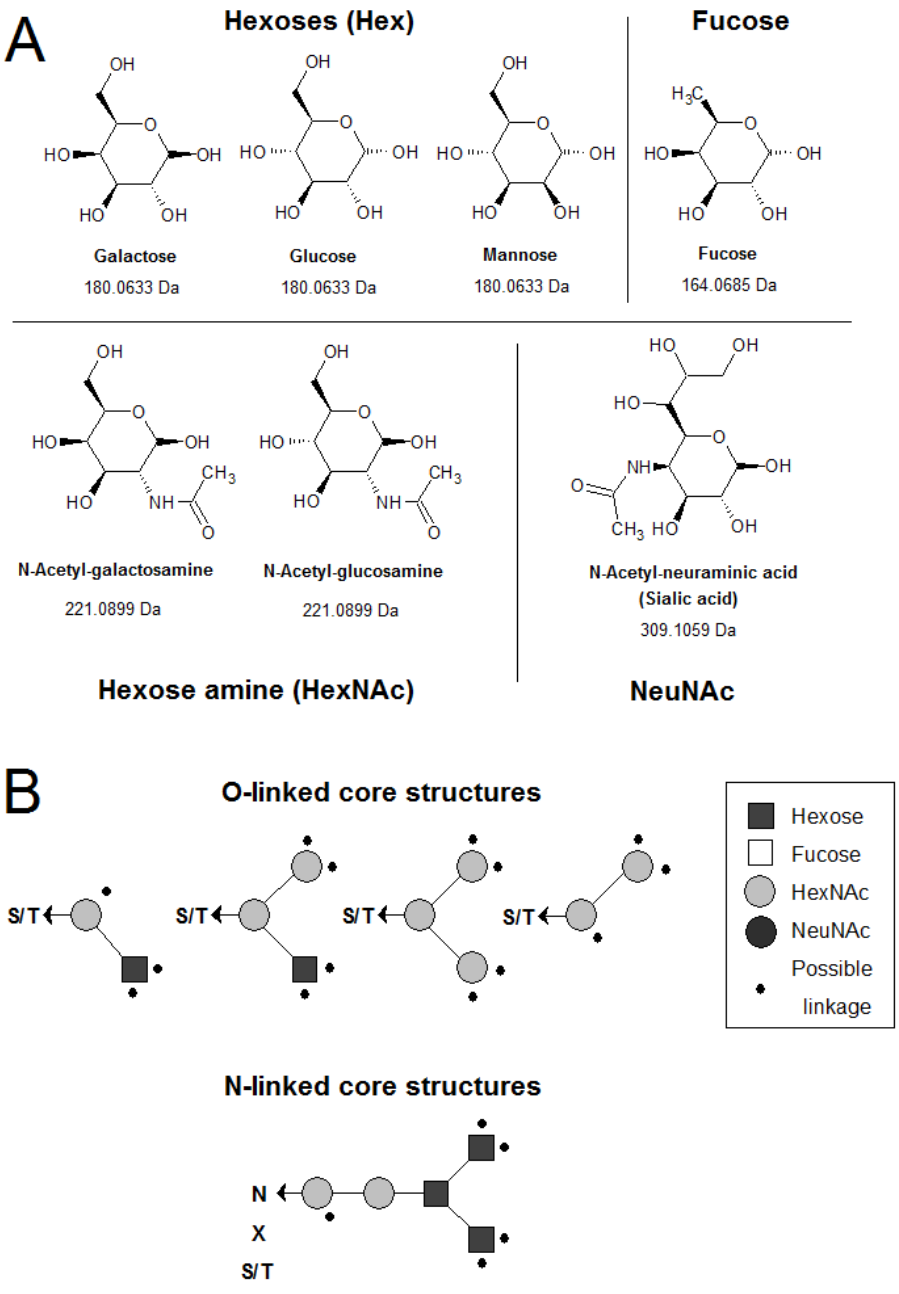

Fig. 1. Overview of saccharide makeup und structure

The general accepted solution to this problem is to separate the oligo-saccharide side chain from the peptide by using endoglycosidases specific to N-linked, and strong bases for both side chains [3] [4]. The resulting components are then analyzed separately, with a single mono-saccharide remaining attached to the peptide, which identifies the modified amino acid in MS experiments. This method works best on samples of low complexity where it is clear which oligo-saccharide originated from which peptide, but it is unsuited for large-scale proteomic studies. Yet glycosylations are ideally suited to be analyzed by MS workflows. First of all, they are readily ionized and thus result in good signal to noise ratios in the mass spectra. Secondly, the mono-saccharides produce distinct signals in the low mass region of such a spectrum. These identify spectra of glycosylated peptides with absolute certainty. And thirdly, the signals stemming from amino acid and saccharide residues in conjunction have a characteristic atomic mass that sets them apart from pure amino acid residues of similar mass. The reason for this is the greater amount of oxygen contained in saccharides which weighs in at just under sixteen Dalton, while every other atom prevalent in amino acids has a decimal value just above a full Dalton (C,H,N,S) [5]. It will be these criteria that are exploited by the program introduced herein in order to identify the peptides and their glycosylations from the same spectrum. This paper will demonstrate the feasibility of this new approach that does not rely on a reference database for glycosylated peptides like UniProt[14] or Glycosuite[15].

2 Methods

2.1 Sample preparation and acquisition

3x 2.5ml urine were desalted in turn with a PD-10 column, united, and then concentrated to 400µl with a speedvac and prepared according to the "glycoprotein isolation kit, WGA" from Thermo Scientific with minor adjustments. Incubation time was increased to 30min, the elution time to 15min. Subsequent digestion was performed over night with 20µg trypsin. Two MS measurements were performed in turn, one injecting 1µl onto the loading column of a Dionex Ultimate 3000 HPLC, the other 3µl. The samples were separated over 60 min on a reverse phase LC at 350µl/min flow rate with 80% Acetonitrile as mobile phase. Ions were introduced into the Thermo Fisher MS Orbitrap Velos device via an electrospray ion source. The FTMS Orbitrap was used as the detector for all scans. Survey scans were performed at a resolution of 60000 in the range of 700 to 2000 and the 5 most intense signals were chosen with an isolation width of 4 Da for MS^2 fragmentation in the HCD collision cell at normalized collision energy of 40 for 40ms. The minimum signal threshold was set at 10000 and the default dynamic exclusion list was enabled, removing signals for 90 sec from the selection process after having been measured twice. Concluding MS^2 scans were performed at a resolution of 7500 and a range of 100 to 2000. All preprocessing steps like baseline estimation and noise removal were performed by Xcalibur 2.7 at the time of acquisition with the device inherent parameters.

2.2 Program setup

The program was created using Java programming language from Oracle and Eclipse as programming environment. Readw.exe [6] from the trans-proteomic pipeline [7] was used in conjunction with files from ThermoFinnigan for centroidization and to convert Thermo Fisher RAW files into mzxml. Mzxml in turn was converted to mgf using JRAP [8]. Alternative mzML parsing is provided by jmzML [9]. Settings are stored as xml file using the Apache commons parser [10]. The program requires the JRE and has been tested on Windows 32 and 64bit as stand-alone application. Note that the readw.exe only works on 32bit windows operating systems, and thus 64bit machines can only process mzxml, mzml and mgf files.

2.3 Spectra preparation and identification of glycosylated spectra

All signals of a deconvoluted spectrum were checked for isotopic C13 satellites allowing for 20ppm mass tolerance, and the appropriate charge state was deduced from the distance between the isotopes. All following steps were then performed on the mono-charged mass of the most intense isotopic signal. All spectra were searched in turn for any two of the following marker oxonium-ion masses: dehydro-N-acetyl hexose (204.079 Da), a dehydro di-saccharide of N-acetyl hexose and a hexose (366.132 Da) or di-dehydro sialic acid (274.085 Da). The search was performed with a mass tolerance of 20ppm and a minimum marker signal intensity of 10% of the most intense signal in the spectrum. Every signal that was found during identification and, if present, an additional mono-dehydro sialic acid (291.095 Da) signal were marked as signals of saccharide origin. All spectra that met the identification criteria were then considered glycosylated and passed on to the following processing steps.

2.4 Signal makeup

Any mass from top to bottom, beginning with the total mass of the modified peptide, was then considered as the starting point of a tree of saccharide masses. In order to qualify as such a seed, all mono-saccharide masses from **Fig. 1** A where in turn checked to find a smaller signal in the estimated mass range of starting signal minus saccharide mass. The steps are repeated from these seeds onward until no further signals are found that can be explained by the loss of a saccharide. There are additional criteria that influence the signal selection process:

Firstly, the peaks are only considered if the mass delta between the actual and expected decimal fractions of a peptide (Mi) [5] increases from higher peak (Pi) mass to lower peak mass (Pi+1), denoting the loss of an oxygen-rich monosaccharide from the remaining molecule. The logic allows for a small mass-dependent inaccuracy (Δ) of 15ppm (1).

$$f(x) = \begin{cases} true, & M_i - P_i < M_{i+1} - P_{i+1} - \Delta \\ false, & otherwise \end{cases} \quad (1)$$

Secondly, the program allows for a shift in the most abundant isotope form, searching for a second peak at +/- 1.00335 Da distance with 15ppm tolerance if no primary peak can be found. Such shifts are marked, and block a further shift into the same direction, which is highly unlikely given the standard distribution of C13 isotopes. It may, however, shift back to the original isotope pattern once to accommodate for inaccuracies in the measured intensities.

Thirdly, signals without an isotope pattern that denote the charge state are considered to be of any charge state up to the charge state of the mother mass. Such signals are called jokers, and may be assigned only one charge during the enlargement of the saccharide trees. Signals that pass all criteria and thus carry a glycosylation are marked and the saccharide tree is stored for further analyses.

2.5 Typecasting spectra

The kind of glycosylation found in any given set of spectra can be determined depending on the completeness of the saccharide tree. Four different core structures as introduced in **Fig. 1** B and three further N-linked sub types are known to be expressed in human tissue. At least four saccharides from the start of the tree are necessary to discern between N and O-linked glycosylations. The more complex N-linked core structure is checked first. This is done by a pattern recognition that performs a depth-first traversal of the saccharide tree and checks first if there are one or two leading HexNAc saccharides, and then if there are three consecutive hexoses following that. The traversal for any sub-tree is aborted if the child does not match expectations, and the next child is considered until the pattern is found, or no children are left. There is some allowance for a single extra saccharide attached to the leading HexNAc, and either the last hexose or the first and its connected HexNAc may be substituted by a Hex-HexNAc di-saccharide.

If the algorithm fails to find the key elements for an N-linked core structure, it checks the less complex O-linked core structures in the following order: 2 HexNAc and a hexose, 3 HexNAc, 2 HexNAc, HexNAc and a hexose. The leading HexNAc may be missing in all cases.

2.6 Determining true peptide mass

All signals that have lost a saccharide are removed from the spectrum. The remaining signals are then considered from top to bottom. If they have a relative intensity of 5% of the most intense signal ever present in the spectrum, the algorithm tries to fit any combination of saccharide masses into the mass delta of this signal and the total mother ion mass. This is achieved by beginning with just one saccharide, trying each in turn before adding a second saccharide and checking all possible mass combinations of these. The method stops adding more saccharide elements once it has found one combination that fits, or multiple combinations of the smallest mass have become larger than the mass delta it was trying to fit. Signals that can be explained by the loss of a combination of saccharides are stored. All spectra are then grouped by their

mother mass and charge state, and the stored signal intensities are added up. The four most intense combined signals are then chosen in turn and proposed as the true peptide mass and new mother mass of the spectrum retaining the charge the spectrum originally carried, resulting in up to four new spectra for each that was identified as containing saccharides. The spectra are then stored separately and can be submitted to the Mascot server for identification.

2.7 Mascot search

The Mascot search (V 2.3.02, Matrix Science) was submitted via Mascot Daemon against a UniProt database (Release 2012_02) restricted to human proteins with a peptide mass tolerance of 20ppm, a fragment mass tolerance of 0.4 Da and no missed cleavages. Mother ion charge was limited to up to 4 charges. Allowed modifications were Carbamidomethyl (fixed) and Oxidation (M) (variable).

3 Results

8615 and 9577 spectra, respectively, were processed from the 1µl and 3µl samples. About a third of these were identified as glycosylated and resulted in more than 11000 proposed spectra per sample (**Table 1**). Roughly 34% of all spectra were found to carry a glycosylation. Only 5% of these could be identified as carrying an N-linked glycosylation tree. The majority is estimated to carry the less complex O-linked glycosylation. Less than 20% resulted in an unknown type of glycosylation. The final number of spectra proposed to Mascot by the program is nearly equal to four times the amount of glycosylated spectra, meaning that almost every spectrum had enough signals that qualified it as one of four new possible true peptide masses. Only 13 spectra in total could not be assigned a new mass and thus resulted in no new spectra (not shown).

Sample	Original Spectra	Glycosylated	N-linked	O-linked	Unknown	Proposed Spectra
1µl	8615	3086	139	2439	508	11598
3µl	9577	3255	207	2475	573	12328

Table 1. Overview of glycosylated Spectra

The results of the identification by Mascot are summarized in **Table 2**. The identity score cut offs for the 1µl and 3µl samples were 19 and 20, respectively, and the significance (homology) threshold was a p-value of 0.05. The false discovery rate (FDR) of Mascot was 5.13 (7.5) above identity (above homology) for the 1µl sample, and 1.83 (5.17) in the 3µl sample. The table shows all proteins that were found in both samples by Mascot with their respective score. Shown next are the number of matching spectra that were associated with this protein in total and the number of unique sequences. The affirmed glycosylations are the number of unique sequences already known and found in this analysis. The number in parenthesis for all three groups is

the number of hits above homology. The numbers of predicted and known glycosylations are taken from the UniProt entry and show the total number of listed glycosylations followed by the number of confirmed ones in parenthesis. Asterisks mark entries where more than one glycosylation occurs on the same tryptic peptide and the location is not unambiguous. Values were taken from the protein overview tab of Mascot and UniProt Release 2012_02.

Uniprot Name	Score		Matches		Sequences		Affirmed Glycosylations		Uniprot
	1µl	3µl	1µl	3µl	1µl	3µl	1µl	3µl	
HEG1	140	326	22(7)	27(17)	5(2)	5(3)	1(0)	1(0)	9(3)
CSPG2	179	267	17(7)	21(13)	4(2)	6(3)	0	0	23(1)
CSF1	275	137	14(9)	8(7)	2(1)	1(1)	1*(0)	0	4(2)
X3CL1	45	263	14(3)	26(16)	3(1)	4(3)	2(1)	3(2)	4(3)
EGF	140	125	9(6)	9(5)	3(1)	2(1)	0	0	10*(1)
FBLN2	188	40	8(6)	3(3)	2(1)	1(1)	0	0	3*(1)
YIPF3	102	89	6(5)	4(3)	1(1)	1(1)	1(1)	1(1)	1(1)
NCAN	79	108	9(4)	5(4)	2(1)	2(1)	0	0	4(0)
NID1	46	112	5(2)	14(7)	3(2)	3(2)	0	0	1(0)
IGF2	59	70	4(3)	11(7)	1(1)	1(1)	2*(2)	2*(2)	3(3)
PGCB	107	21	3(3)	1(1)	1(1)	1(1)	0	0	4(0)
FA5	79	41	4(4)	2(2)	1(1)	1(1)	0	0	26(5)
APOF	45	57	3(2)	2(2)	1(1)	1(1)	1(1)	1(1)	2(1)
CF072	26	48	2(1)	2(2)	1(1)	1(1)	1(1)	1(1)	4(0)
VASN	18	52	1(1)	2(2)	1(1)	1(1)	0	0	5(3)
IC1	46	22	5(2)	4(1)	1(1)	1(1)	2*(2)	2*(2)	15(0)
GOLM1	30	30	4(2)	4(3)	2(1)	1(1)	0	0	3(1)
P3IP1	25	29	4(2)	4(2)	1(1)	1(1)	1(1)	1(1)	2(1)
OSTP	27	23	5(3)	6(3)	2(1)	3(3)	0	0	7(7)

Table 2. Comparison of identified and known glycosylations

The 3µl sample proved to be the more reliable and was consecutively used for further analyses of the saccharide trees and glycosylation sites. 7 of the identified 19 proteins were detected with at least one peptide with a known glycosylation site above homology. All of these sites, except the one from CF072, were related to O-glycosylations. All of them were identified by the program as O-glycosylated at least once per sequence.

Three spectra are shown in **Fig. 2**. They exemplify different O-glycosylations that were correctly identified by the program, and demonstrate the potential to reliably

identify the glycosylation site from glycosylated spectra. The notable exception is CF072_HUMAN, which is predicted to have an N-linked glycosylation at the identified site 191, but was identified as O-type. The identified spectra in question only contain unlinked Hex-HexNAc and Hex, and thus do not permit the direct identification of an N-linked glycosylation, although the peptide could be identified with an adjusted mono-charged mother mass of 1461.77 Da (down from 1206.06, z=2).

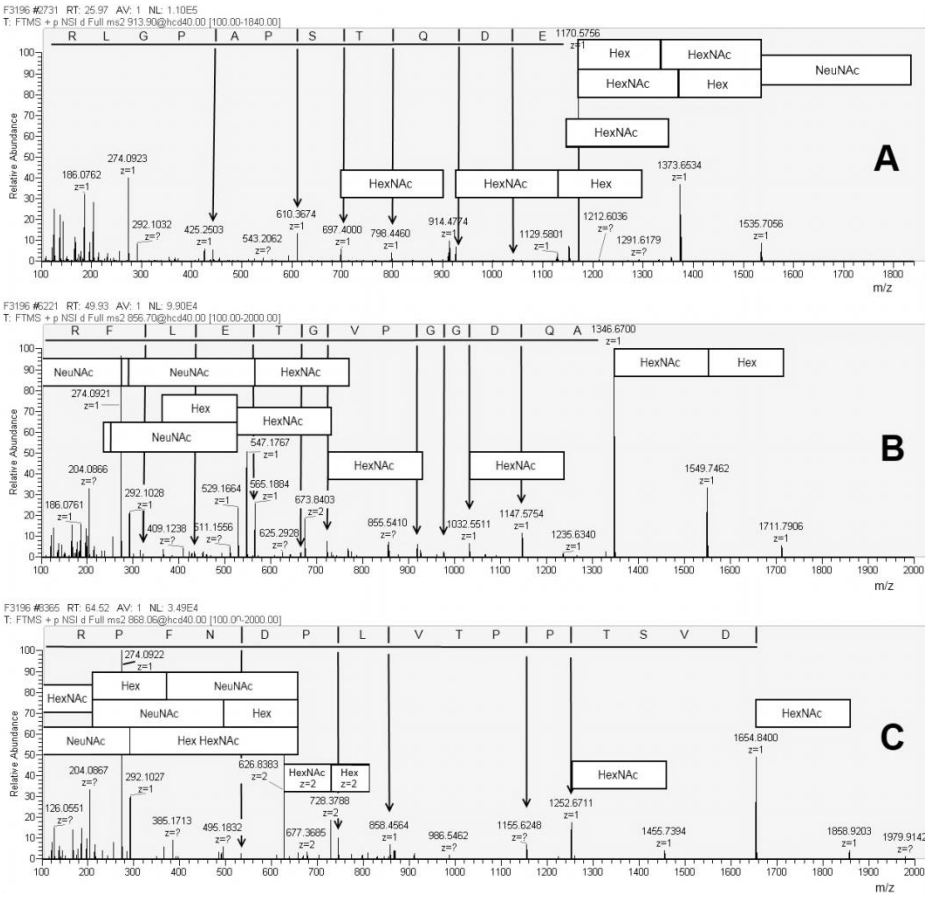

Fig. 2. Saccharide tree and peptide y-ion series

(A) Is taken from P3IP1_HUMAN. A consecutive saccharide tree links the total mass of the mother ion over three steps with the true peptide mass of 1170.58 Da. Three further saccharide trees were found that stem from different dissociation events. One of them is linked directly to S39, the reported glycosylation site. Spectrum (B) shows a glyco-peptide that was associated with site T183 of Protein X3CL1_HUMAN. The lower half of the spectrum shows several saccharide ions from the b-series, in one case extending the mass 274 used for identification by NeuNAc and HexNAc. The true peptide mass 1346.67 Da is again linked to the total mother mass. (C) Shows two

saccharide trees in the lower mass spectrum. They originate from the same fragment, which dissociated from the peptide and were found complementary forward and backward. Note the double charged di-saccharide that was found and extends the mono charged saccharide that is attached to 1252.67 Da. Neither the total mother mass nor the actual site is directly attached to a saccharide. It was nonetheless identified as IGF2_HUMAN, which could be modified either at T96 or at T99. The HexNac-Hex tree connected to proline (P) suggests that it is site T96. Spectra were taken from Thermo Fisher Xcalibur – QualBrowser [13].

4 Discussion

This paper describes a new approach to analyze peptides together with their glycosylation. As the current standard is to analyze both separately, there is just one other project that uses a similar approach [12]. That approach uses the ProteinScape 3.0 and GlycoQuest programs in conjunction. It compares the unadjusted glycosylated spectra with a list of theoretical spectra of dissociated glyco-peptides, very similar to the database-based approach performed by Mascot. Although this method is less likely to provide false positives, it will not be able to identify new glycosylations. Indeed the high number of predicted glycosylations apparent from **Table 2**, especially with regards to N-linked glycosylations, demonstrates there is need for an approach similar to de-novo sequencing for peptides.

The example of an unidentified N-linked glycosylation described above demonstrates the problems when identifying large saccharide trees in combination with large peptides in mass spectrometric studies. The variety of possible dissociation events and multiple fragmentations within the tree means that no specific fragment reaches a detectable quantity in the highly accurate, but not as sensitive Orbitrap detector [11]. This results in detectable amino acids and mono-saccharides, but no larger consecutive saccharide elements. Peptide identification itself is based largely on correct total mass and some preferred dissociation locations like proline. The situation may be improved by allowing for di- and tri-saccharides during peak analysis, however, this could decrease the specificity by increasing the chance of false positives. This problem may, in turn, be addressed by using collision-induced dissociation (CID), which generates only two fragments per ion and thus can generate a more gradual distribution of signals. The results may be even better when coupled to the more sensitive but less accurate ion trap detector. On the other hand, this would produce even less pure amino acid ions, resulting in less ion intensity. As is, many of the 3000+ original spectra did not lead to peptide identification, even when allowing four possible interpretations of the remaining non-glyco signals. The reason for this may be found in the more readily ionized saccharides. The remaining amino acid part is less likely to be ionized, provides weaker signals when fragmented, and eventually vanishes into the background noise. A possible solution to this is to perform a third dissociation and detection cycle (MS^3) of the most likely true peptide mass. The evident workflow for this would be as follows:

After the first data acquisition cycle, the peptides would have to be analyzed by the program and the mother mass, estimated true peptide mass and retention time recorded. A second data acquisition cycle would follow the first, using the recorded data as an inclusion list for MS^3 experiments. The results would then be sent to Mascot for identification. It is, however, presently unclear whether such a general approach is possible and would result in spectra of the required intensity.

As demonstrated above, the basic principle works. The program uses a shotgun approach to reliably identify O-linked glycosylations and their site – but further development is necessary to extend this to N-linked glycosylations, which is already underway at the time of this writing.

The program is available from the author upon request.

5 References

1. K. Marino, J. Bones, J. Kattla and P. Rudd: A systematic approach to protein glycosylation analysis: a path through the maze. *Nat Chem Biol,* 6, pp. 713-723, 2010.
2. D. Perkins, D. Pappin, D. Creasy and J. Cottrell: Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis,* 20(18), pp. 3551-3567, 1999.
3. T. Patel, J. Bruce, A. Merry, C. Bigge, M. Wormald, A. Jaques and R. Parekh: Use of hydrazine to release in intact and unreduced form both N- and O-linked oligosaccharides from glycoproteins. *Biochemistry,* 32, pp. 679-693, 1993.
4. S. Pan, R. Chen, R. Aebersold and T. Brentnall: Mass Spectrometry Based Glycoproteomics - From a Proteomics Perspective. *Mol Cell Proteomics,* 10(1), 2011.
5. E. Dodds, H.J. An, P.J. Hagerman and C.B. Lebrilla: Enhanced peptide mass fingerprinting through high mass accuracy: Exclusion of non-peptide signals based on residual mass. *J Proteome Res.,* 5, pp. 1195-203, 2006.
6. ReadW, http://tools.proteomecenter.org/wiki/index.php?title=Software:ReAdW.
7. P. Pedrioli: Trans-proteomic pipeline: a pipeline for proteomic analysis. *Methods Mol Biol.,* 604, pp. 213-38, 2010.
8. JRAP Extended, http://javaprotlib.sourceforge.net/packages/io/jrap/.
9. JMZML, http://code.google.com/p/jmzml/.
10. The Apache XML Project, http://xerces.apache.org/xml-commons/.
11. M. Scigelova and A. Makarov: Orbitrap Mass Analyzer - Overview and Applications in Proteomics. *Practical Proteomics,* 2, pp. 16-21, 2006.
12. K. Neue, A. Kiehne, M. Meyer, M. Macht, U. Schweiger-Hufnagel and A. Resemann: Straightforward N-glycopeptide analysis combining fast ion trap data acquisition with new ProteinScape functionalities. http://www.bdal.com/library/literature-room/detail-view/article/straightforward-n-glycopeptide-analysis-combining-fast-ion-trap-data-acquisition-with-new-proteinsca.html.
13. Xcalibur Thermo Scientific. http://www.thermoscientific.com/ecomm/servlet/productsdetail_11152_L11240_80588_11961721_-1. [20 04 2012].
14. UniProt, http://www.uniprot.org/
15. C.A. Cooper, H.J. Joshi, M.J. Harrison, M.R. Wilkins, N.H. Packer: GlycoSuiteDB: a curated relational database of glycoprotein glycan structures and their biological sources. 2003 update. *Nucleic Acids Res.* 31(1), pp. 511-3, 2003