universität
wien

MASTERARBEIT

Titel der Masterarbeit

„FindGlycoPeptides – an Open Source Program for High-Throughput N-Glycopeptide Identification in Large LC-MS/MS Data Sets“

verfasst von

Nikolaus Voulgaris BSc

angestrebter akademischer Grad

Master of Science (MSc)

Wien, 2015

Studienkennzahl lt. Studienblatt:	A 066 862
Studienrichtung lt. Studienblatt:	Masterstudium Chemie
Betreut von:	Univ.-Prof. Dr. Andreas Rizzi

Danksagung

Zunächst möchte ich mich bei Univ.-Prof. Dr. Andreas Rizzi herzlich bedanken, der mir diese Masterarbeit ermöglicht hat und mich ermutigt hat das Thema selbstständig zu erarbeiten, meine Fähigkeiten einfließen zu lassen und neue Wege zu gehen.

Weiters danke ich Dipl. Nat. Claudia Michael, die mir die praktischen Aspekte der Glykoanalytik näher brachte sowie meinen weiteren Laborkollegen und den Kollegen im MSC, allen voran Anna Fabisikova MSc, die mich bei den Messungen unterstützt haben und bei technischen Problemen weiterhelfen konnten.

Erwähnen möchte ich auch Univ.-Prof. Dr. Christoph Flamm und Univ.-Prof. Dr. Christopher Gerner, die mich inspiriert haben mir Fähigkeiten im Bereich des Programmierens anzueignen.

Abschließend bedanke ich mich herzlich bei meinen Eltern, die mich auf jede erdenktliche Weise unterstützt haben sowie meinem Bruder, mit dem ich mich über viele technische Aspekte von Perl austauschen konnte.

Abstract

Advances in mass spectrometric (MS) techniques made analysis of protein glycosylation on glycopeptide level feasible, but data analysis is a severe bottleneck. In this work, an open source program for high-throughput identification of glycopeptides in large LC-MS/MS data sets, *FindGlycoPeptides* (*FGP*), was developed, using the *Perl* programming language. *FGP* matches the peptide portion based on the the Y-series ions of low accuracy CID spectra. It calculates decoy based false discovery rate (FDR) estimates, uses an empirical scoring function to rate the assignments and provides semi-quantitative information by spectra counting. The program runs on various operating systems and uses the open MS data format mzXML, allowing the analysis of data originating from different instruments.

Test data sets of tryptic digests of several standard glycoproteins, focusing on bovine α1-acid glycoprotein (AGP), mixtures of these and biological samples, derived from MCF-7 cell supernatant, were acquired using nano-RP-LC-Orbitrap-MS/MS with standard proteomics methodologies. Up to 1500 spectra could be assigned in a single run of a bovine AGP digest, covering more than 100 distinct glycopeptides that could be verified. The performance was compared to a similar freely available program, *GlycoPeptideSearch*, and demonstrated to be supererior both regarding number of hits and false discovery rate. The spectra counts of the various glycopeptide species were compared to the peak heights and integrals and provided similar results, which were obtained with no additional effort.

Analysis of other single glycoprotein digests (human AGP, bovine Fetuin and Asialofetuin, rabbit IgG and chicken Ovalbumin) yielded fewer assignments, since the experimental methodologies were not optimized for glycoproteomics. Assignment rates with digests of glycoprotein mixtures were similar to those of the single proteins. However, *FGP* is suitable mainly for targeted glycoproteomics, as with very complex samples such as the SDS-PAGE fractioned cell supernatants with more than 20 possible glycoproteins, it fails to provide useful results, due to the vast peptide search space. In such cases additional information of the peptide sequences must be incorporated, e.g. by ETD spectra.

Zusammenfassung

Fortschritte im Bereich der Massenspektrometrie ermöglichen die Analyse der Glykosylierung von Proteinen auf Glykopeptidebene. Ein Schwachpunkt ist aber weiterhin die Datenverarbeitung. In dieser Arbeit wurde ein open source Programm für die automatische Identifikation von Glycopeptiden in großen LC-MS/MS Datensätzen entwickelt. *FindGlycoPeptides* (*FGP*), welches in *Perl* geschrieben wurde, identifiziert den Peptidteil anhand der Y-Ionen in CID-Spektren niedriger Massengenauigkeit. Die Zuordnungen werden mittels einer empirischen Bewertungsfunktion gewertet und die False Discovery Rate wird über die Zuordnung von Decoypeptiden abgeschätzt. Weiters zählt das Programm die jeder Spezies zugeordneten Spektren und liefert dadurch semiquantitative Informationen. Das Program läuft auf unterschiedlichen Betriebssystemen und benutzt das offene Datenformat mzXML, womit Daten unterschiedlicher Gerätehersteller analysiert werden können.

Als Testdatensätze wurden verschiedene Glykoproteine, mit dem Fokus auf bovinem saurem α1-Glykoprotein (AGP), einzeln und in Mischungen sowie biologische Proben, die aus MCF-7 Zellüberständen gewonnen wurden, mit Trypsin verdaut und den Standardmethoden der Proteomik entsprechend Standardmethoden mit nano-RP-LC-Orbitrap-MS/MS gemessen. Bis zu 1500 Spektren konnten in einem AGP Datensatz zugewiesen werden, die mehr als 100 unterschiedliche Glycopeptide abdeckten, welche durch manuelle Überprüfung bestätigt werden konnten. Die Ergebnisse wurden mit deinen eines ähnlichen, frei zugänglichen, Programms, *GlycoPeptideSearch*, verglichen, wobei sich zeigte, dass *FGP* sowohl in Bezug auf die Anzahl der Identifikationen als auch der False Discovery Rate deutlich leistungsfähiger war. Die Anzahl der jedem Glykopeptid zugeordneten Spektren wurde mit der maximalen Peakhöhe sowie dem Integral verglichen. Es zeigte sich, dass die drei Methoden zu ähnlichen Ergebnissen führten, wobei das Zählen der Spektren keines zusätzlichen Aufwandes bedurfte.

Die Analyse von HPLC-MS/MS Datensätzen von Verdauen anderer einzelner Glykoproteine (humanes AGP, bovines Fetuin und Asialofetuin, Hasenimmunoglobulin G und Hühnerovalbumin) brachte weniger Zuordnungen, da die experimentelle Methodik nicht für Glykoproteomik optimiert wurde. Ähnlich viele Zuordnungen wurden bei der Analyse von Mischungen aus fünf Glykoproteinen gefunden. Allerdings liegen die Stärken von *FGP* vor allem im Bereich der zielgerichteten Experimente (targeted glycoprotemics). Bei sehr komplexen Proben, wie Zellüberständen, die mit SDS-PAGE fraktioniert wur-

den, in denen 20 oder mehr mögliche Glykoproteine vorkommen, können keine brauchbaren Ergebnisse erzielt werden, da der Peptidsuchraum zu groß ist. Bei solcherlei Proben müssten zusätzliche Informationen über die Peptidsequenzen herangezogen werden, zum Beispiel durch ETD Spektren.

Contents

Part I.

Theory

1. Theoretical Background

1.1. Protein Glycosylation

Glycosylation is the most complex and one of the most frequent posttranslational modifications (PTMs) of proteins in eukaryotes. Enzymatic glycosylation of proteins can be divided into three groups. With O-glycosylation the hydroxy-groups of serine or threonine are modified. The attached carbohydrate chain is typically rather short and unbranched and there is no strict common core structure. N-glycans in contrast, which are linked to the amide nitrogen of asparagine in context of the consensus sequence Asn-Xxx(not Pro)-Ser/Thr (or very rarely Cys), are branched oligosaccharides, with a core structure consisting of three mannose and two N-acetyl-glucosmanine residues. Besides those two common types of enzymatic glycosylation there is the more exotic C-glycosylation of tryptophan, which consists of a single mannose unit only.

Glycosylation serves various biological functions. Most obvious is the increase in solubility, caused by the physicochemical proper

ties of oligosaccharides. This effect is particularly pronounced when charged sugars like sialic acids are present. Attached oligosaccharides are further important in preventing enzymatic proteolytic degradation of the protein and they affect the local polypeptide structure, aiding in this way the formation of a correct fold. Furthermore, N-gylcosylation plays an important role in the endoplasmatic reticulum (ER) protein quality control via the calnexin/calreticulin cycle [1], and it is involved in protein trafficking and cell signaling [2].

N-linked Glycosylation Of the three families of enzymatic glycosylation, the N-linked type is characterized in most detail. The occurence in the context of the Asn-Xxx-Ser/Thr consensus sequence allows the prediction of possible glycosylation sites. All N-glycans have the same pentasaccharide core structure. Furthermore, there are enzymes that hydrolyze the N-glycosidic bond between the amide-N of asparagine and the glycan, releasing the intact oligosaccharide. These features facilitate the analysis.

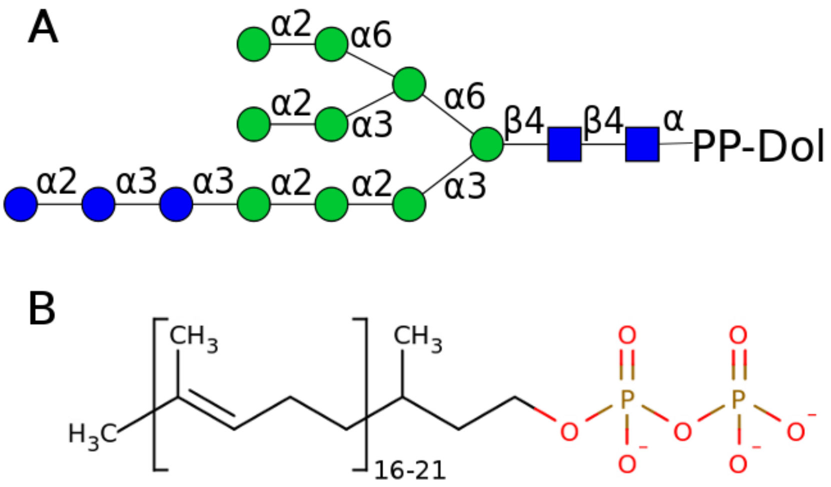
N-glycans are assembled in the endoplasmatic reticulum (ER)[3] as membrane bound tetradecasaccharide with the composition $(\mathrm{Glc})_3(\mathrm{Man})_9(\mathrm{GlcNAc})_2$, which is attached to dolichol pyrophosphate, a polyprenol. This pre-assembled oligosaccharide structure, which is shown in Fig. 1.1, is transferred to the nascent polypeptide chain by the oligosaccharyltransferase complex. The glucose residues are cleaved off inside the ER if the protein is correctly folded, and this process induces exportation from the ER. Before exportation one mannose unit might be cleft off as well, yielding N-glycans with the composition $(\mathrm{Man})_8(\mathrm{GlcNAc})_2$. This initial fabrication pathway is conserved in all eukaryotes [3]. Symbolic representation of monosaccharides and the abbreviations used in this work are listed in Table 1.1 .

Table 1.1.: Symbolic Representation of Monosaccharides

Symbol	Monosaccharide	Abbreviation	Letter
🟡	Galactose	Gal	H*
🔵	Glucose	Glc	H*
🟦	N-Acetylglucosamine	GlcNAc	N
🟢	Mannose	Man	H*
◆	N-Acetylneuraminic acid	Neu5Ac	A
◇	N-Glycolylneuraminic acid	Neu5Gc	G
🔺	Fucose	Fuc	F

The one letter codes will be used in this work in context of mass spectrometry. With the experimental methodology employed the different hexoses can not be discriminated. The letter S denominates any sialic acid (A or G).

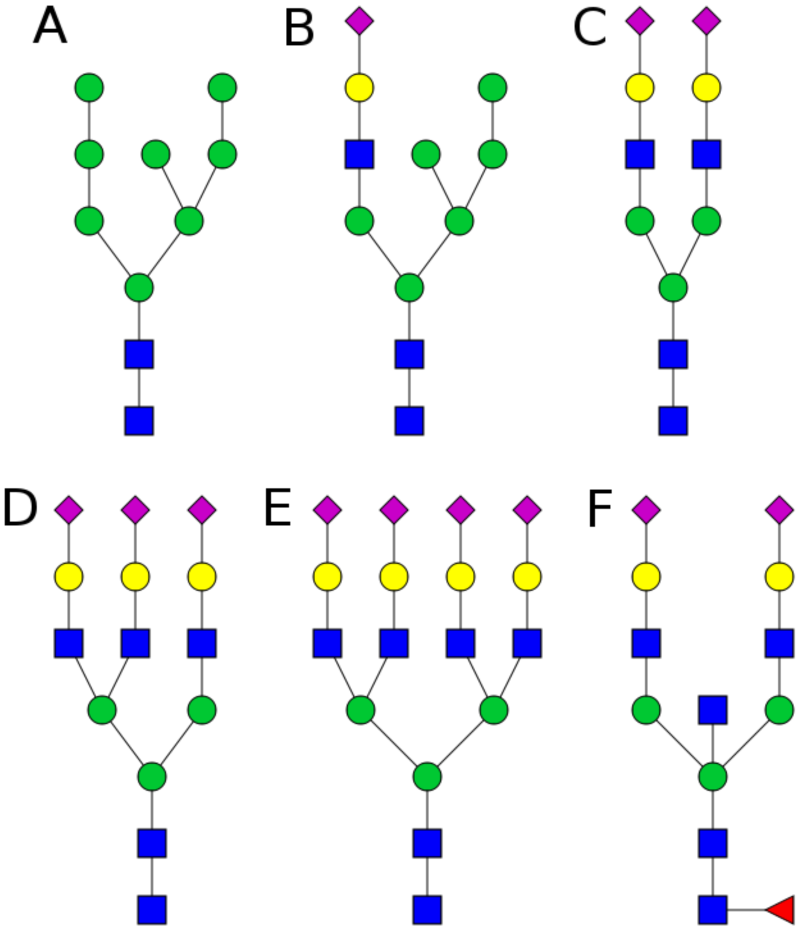
Figure 1.1.: Tetradecasaccharide Precursor



A The tetradecasaccharide precursor $Glc_3Man_9GlcNAc_2$, which is assembled at the ER membrane, attached to dolichol pyrophosphate; **B** dolichol pyrophosphate

The glycans are further processed in the Golgi apparatus. First, the mannose units are cleaved off by Golgi mannosidases until only the pentasaccharide core structure remains, a process which is known as trimming. Afterwards, different sugars are attached to the trimmed core. In humans, the most common sugars in N-glycans are galactose, N-acetylglucosamine, fucose and N-acetylneuraminic acid. In other mammals, N-glycolylneuraminic acid is frequently found in N-glycans. Humans have no active enzymes for the synthesis of this sialic acid, but it can be incorporated from external sources[4]. This is especially pronounced in cancer cells which have a high demand for saccharides due to their increased metabolism. Based on the extent of removal and addition of sugars in the Golgi apparatus (trimming), N-glycans can be categorized into three types. High-mannose types are composed of mannose only, except the initial two core GlcNAc. Usually the number of Man residues is less than 8 in most higher eukaryotes, whereas in yeast a high number of mannose units is added in the Golgi, yielding mannan oligosaccharides. In complex type glycans both branches were trimmed to the core and subsequently mounted with any number of the sugars available for the enzymatic machinery. In mammals the attachment of lactosamine (GlcNAc-Gal) repeats is common with up to two arms on each branch, yielding bi- , tri- or tetraantennary glycans, with one or more lactosamine repeats on each antenna, which are often terminated by sialic acids. Monoantennary N-glycans, where additional sugars are attached only to one arm of the pentasaccharide core, are rarely observed, for instance in human chorionic gonadotropin, as reported by Valmu et al [5]. Frequently found features are also fucosylation, often at the first GlcNAc (core fucosylation) and sometimes bisecting N-acetylglucosamines. In hybrid oligosaccharides one of the core branches is trimmed completely, with subsequent reattachment of complex-type oligosaccharides, while the other branch still bears some initial mannose residues. An overview of the different types of N-glycans is shown in Fig. 1.2 .

Figure 1.2.: Glycan Types



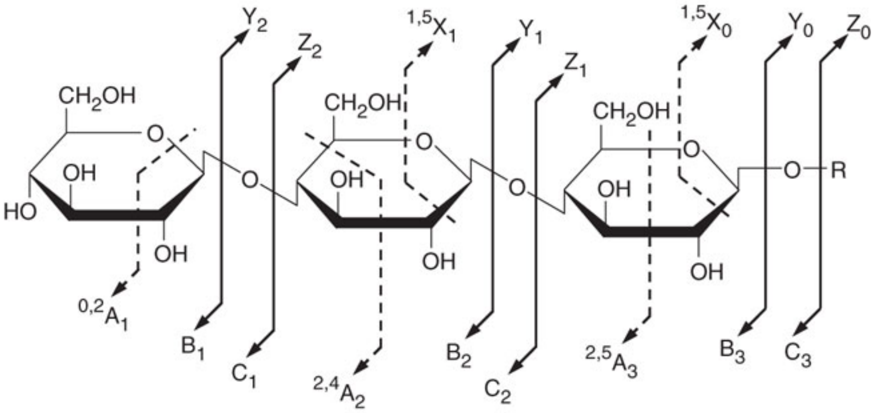
A high-mannose, **B** hybrid, **C** complex biantennary, **D** complex triantennary, **E** complex tetraantennary, **F** complex biantennary with core fucosylation and bisecting GlcNAc

Glycosylation Analysis Protein glycosylation analysis can be performed on different levels, each supplying different information [6]. Analysis of intact glycoproteins gives insight to the extent of glycosylation, analysis of the free glycans gives detailed structural information like branching isomerism and linkage positions while from the glycopeptide level information about site-specificity of multiply glycosylated proteins can be inferred.

Classic methods of glycoprotein analysis rely either on specific binding to various different lectins or antibodies or on the specific cleavage by different glycosidases followed by chromatographic separation of the reaction products. In the last years mass spectrometric (MS) techniques have become the most important tool, so that today analysis usually relies on this method, most commonly in combination with separation techniques like capillary electrophoresis (CE) for intact glycoproteins or different chromatographic systems for glycopeptides and free glycans.

This work will focus on the analysis of glycopeptides to infer information about relative site occupancy. With this strategy, the glycoproteins are typically digested by use of specific proteases like trypsin, similar to proteomic approaches. Often, the resulting peptides are quite long, leading to certain problems regarding the ionization yields, separation and fragmentation yield. Unlike to typical proteomics strategies, where protein identities can be inferred from a few peptides per protein and sequence coverage is insignificant in most cases, in glycoproteomics all glycosylation sites should be covered to get detailed information about site-specific occupation. Therefore, unspecific proteolysis is sometimes used, employing for instance proteinase K [7], or a mixture of different specifically cleaving enzymes. Since MS analysis of glycopeptides suffers from ionization suppression in the presence of non-glycosylated peptides and since ion intensity is distributed over several charge states and over several glycoforms present (some of which are very low abundant), separation or enrichment of glycopeptides before MS detection is crucial. Common enrichment methods utilize lectins with broad specificity like concanavalin A (binding to mannose) or immobilized boronic acid [8] which binds covalently to cis-diols at basic pH and can be released under acidic conditions. It is also possible to enrich only a subclass or fractionate the glycopeptides with more specific lectins [9], like sambucus nigra agglutinin which binds preferentially to sialic acid linked α-2,6 to a terminal galactose [10]. Commonly, the lectins are bound to a solid support like agarose and are used in the way of affinity chromatography (AFC). Stationary phases for general glycopeptide separation and enrichment based on their physicochemical properties often used are different hydrophilic interaction chromatography (HILIC) [11] materials and porous graphitic carbon (PGC) [12] as well as cellulose based columns [13]. For acidic glycopeptides TiO2 [11] can be used as well as cation exchange chromatography, where the sialylated glycopeptides are found in the flow-through fraction at low pH [14].

Figure 1.3.: Nomenclature of Glycan Fragmentation



Nomenclature of glycan fragmentation. With low energy CID of glycopeptides (R = peptide) in positive ion mode, B- and Y-type ions are predominantly observed. [15]

Like in proteomics, tandem MS is employed for detection and identification. Fragmentation of glycopeptides using low-energy collision-induced dissociation (CID) or similar activation methods like higher-energy collisional dissociation (HCD) usually leave the peptide backbone intact while primarily dissociating the glycosidic bonds. The generated fragment ions are termed B- and Y-series ions, with the Y-ions being those covering the reducing end of the oligosaccharide, that is linked to the peptide backbone in case of glycopeptides [15]. A scheme of the glycan fragmentation nomenclature is shown in Fig. 1.3 . From the Y_1-ion (GlcNAc+peptide), which is present, though with varying intensities, in virtually all glycopeptide fragment spectra the peptide mass can be inferred. The B-ions found in a glycopeptide spectrum are important indicators for the presence of an oligosaccharide chain, with the HN^+ ($[HexHexNAc+H]^+$) peak at m/z 366.14 being the most important one, because it is common to all N-glycans and usually quite intense. Other important B-ions are those at m/z 512.20 (HNF^+), m/z 657.23 (HNA^+) and m/z 673.23 (HNG^+) which are indicative for the presence of fucose, N-aceylneuraminic acid and N-glycolylneuraminic acid respectively. Detailed glycan structure cannot generally be deduced from glycopeptide fragment spectra, although there have been approaches to that task [16].

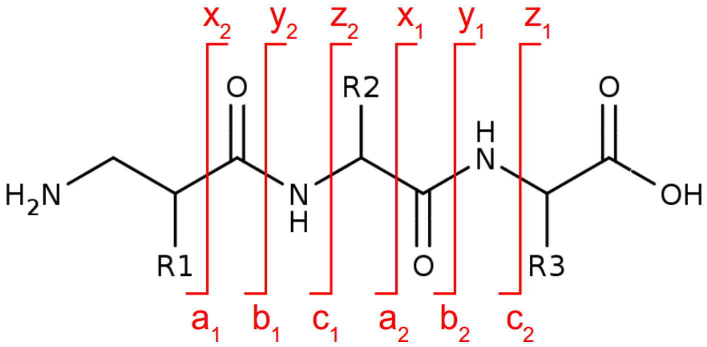
Since fragmentation of the peptide backbone is rarely observed, CID spectra provide no information on the peptide sequence so identification solely depends on its mass. High mass accuracy of the MS data is thus essential for this kind of analysis necessitating the use of high resolution mass spectrometers like quadrupole-time-of-flight (qTOF), Orbitrap or ion cyclotron resonance fourier transform mass spectrometry (ICR-FTMS) instruments, at least for the determination of the precursor mass spectra. The higher the mass accuracy available, the less restrictions are needed for the search space, permitting so the analysis of more complex samples. However, the analysis of biological samples can quickly lead to ambiguous glycopeptide assignments, particularly if the fragment spectra are acquired with low mass accuracy only. To avoid this limitation, electron transfer dissociation (ETD) can be used to produce peptide fragment spectra in which the peptide bonds are predominantly dissociated, providing so information on the amino acid sequence in the peptide [17].

1.2. Bioinformatic Analysis

1.2.1. Proteomics

Proteomic bottom-up type experiments rely on effective and robust high-throughput data analysis methods for peptide identification. Typically, the sample, consisting of many hundreds and more proteins present at different abundance levels, is digested by trypsin, often after fractionation, e.g. by sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE), isoelectric focusing or ion exchange chromatography (IEC). The digests are further separated by reversed phase high performance liquid chromatography (RP-HPLC) and finally detected by electrospray ionization (ESI) MS [18]. Peptides are subsequently fragmented, usually employing data-dependent MS methods like "Top-6", where the six most intense ions are selected for activation, in combination with blacklists for non-interesting ions, like keratin peptides, which are excluded from fragmentation. Most commonly, low energy CID (10 - 100 eV) is used for fragmentation and this leads to dissociation of the peptide bonds yielding b- and y-series ions, which may further lose neutral molecules like water or ammonia depending on the peptide composition. Another fragmentation technique often used is ETD, yielding exclusively c- and z-series ions with the advantage of attaining similar intensities for each fragment, facilitating de-novo sequencing. In addition there are no losses of water or ammonia or weakly bound groups like phosphate. An illustration of the peptide fragmentation nomenclature is shown in Fig. 1.4 .

Figure 1.4.: Peptide Fragmentation



Schematic fragmentation of a tripeptide. With collisional activation (CID, HCD) b- and y-type ions are observed predominantly, often accompanied by neutral losses. With ETD c- and z-type ions are observed exclusively.

To identify the proteins present in the sample, computer programs are used to find fragments of proteotypic peptides. Usually, three different search strategies are used for this task, namely a peptide mass fingerprint in combination with a sequence query of *in silico* digested peptides, generated from a supplied sequence list, aimed at finding candidate peptides with masses comparable to the precursor ion. From this list of potential peptides, the right one is determined by tandem MS (MS/MS) ion matching, done by comparing the measured fragment spectrum with the theoretical ones. The best matches are returned and scored based on statistical models.

The algorithms employed for matching MS/MS data to peptides can be categorized either as heuristic or as probabilistic algorithms [19]. In heuristic algorithms (like *X!Tandem* and *SEQUEST*), the experimental spectrum is correlated with the theoretical spectrum and a score is calculated based on the similarity which is essentially determined by the number of shared peaks. In probabilistic approaches (like in *MASCOT*) on the other hand, the fragmentation process is modeled and the probability is calculated that a certain peptide sequence gave the spectrum by chance.

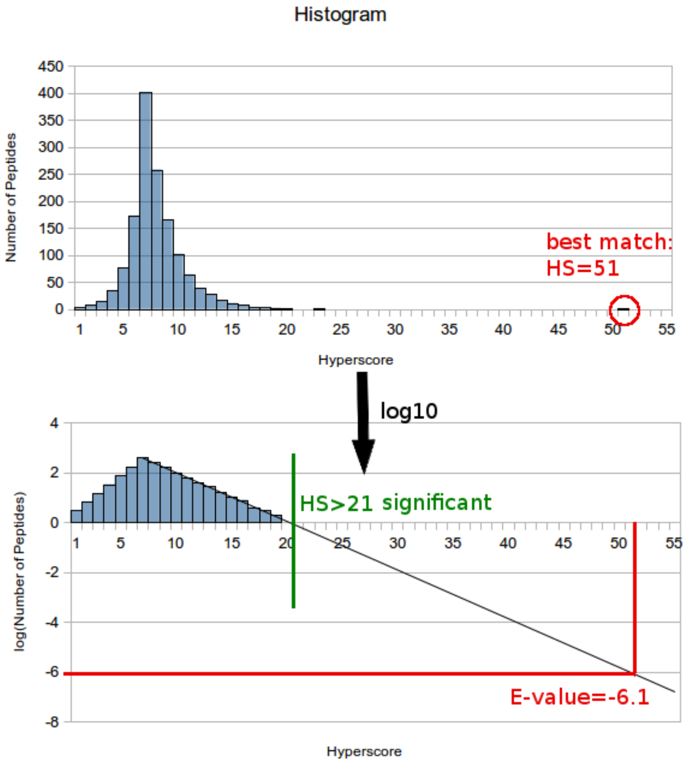
X!Tandem *X!Tandem* [20] is an open source program for the task of matching tandem mass spectra to amino acid sequences in peptides. In *X!Tandem*, model spectra consisting of all possible b- and y-ions, each having the same intensity, are calculated for all peptides that are within the specified precursor mass tolerance window. Peaks found both in the theoretical and experimental spectra within a specified mass tolerance are used to calculate a "hyperscore", which is the "dot product" of both spectra, equivalent to the sum of the intensities of the matched peaks, multiplied with the factorials of the number of assigned b- and y-ions

$$HS = \left(\sum_{i=0}^{n} I_i P_i \right) * N_b! * N_y!$$

with I_i the experimental intensity of the b- or y-ions and P_i a binary variable, being 1 if the peak was predicted and 0 otherwise.

This hyperscore is calculated for all peptides in the list that may match the MS^1-spectrum, i.e. having a similar parent ion mass. The peptide with the highest hyperscore is assumed to be correct. For a statistical score, a histogram of all the hyperscores is made plotting the number of peptides found with a given hyperscore (y-axis) against the hyperscore (x-axis). The y-axis of this histogram is log-transformed and the right (decreasing) side is fitted with a straight line. Significant matches have a hyperscore greater than that where this line intersects zero log(number of peptides). By extrapolating the straight line, the "expectation value" (negative y-axis) is calculated [21] which is reported by the program as score. Its calculation is illustrated in Fig. 1.5 . This "expectation score" is a measure for how much the best scoring peptide differs from the random matches. It depends on the sequence list used for searching. The hyperscore, in contrast, is determined by the match only and is therefore independent from the other possible peptides. If there are multiple experimental spectra of the same peptide, only the highest scoring spectrum is kept for the calculation of protein scores.

Figure 1.5.: X!Tandem Expectation Score



All peptides from the database having a mass similar to the precursor (i.e. within the specified mass tolerance) are matched on MS^2 level. The peptide with the highest hyperscore is assumed to be correct while the scores of the other peptides are assumed to be random. When the numbers of peptides within a certain hyperscore range are plotted against the hyperscore, a histogram, where the score frequency decrease exponentially around a maximum value, is expected. Thus log-transformation of the x-axis results in a straight line. The hyperscore of the highest scoring peptide is projected onto the extrapolated line, yielding the expectation value, which corresponds to the x-value of that line.

SEQUEST *SEQUEST* [22] uses a similar scoring scheme. The "preliminary score" which is similar to the "hyperscore" of *X!Tandem*, is the sum of the matched intensities with consideration of the continuity of the ion series and peptide length. The 500 best scoring amino acid sequences are used for cross-correlation analysis. *In silico* calculated (synthetic) spectra are generated from the sequences, containing the expected m/z values as well as a simple intensity component, considering the b- and y-ions with a intensity of 50, their isotopic peaks with a intensity of 25 and their loss of water, ammonia or carbon monoxide with a intensity of 10. The experimental spectrum is divided to ten equally large regions, where the highest intensity is normalized to 50. These two spectra are compared by cross correlation.

$$R_\tau = \sum_{i=0}^{N-1} I_i^{exp} I_{i+\tau}^{calc}$$

The index i numbering the peaks present in both spectra, the experimental and the synthetic one. τ is a displacement value in the m/z space.

The correlation function, R_τ, maximizes at the displacement value $\tau = 0$, if the two signals are the same. For the final score the average value of the correlation function over the range $-75 < \tau < 75$ is subtracted from the value at $\tau = 0$, yielding the correlation parameter, XCorr. A measure for the discrimination of the best match from the lower scoring, random matches (similar to the "expectation-value" of *X!Tandem*) is the difference between the normalized correlation parameters, C_n (highest XCorr value set to 1), of the first and second ranked peptide, ΔC_n. The match is usually correct if ΔC_n is greater than 0.1 .

Mascot A widely used program for fragment ion searching is *Mascot*. For scoring a probabilistic model is used [23] which is not described in detail in peer reviewed literature, because the algorithms are proprietary. The program reports a score that reflects the probability that the match has arisen by chance and correlates to $-10\log_{10}(P)$. A score 80 for example translates to a probability of 10^{-8} that the match was random.

1.2.2. Glycoproteomics

While high-throughput search algorithms for peptides are quite sophisticated today, automated analysis of glycopeptide or released glycan data sets is still in its early stages. There are several difficulties complicating glycopeptide analysis besides the general problem of ionization suppression. In most cases, the intact mass gained from MS^1 alone is not sufficient to resolve the glycopeptide composition, not even with high resolution and high mass accuracy spectra available. Therefore, identification must be conducted based on MS^2 spectra [24]. These, however, are often acquired in low resolution with isotopic peaks insufficiently resolved. This can lead to high mass errors and difficulties discerning the charge state. These problems are enhanced by the fact that glycopeptides are usually large analytes, compared with typical non-glycosylated peptides, so that higher isotopic peaks are dominant and the spectra are often composed of multiple charge states. Glycopeptides often suffer from poor ionization efficiency and the intensity is spread between the various glycoforms which often span a wide concentration range. With ESI, the signal strength is further distributed to various charge states, leading to very low intensities for lower abundant glycans, often near or below the limit of detection/quantitation. Fragment spectra of such analytes usually have low quality with unsatisfactory signal to noise ratios. These difficulties are even more pronounced with large peptides, while large glycans lead to complicated fragment spectra where assignment of the individual ions can be difficult. An automatic analysis program must therefore consider many different possibilities and assess their plausibility.

Low intensity signals can also complicate data processing methods like deisotoping and denoising. An ideal program should therefore be error tolerant and able to process data from various sources. Such an ideal program must further be able to perform high-throughput analysis, i. e. to analyze whole LC-MS data sets, be it in multiple vendor specific binary data formats or open formats like mzXML or mzML. Programs that can handle only single text-based peak lists are not suited for high-throughput analysis. The software should also support multiple platforms, like Windows, Unix and its derivatives, and be freely available to facilitate widespread use or even open source to enable community based improvements and custom implementations.

Since glycoprotein analysis is conducted on different levels, it would be desirable for a program to integrate these different types of experiments, i.e. to combine free glycan tandem MS for glycan structure analysis with glycopeptide analysis for specific glycosylation site determination, and to include peptide sequence analysis using ETD spectra.

Assuming that a reliable identification can only be achieved based of MS^2 spectra, an exhaustive analysis of all glycopeptides usually requires multiple LC-MS/MS experiments, since not all of them will be isolated for fragmentation in a single run. The efficiency of isolating the missing lower abundant glycopeptides can be enhanced by supervised mass spectrometric methods, e.g. by blacklisting identified analytes. A program that assists in creating such blacklists would be desirable.

Several programs for glycopeptide analysis exist, but all of them have specific shortfalls.

GlycoMod Tool A widely used tool is *GlycoMod* [25] from ExPASy (http://web.expasy.org/glycomod/), which calculates from the amino acid sequences of peptides originating from a considered target protein all those combinations with oligosaccharide structures which match to the experimentally determined precursor mass (in the MS^1 spectrum), within a specified mass tolerance. The list yielded by this procedure can be very long and even if there is a possibility to list those compositions which exist in a glycan database first, the choice of the right one can be cumbersome. Since there is no interface for the consecutive transmission of multiple masses, analysis of complete HPLC-MS runs with *GlycoMod* alone is a very tedious task.

GlycoSpectrumScan *GlycoSpectrumScan* (http://glycospectrumscan.org) [26] takes the target protein sequence(s), the (short) list of expected glycan compositions (this is an essential difference to *GlycoMod*) and MS^1 spectra in form of peak lists and assigns all peaks in these spectra matching within a given tolerance limit as potential glycopeptides. It can assign both N- and O-glycans. However, identification is not confirmed by fragment spectra and the program cannot handle whole HPLC-MS data sets, requiring averaging, data preprocessing and conversion to a peak list.

GlycoPeptideSearch *GlycoPeptideSearch* (*GPS*) [27] is one of the few freely available programs that can perform high-throughput analysis, taking LC-MS data in mzXML format. It is platform independent and uses MS^2 for identification, where the peptides are matched based on the Y-ions. For assignment of the corresponding glycan it relies on glycan structure databases, of which some are supplied with the program. This, however, is a double edged sword. No information about the expected glycan compositions needs to be supplied and possible glycan structures are provided for the glycopeptides assigned. On the other hand these databases do not cover all possible glycan compositions, and supplementation of the database with glycans found by free glycan experiments (or the setup of other custom SQL databases) is a complicated task. Furthermore, the program is closed source, making it somewhat inflexible regarding the optimization of search parameters. Therefore the number of assigned spectra is not very high, depending on the data source, limiting thus its value to a first glance that must be followed by a more detailed manual examination. Besides, the results output, which can be written either to an *Excel* spreadsheet or a plain text file, is, while providing all essential information about the assigned scans, not very well-arranged and needs to be parsed for a descriptive aggregation.

GlycoMiner A different approach is used in the *GlycoMiner* [28] software, which is only available for windows computers. In this program, the peptide mass is directly derived from the MS^2 spectra and the supplied sequence information is only used for confirmation. It does so by assignment of the Y ion ladder, where the fragments are separated by monosaccharides. Starting at the fragments with highest mass, the series ends either at the non-glycosylated peptide (Y_0) or the peptide with a residual GlcNAc attached (Y_1). The resulting mass is then searched in a sequence database for confirmation. The assigned Y ions are also used to select the appropriate glycan if multiple glycans match to the calculated mass.

While this approach results in very confident results, it requires spectra of singly charged fragment ions. With ESI-MS multiply charged ions are observed, so the MS^2 spectra are usually composed of higher charge states. This requires transformation to singly charged spectra, which can only be attained if the isotopic peaks are resolved, prohibiting the analysis of ion trap data. Furthermore, the open data format mzXML is not supported, requiring ASCII peak list based formats, e.g. pkl, or the proprietary *ProteinLynx* xml format as used by Waters.

Part II.

N-Glycopeptide Analysis by LC-Orbitrap-MS and Development of a Program for High-Throughput Analysis

2. Introduction

Protein glycosylation is one of the most common posttranslational protein modifications in mammals. The understanding of the various functions is only at the beginning. Aberrant glycosylation is linked to several pathological states like inflammation [29] or cancer [30].

Usually glycosylation analysis is performed on the level of glycans chemically or enzymatically cleaved off the proteins, utilizing different separation techniques as well as tandem MS in order to obtain structural information and differentiate between the multitude of possible isomers. However, upon releasing the glycans the information about their origin is lost, so no conclusions about the site specificity of multiply glycosylated proteins can be made. Furthermore, the presence of many proteins in the sample can affect the observed glycan mixture, especially regarding low abundant species, requiring high purity samples of the target protein. Since most enrichment and purification techniques are not strictly specific, leading to co-enrichment of other glycoproteins, this goal is usually difficult to realize. Therefore, often glycan analysis needs to be complemented by glycopeptide analysis in order to draw conclusions of high confidence.

Glycopeptide analysis is usually conducted using a proteomics-type approach, employing specific (sometimes also unspecific) digestion followed by HPLC-ESI-MS/MS. The aim to cover all glycosylation sites and all glycoforms present, which usually occur in a wide concentration range covering several orders of magnitude, makes glycopeptide analysis a very challenging task. This is exacerbated by their physicochemical properties, hampering ionization, so advanced experimental methodologies as well as powerful instrumentation is required.

A severe bottleneck remaining, however, is data analysis. While in the field of proteomics a wide array of sophisticated algorithms is established which are usually fully integrated into the workflows and which provide identifications, statistical assessment and even quantification of large data sets with relatively little effort by the user, comparable solutions for glycopeptides are only at the initial stages. The widely used software tool, *GlycoMod*, takes an experimental mass and returns a list of possible glycopeptides fitting within a specified mass tolerance. The correct glycopeptide must then be chosen by manual inspection of the fragment spectrum. That way, analysis of large data sets is a very tedious task. There is a number of other free, mostly web-based, tools aimed at facilitating data analysis, but only very few of them are capable of high-throughput analysis of whole HPLC-MS/MS data sets. One of those is *GlycoPeptideSearch* (*GPS*) which uses fragment spectra to identify the peptide part of the glycopeptide and uses one of several supplied databases for assignment of the corresponding glycan. While this approach requires no *a priori* knowledge of the possible glycans, it restricts the number of identifications and makes it difficult to integrate the results of free glycan experiments. Furthermore, the program lacks flexibility regarding parameter optimization, since its source code is not available.

In the present work a new program for high-throughput analysis of HPLC-ESI-MS/MS data of glycopeptides is developed and presented which is named *FindGlycoPeptides* (*FGP*), written in *Perl* programming language. It takes the (low resolution) CID spectra of glycopeptides in the open mzXML format and a list of protein sequences from established data bases as well as the supplied list of expected glycans, obtained from free glycan experiments or literature search. Then it assigns the N-glycopeptides structures/sequences by maximizing the sum of measured intensities matching with potential (calculated) Y-ions and rates the match with an empirical score. Decoy peptides are matched as well to estimate the false discovery rate.

It was tested with several glycopeptide data sets acquired by nano-RP-HPLC-Orbitrap-MS/MS using standard proteomics methodology. The samples ranged from standard glycoproteins (focused on α1-acid-glycoprotein (AGP) a heavily glycosylated acute phase protein known to have altered glycosylation patterns associated with pathological processes [29]), to glycoprotein mixtures and biological samples, like the secretome of MCF-7 cells. The assignments obtained by use of this new *FGP*-program were examined in detail and compared with those obtained by the *GPS*-program.

3. Experimental

3.1. Chemicals

The investigated test glycoproteins, i.e., α1-acid-glycoprotein (AGP), Fetuin, Asialofetuin (all bovine), rabbit Immunoglobulin G (IgG) and chicken Ovalbumin were bought from Sigma-Aldrich (St Louis, MO, USA). The proteolytic enzymes porcine trypsin (proteomics grade) and proteinase K were bought from Sigma-Aldrich, PNGase F for de-N-glycosylation from Roche (Hoffmann-La Roche, Basel, Switzerland). All chemicals used were of highest available quality. Sodium dodecyl sulfate (SDS), ammonium persulfate (APS), dithiothreitol (DTT), 3-[(3-cholamidopropyl)dimethylammonio]-1-propanesulfonate (CHAPS), thiourea and potassioum hexacyanoferrate, hydrochloric acid (HCl), tetramethylethylenediamine (TEMED), glycerine, formaldehyde, iodoacetamide (IAA) and ammonium bicarbonate (ABC) were also purchased from Sigma-Aldrich. Acrylamide, piperazine-di-acrylamide (PDA), glycine, bromophenol blue, sodium thiosulfate pentahydrate, silver nitrate, sodium carbonate, calcium chloride dihydrate and urea were bought from Merck (Merck KGaA, Darmstadt, Germany), as well as all organic solvents, comprising methanol, ethanol, acetic acid, formic acid (FA) and acetonitrile (ACN). Doubly distilled and ultra-high quality (UHQ) water were produced in-house using Millipore (Merck KGaA, Darmstadt, Germany) filtration systems. Solvents used for sample pretreatment steps like electrophoresis were analytical grade and MS-grade was used for procedures like digests, were the sample was directly used for analysis afterwards. Tris(hydroxymethyl)aminomethane (Tris) and Tris-HCl from Roth (Carl Roth GmbH, Karlsruhe, Germany). Bradford reagent and the molecular weight marker (SeeBlue Plus 2) were purchased from Bio-Rad (Bio-Rad Laboratories, Berkeley, CA, USA).

3.2. Instrumentation

Gel electrophoresis was conducted with a Mini-PROTEAN Tetra Cell from Bio-Rad using an Amersham (GE Healthcare, Chalfont St Giles, UK) EPS-301 power supply unit.

For chromatographic separation of the digests, a Dionex Ultimate 3000 UHPLC+ system was used (Thermo Fisher Scientific Inc., Waltham, MA, USA). The columns used for peptide mapping were Thermo Acclaim PepMap RSLC, packed with C_{18} material with a diameter of 5 µm and 100 Angstrom pore size. The inner diameter of the columns was 75 µm and the length either 15 or 50 cm. In both cases the same trapping column was used, a Thermo Acclaim PepMap 100 Nano-Trap, with C_{18} material of 5 µm diameter and 100 Angstrom pore size, an inner diameter of 100 µm and a length of 2 cm.

The outlet was interfaced with a hybrid mass spectrometer, the Thermo Orbitrap LTQ Velos ETD, using a Thermo Nanospray.

3.3. Sample Preparation

Stock solutions of five standard glycoproteins were made by dissolving in UHQ water. They had the following concentrations: bovine AGP 5 mg/mL, Fetuin from fetal calf serum (FCS) 6.7 mg/mL, Asialofetuin from FCS 1.2 mg/mL, rabbit IgG 2 mg/mL and chicken Ovalbumin at 5 mg/mL. They were aliquoted and stored at -18 °C.

The standard glycoproteins were either purified by SDS-PAGE and in gel digested with trypsin or digested in solution.

Gel Electrophoresis The SDS-PAGE was conducted with a discontinual buffer system. The resolving gel was made by 12 % acrylamide with PDA as crosslinker in a 375 mM Tris-HCl buffer with a pH of 8.8. SDS was added to a concentration of 0.1 % and the polymerization was started by addition of APS and TEMED to a final concentration of 0.045 % and 0.075 % respectively. The stacking had an acrylamide concentration of 4 % in 125 mM Tris-HCl buffer with a pH of 6.8 with 0.1 % SDS, 0.1 % TEMED and 0.05 % APS. Both components were allowed to polymerize for at least 30 min. A layer of water was added on top of the resolving gel for polymerization to avoid desiccation. The gel was loaded with 20 µg of protein sample, diluted to 20 µL and mixed with 5 µL of loading buffer, containing 50 % glycerine, 10 % SDS and 0.25 % bromophenol blue in 250 mM Tris-HCl buffer, pH 6.8, per lane. The running buffer contained 2.5 mM

Tris, 19.2 mM glycine and 0.1 % SDS. The separation was conducted with a maximum current of 20 mA or a maximum voltage of 250 V for about 40 min.

Silver Staining The separated proteins were fixated by panning in 50 % methanol with 10 % acetic acid for at least 90 min. The gels were stained by silver. First the fixated gels were washed with 50 % methanol and water, then sensitised with 0.02 % $Na_2S_2O_3$ and, after washing again with water, panned for about 10 min in 0.1 % $AgNO_3$. The gels were washed again with water and developed with a freshly made developing solution containing 3 % Na_2CO_3 and 0.05 % formaldehyde. The colouring reaction was stopped by addition of 2 % acetic acid, in which the gels were also stored, cooled to 4 °C.

The colored bands were excised, cut to small pieces and destained with a 300 µL of a mixture of 15 mM $K_3[Fe(CN)_6]$ and 50 mM Na_2CO_3 each by vortexing for some minutes. Some of the gel pieces had a blue color afterwards which probably resulted from Fe^{2+} ions originating from the scalpel used for cutting. The destained gel pieces were washed with 200 µL of a washing solution, containing 50 % methanol and 1 % acetic acid, repeatedly by vortexing for 10 min and changing the solution afterwards.

Reduction and Alkylation The destained gel pieces were equilibrated with 200 µL of 50 mM ABC buffer. They were reduced with 200 µL of 10 mM DTT in ABC buffer for at least 30 min at 56 °C. Afterwards they were washed with ABC buffer and alkylated by 200 µL of 50 mM IAA in ABC buffer for at least 20 min at room temperature in the dark. The reduced and alkylated gel pieces were washed again with ABC buffer and subsequently dried with ACN using 200 - 500 µL depending on the amount of gel. Finally they were completely dried in a vacuum centrifuge. They were stored at -18 °C if they were not digested immediately.

In-Gel Digestion with Trypsin 15 µL of porcine trypsin solution with a concentration of 12.5 ng per µL in 25 mM ABC were added to the dried gel pieces. They were kept cool for about 30 min, then 25 µL of 50 mM ABC was added. The reaction mixture was heated to 37 °C overnight, aiming for a reaction time of at least 18 h. The amount of trypsin was 187.5 ng, corresponding to a ratio of 1:107 by weight assuming that the spot contained the whole 20 µg that were loaded on the gel.

40 µL of ABC puffer were added and the gel pieces were sonicated for about 15 min. The peptide solution was removed and the gel was extracted with 40 µL of 50 % ACN

with 0.5 % FA twice, using sonication. The combined solutions were dried completely in a vacuum centrifuge and stored at -18 °C.

In Solution Digestion with Trypsin Some samples were digested with trypsin directly in solution. A mixture of the standard glycoproteins, containing 20 µg each, resulted in 41.2 µL protein solution. 10.3 µL of 200 mM ABC and 0.5 µL of 500 mM DTT, reaching a concentration of 5 mM, were added and the mixture heated for two hours to 60 °C for denaturation and reduction. After cooling, 3 µL of 250 mM IAA were added to a concentration of 15 mM and the mixture was left in the dark at room temperature for about three hours. The resulting alkylated and reduced protein solution (55 µL in 50 mM ABC) was diluted with 35 µL ABC, 9 µL ACN and 1 µL 100 mM $CaCl_2$were added. The reaction was started by addition of 1.5 µg trypsin in 6 µL 1 mM HCl, resulting in a enzyme:substrate ratio of 1:66.7 by weight. The mixture was incubated at 37 °C for about 18 hours. The reaction was quenched by addition of 0.5 µL FA and the mixture was dried with a vacuum centrifuge. Finally, the peptides were redissolved with 200 µL 2 % ACN in 0.1 % FA.

A similar procedure was employed for bovine AGP with the following amounts: 40 µg AGP in 50 mM ABC, 5 mM DTT for reduction 15 mM IAA for alkylation, 9 % (v/v) ACN and 0.5 µg trypsin (1:80). The final volume of the mixture was 40 µL, $CaCl_2$was omitted this time.

In-Gel Deglycosylation A mixture of 5 µL PNGase F solution (1 u/µL) and 15 µL 50 mM ABC was added to the dried gel pieces and kept cold for about 30 min. Afterwards 20 µL of 50 mM ABC were added and the mixture was incubated at 37 °C for a minimum of 18 h. The released glycans were extracted with 40 µL of 50 % ACN with 0.5 % FA, using sonication, and the gel was dried in a vacuum centrifuge for subsequent proteolysis.

In-Gel Digestion with Proteinase K A 10 mg/ml stock solution of Proteinase K was made using 50 mM ABC (pH 8.1) with addition of $CaCl_2$ at a concentration of 2 mM. For unspecific in-gel digest the stock solution was diluted to 1 mg/ml with 50 mM ABC. Dried gel pieces were soaked with 20 µL of the enzyme solution and kept cold (4 °C) for 10 min. 20 µL of 50 mM ABC was added and the mixture was heated to 37 °C, either for 20 min (peptide lengths of 12 - 18 residues) or 40 min (5 - 10 residues). The reaction was quenched by addition of 2 µL 90 % FA. The peptides were extracted with 40 µL 50 mM ABC and twice with 50 % ACN / 0.5 % FA, using sonication for around 10 min. The eluted peptides were dried in a vacuum centrifuge.

Biological Samples MCF-7 cells were cultivated protein free over night, one culture having been inflammatorily activated with IL-1β. 6 mL of supernatant were taken from the activated and control culture each, microfiltered (0.2 µm) and mixed with cold ethanol. The proteins were left to precipitate at -20°C for a week. They were sedimented and dried in vacuo. 400 µL sample buffer, containing 7.5 M urea, 1.5 M thiourea, 4% CHAPS, 0.05% SDS and 100 mM DTT were added in portions as well as solid urea nearing saturation. The samples were left over night at 4°C and repeatedly vortexed and sonicated. The protein concentration was estimated by mixing 1 µL sample with 199 µL water and 50 µL Bradford reagent and was around 1 mg/mL. Electrophoresis was carried out analogously to the standard glycoproteins with longer separation. Six fractions were cut from the gels: heavy (apparent molecular weight $>$ 150 kDa), I (100 - 150 kDa), II (65 - 100 kDa), III (50 - 65 kDa), IV (40 - 50 kDa) and light (15 - 40 kDa).

3.4. nano-RP-LC-nano-ESI-Orbitrap-MS/MS

Peptide mapping was conducted using water-acetonitrile gradients carried out in several steps. Eluent A was 0.2 % FA in MS-grade water, eluent B was a mixture of 80 % ACN, 20 % water and 0.08 % FA (all v/v %). 5 µL of sample, solved in 2 % ACN/ 0.1 % FA was loaded with a flow of 2 µL/min. Depending on column length, different gradients were used. For the 50 cm column a 70 min stepped gradient was employed, starting with 2 % B for the first 10 min, 2 - 7 % B in 1 min, 7 - 35 % B from 11 to 49 min, 35-40 % B from 49 to 52 min, steady 40 % B in min 52 - 54, 40 - 80 % B in min 54 - 56, steady 80 % B for the following 4 min, followed by a decrease back from 80 to 2 % B in min 60 to 63 and washing with 2 % B for the last 7 min. With the 15 cm column, the gradient was shorter, completing the cycle in 60 min: 2 % B in the first 10 min, 2 - 7 % B in min 10 - 11, 7 - 35 % B in min 11 - 40, 35 - 40 % B in min 40 - 42, steady 40 % B in min 42 - 44, then 40 - 80 % B in min 44 - 46 and steady 80 % B in min 46 - 50, followed by decreasing B from 80 - 2 % in min 50 - 53 which was held steady for the last 7 min. The gradient within all steps was linear. The flow rate was held constant with 0.3 µL/min. The temperature of the columns was kept at 35 °C, the samples were kept at 6 °C. Blank runs for washing, injecting 10 µL of 2 % ACN/ 0.1 % FA, was performed after 2 to 6 samples, depending on sequence. Mass standards containing 10 fmol of a mixture of peptides, dissolved in the same solvent, were injected at least once during the batch.

A data dependent Top 6 method, acquiring CID spectra with the LTQ ion trap, was used as standard method for Orbitrap-MS/MS. Recording was initiated by contact closure from the HPLC system after 10 minutes, producing run times of 60 min and 50 min for the 70 min and 60 min gradients, respectively. The MS^1 spectra were acquired with the FTMS detector in the mass range 400 - 1400 *m/z* with a resolution power of 60000 (at 400 *m/z*). Fragmentation was conducted by CID in the LTQ ion trap with a normalised collison energy of 35 eV, an isolation width of 3 Da, default charge 2, activation Q of 0.25 and activation time of 30 ms. MS^2 scans were acquired by the linear trap, producing low resolution spectra (centroided peaks) with the mass range determined automatically based on precursor mass. The fragmentation was performed on the 6 most intense peaks in decreasing intensity, followed by the next FTMS full scan. A lock mass of 445.120025 *m/z* was used. Singly charged ions and undefined charge states were excluded from activation. Dynamic exclusion was enabled with a repeat count of 1, repeat duration of 30 s, exclusion list size of 500, exclusion duration of 60 seconds and exclusion mass width of 5 ppm both for lower and higher masses. All FTMS spectra were recorded in profile mode, whereas the peaks of the LTQ spectra were recorded as centroids, producing line spectra.

Alternative fragmentation methods were tested as well. First, the normalised collision energy for CID in the LTQ was changed from 35 eV to 25 and 45 eV, respectively. No significant difference was seen between the 25 and 35 eV settings, while at 45 eV considerable increase of noise was observed. Second, CID by activation in the HCD trap was carried out with subsequent mass analysis by FTMS. Here, normalised collision energies of 25, 35 and 45 % were used and an isolation width of 2 Da, default charge state 2 and activation time 0.1 ms. Fragment spectra were recorded with a resolution of 7500 (at 400 *m/z*) with automatic mass range selection, and all other options being the same as in the standard LTQ method. Very few Y-ions were obtained, the peak intensities were low and the noise signals high. Third, acquisition of CID spectra from the LTQ in FTMS mode with resolution of 7500 was tested. Again, peak intensities were very low and number of signals few. Thus, none of these modified methods provided satisfactory results and were not used for further tests.

3.5. Data Analysis

The raw files from the data acquisition software provided by Thermo Fisher Scientific were converted to mzXML format with the *msConvert* tool from *Proteowizard* [31] on a windows pc. The binary accuracy was set to 64 bit and *zlib* compression was used. No peak picking was performed.

For protein identification *OpenMS [32]* was used, utilizing the *X!Tandem* search engine. The following options were used: precursor mass tolerance 10 ppm, fragment mass tolerance 0.3 Da, precursor charge 1-4, fixed modification: carbamidomethylation of cystein, variable modification: oxidation of methionine, maximum missed cleavages 2, cleavage sites [RK]{P} (trypsin specifity) and minimum fragment *m/z* of 150. The sequence database used was the fasta file with the proteome of the respective species obtained from *UniProt*.

For the glycopeptide analysis, the self developed *Perl* program *FindGlycoPeptides* (*FGP*) was used. Default parameters were: (i) precursor mass correction of up to ±2 Da, (ii) peptide mass range of 500 to 3500 Da, (iii) maximum missed cleavages sites per peptide: 2, fixed peptide modifications: carbamidomethyl cystein, variable modification: methionine oxidation, cleavage sites according to trypsin specificity, (iv) the presence of an oxonium ion at *m/z* 366 with a minimum intensity of 10 % for a scan to be considered as a glycoscan, (v) MS^1 mass tolerance 10 ppm, MS^2 mass tolerance 750 ppm, (vi) a threshold of 150 % base peak intensity for the sum of all matched Y-ions for a peptide to be considered further. Only the best fitting peptide was kept for each scan if multiple proteins were matched. The hits were refined and the false discovery rate was calculated. The threshold score was 20.

For comparison, the freely available software *GlycoPeptideSearch* (*GPS*) was used with the following default settings: (i) consideration of up to two ^{13}C Peaks, (ii) peptide mass range 500 - 3500 Da, (iii) maximum missed cleavage sites: 2, fixed modification: carbamidomethyl cystein, variable modification methionine oxidation, tryptic peptides, (iv) presence of at least 1 oxonium ion with intensity greater 10 %, minimum 2 peptide fragments with 5 % intensity, sialic acid intensity treshold of 10 %, fucose intensity treshold of 7 %, (v) MS^1 mass tolerance 0.05 Da, MS^2 mass tolerance 1 Da and maximum fragment charge of +3, (vi) maximum cluster score 20. There was no constraint of the glycan composition. The glycan data base used was the included *GlycomeDB*, either mammalian or human, depending on the sample.

Quantification by peak integration was performed with *mzMine* [33]. The data was imported in mzXML format. Mass detection was performed for MS^1 scans with an intensity treshold of 10^4 and the chromatogram builder was used with a minimum time span of 0.5 minutes, a minimum height of $2*10^4$ and a mass tolerance of 0.1 *m/z* respectively 10 ppm. The signals were deisotoped using the isotopic peaks grouper function, with the same mass tolerance, a retention time tolerance of 1 minute, a maximum charge of

+6 and the lowest m/z as representative isotope.

For manual data evaluation and quantification by peak height, the *Xcalibur* software package from Thermo Fisher Scientific was used. The peak height was determined as the highest intensity of the M+1 peak.

4. Development of a Program for High-Throughput Glycopeptide Search

4.1. General Concept

With CID of glycopeptides, fragmentation occurs mainly in the glycan chain predominantly with dissociation of the glycosidic bonds, yielding B- and Y-ions. Y-ions include the peptide part and give information about the peptide mass and, with a restricted peptide space common in targeted glycoproteomics, its sequence. Therefore, the peptide can be assigned by its Y-ions. With the peptide identified, the glycan mass can be calculated from the precursor mass and its composition can be determined.

A program which matches peptides by their Y-ions must perform three principal tasks if it is supplied with the HPLC-MS/MS data and the protein sequence(s). First, it has to create a list of peptide masses to be searched, by making an *in-silico* digest and filter the peptides which may be glycosylated. For N-glycans this means that the consenus sequence NXS/T must be part of the peptide (except in the case where the protease used cuts within this sequon, e.g. if X is K or R in the case of trypsin). Second, if the data format uses encoded binary data, the data must be decoded to obtain a peaklist which can be searched for the masses of interest. Popular open formats for HPLC-MS/MS data are mzXML and its successor mzML. Finally, the matching must be performed, that is, the peaklist is searched for masses indicating a particular peptide.

For this work a program that performs these tasks has been developed in the *Perl* programming language. It takes multiple protein sequences in form of fasta-files, encoded HPLC-MS/MS data in mzXML format and a list of glycan masses as input and tries to match the peptide-glycan pair for each MS^2 scan having the indicative B-ion at *m/z* 366.1 .

4.2. Core Algorithm

Before any matching can be performed, a list of possible glycopeptides must be created. To do this, a fasta file with one or more protein sequences is read in and parsed. The trypsin cutting sites are determined and the sequence string is cut to substrings according to trypsin specificity with zero, one and two missed cleavages. The resulting peptides are moved to the target array if they include the N-glycosylation consensus sequence, otherwise they are added to the decoy array. It turned out that the special case, where trypsin cuts next to the glycosylated asparagine can not be neglected. Since the check for the NXS/T would fail for this peptides, they are checked for C-terminal NK/R and moved to the target array if the next residue in the protein sequence would be S or T.

The mzXML file is then read line by line and skipped forward to the specified first scan number (or the first encoded scan data if no scan number was specified). MS^1 scans are not considered at the moment, so they are skipped. Base peak intensity, total ion current, precursor m/z and charge are extracted as well as the actual encoded scan data, which is deflated and decoded to a peaklist. First, the indicative (HN^+) oxonium ion at m/z 366.1 is searched. If the intensity is above the threshold (10 % of base peak by default), the Y_1-ions, that is, the intact peptide with one residual GlcNAc, of each peptide in the list are searched in charge states one, two and three. A list of all possible Y_1 ions having an intensity of more than 3 % is generated. For these candidate glycopeptides the Y-ions with a generic biantennary glycan with two sialic acids (H5N4A2 or H5N4A1G1 if Neu5Gc is expected) are calculated and matched with the spectrum. The cumulative matched intensity, which is the sum the (relative) intensities of all calculated Y-ions found in the spectrum, is calculated

$$CI = \sum_{Y} I_Y^{exp} \qquad \forall Y \in exp \wedge calc$$

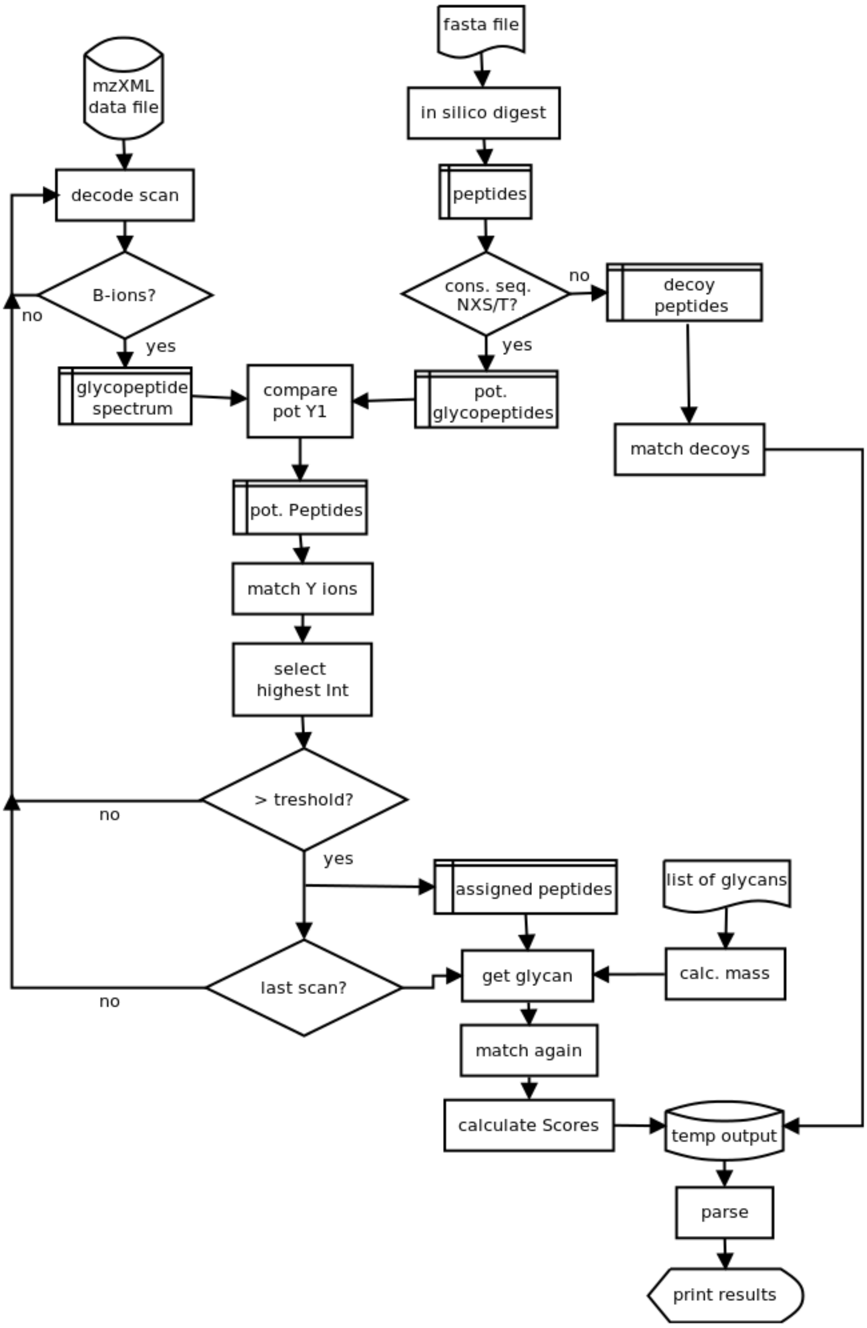
The peptide with the highest cumulative matched intensity is kept as well as the matched m/z / intensity pairs. Although this approach is quite trivial from a computational point of view, it has proven powerful in the task of correctly identifying N-glycopeptides by Y-ions. Also the cumulative Y-ion intensity is a good measure for the quality of the assignment, especially in relation to the total ion current. This matching algorithm is performed on each MS^2 scan until the last specified scan has been analysed. The approach is similar to *GlycoPeptideSearch,* where the peptide portion is identified by the presence of at least two of the four Y_{0-3}-ions, termed “intact peptide” fragments.

In the present data, however, only the Y_1-ion was observed reliably, making the peptide identification filter as implemented in *GPS* rather strict. In *FGP* on the other hand, only the Y_1-ion is required (with a default minimum intensity of 3 % whereas the treshold in *GPS* is 5 %). If more than one peptide is found, all possibilities are reported in *GPS*, while in *FGP* only the one with the highest cumulative intensity of Y-ions will be considered further.

Afterwards, the data is written to a temporary file. The program loops through each scan which was identified as a glyco-scan, writing a new line with the respective scan index, precursor *m/z* and precursor mass. If the cumulative matched intensity of the best peptide for that scan was above the specified treshold (150 % base peak by default), the glycan mass is calculated and searched in the list provided. Since the program relies on the precursor mass determined by the MS software, which often picks the wrong isotopic peak, glycans with masses differing by 1*1.003 or 2*1.003 Da are also searched by default. These "wrong" peaks are often the highest isotopic peaks and are thus selected, however, they are not identical with the monoisotopic peak considered by the calculation. While in the majority of cases, the determined mass corresponds to the first or second C13 isotopic peak, requiring a negative correction. It was observed that the other way can not be neglected in all cases, and is considered as well by default. As several combinations of monosaccharide units have similar masses, like 2 * Fuc = 292 and 1 * Neu5AC = 291; 1 * Hex + 1 * Fuc = 308 and 1 Neu5Gc = 307; the problem of assigning the correct monoisotopic peak is essential. In cases where the HNF^+ Y-fragment ion at *m/z* 512.2 was found, only fucosylated glycans out of the glycan list are considered. By default, the spectrum is matched again with the assigned glycan and a score is calculated. All relevant data is written to the file, which is parsed to the final output later.

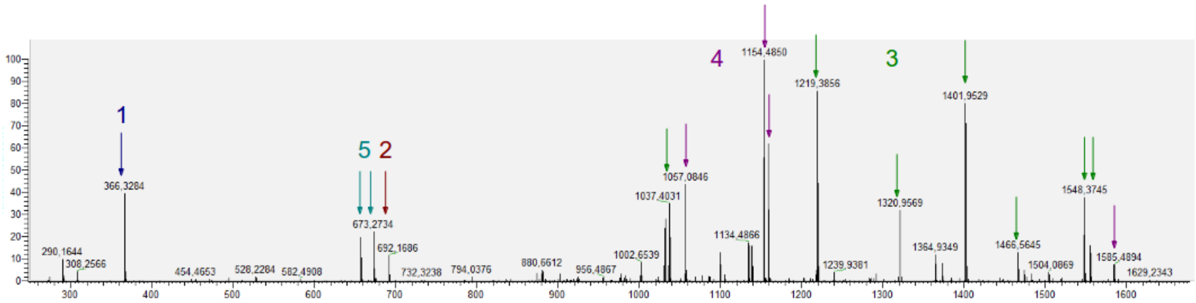
A flowchart illustrating the basic steps performed by *FGP* is shown in Fig. 4.1 and an example spectrum showing the different stages of matching is shown in Fig. 4.2 .

Figure 4.1.: Basic Function of Matching Algorithm



Basic function of *FindGlycoPeptides*: at program start a list of target and decoy peptides are created from the supplied protein sequences (fasta file), by creating substrings according to enzyme specifity (in silico digest). Those peptides having an NXS/T motif are pushed onto the target array, while the rest is pushed to the decoy array. In the main loop each fragment scan is decoded, one ofter another, checked for the indicative ion at m/z 366.1 and, if found, the possible Y_1-ions. From these candidates, the peptide with the highest matched Y-ion intensity (matched with generic biantennary glycan) is assigned. After the last scan was checked, the glycans are searched for the assigned peptides and the spectra matched again, a score is calculated and the results are printed. Decoy matching, which is not shown in detail, is performed the same way.

Figure 4.2.: Stages of Glycopeptide Matching



Matching of a glycopeptide is done in 4 stages. 1. check for glycosignature: HN^{+}(366.14)
2. Y_1ion: pN^{2+}(691.87)
3. Y-ions resulting from fragmentation of default glycopeptide pH5N4A1G1: $pH3N2^{2+}$(1036.49), $pH4N3^{2+}$(1219.05), $pH4N4^{2+}$(1320.59), $pH5N4^{2+}$(1401.62), $pH5N4A1^{2+}$(1547.17), $pH5N4G1^{2+}$(1555.17)
4. Refinement: rematching with the assigned glycan - additional Y-ions: $pH6N5^{3+}$(1056.46), $pH6N5A1^{3+}$ (1153.49), $pH6N5G1^{+}$(1158.82), $pH6N5^{2+}$ (1584.19)
5. checking the sialic acid B-ions: HNA^{+}(657.23) and HNG^{+}(673.23)

4.3. Handling of multiple proteins

In the most simple case which, however, is typical for targeted-glycoproteomics, the sample is comprised of a known single pre-isolated/purified glycoprotein only. In this case, a single protein sequence provided to the program would, in theory, be enough to assign all spectra originating from N-glycopeptides, given that the glycan list contains all glycans that occur with this protein and all possible peptide modifications are considered. In practice, however, even purified proteins are often accompanied by other proteins copurified (at least traces from them). With glycoproteins these are often glycoproteins as well. In biological samples, the number of glycoproteins, as well as non-glycosylated ones, can be quite high, depending on the sample pretreatment in terms of fractionation and enrichment steps. Consequently, a program for glycopeptide finding should be capable of handling multiple protein sequences.

In the program *FindGlycoPeptides* a standard fasta file, containing a virtually unlimited (technically, the upper limit is the maximum number of entries in a hash list in perl) number of protein sequences is provided. The fasta headers can be arbitrary, with the string following the '>' denoting the protein name. The sequence (one letter code, uppercase) is read from the following lines and is terminated by the next line starting with '>'. The target glycopeptide database consists of a 2-dimensional hash array with a subarray for each protein's glycopeptides. In contrast, the decoy peptides are stacked onto a single array, since information about their origin is not needed. When matching the spectra, the best-match peptide is determined for each protein individually. The number of proteins of which the best peptide is kept can be specified, which is one by default. Since at this stage the only criterium for a good match is the cumulative matched intensity, it is possible that the best-match peptide is wrong and the right one dropped. If peptides of different proteins are kept, the false ones may be eliminated at a later stage, for example by a greater precursor mass error.

Since the sequence of the peptide backbone can not be determined unambiguously from the fragment spectrum, it is obvious that passing a whole proteome database to the program, like in proteomics experiments, is futile. The search space must be restricted as much as possible, which is particularly true for spectra with low mass accuracy, such as those produced by ion trap mass spectrometers. Useful results will be obtained only if the supplied sequence database can be reduced to a few glycoproteins, meaning that information about the composition of the sample has to be known before running the program. This can be attained for instance by running a proteomic analysis.

4.4. Peptide Modifications

Since most glycoproteomic workflows deploy carbamidomethylation of cysteine, this modification is considered as a fixed modification accounting for each Cys the mass of 160.03065 Da (Cys-CAM) instead of 103.00918 Da. The most abundant variable modification is the oxidation of methionine, increasing the mass by 15.9949 Da, and this is considered by default. If a peptide contains multiple methionines, a recursive algorithm, in which a single Met gets oxidised in each step, adds the oxoforms to the peptide array. All variable modifications are represented by lower case letters and are treated internally as additional amino acids with distinct masses.

Optional variable modifications considered are the formation of pyroglutamic acid from N-terminal glutamine with a mass shift of - 17.0265 Da and phosphorylation of serine, threonine and tyrosine with a mass shift of + 79.9663 Da, however, restricting the number of possible phosphorylated residues per peptide to one.

4.5. Output

After writing all relevant data to the temporary files, these are parsed to produce the final output. Each line of the temporary results file is assessed based on different criteria and allocated to the appropriate table. With default behaviour four tables are created. The first table lists the scans that are considered being assigned correctly. By default, a "correct" assignment is assumed if the score is higher than the specified threshold and the assignment was not rejected in the refinement step. If one uses the option, not to decide on the basis of a score (i.e., scoring deactivated), the "quality" of matching is judged by MS^1 mass deviation and the sum of the Y-fragment intensities matched. All relevant data is printed, including scan index, precursor mass, protein, peptide, glycan and their respective masses, precursor mass correction, mass deviation, matched intensity and a list of the top five matched Y-ions in the format mz:int:ion:charge. Furthermore, the intensities of the four most important B-ions, the subscores and the total score. There are options to produce more compact output or to list all matched Y-ions.

The second table contains all scans that were rejected because the glycan contains sialic acids, for which no B-ions were found. The third table contains the scans for which a peptide was matched with intensity above a threshold, but did not qualify for the first table. This also includes scans, for which no glycan was found. The fourth table lists all scans that have a glycosignature but no matching peptide.

The output starts with the values of the program parameters that were used for the analysis. Following are the tables listing the analysed scans. After those, a summary is produced, listing the glycans found for each peptide and glycosylation site and the number of assigned scans. Finally, a short statistic is printed with the number of assignments and different species, as well the false discovery rate, FDR, estimated and the computation time.

The output is printed to the standard output (STDOUT) of the terminal and can be redirected to a text file. Columns are separated by tabstops and lines by linebreaks (Unix: \n, Windows: \r\n), which is compatible with any standard spreadsheet software. While MS data being processed, the progress in data analysis is printed to a terminal (to standard error, STDERR).

4.6. Refinement and Scoring

To increase the confidence in the results a refinement strategy was added where the information gained on glycan species and peptides is used to remove unfitting hits from the result list. The Y-ion matching, which was initially done assuming a generic biantennary glycan with two sialic acids (H5N4A2 or H5N4A1G1), is repeated with the set of calculated Y-ions generated on basis of the assigned (matching) glycan, employing an algorithm where one saccharide is removed in each step. As a consequence, the matched intensity increases for larger glycans. Furthermore, the B-ions at *m/z* 657 (HNA^+) and 673 (HNG^+) are checked. If the assumed glycan has a sialic acid without the corresponding B-ion, the assignment is rejected.

It turned out that with low accuracy fragment ion scans and the consequence thereof, i.e. to allow matching Y-ions with high mass tolerance, random matches with high cumulative intensity can occur. The confidence of an assignment can therefore not be assessed by this measure alone. An ideal scan has a high overall intensity and low noise. The precursor mass should correlate to the monoisotopic mass determined by the instrument software. A confident assignment has a minimal mass error and a high proportion of the total intensity can be explained by Y- and B-ions. The B-ions should match the glycan and the assigned peptide should elute in a time range, where a high number of other scans were assigned to the same peptide - at least with RP-HPLC, where the glycan has little influence on the retention time. In Fig. 4.3 two example scans are compared in respect to their quality features and the resulting scores.

In accordance with these criteria, a scoring function was introduced composed of two parts, the base score, S_B, and the penalty, P. The base score consists of three terms, each ranging from 0 to 100 and multiplied with an weighting factor, w. The terms are (i) the score for the cumulative Y-ion intensity relative to the base peak, $S_{ciY_{BP}}$; (ii) the score for the total Y-ion intensity in proportion to the total ion current without the most frequent B-ions, $S_{ciY_{IC}}$; and (iii) a score rating the noise of the spectrum, S_{Noise}.

$$S_B = S_{ciY_{BP}} * w_{ciY_{BP}} + S_{ciY_{IC}} * w_{ciY_{IC}} + S_{Noise} * w_{Noise}$$

Within this program testing, the weighting factors w were chosen as 0.2, 0.6, and 0.2, respectively.

From this base score the penalty, P, is subtracted, where P is composed of several contributions, i.e., the penalty for the precursor mass correction, $P_{massCorr}$, the penalty for non-fitting B-ions, P_{Bions}, the penalty for the mass deviation in the MS^1 scan, $P_{massDev}$ and a penalty for modified peptides, P_{pepMod}. Furthermore there is a term taking into account the chromatographic separation, $P_{Elution}$, which is given positive values if the assigned peptide elutes outside the chromatographic peak of this particular peptide and negative values near the elution maximum. These terms combine additively with equal weighting factors ($w = 1$)

$$P = P_{massCorr} + P_{Bions} + P_{massDev} + P_{pepMod} + P_{Elution}$$

the final score is therefore

$$S_F = S_B - P$$

The individual scoring functions for the base score are linear functions between the lower and the upper limit and are scored from 0 to 100 points. The threshold value for $S_{ciY_{BP}}$ is 100 % of base peak intensity. Values above 600 % (upper limit) are scored with 100. The weighting factor is 0.2. The threshold value for $S_{ciY_{IC}}$ is 5, the upper limit (above which the score is 100) is 70 % of total non-oxonium ion current; weighting factor is 0.6. The limits for the linear scoring function, S_{Noise} , of the noise index I_N are 0.75 and 0.25, respectively. It has a weighting factor of 0.2 . The noise index is hereby defined as $1 - \frac{IC_{Top20}}{TIC}$.

The contributions to the total penalty are defined as follows: $P_{massCorr}$: 5 points for each mass unit of precursor correction to lower masses and 10 points for correction to higher masses. P_{Bions}: 7 points penalty if the assigned glycan has a fucose residue but zero intensity for the B-ion at m/z 512; there is a bonus of 5 points if the intensity of this ion is higher than 2 % of base peak; 5 points penalty if the B-ions for Neu5Ac (657 m/z) or Neu5Gc (673 m/z) are found in the spectrum with an intensity of more than 1 % of base peak and the assigned glycan lacks the corresponding sialic acid. Similarly, 5 points are added if the number of Neu5Ac and Neu5Gc in the assigned glycan is the same but the intensity of the lower B-ion is less than 50 % of the higher one. For the mass error term $P_{massDev}$ there are three options: if the mass error in the MS^1 spectrum is less than 10 % of the set tolerance (10 ppm by default), 3 bonus points are added, between 10 % and 100 % of the specified threshold, the penalty rises linearly with the absolute value of the error up to 10, which is given if the error is exactly the tolerance. For higher errors, the penalty is doubled. The peptide modification penalty, P_{pepMod}, is assigned to a value of 15 for any phosphorylated peptide, so that only high confidence hits of supposedly phosphorylated peptides lie above the threshold. Furthermore, a penalty of 3 is given for each oxidized methionine.

For $P_{Elution}$, the occurrences of each peptide are counted in windows of 100 scans, which are shifted 10 scans forward for each time step, similar to a moving average signal filter. This procedure is based on the preliminary peptide identifications. Two values are considered for $P_{Elution}$, the maximum number of occurrences of that peptide in any 100 scan window and the maximum in any 100 scan window including the particular scan. If these two values are the same, that is, the assigned peptide eluted around its elution maximum, a bonus of 5 points is given. If the local count is more than 50 % of the global maximum and greater than 5, 3 bonus points are given. 3 points penalty are assigned if the local count is less than 20 % of the maximum and 10 penalty points if the peptide eluted isolated. This way, outliers are removed in datasets of purified glycoproteins, where each peptide is found in a high number of scans. In complex samples, however, where each peptide is assigned a few times only, often separated by more than 100 scans, it can lead to high penalties for valid assignments and should be turned off.

The scoring function in general is provisional and the weights, threshold values and penalties may be optimized for the final release of the software and even more sophisticated measures may be added.

4.7. False Disovery Rate Estimation

The basic approach for false discovery rate estimation has been taken from Chandler et al [27], therefore, FDR is calculated the same way as in the program *GlycoPeptideSearch* (*GPS*). Shortly, peptides lacking the NXS/T consensus sequence are used as decoys. Each time a scan with glycosignature is found, the program tries to match the scan to a decoy peptide, using the same criteria as for the target peptides. Results are written to another temporary file and can be viewed in detail after the program finished. The different number of target and decoy peptides is corrected and the FDR is calculated as follows. The probability for matching a random spectrum with a random decoy glycopeptide, p, is given by

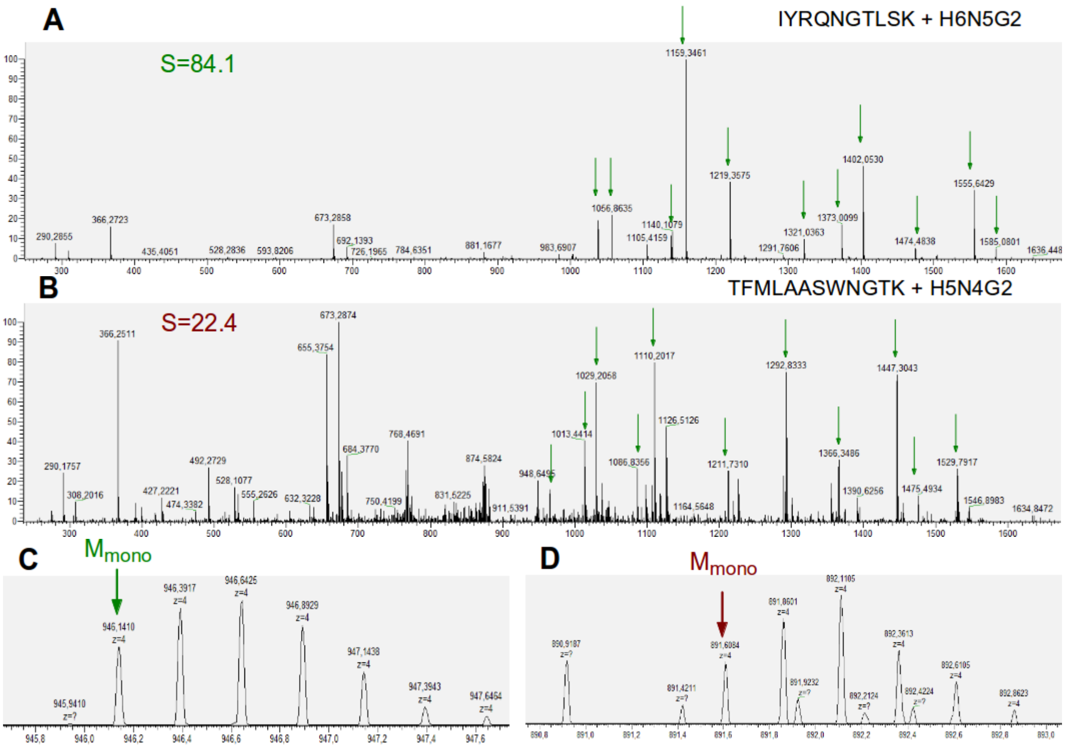
$$p = \frac{1}{N}\frac{S_D}{P_D}$$

with S_D the number of spectra matched to decoy glycopeptides and P_D the number of decoy peptides. The probability, that no target glycopeptide matches randomly to a spectrum is therefore $q = (1-p)^{P_T}$ with the number of target peptides P_T. It follows that the FDR can be estimated by the expectation value of random matches $E_{rand} = N(1-q)$ divided by the number of spectra matched T, so

$$FDR = \frac{N(1-q)}{T}$$

It must be noted, that the estimate is based on the total number of assigned scans, rather than the number of different glycopeptides.

Figure 4.3.: Scoring of Good and Bad Scans



Comparison of the scoring of a good and a bad scan in the same data set from a bovine AGP digest.

A Fragment scan of peptide IYRQNGTLSK with glycan H6N5G2. The matched Y-ions are marked by arrows, these are: pH3N2^{2+}(1036.49), pH6N5^{3+}(1056.46), pH5N4G2^{3+}(1139.48), pH6N5G1^{3+}(1157.82), pH4N3^{2+}(1219.05), pH4N4^{2+}(1320.59), pH3N2^{2+}(1036.15), pH4N3G1^{2+}(1372.60), pH5N4^{2+}(1401.62), pH4N4G1^{2+}(1474.14), pH5N4G1^{2+}(1555.17) and pH6N5^{2+}(1584.19). 60.5 % of the total intensity were matched to the Y-ions and the noise index was 0.403. The sub-ppm MS^1 mass error (-0.21) and the absence of the Neu5Ac indicative ion at m/z 657.23 give 3 bonus points each. The total score is 84.1, making it one of the best scans of the data set.

B Fragment scan of peptide TFMLAASWNGTK with glycan H5N4G2. The matched Y-ions are: pH4N3G1^{3+}(964.40), pH2N2^{2+}(1028.96), pH5N4G1^{3+}(1086.11), pH3N2^{2+}(1109.99), pH3N3^{2+}(1211.53), pH4N3^{2+}(1292.56), pH3N3G1^{2+}(1365.08), pH4N3G1^{2+}(1446.10), pH5N4^{2+}(1475.12) and pN1^{1+}(1529.74). Only 13.9 % of the total intensity is covered by matched Y-ions because of high noise, with a noise index of 0.791 and fragments of other species that were also selected. The isotopic peaks of the signal at m/z 655 caused a B-ion penalty of 5 points (peak at m/z 657 for glycan without Neu5Ac). The total score is just slightly above the default threshold.

C Precursor isotope cluster of fragment scan A

D Precursor isotope cluster of fragment scan B, being at the end of the chromatographic peak, with probably two coeluting species in the same mass range, which cause a high number of unmatched fragments.

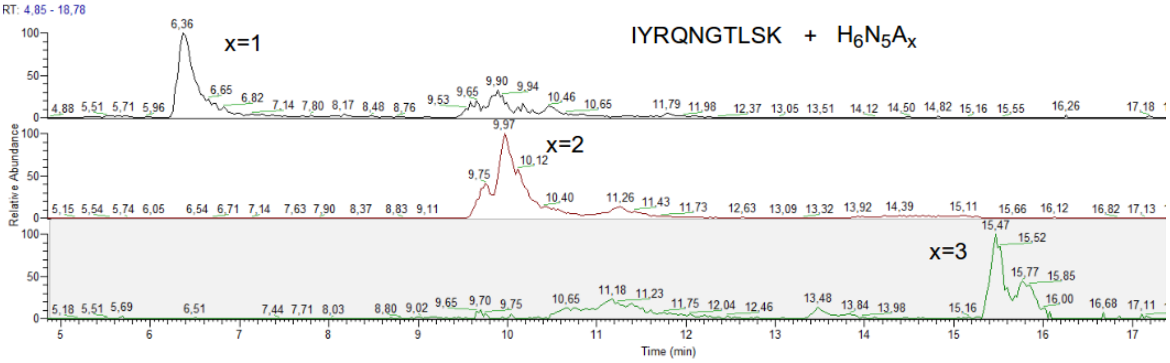
5. Evaluation of FindGlycoPeptides for Finding N-glycopeptides in Protein Digests

The effects of different parameters on the number of assignments and on the false discovery rate were tested thoroughly with datasets acquired from purified standard glycoproteins giving particular emphasis on bovine AGP. The performance and usability of this new program *FindGlycoPeptides* (*FGP*) is compared with the freely available program *GlycoPeptideSearch* (*GPS*). In both instances, the real false discovery rate was determined by manually validating each of the predicted peptide-glycan pairs. To evaluate the performance attainable with more complex samples, a digest of a mixture of five (known) glycoproteins was investigated. Finally, the program was tested with an even more complex sample, i.e., the digest of the supernatant of MCF-7 cells, and in this case the sample composition was unknown.

5.1. Bovine AGP

With this purified test protein, first, the influence of fragmentation regime and fragment ion analysis was tested. Datasets produced by high-energy CID in the HCD-cell of the LTQ-Orbitrap instrument were of little use for this automated data analysis because they lacked most Y-ions which are crucial for the identification algorithm. In most cases, only the singly charged Y_1-ion was found, and with larger peptides, this signal was often outside of the recorded mass range. Besides, intensity was usually very low and a lot of artifacts occured in the spectra. With CID fragmentation in the LTQ-ion trap and subsequent analysis by high resolution FTMS, the intensity of the Y-ions was very low in most scans (under the given scanning-time frame) and peaks from chemical and/or electronic noise were prevalent overcompensating the high mass accuracy by far. Consequently, the best results were obtained by low mass accuracy IT-MS fragment scans. Regarding the chromatographic separation, two methods were used: a 50 cm RP-type column with an elution gradient over 70 min, and a 15 cm column of the same type with a 60 min gradient. On the C_{18} reversed phase column used, the elution times of the glycopeptides were determined mainly by the peptide portion, with larger peptides tending to elute later. Regarding the glycan portion, the number of sialic acid residues had a significant effect with an increase in retention time by 3 to 7 min for each additional sialic acid (under low pH conditions with 0.2 % formic acid in the eluent). Neu5Ac was retained slightly longer than Neu5Gc. An example of retention time shift upon the presence of different numbers of sialic acids is given in Fig. 5.1.

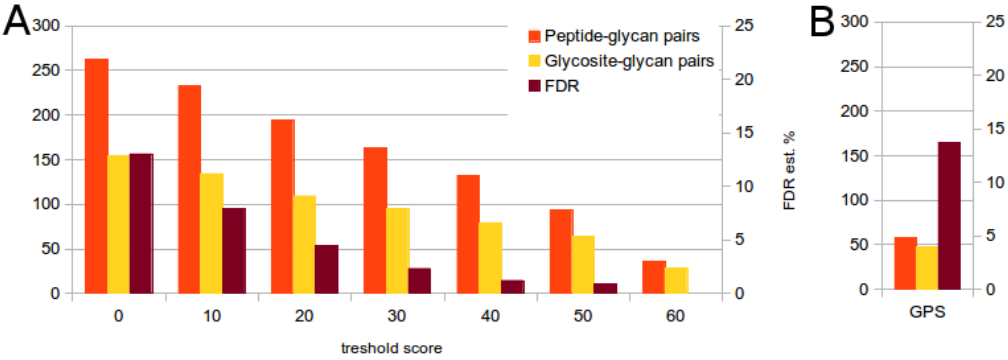
Figure 5.1.: Retention of Sialylated Glycans



Retention times of the peptide IYRQNGTLSK with a threeantennary glycan with one (6.36 min), two (9.97 min) and three Neu5Ac (15.47 min) under acidic conditions (0.2 % formic acid).

A "high-quality" MS dataset, gained from 3 different HPLC-ESI-MS/MS runs using a 50 cm column and loading 50 % of a digested gel band (20 µg), was analyzed automatically with *FindGlycoPeptides*. The effect of different internal parameter settings for the score function was tested, which is the most relevant parameter for balancing the number of assignments with the false discovery ratio. Different problems might need different strategies. For instance, with a low score threshold, e.g. 0 or 10, a high number of assignments can be attained, including lower abundant ones, however, these assignments have to be validated manually. On the other hand, a very high threshold, e.g. 50, results in few but highly confident assignments including predominantly the highly abundant glycans. Thus, for glycopeptide discovery a threshold of 20 is suitable. With this setting and an FDR of less than 5 % (calculated on the basis of the number of spectra assigned) more than 100 distinct glycopeptides (meaning all combinations of glycosites and glycans, taking into account peptides with different numbers of missed cleavages) were found in this dataset. With a threshold of 40, still around 80 distinct glycopeptides were assigned with an estimated FDR near 1 %. With higher thresholds the number of hits decreases sharply, albeit for the reward of very low false discovery rate, which is near zero at a threshold score of 60. This is a very good performance compared to *GlycoPeptideSearch* (*GPS*), which returned few hits with default settings. Some more assignments can be produced by *GPS* if the fragment ion tolerance is set to a high value (1 Da) resulting, however, in a very high FDR. The results are shown in Fig. 5.2 .

Figure 5.2.: Assigned Bovine AGP Glycopeptides Depending on Score Threshold



A Number of assigned glycopeptides (peptide-glycan pairs) and distinct glycopeptides (glycosite-glycan pairs) as well as FDR estimates (right axis) with using different threshold scores. **B** The respective values for the same sample analyzed by *GlycoPeptideSearch*.

In the runs with the 50 cm column, more than 12000 fragment scans were obtained with each run, around 4800 of which had the glycosignature HN^+ at m/z 366. With a threshold score of 20, around 1500 of these scans could be matched with an AGP glycopeptide, with FDR estimates of around 3.5 %, based on the number of matched scans. The results were validated manually, and under the assumption that all assignments of the same species are correct if one of them could be confirmed, the number of false matches ranged from 8 to 17, giving a FDRs between 0.56 and 1.12 %. Considering oxidized methionine forms as different peptides, the number of peptide-glycan combinations in all 3 samples combined was 251. The number of distinct glycopeptides, that is, combinations of glycosites and glycans, ranged from 107 to 121, with 135 distinct glycopeptides found in 3 samples combined. 16 of those 135 could not be verified, amounting to a false discovery rate of 11.9 % with all samples combined or 6.6 to 8.4 % in each individual run. Each glycosite was found in two main peptides (different numbers of missed cleavages or methionine oxidation), with very few assignments made for other peptides, of which a high number was falsely assigned. When only the two main peptides of each glycosite are considered, the number of distinct glycopeptides decreases to 128 of which 10 were falsely assigned, giving a FDR of 7.8 % or 2.9 to 6.7 % for the individual samples. An overview of the numbers of spectra and glycopeptides assigned in the bovine AGP runs is shown in Table 5.1 .

Table 5.1.: Overview of bAGP Glycopeptides Assigned by *FGP*

A

	I	II	III
spectra assigned	1423	1520	1547
false	8	17	15
FDR spectra	0.56	1.12	0.97
distinct glycopeptides	107	107	121
false	7	9	8
FDR distinct	6.54	8.41	6.61

B

Glycosite	I	II	III	tot
2	21	20	24	24
3	14	14	17	17
4	40	40	45	48
5	25	24	27	30
all	**100**	**98**	**113**	**119**

A Overview of the assigned spectra and distinct glycopeptides in the three bAPG samples with an *FGP* threshold score of 20, as well as the respective numbers of false assignments and the corresponding false discovery rates. **B** Numbers of confirmed distinct glycopeptides on each glycosite in each sample and in total.

Spectra Counting With around 1500 glycopeptide spectra assigned for each run, a rough estimate of the relative abundance of individual glycan species can be inferred by spectra counting. Label free quantification of glycopeptides is complicated by several problems. The peak height can be used if the glycan moiety does not affect ionization efficiency and trypsin cutting, both of which are not strictly met. The mass range covered is large and the intensity is spread to different charge states, again depending on the glycan portion. The sensitivity of the mass spectrometer can depend on the mass to charge ratio. Furthermore, the ionization can be suppressed due to co-eluting species, and this can affect as well glycoforms that are separated by the chromatographic system, e.g. species with different numbers of sialic acids. Instability of the electrospray might be a general problem, especially with the nanospray in use. These difficulties also affect peak area methods, which are further challenged by computational problems. For instance, a number of peaks was not found by *mzMine* if they occurred in a dense region of the mass spectrum, and this happened not only for low intensity peaks. The deisotoping within *mzMine* (taking the areas of all isotopic peaks for quantification) was not very reliable as well, as often the monoisotopic peak was missed. This problem is aggravated by the fact that, at least with data obtained by the Orbitrap analyzer, isotope clusters deviate significantly from the theoretical shape if intensity of the signal is low. Consequently, no accurate quantification can be expected by this method. The count of matched spectra of the same glycopeptide on the other hand is a measure for the peak width, which is also depending on the analyte amount and is affected by ionization processes to a lesser extent. It is obvious, however, that spectra counts are an inherently inaccurate measure and are heavily affected by the settings of the MS instrument. Besides, the prerequisite that all glycans are identified with same efficiency by the software is not strictly met. Nevertheless, the number of matched spectra can give a rough estimate of the relative abundances of the different glycan-species present in each glycopeptide, and is an information obtained with no additional effort.

The number of assignments for each species was compared to the highest intensity of the M+1 peak and the peak area, determined by *mzMine* after deisotoping. Each glycosylation site was investigated separately, considering the peptide with the highest number of matches and the most intense charge state. For site 2 this was WFYIGSAFRNPEYNK (1 missed cleavage, 1890.91 Da) with z=4, for site 3 EYQTIEDKCVYNCSFIK (1 mc, 2195.99 Da) with z=4, for site 4 QNGTLSKVESDREHFVDLLLSK (2 mc, 2514.31 Da) with z=5 and for site 5 TFMLAASWNGTK (0 mc, 1325.65 Da) with unmodified methionine and 2 relevant charge states, z=3 and z=4, with the maximum intensities of both charge states added. For the peak areas of the last peptide, the most intense charge state was used, i.e. z=3 for small glycans, $M_r < 2236$ (corresponding to H5N4G2), and z=4 for glycans with three or more antennas. The values of the 3 samples (technical repeats) were averaged and normalized by setting the most abundant glycan to 1. The glycans were sub-summarized to glycan classes based on the number of lactosamine units (HexHexNAc, supposedly GalGlcNAc) and sialic acids. Since no reliable unambiguous structure can be inferred from glycopeptide CID spectra alone, the number of antennae was assumed to equal the number of lactosamine units, e.g. a glycan with the composition H5N4 would be a biantennary glycan. A few glycans had 5 lactosamine repeats (H8N7), which were counted as tetraantennary. The corresponding tables are attached as supplemental data.

With all 3 measures being far from accurate, they, however, provide a similar picture. On each glycosite, the main glycan is biantennary with 2 sialic acids, with the exception of glycosite 3, where biantennary glycans with 3 sialic acids have the highest peak area, when the main peptide is considered. This, however, is probably an artifact, since the alternative peptide CVYNCSFIK (0 mc) showed a different pattern.

Glycosite 1 For glycosite 1, no confident assignment of a spectrum to a glycopeptide of bovine AGP could be made. There were, however, a few spectra indicating the presence of Bi-Ant-SA2 ((biantennary glycan with two sialic acids) glycans but with scores below the threshold. Glycosylation of the site could be confirmed by deglycosylation of a proteinase K digest. Peptides (length between 5 and 10 residues) covering this site with N changed to D were found, whereas no unmodified ones were found, verifying that the site was glycosylated indeed.

Glycosite 2 On the main peptide of glycosite 2, 21 different glycan compositions were found, with biantennary being the major class with more than 80 % incidence, followed by triantennary glycans with 10 to 15 % and around 1 % tetraantennary. All found glycans have sialic acids, and no fucosylated species were found. From the Bi-Ant-SA2 class, H5N4A1G1 is the most abundant one, and H5N4G2 more abundant than H5N4A2. Bi-Ant-SA3 glycans are also very common. The distribution of sialic acids here seems opposite to the SA2 class, with H5N4A2G1 being the most abundant species, and H5N4A3 more common than H5N4G3, although the ratio of these differs greatly with the quantitative measures. Bi-Ant-SA1 is low abundant, with H5N4G1 50 to 90 % higher than H5N4A1. Among the triantennary glycans, those with 3 sialic acids are the most frequent ones, closely followed by those with 2 sialic acids, while 1 and 4 SAs occurred rarely. Only 3 tetraantennary species were found, H7N6A1G1, H7N6A2G1 and H7N6A1G2, with the last one being the most and the first one the least frequent one.

Glycosite 3 Glycosite 3 has the lowest glycan diversity of the 4 accessible glycosylation sites, with only 18 different glycans found on the main peptide. Biantennary glycans are dominant with 95 % relative abundance and about 2.5 % of mono- and triantennary glycans, each. One non-sialylated glycan was found, H5N4, but no fucosylated one. The relative abundances of the different sialoforms varied significantly depending on the quantitation method used. The agreement was better with the alternative peptide CVYNCSFIK (0 mc, 1189.53 Da) for which only biantennary glycans were found. According to the values generated from this peptide, most of these species have 2 sialic acids, Bi-Ant-SA2, (63-77 %), followed by Bi-Ant-SA3 (20-31 %) and Bi-Ant-SA1 (3-6 %). The distribution of Neu5Ac and Neu5Gc seems to be random. A notable feature of the peptides comprising glycosite 3 is the high degree of methylation. The monomethylated peptides elute slightly after the unmodified ones, still concurring with the tails of the peaks of the glycopeptides having Neu5Gc. Dimethylated peptides have also be observed although with low abundance. As consequence, the sialoform clusters have complex intensity distributions, making the determination of the monoisotopic masses difficult. Moreover, quantification by the peak integrals is flawed.

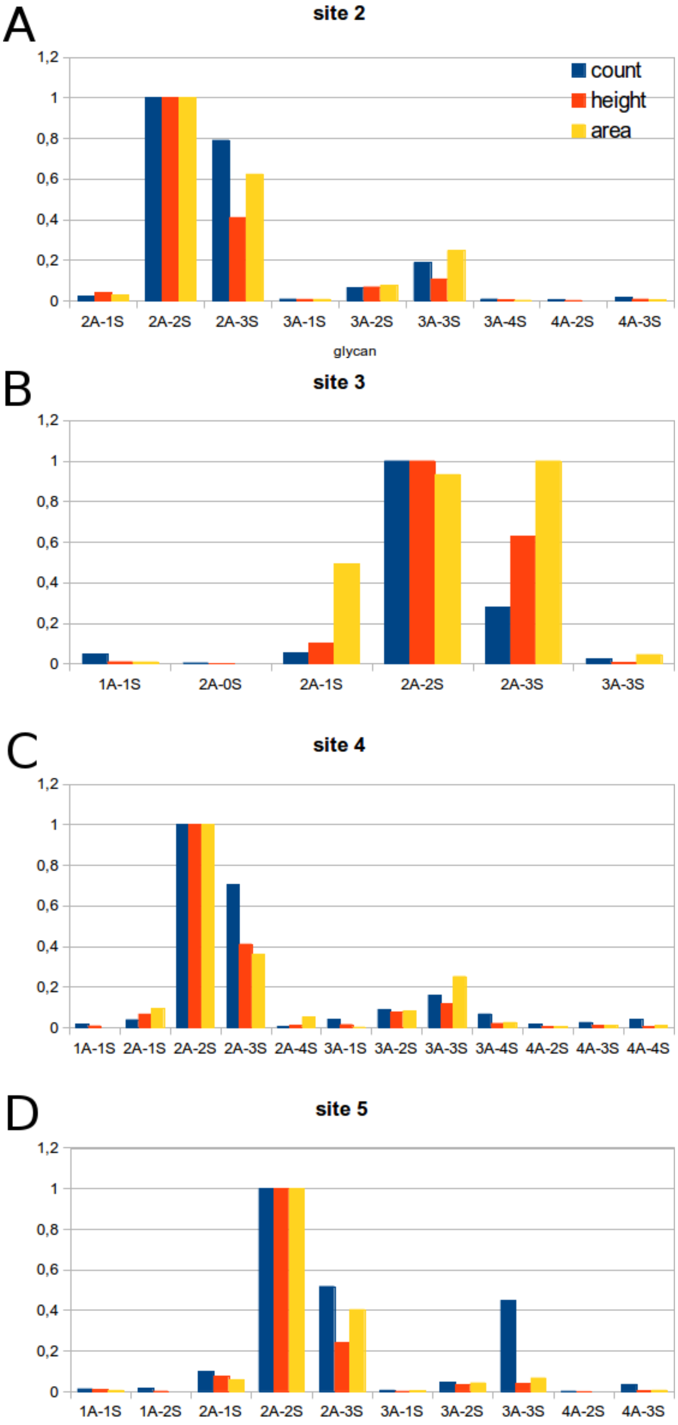
Glycosite 4 Glycosite 4 has the most diverse glycan attachments. The peptide with the most spectra assigned was QNGTLSKVESDREHFVDLLLSK (2 mc, 2514.31 Da) with a total of 692 matches spectra and 33 assigned glycans. A total of 483 spectra was matched for the alternative peptide IYRQNGTLSK (1 mc, 1178.65 Da), covering even 45 glycans. The predominance of different charge states, however, hindered comparisons of the relative abundances in the latter. While in the first long peptide z=+5 was the dominant charge state, in the shorter one with only one missed cleavage both z=+3 and z=+4 were relevant. 48 glycans were found on both peptides combined, amongst them one non-sialylated (H6N5) and several fucosylated species. Again, biantennary glycans are dominant, with Bi-Ant-SA2 being the most abundant class, followed by Bi-Ant-SA3. Bi-Ant-SA1 and even one Bi-Ant-SA4 were found as well. Looking at the triantennary glycans, which have an abundance of about 15 %, there are slightly more tri-sialylated glycans than bi-sialylated ones, with a minor fraction having one or four sialic acids. To a low degree, mono- and tetra-antennary glycans also occurred. The distribution of sialic acids appears to be random for species with one or two SAs, whereas on species with three or four sialic acids a bias towards Neu5Ac was observed. Glycosite 4 is the only site, where considerable levels of fucosylated glycans were found, even though *FindGlycoPeptides* assigned only few spectra. The peak heights were between 1 and 12 % of the non-fucosylated ones. The glycan series Bi-Ant-SA2F1 has about 6 % of the intensity of the non-fucosylated series, making it more abundant than all mono- and tetraantennary glycans combined. Only a single spectrum of H5N4G2F1 was assigned though, because the z=+3 spectra had a very low *m/z* 366.1 peak and the z=+4 spectra missed the Y_1-peak (peptide plus first GlcNAc). Furthermore, monoantennary bisialylated glycans and species with bisecting GlcNAc were found, assuming that compositions with the same number of N and H fall into that class. However, the intensities of these peaks were very low and just above detection limits.

Glycosite 5 Site 5, on which 30 glycancs were found, is more troublesome regarding quantification. The charge states +3 and +4 are both significant, with +3 being more intense, depending on the glycan composition. For large glycans, with masses higher than that of H6N5A3 (2861 Da), the +3 charge state is outside of the recorded mass range (400

- 1400 m/z). For small glycans, on the other hand, charge state +4 is below detection limit. Therefore, the intensities of both charge states were added for quantification. For large glycans the intensity of the +3 state was estimated to be as high as the +4 state, based on the +3 ion with the highest mass found. The agreement between the quantification approaches is rather low in regards of this glycosylation site. Again, Bi-Ant-SA2 are most common, followed by Bi-Ant-SA3. For Tri-Ant-SA3 however about 45 % as many spectra as for Bi-Ant-SA2 were found, while looking at peak heights, the ratio was only 4 % and 7 % in terms of peak areas. This difference may be partly caused by massive peak tailing, a consequence of the high retention, with elution extending to the end of the chromatogram. For the minor glycans the agreement is good.

It can be concluded that spectrum counting, while not being accurate by any means, can provide an estimate of the quantitative glycan distribution that largely agrees with the peak heights without any additional effort. An overview of the relative glycan abundances, determined by the three different methods, for each glycosylation site is shown in Fig. 5.3 and the relative frequencies of glycan types found on each glycosylation site are shown in Table 5.2 .

Figure 5.3.: Relative Glycan Composition



Relative glycan abundances at each glycosylation site, determined by spectrum counting, maximum peak height of M+1 peak and deisotoped peak area. Average of 3 samples. **A** glycosite 2, peptide WFYIGSAFRNPEYNK, **B** glycosite 3, peptide EYQTIEDKCVYNCSFIK, **C** glycosite 4, peptide QNGTLSKVESDREHFVDLLLSK, **D** glycosite 5, peptide TFMLAASWNGTK. No reliable assignments were made for glycosite 1.

Table 5.2.: Glycan Compositions Found on each Glycosite

	GS 2	**GS 3**	**GS 4**	**GS 5**
Mono-Ant-SA1		1.5±1.6	0.7±0.6	0.9±0.7
Mono-Ant-SA2				0.5±0.5
Bi-Ant-SA0		0.2±0.4		
Bi-Ant-SA1	1.7±0.7	10.8±10.4	3.4±1.9	8.0±4.9
Bi-Ant-SA2	52.7±6.2	53.1±18.0	52.2±6.5	54.5±12.3
Bi-Ant-SA3	31.2±5.7	33.2±12.5	24.7±5.9	20.9±3.1
Bi-Ant-SA4			1.3±1.6	
Tri-Ant-SA1	0.4±0.1		1.2±0.7	0.3±0.3
Tri-Ant-SA2	3.8±1.3		4.2±1.5	2.0±0.9
Tri-Ant-SA3	9.2±3.5	1.4±1.1	9.1±3.4	12.3±8.0
Tri-Ant-SA4	0.5±0.4		1.9±1.0	
Tetra-Ant-SA2	0.3±0.2		0.4±0.3	0.1±0.1
Tetra-Ant-SA3	0.6±0.4		0.8±0.4	1.3±1.5
Tetra-Ant-SA4			0.9±0.9	

Relative glycan frequencies found on each bovine AGP glycosylation site in percent (average ± SD) . The numbers are based on the three mentioned quantitation methods with three technical repeats with the mentioned main peptides considered only.

Comparison with GlycoPeptideSearch For the purpose of comparing the peak assignments attained by the two search programs, i.e. *FindGlycoPeptides* and *GlycoPeptideSearch*, respectively, the data of one HPLC-ESI-MS/MS run was analyzed with latter program in detail. For *FGP* the parameter setting discussed above was used (threshold score of 20), with the software *GPS* the fragment mass tolerance was set to 1 Da. *GPS* matched 254 spectra to a single target glycopeptide with an estimated FDR of 13.7 %, as calculated by the program. 183 of the matches were in agreement with those found by *FGP*. In 31 scans matched by *GPS* only, no glycosignature was found by *FGP* since in the latter the spectra are checked only for the HN^+ fragment at *m/z* 366. Three more spectra passed the glycosignature filter but no peptide was matched, and 21 scans scored lower than 20 and were consequently not reported as assignments. 16 spectra were reported as hits by both programs, but matched with different peptides. In all of these 16 cases the *FGP* assignments were correct. In 5 of these, *GPS* assumed oxidised methionine in the peptide TFMLAASWNGTK combined with the glycan composition H5N4A1G1, but it proved to be the non-oxidised peptide with the glycan species H5N4G2. *GPS* seems to favor Neu5Ac compared to Neu5Gc, thus assigning very few H5N4G2 glycans. In at least one of those scans, both species seem to be coeluting, with the former as minor component of less than 10 %. In the other 10 spectra, *GPS* matched the peptide EYQTIEDKCVYNCSFIK with the glycan composition H6N4A1F1 and H5N4A1F2, a result that was not in agreement with the most abundant Y-ions as well as the B-ions. In 9 of these cases, *FGP* correctly matched them to WFYIGSAFRNPEYNK with H5N4A1G2 and H5N4A2G1. In the last of these 16 scans, both programs correctly reported the peptide EYQTIEDKCVYNCSFIK, but differed in the glycan composition. *GPS* assigned H6N4A1F1 (2221.74) with no precursor mass correction (Δm -9.97 ppm), while *FGP* assigned H5N4A1G1 (2220.77 Da) with a precursor mass correction of -1 (Δm -5.85 ppm).

The corresponding Y-ions found strongly support the glycan suggested by *FGP*. It must be noted that the peptides EYQTIEDKCVYNCSFIK (2195.99 Da) and WFYIGSAFRNPEYNK (1890.91 Da) are coeluting for a certain time span and produce some Y-fragment ions that are difficult to distinguish with the mass accuracy available, for example H5N4A1G1 on WFYIGSAFRNPEYNK (z=3, 1371.6 Da) and H5N4A1 on EYQTIEDKCVYNCSFIK (z=3, 1370.9 Da). In the spectrum an intense signal is found at 1371.9 which presumably arises from the superimposition of the monoisotopic peak of the latter species with the C13 peak of the former one. The different results of both programs arise from the fact that *FindGlycoPeptides* tries to maximize the matched intensity and thus selects the more abundant species if two or more glycopeptides are fragmented at the same time, while in *GPS* assignment is based solely on presence of the Y_0 - Y_3-ions.

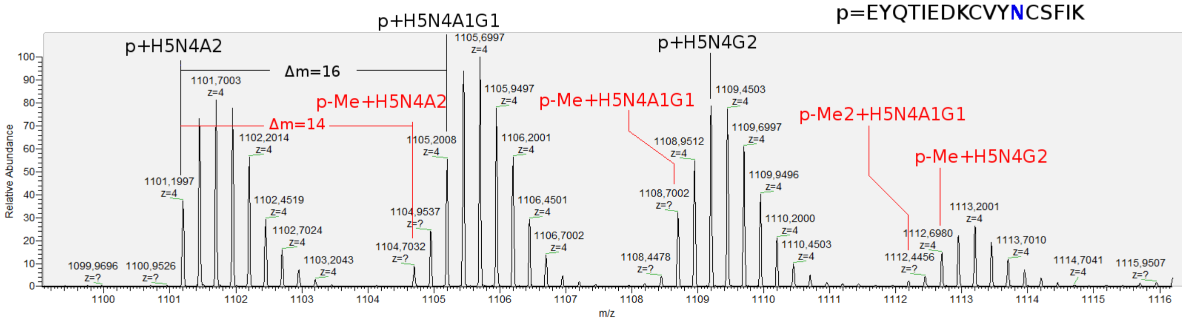
With *GPS*, the 254 matched spectra cover 58 peptide-glycan pairs on 13 different peptides, considering a peptide with oxidised methionine as distinct, otherwise the number decreases to 53 pairs on 12 different peptides. 38 of the 58 peptide-glycan pairs could be confirmed by manual inspection, 36 of which were found by *FindGlycoPeptides* as well. The two species not found carry a glycan which was not part of the glycan list used by *FGP* (H5N4A4 on NPEYNKSAR and EYQTIEDKCVYNCSFIK). The other 20 assigned species could not be verified and are either uncertain or wrong. This equates to a false discovery rate of 34.5 % based on the number glycopeptides. The number of different glycosite-glycan pairs that were found was 47 and 27 of them proved correct. The according rate of false positives is 40.4 %. A summary is shown in Table 5.3 .

Table 5.3.: Comparison of *FGP* and *GPS*

	FindGlycoPeptides	**GlycoPeptideSearch**
Spectra Assigned	1520	254
FDR Estimates	4.4 %	13.7 %
Peptide-Glycan Pairs	194	58
Glycosite-Glycan Pairs	107	47
Correct GS-Glycan Pairs	98	27
Percentage Correct	91.6 %	57.4 %

Peptide Modifications A common peptide modification in the AGP sample was the cyclisation of N-terminal glutamine with loss of ammonia, leading to pyroglutamate associated with a mass shift of -17. This was observed with all peptides starting with Q, indicating that this modification occured in course of the sample preparation. Another amino acid modification found was methylation, as indicated by a mass shift of +14 Da. N-methylation of lysine and arginine is the most frequent type of enzymatic methylation, potentially concerning every tryptic peptide and resulting in a mass shift of +14 Da for every methyl group. Because of the drastic increase of the search space, consideration of all possible methylation forms is not reasonable. Each lysine residue can carry up to 3 methyl groups, each arginine two. In practise, the isotope clusters of a methylated glycopeptide coincide with those of the unmethylated form, where a Neu5Ac is exchanged for a Neu5Gc (or a desoxyhexose for a hexose). This can complicate monoisotopic mass determination of the glycolyl form. Moreover, different species will be isolated for fragmentation, producing mixed fragment spectra. Most noticeable is the presence of the HNF$^+$ B-ion peak at *m/z* 657 in a spectrum of a glycopeptide which lacks Neu5Ac in the unmethylated form. Because of these difficulties, methylated peptides are not considered at all in *FindGlycoPeptides*. The faulty precursor mass determination is corrected by the program and the presence of unfitting peaks is not prohibitive for the successful assignment of a spectrum. Both aspects, however, decrease the score, so that the affected spectra result in scores below the threshold. A mass spectrum which indicates the presence of methylated bovine AGP peptides is shown in Fig. 5.4 .

Figure 5.4.: Mass Spectrum with Evidence of Peptide Methylation



MS^1 spectrum of the glycosylated bovine AGP peptide EYQTIEDKCVYNCSFIK (GS3, 1 mc, 2195.992 Da; RT: 30.16 min). The monoisotopic peaks of the unmethylated glycopeptides (H5N4A2 - 4400.764 Da, H5N4A1G1 - 4416.759 Da, H5N4G2 - 4432.754 Da), which are labeled in black, are separated by 16 Da. The monoisotopic peaks of the methylated glycopeptides, labeled in red, have masses which are 14 Da higher than their unmethylated counterparts.

Other Glycoprotein Components The commercial AGP preparation contains other glycoproteins as minor components. The protein composition of one data set was determined by *X!Tandem* and the sequences of those glycoproteins scoring better than -50 (expectation value) were extracted from the proteome database and used for a multi-protein analysis with *FindGlycoPeptides*. The signal peptide of AGP was removed in the sequence database, all other sequences were left the way they occured in the uniprot proteome. From the 12 sequences used for searching, a total of 70 spectra, which were checked manually, could be assigned to glycopeptides of 6 proteins other than AGP. More than one peptide was found for 3 of these. The estimated false discovery ratio per scan was 7.14 %. This demonstrates the capabilities of the program to find minor species in a relatively complex sample, provided that they were isolated for fragmentation. The found glycopeptides are listed in Table 5.4 .

Table 5.4.: Found Glycopeptides of Other Proteins Found in AGP Samples

Peptide	**Glycan**	**Assigned Spectra**
Leucine-rich alpha-2-glycoprotein 1		
NHTR	H5N5A2	1
	H5N5A1G1	2
	H5N5G2	1
	H5N5A3	1
	H5N5A2G1	1
Serpin A3-1		
TPFDPKHTEQAEFHVSDNK	H5N4A2	1
	H5N4A1G1	3
	H5N4G2	2
Serpin A3-5		
HTEQAEFHVSKNK	H5N4A2G1	6
	H5N4A1G2	1
SLINDYVKNK	H5N4G2	2
Serpin A3-6		
VHCLPENVTPEEQHK	H5N5A2G1	1
HTEQAEFHVSDNK	H4N4A1G1	1
	H5N4A2	2
	H5N4A1G1	3
	H5N4G2	4
	H5N4A3	3
	H5N4A1G2	4
	H5N4A2G1	2
	H5N4G3	2
TRFDPKHTEQAEFHVSDNK	H4N4A1G1	2
Serpin A3-7		
TPFNPNHTYESEFHVSQNER	H5N4A2	1
	H5N4A1G1	1
	H5N4A1G1F1	1
	H5N4G2	3
	H5N4A3	1
	H5N4A1G2	1
LINEYVKNKTHGK	H5N4A1G1	1
	H5N4G2	1
	H5N4A3	2
	H5N4A2G1	2
	H5N4A1G2	1
LINEYVKNK	H5N4A2	1
	H5N4A1G1	2
	H5N4G2	3
	H5N4A1G2	2
Transthyretin		
SLGISPFHEFAEVVFTANDSGPR	H5N4A2G1	2

Glycopeptides from minor glycoprotein compounds of the bovine AGP sample. All assignments were verified.

5.2. Human AGP

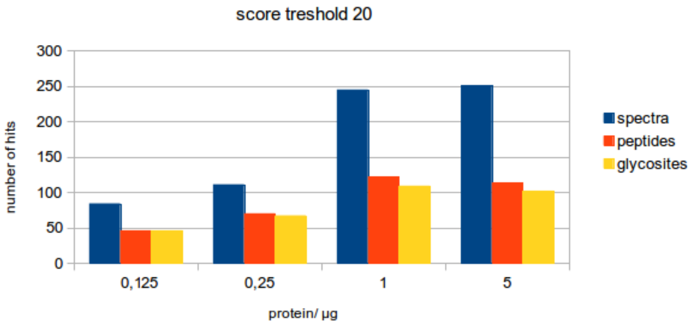
An analogous study as reported for bovine AGP was carried out with human AGP. The human AGP (taken as standard protein purchased from a supplier) was digested with trypsin in solution and measured using the standard method with the 15 cm column. 4 different amounts of digest were used for analysis to determine the correlation between protein amount and number of identified glycopeptides as well as false discovery rate.

There are two different genes for human AGP coding for two protein variants, AGP1 and AGP2, which are largely homologous. Two of the glycosylation sites (2 and 3) are present in peptides of identical sequence in both AGP variants and can thus not be distinguished in a digest of both proteins. The peptides containing sites 1, 4 and 5 have different sequences in the two AGP variants. The ratio between the two proteins is unknown, but in the investigated samples AGP2 is present in significant amounts.

Effect of Sample Load on Assigment Quality HPLC-ESI-MS/MS runs were made loading 0.125, 0.25, 1 and 5 µg of protein digest. The number of recorded MS^2 spectra was rising with higher sample amount, ranging from 5300 to 9900. The number of spectra with glycosignature (at *m/z* 366) on the other hand had a maximum at 1 µg of sample with about 3400 scans. The percentage of fragment scans having a glycosignature was around 40 % for 0.125, 0.25 and 1 µg and decreased to under 30 % with 5 µg. This might be a consequence of ionization suppression, which has greater effect with high sample load. Consequently, the number of assignments also reached a maximum with 1 µg with 245 assigned spectra and 122 peptide-glycan pairs, with a threshold score of 20, compared to 251 assigned spectra but only 113 distinct peptide-glycan pairs at 5 µg. The number of matched decoy peptides on the other hand also decreased, so that the estimated false discovery rate decreased from 15 to 12 % at a threshold score of 20 and from 3.5 to 0.75 with a threshold score of 40. *GlycoPeptideSearch* estimated an FDR of 29 and 18 %, respectively. The different numbers of assigned spectra, peptide-glycan pairs and glycosite-glycan pairs (distinct glycopeptides) are illustrated in Fig. 5.5 .

In all runs combined, a total of 115 validated distinct glycopeptides were found. In comparison with the bovine protein, the glycan distribution is markedly different. The most obvious difference is, as expected, the absence of Neu5Gc. The glycan abundances are only a rough estimate, based on the number of matched spectra and the number of samples the species were found in. According to our findings, the glycans are larger than in bovine AGP, with triantennary (H6N5X) being the most abundant glycan on most glycosylation sites, followed by tetraantennary (H7N6X). Mono- and biantennary glycans were quite rare. H5N4A3, which was common on bovine AGP was found on one glycosite only. On the other hand, a high number of fucosylated glycans was identified and also some bifucosylated ones. In many of those fucosylated species, no pNF or similar fragments were found at all, indicating that the fucose is not attached at the core. This total absence might be seen as indicator that fucose-migration processes during CID in the positive ion mode as reported by Wuhrer et al [34] were not active in the sense that migration from the antenna towards the core position took place. There was also a high number of glycan compositions with the same number of hexoses and N-acetylhexoses which might be indicative of bisecting GlcNAc. Due to the higher variety of glycan compositions and the higher number of discriminable glycosites, the total number of distinct glycopeptides found was similar to bovine AGP, even though there were no sialoforms with Neu5Gc.

Figure 5.5.: Effect of Different Sample Amounts



Number of assigned spectra, glycopeptides and distinct glycopeptides with different sample loads of human AGP with a score threshold of 20. A threshold score of 40 provides a similar picture with about half as many assignments.

Comparison with GlycoPeptideSearch The results that *GlycoPeptideSearch* produced for two runs (0.125 µg and 5 µg) were compared in detail with the results from *FindGlycoPeptides.*

0.125 µg With 0.125 µg of digest, *GPS* assigned 75 spectra to a single peptide, covering 48 peptide-glycan paris and the same number of distinct glycopeptides (glycosite-glycan pairs). The estimated false discovery rate was 38 %. *FindGlycoPeptides* assigned 84 spectra with a threshold score of 20, covering 46 peptide-glycan and glycosite-glycan pairs with an FDR estimate of 14.8 %. (The definition of FDR in both programs is identical.)

29 spectra produced the same hits in both programs, all of which could be verified. In 19 of the *GPS* hits, no glycosignature was found by *FGP*, because only the HN$^+$ ion at *m/z* 366 is considered with *FGP*. Ten of those assignments could be verified, the other nine are either uncertain or wrong. In eleven of the *GPS* assignments, of which six could be verified, *FGP* could not match any peptides. In those false negatives, the glycan was rather big, ranging from H6N5A3 to H8N7A4F1, so that the matched intensity with the Y-ions of the default glycan (H5N4A2) was not high enough for the scan to be considered further. To overcome this problem, the minimum matched intensity could be reduced for better coverage of those species, with a higher number of false positives as trade-off. In nine cases, *FGP* found other peptides than *GPS*, with all of them scoring below the threshold. Two of those turned out to be correct. In the last seven *GPS* hits, of which four were correct, *FGP* found the same peptide, but the score was too low for them to be reported as successfully assigned, either because assigning other glycans or as result of high noise levels.

Overall, 34 of the 48 peptide-glycan pairs reported by *GPS* could be confirmed, amounting to 29.2 % of false positives. *FGP* in comparison returned 46 peptide-glycan pairs, of which 42 were correct, resulting in 8.7 % of false positives. 20 pairs were found by both programs, 14 were found by *GPS* only whereas 22 were only assigned by *FGP*. The total number of distinct glycopeptides covered by both programs was 56, so both programs delivered results complementary to some extent.

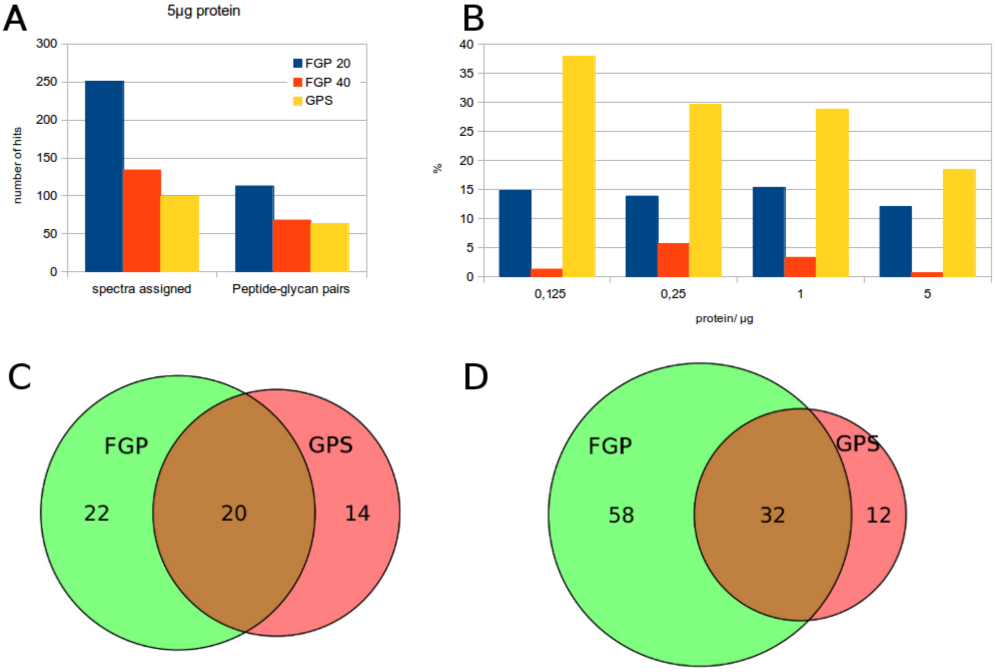
5 µg With 5 µg of digest, *GPS* assigned 99 spectra to a single peptide, covering 62 peptide-glycan pairs or 52 glycopeptide-glycan pairs with an FDR estimate of 18.5 %. 79 of the 99 assigned spectra could be validated, so the real FDR is 20.1 %. The respective numbers for *FGP* are 251 spectra, 113 peptide-glycan pairs, 102 glycosite-glycan pairs and 12.1 % estimated FDR.

Both programs reported the same results for 40 spectra, all of which could be confirmed again. No glycosignature defined in *FGP* could be found in 26 of the *GPS* hits, though 22 of them were correct. No peptide was matched by *FGP* in nine of the *GPS* hits, of which seven were correct. Again, the spectra in concern resulted from large glycans, ranging from H6N6A2F1 to H9N8A4F1. In 21 spectra, of which only six could be confirmed, *FGP* matched other peptides and scored below the threshold. Four of these six correct assignments made by *GPS* concerned the peptide SVQEIQATFFYFTPNK with a mass of 1918.951 Da, which was mistaken as QNQCFYNSSYLNVQR having a mass of 1919.863 Da by *FGP* assuming in this case a precursor correction of +1 Da. The fragments of both peptides are difficult to distinguish with the available mass accuracy of the ion trap. A certain distinction can only be made by taking the exact precursor mass into account. One option would be to analyze the isotope cluster and choose the correct monoisotopic peak. The other option would be to select the glycopeptide ion exhibiting the minimum mass deviation. In a particular scan the glycopeptide assigned by *FGP* (QNQCFYNSSYLNVQR with H6N5A2, 4489.768 Da) has a mass deviation 23.24 ppm from the supposed monoisotopic peak, whereas the correct *GPS* assignment (SVQEIQATFFYFTPNK with same glycan, 4488.856 Da) departs only 3.55 ppm from the real monoisotopic peak. However, the program is not yet designed to look for other possible species if the initial match is rejected. The high mass error leads to a score below the threshold, so the wrong hit is not reported as successful hit. Thus, false positives are avoided anyway.

In the remaining three spectra, both programs assigned the same peptide-glycan combination, but *FGP* did not report them as successful because the scores were too low as a result of high noise. All of them could be verified.

Looking at the different peptide-glycan pairs found by *GPS*, 46 of the 62 could be confirmed, so the rate of false positives is 25.8 %. These peptides represent 56 glycosite-glycan combinations, with 44 of them having at least one correct assignment, resulting in false positive rate of 21.4 % in this respect. *FGP* produced 113 different peptide-glycan pairs, of which 95 could be confirmed (15.9 % false positives). The number of distinct glycosite-glycan pairs was 102, of which 90 were confirmed for at least one peptide (11.8 % false positives). Five of them could not be confirmed for all peptides, i.e. the assignements could not be verified on peptides with different numbers of missed cleavages. The total number of glycosite-glycan pairs covered by both programs was 102, with 33 of them found by both, although one of was a false match by *GPS* whereupon it was correctly matched to another spectrum by *FGP*. Except from this case, all concurring matches were confirmed as being correct. 77 glyosite-glycan pairs were found by *FGP* only and 58 (75.3 %) were correct. 26 pairs were found by *GPS* exclusively, of which only 12 (46.2 %) could be verified. Fig. 5.6 shows a comparison between *FGP* and *GPS* in terms of FDRs as well as the number of spectra assigned.

Figure 5.6.: Comparison Between *FindGlycoPeptides* and *GlycoPeptideSearch*



A Number of spectra assigned and distinct glycopeptides found in a sample of 5 µg of human AGP. Comparison of different *FGP* threshold scores and *GPS*. **B** FDR estimates with FGP threshold scores of 20 and 40 as well as with *GPS*. With a threshold score of 40 *FGP* returns a similar number of assignments compared with *GPS*, but with much better FDRs. **C** Venn diagram representing the overlap of distinct glycopeptides between *FGP* (threshold score 20) and *GPS* in 0.125 µg human AGP. **D** Same as C with 5 µg human AGP (only confirmed hits considered for Venn diagrams).

5.3. Other Standard Glycoproteins

Bovine fetuin and asiolofetuin (from fetal calf serum), chicken ovalbumin as well as rabbit IgG were cleaned by SDS-PAGE, digested with trypsin (in gel digest) and measured using the standard method with the 15 cm column. In comparison with AGP, these proteins have a limited variety of different glycopeptides. While bAGP has five glycosites, the rabbit IgG chain c has only one, ovalbumin has two and the fetuins have three, each. Besides, AGP is by far the smallest of those proteins, producing a large portion of glycosylated peptides, while digests of the others contain mainly non-glycosylated peptides, inducing problems with ionization suppression if the glycopeptides are not enriched. Furthermore, some glycosites are located within large tryptic peptides and this leads to additional difficulties in separation, to lower signal intensity and higher noise. Glycopeptides which have a peptide portion with a mass higher than 2000 are problematic in this respect, "ideal" peptides have a mass of less than 1000. The first glycosite of bovine AGP, for example, is covered by a peptide with at least 2604 Da (if no cleavage site is missed), and therefore, no associated glycopeptide could be assigned reliably. However, another bovine AGP peptide with a molecular weight of 2514 Da produced intense signals and a high number of fragment scans which could be assigned, demonstrating that ionization highly depends on the peptide sequence. An overview of the investigated glycoproteins is shown in Table 5.5 .

Table 5.5.: Overview of Five Standard Glycoproteins

Protein	**Pot. Sites**	**Found Sites**	**M$_r$ (avg)**
Alpha-2-HS-glycoprotein (Asialofetuin)	3	2	36353
Fetuin-B (Fetuin)	3	1	40846
Ovalbumin *	2	1	42750
rabbit Ig gamma chain C region	1	1	35404
bovine AGP	5	4	21253

* in ovalbumin, both glycosites are on the same tryptic peptide

5.3.1. Fetuin and Asialofetuin

For the analysis of these two glycoproteins (i.e., Fetuin-A (Asialofetuin or Alpha-2-HS-glycoprotein) and Fetuin-B (Fetuin)) two different commercial standard protein samples were used. However, both proteins were found in each sample. In the Fetuin-B sample, the number of scans assigned to Fetuin-A was even higher than those assigned to Fetuin-B. The difference of the glycan species found in both samples was striking: In the Fetuin-B sample only glycans having sialic acids were found, while in the Asialofetuin sample most glycans were not sialylated. This observation is probably linked to the purification process of the proteins, which obviously includes sialic acid targeting affinity materials.

The Asialofetuin run had around 10500 fragment scans, and in 2200 of them a glycosignature was found. Only 51 thereof were assigned with a score above 20, covering 19 peptide-glycan pairs and 17 glycosite-glycan pairs, 14 (82 %) of which could be verified. The FDR estimate of was 5.1 %. The Fetuin run produced 10800 fragment scans, 1200 with glycosignature, of which 73 could be matched to a glycopeptide above the threshold score of 20. These covered 27 peptide-glycan pairs and 23 distinct glycopeptides, of which 19 (83 %) could be confirmed. The FDR estimate was 2.3 % for this run.

Considering only confirmed hits, 32 distinct glycopeptides were found in both samples, covering 3 glycosylation sites. Additional glycopeptides were detected analyzing the mixture of 5 glycoproteins, which was digested with trypsin in solution, increasing the total number to 47.

With the species found analyzing the in-solution digest of the protein mixture, 19 glycopeptides were found for Fetuin-B, all belonging to the third glycosylation site. Most common were triantennary glycans, most of them sialylated, with Neu5Ac strongly prevailing. A few spectra were assigned to the peptide PSSLLSLDCNSSYVLDIANDILQD-INRDR (3305.6 Da), covering the first glycosylation site of Fetuin-B, but none of them could be confirmed. No evidence at all was found for the presence of glycans on the second potential glycosylation site, though this peptide (1529.8 Da) lies in a mass range convenient for analysis. One possible reason could be that cysteine is part of the consensus sequon NCT, impeding carbamidomethylation due to steric hinderance, but this was not observed with bovine AGP, where glycans were found on peptides having an NCS sequon.

28 distinct glycopeptides originating from Asialofetuin were found, 17 on the second and 11 on the third potential site. The glycan distribution is similar to Fetuin-B, with a high number of triantennary glycans. H6N5 and H6N5A3 were the most abundant saccharides on both sites, which was also confirmed by peak intensities. Compared to Fetuin-B, the number of unsialylated glycans was higher, as was their abundance. No glycopeptides were found for the first site, which was expected since the smallest tryptic peptide has 31 residues and a mass of 3670.8 Da, impeding analysis with the present experimental setup.

The peptides that were glycosylated and the number of assigned spectra are listed in Table 5.6 and The glycans found on the peptides of Fetuin-A and Fetuin-B are listed in Table 5.7 .

Table 5.6.: Glycosylated Peptides Assigned to Fetuin A and B

Protein	**found Peptides**	**GS**	**MC**	**M$_R$**	**Fet**	**Asf**
Fet A	KLCPDCPLLAPLNDSR	2	1	1867.933	14	7
Fet A	VVHAVEVALATFNAESNGSYLQLVEISR	3	0	3015.571	30	22
Fet A	LCPDCPLLAPLNDSR	2	0	1739.838	5	7
Fet B	GENATVNQRPANPSK	3	1	1581.791	22	10
Fet B	GENATVNQR	3	0	987.478	2	

Overview of the found glycosylated peptides in the fetuin and asialofetuin samples with the numbers of assigned spectra. In Fetuin-A two of three potential glycosylation sites were found with confidence, in Fetuin-B only a single one. The right columns indicate the number of assigned spectra in the Fetuin (Fet) and Asialofetuin (Asf) samples respectively.

Table 5.7.: Glycosylated Peptides Correctly Assigned to Fetuin-A (Asialofetuin) and Fetuin-B

FetuA	**A**	**B**	**C**
GS 2			
H4N3	1		1
H4N4			8
H5N4	6		21
H5N4A1			2
H5N4A1G1			1
H5N4A2		6	14
H5N4A3			4
H5N5	1		4
H6N5	5		15
H6N5A1	1		4
H6N5A2		2	8
H6N5A2F1			5
H6N5A2G1			14
H6N5A3		11	17
H6N5A3F1			1
H6N5A3G1			7
H6N5A4			4
GS 3			
H5N4	4		9
H5N4A2		2	
H5N5	2		3
H6N5	13		46
H6N5A1	2		1
H6N5A2		4	1
H6N5A2G1		6	
H6N5A3		13	
H6N5A3F1		1	
H6N5F1			2
H7N6	1		1

FetuB	**A**	**B**	**C**
GS3			
H5N4	2		2
H5N4A1		1	
H5N4A1G1		1	
H5N4A2		3	6
H5N4A3		1	
H5N5	2		
H5N5A2		1	
H5N5A3F1		1	
H6N5	5		4
H6N5A1		1	5
H6N5A1G1		1	
H6N5A2G1		5	
H6N5A3	1	6	8
H6N5A3F1		3	
H6N5A1G2			2
H6N5A2			1
H6N5A2F1			1
H6N5A4			7
H6N5F1			2

The numbers indicate the number of assigned spectra in the respective samples: **A** Asialofetuin (in gel digest), **B** Fetuin (in gel digest), **C** mixture of 5 glycoproteins (in solution digest, 5 runs combined). Note that each of the three peptides were found in both samples, but the actual overlap of glycopeptides was low, as only H6N5A3 on Fetuin-B glycosite 3 was found in both Fetuin-A and Fetuin-B samples. Preparation B is dominated by sialylated glycans, whereas only few of them were found in preparation A. The high number of assigned glycopeptides in the protein mixture can probably attributed to the differing sample preparation (in-solution digest instead of in-gel digest) and the higher number of samples measured

5.3.2. Rabbit IgG

The HPLC-MS run with the tryptic digest of rabbit IgG produced 11400 fragment scans of which only 250 had a glycosignature. This is not surprising, as Immunoglobulin G is a large and complex protein with a high number of peptides due to the variable regions, and there is only one potential N-glycosylation site in the constant region of the heavy chain. 37 of the scans could be assigned with a score higher than 20, covering 12 distinct glycopeptides, featured by the peptides EQQFNSTIR (1121.55 Da, 0 mc, 34 spectra assigned) and TARPPLREQQFNSTIR (1913.03 Da, 1 mc, 3 spectra assigned). The estimated FDR was 3.0 %, but despite this nonzero value all glycans assigned could be confirmed. 6 additional glycans were found in the glycoprotein mixture (probably because of in-solution digestion), increasing the total number to 18. The glycans were mostly small, such as H3N3, and Neu5Gc was the dominant sialic acid. Only one glycan species containing Neu5Ac, i.e. H5N4A1F1, was found with an intensity of about 7 % compared to H5N4G1F1. The glycans found on rabbit IgG as well as chicken Ovalbumin are listed in Table 5.8 .

Table 5.8.: Found Glycopeptides and Assigned Number of Spectra for Rabbit IgG and Chicken Ovalbumin

Rabbit IgG		
Glycan	**Pure**	**Mixture**
H3N3	2	2
H3N3F1	1	
H3N4		1
H4N3	5	4
H4N3G1	2	
H4N4	3	1
H4N4G1	6	
H5N4	2	1
H5N4A1F1		2
H5N4G1	8	6
H5N4G1F1		4
H5N4G2	3	5
H5N4G2F1		1
H5N5	1	3
H6N4		1
H6N5		2
H6N5G1	1	2
H6N5G1F1	3	

Chicken Ovalbumin	
Glycan	**Pure**
H4N4	1
H3N4	1
H6N2	4
H5N4	1
H7N2	2
H6N5	1
H5N2	4
H4N5	2

Left table: rabbit IgG glycopeptides found in the IgG (in-gel digest) sample and in the glycoprotein mixture (in-solution digest, 5 runs combined). Right table: chicken Ovalbumin glycopeptides found in the ovalbumin sample (with a threshold score of 0). No ovalbumin glycopeptides were found in the mixture.

5.3.3. Chicken Ovalbumin

The RP-HPLC-ESI-MS run of the tryptic digest of chicken ovalbumin gave 11617 fragment spectra, but only 77 of them exhibited the oxonium ion at m/z 366. With the default threshold score of 20 not a single hit was produced by *FindGlycoPeptides*. After setting the threshold to 0, 16 spectra were assigned, covering 8 distinct glycopeptides with an estimated false discovery rate of 2.6 %. Only one peptide was found, without missed cleavages but with up to two oxidized methionins.

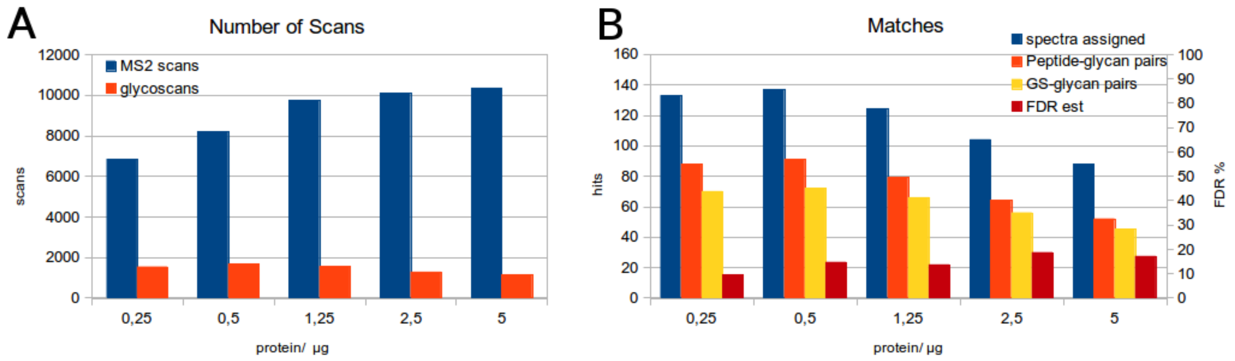
The two possible N-glycosylation sites of chicken ovalbumin are not separated by a potential trypsin cleavage site, so no statement can be made about the occupancy. However, since only small glycans were found it can be assumed that only one site was glycosylated. One the other hand, it is possible that such a large peptide (3292.6 Da) having two glycans will not become sufficiently ionized.

5.4. Mixture of Standard Glycoproteins

A mixture of 5 standard glycoproteins, composed of equal weights of bovine AGP, Fetuin and Asioalofetuin, chicken Ovalbumin and rabbit IgG, was digested with trypsin in solution and measured in different protein amounts, using the standard method with the 15 cm column.

With this complex sample the number of assignments is strongly dependent on the total protein amount loaded. In the single protein samples of human AGP, the number of assigned spectra increased with the sample amount and reached a plateau around 1 µg. 5 µg resulted in a similar number but a lower FDR estimate. In this sample the picture was different: while the total number of fragment scans increased with sample load, the number of glycoscans as well as assigned spectra reached a maximum already at 0.5 µg and considerably decreased with higher protein amounts with a slightly increasing FDR estimate. This is probably a consequence of ionization suppression, which increases if greater amounts of non-glycosylated peptides are present. The effect of different sample loads on the number of glycoscans and assignments is shown in Fig. 5.7 .

Figure 5.7.: Effect of Different Protein Amounts of Glycoprotein Mixtures



Effects of different sample amounts (digest of 5 glycoproteins) loaded. **A** total fragment scans and glycopeptide fragment scans (having the indicative HN^+ oxonium ion at m/z 366.1); **B** number of spectra assigned, glycopeptides (peptide-glycan pairs) and distinct glycopeptides (glycosite-glycan pairs) found as well as FDR estimates (threshold score 20). Optimal sample amount seems to be between 0.25 and 1.25 µg.

Most assignments resulted when using 0.5 µg of protein digest. The HPLC-MS run had 8210 fragment spectra, of which 1672 (20.4 %) exhibited a glycosignature. Of those, 137 were assigned to 91 peptide-glycan pairs or 72 distinct glycopeptides with an FDR estimate of 14.9 %. In contrast, only 1127 glycoscans out of 10352 fragment scans (10.9 %) resulted from the run with 5 µg digest loaded, with 88 spectra assigned to 52 peptide-glycan combinations or 45 distinct glycopeptides and a slightly higher FDR estimate of 17.1 %.

In all runs combined, a total of 118 distinct glycopeptides was found, belonging to four different glycoproteins. 55 glycopeptides originated from bovine AGP, 28 from Asialofetuin, 18 from rabbit IgG and 17 from Fetuin-B. As expected, no glycopeptides of ovalbumin were found, since there was no assignment with a score greater than 20 in the pure protein sample. Not all of the assigned species could be confirmed. Similar to the Fetuin samples, a high number of Fetuin-B glycopeptides were false matches, since 7 different glycan species were reported for the first glycosite, all of them being presumably wrong. Assignments were done matching the peptide PSSLLSLDCNSSYVLDIANDILQDINRDR, but actually, the peptide RPTGEVYDIEIDTLETTCHVLDPTPLANCSVR containing the first glycosylation site of Asialofetuin seems to be more likely. The mass of the latter peptide is 365.1409 Da smaller than that of the first peptide, a difference which is very similar to the mass of HexHexNAc (365.1322) which is just 0.0087 Da lower. It must be noted that proteolytic cleavage N-terminal to proline would have to occur to yield former peptide. According to generally accepted rules regarding trypsin specifity ([RK].[^P], i.e. between R or K and every amino acid except proline) such cleavage should not be found. However, statistical analysis of tryptic peptides revealed that [RK].P cleavage does occur, although with lower frequency [35]. Thus the possiblity can not be ruled out definitely but such peptides may be penalized to reflect the lower probability of [RK].P cleavage.

Of the 118 distinct glycopeptides found in the five samples, 97 (82 %) could be verified. Of these, AGP glycopeptides are the most various with 49, while the three others combined add up to 48, 24 originating from Asialofetuin, 14 from rabbit IgG and 10 from Fetuin-B. In relation to the number of species found in the pure protein samples, the biggest decrease concerned AGP where 135 distinct glycopeptides were assigned in the 3 pure protein samples. Thus, only 36.3 % as many were found for AGP when present in the mixed protein sample. In the pure AGP samples a better chromatographic separation was achieved when using a longer column and an adapted gradient and by this the number of glycan species assigned was higher. For the other proteins, the number of glycopeptides found was similar to those in the pure protein samples. For Fetuin-B, fewer glycan species were found in the mixture, for Asialofetuin and IgG a higher number of species was assigned. However, it must be noted, that the analyses of the single glycoproteins and the glycoprotein mixture were conducted using different experimental methodologies, i.e, in-gel digest for the former and in-solution digest for the latter samples. Moreover, the results of five HPLC-MS runs were combined in the case of the mixture, whereas the single proteins were measured only once. The results show that the program can also handle moderately complex samples, given that the optimal amount was loaded, although the larger search space increases the false discovery rate significantly.

GlycoPeptideSearch reported less than half as many assignments as *FindGlycoPeptides*, with FDR estimates ranging between 35 and 80 %. The results were not evaluated in detail.

Free Glycan Analysis The glycopeptide analysis of the mixture of these five glycoproteins was complemented by the analysis of the free glycans, released by PNGase F, labeled with aniline and measured by an ESI-QqTOF mass spectrometer in positive ion mode. The 36 glycans found by this method were added to the list for the program. The overlap between these experiments was poor, with only 16 of the PNGase glycans found at the glycopeptide level. The total number of different glycan compositions found by glycopeptide analysis was also 36. These findings demonstrate that both methods are, to some extent, complementary, providing distinctly different pictures. The difference can be explained by the varying ionization efficiencies. Those glycans not found in the PNGase experiment are mostly sialylated with up to 4 sialic acid residues and the acidic sugars may impede the formation of positive ions. Additionally, sialic acids are not very stably bound and prone to dissociation upon ionization. Those glycans found only in the PNGase experiment are mostly typical for Ovalbumin, of which no glycopeptide was detected because of the high mass of the tryptic peptide, at which two glycans might be attached. An advantage of the analysis on glycopeptide level is the possibility to assign the glycan to a specific glycosylation site or protein. Most of the PNGase glycans found were present in multiple proteins, limiting the value of released glycan experiments with protein mixtures. Table 5.9 shows a list of all glycans found on glycopeptide level, on free glycan level and in both experiments.

Table 5.9.: Glycans Found in Glycoprotein Mixture

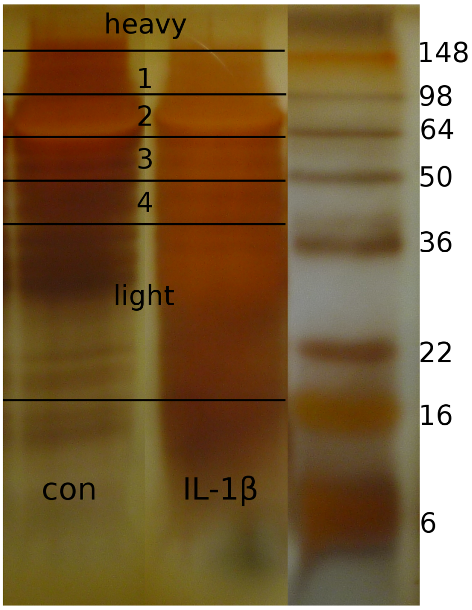
peptide + free glycan		free glycan only	peptide only
composition	proteins	composition	composition
H3N3	IgG	H3N2	H5N4A1F1
H3N4	IgG	H3N5	H5N4A1G1F1
H4N3	IgG, FetA	H3N6	H5N4A1G2
H4N4	IgG, FetA, AGP	H3N7	H5N4A2G1
H5N4	IgG, FetA, FetB, AGP	H3N8	H5N4A3
H5N4A1	FetA, AGP	H4N2	H5N4G2F1
H5N4A1G1	FetA, AGP	H4N5	H5N4G3
H5N4A2	FetA, FetB, AGP	H4N6	H6N4
H5N4G1	IgG, AGP	H4N7	H6N5A1G2
H5N4G1F1	IgG, AGP	H4N8	H6N5A2F1
H5N4G2	IgG, AGP	H5N2	H6N5A2G1
H5N5	IgG, FetA	H5N3	H6N5A2G2
H6N5	IgG, FetA, FetB, AGP	H6N2	H6N5A3
H6N5A1	IgG, FetA, FetB	H6N3	H6N5A3F1
H6N5A1G1	AGP	H7N2	H6N5A3G1
H6N5A2	FetA, FetB, AGP	H3N4F1	H6N5A4
		H4N4F1	H6N5F1
		H5N4F1	H6N5G1
		H6N4F1	H7N6
		H7N4F1	H7N6A2

Glycans found in the mixture of 5 glycoproteins. Left: glycans found both on free glycan level and on glycopeptide level, middle: glycans found as released glycans only, right: glycans found as glycopeptides only.

5.5. Biological Samples

When dealing with the supernatant of MCF-7 cell cultures, the sample pretreatment covered the separation of the proteins by means of one-dimensional SDS-PAGE and a fractionation by cutting the gel in six pieces. The four pieces in the middle were equally spaced and the two at the beginning and the end covered the high- and low mass proteins, respectively. The protein composition of each fraction was determined using *X!Tandem*. Based on the assumption that the most abundant proteins give the highest scores and yield most detectable glycopeptides, the 20 best scoring proteins or all proteins with a better score than -30 (E-value), whichever were more, were screened for potential N-glycosylation sites (NXS/T motif). The sequences of the proteins having such site were extracted from the proteome database and used for analysis with *FindGlycoPeptides*. Glycopeptide analysis dealing with such highly complex digests is a challenging problem, especially without any enrichment steps. Since most of the species present in the sample are non-glycosylated peptides or other interfering substances, the number of glycopeptides chosen for fragmentation (Top-6 criterion) is low. This is, because most glycopeptide signals were too weak for fragment scans due to the competitive nature of the ionization process. The number of scans identified as glycopeptide scans varied greatly between the gel-separated fractions, ranging from 15 in the high-M_w fraction to 596 in the 65-100 kDa sample. Comparing the glycopeptides and proteins found in the two samples without (control) and with inflammatory cell activation by IL-1β, a notable difference between these two states was the higher number of proteins in the low M_w fractions of the activated sample. The number of assigned glycopeptides, however, was not much different. (Interestingly, most of the proteins identified as being present in the low-M_w fraction of the IL-1β treated culture (by using *X!Tandem)* were actually high M_w proteins which were probably degraded by proteolytic processes mediated by the induced inflammatory pathways). An overview of the different fractions is shown in Table 5.10 and a picture of the gel in Fig. 5.8 .

Figure 5.8.: Fractionation by SDS-PAGE



Photograph of the SDS-PAGE gel of the cell supernatants with cutting sites of the fractions marked.

Table 5.10.: Overview of the Samples and Glycopeptide Isolation Yield

Sample	M_R / 10^3	Proteins	Glycoproteins	MS^2 Scans	Glycoscans	%
c-heavy	$>$ 150	20	16	4871	23	0.5
c-1	100-150	29	25	6080	120	2.0
c-2	65-100	20	18	8960	551	6.1
c-3	50-65	20	17	5320	146	2.7
c-4	40-50	20	17	4403	40	0.9
c-light	15-40	20	14	3935	37	0.9
i-heavy	$>$ 150	20	15	3558	15	0.4
i-1	100-150	22	21	4417	85	1.9
i-2	65-100	20	19	7845	596	7.6
i-3	50-65	41	35	5491	303	5.5
i-4	40-50	30	24	5266	312	5.9
i-light	15-40	45	40	7121	181	2.5

Overview of the MCF-7 supernatant derived samples: c - control samples, i - samples from cells activated by IL1β . Proteins column: number of sequences extracted, either all with *X!Tandem* expectation score better than -30 or the 20 highest scoring, glycoproteins column: number of those proteins having at least one NXS/T motif, MS^2 scans: number of fragment scans, glycoscans: number of scans with glycosignature at m/z 366.1, last column: percentage of glycoscans.

Analysis with *FGP* was not very productive, resulting in few confident assignments only. In the c-1 sample 13 spectra were assigned that could be verified, comprising 9 distinct glycopeptides of two glycoproteins. 7 glycans were assigned to the peptide VVNSTTGPGEHLR of Thrombospondin-1 (TSP1). Based on this finding, this protein was selected as candidate protein for the analysis of possible glycosylation changes upon inflammatory activation, conducted in another work. Several more possible glycopeptides were found in the elution range of this TSP1 peptide. For these glycopeptides a mass match was found in the MS^1 spectrum (using the program *Glycomod*), however, their presence could not be confirmed by fragment ion scans. All these species were found in the activated sample as well. Comparison of the maximum peak intensities revealed only minor differences, with none of them changing twofold or more.

In total, 28 distinct glycopeptides were assigned in 7 of the samples with a score of more than 20. Glycopeptides of 12 proteins were found, including one, Desmoplakin, for which peptides of two different glycosylation sites were assigned. The presence of Neu5Gc was notable, with two glycans having been assigned and several more scans being found with the indicative B-ion at m/z 673.2 . It was probably incorporated from the culture medium containing FCS. The table listing all assigned glycopeptides can be found in the supplemental data.

In such highly complex samples, the peptide search space is too large for matching the peptides by their Y-ions obtained from low accuracy fragment scans. Some of the proteins have more than 20 potential glycosylation sites, so there are hundreds of potential glycopeptides, especially since missed cleavages must be considered as well. This results in a number of peptides with very similar masses, or with masses differing one or multiple glycan masses. Therefore, an unambiguous peptide assignment is severely challenged, even with manual interpretation. The heuristic approach for the selection of possible glycopeptide sequences may not be well suited for covering all probable glycopeptides present in the sample. One has to keep in mind that the presence of the NXS/T motifs qualify the protein as potentially glycosylated, however, without implying that all these protein sites are actually glycosylated under physiological conditions. Another approach could be considering only proteins known to be glycosylated (with the risk to miss some glycoproteins) or of secreted proteins only, which are known to be glycosylated to a high degree.

For descriptive results, however, the experimental methodology must be modified. The yield of glycopeptide fragment scans could be heavily improved by various enrichment steps. Such affinity enrichment steps, however, carry the risk of inducing a bias regarding the glycosylation patterns. Having access to ETD based fragmentation data providing the amino acid sequences of the peptide backbone, the confidence of the results could be greatly increased because of the definitive peptide assignments.

6. Discussion

The program *FGP,* developed within this work, produced very good results when using high-quality HPLC-MS data. This is particularly valid when applied to the targeted analysis of glycoproteins. Bovine AGP was selected as test protein as it is an ideal protein for glycopeptide analysis, being rather small (184 amino acids without signal peptide) and including five N-glycosylation sites. With good chromatographic separation, more than 100 unique glycopeptides could be found and verified with an acceptable rate of false positives ($<$ 10 % based on the number of distinct glycopeptides, threshold score of 20). *GPS* in comparison found less than half the number of species, with FDRs around 30 %. Moreover, the high number of glycopeptide spectra allowed multiple assignments of the same species giving an approximate quantitative measure for the relative abundances of the different species. This species-specific spectrum-counts agreed in most instances with the peak heights and peak areas of these species and did not require any additional efforts. Furthermore, the confidence of the hits is raised, because false hits were commonly reported only once or twice. Under these mentioned conditions, the program proved very useful for glycopeptide discovery.

The program has the inherent constraint that the expected glycans must be supplied for the automated search. This, however, can be easily done by using data from the literature or by taking the glycan species analyzed after enzymatic release by use of PNGase. If the expected glycans are completely unknown, a few glycopeptides can be assigned manually, e.g. using the *Glycomod* tool, and a list of similar composition made. The glycan list can be of unlimited length, but the glycan masses should differ by at least 2 Da. This difference is not required if one of the glycans is fucosylated, because the program considers the HNF^{+} ion (*m/z* 512.2) for selecting the glycan species. More sophisticated rules for selection of isobaric glycans may be implemented in the future, as well as matching of theoretical glycan compositions without the dependence on a supplied mass list.

The output of the program can be useful even if the program fails to attain an assignment. All scans in which the NH^{+} oxonium ion at *m/z* 366.1 was found are reported, thus creating a list of "glycoscans" which may be assessed manually. In fact, the program could be used for producing a list of selected MS^2 scans having any fragment ion with an intensity over a particular threshold with minor modifications and will be extended in this way for taking other diagnostic B-ions into account. For scans which cover the low mass range, as produced by HCD, the H^+ and N^+ ions could be searched. Similarly, the HNF^+, HNA^+ and HNG^+ fragment ions could be used as indicators for a glycopeptide scan if the HN^+ ion is not present. If glycans not provided in the glycan list occur in the sample, the output table of matched peptides can also be valuable. Confident matches can be searched by the cumulative intensity and the glycan composition can be assigned manually, for example with aid of the *Glycomod* web tool. Exhaustive assignment of all scans having a glycosignature remains a challenge, due to ion clusters like $[M+NH_4]^+$, peptide modifications not accounted for, semi-specific cleavage and glycoproteins present in the sample but not considered for analysis. Furthermore, the program can not handle O-glycopeptides at the moment.

Impact of the HPLC-MS Setup Getting a complete picture of the glycopeptide composition is of course depending on high-quality data. Ionization suppression is a general problem in glycopeptide analysis, which occurs if glycosylated and non-glycsosylated species are eluting at the same time. To avoid this effect a good separation is crucial. The changing to a shorter column with a shorter gradient already had a pronounced impact on the number of assignments. The chromatographic method used was a standard method for high-throughput proteomics, which was not optimized for glycopeptide analysis, where coverage of all glycosylation sites of a protein is desired. Long separation times are also important to allow a higher number of analytes to be fragmented. Furthermore a high dynamic range is required to include low abundant glycoforms. With the Orbitrap-analyzer, the isotope clusters get severely distorted for low intensity signals, up to the point were the monoisotopic mass can not be determined with certainty. Better results could probably be obtained if a HPLC-separation method specifically for glycopeptide analysis is developed, accounting for the specific requirements of such analysis.

The need for good separation increases drastically for more complex samples. With a decreasing portion of glycosylated peptides, the chance of coelution of glycosylated and non-glycosylated peptides increases considerably. This is a particular concern for protein mixtures. As demonstrated, the ratio of fragment scans with glycosignatures strongly decreased with increasing amounts of a glycoprotein mixture digest. With the biological sample from the cell culture supernatants, only a few glycoscans were found at all. Attempts to modify the mass spectrometric method to select a higher number of glycopeptides for fragmentation failed, because the intensity of most of the suspected ions was too low. Fragment scans with low intensity are also difficult to interpret because of a bad signal to noise ratio. Nevertheless, some glycopeptides could be assigned even in the biological samples, and this assignments could be used as starting point for more in-depth analysis. More advanced separation or enrichment techniques, like lectin affinity chromatography or HILIC, must be employed in order to gain useful information on such complex samples.

Mass Accuracy and Ambiguity A fundamental problem with complex samples is the unambiguous assignment of peptides with similar masses, or with masses that differ by one or more saccharide units if the mass accuracy of the fragment spectra is low. With the ion trap used, fragment masses are usually shifted to higher values because the isotope peaks can not be resolved and are coalescing for higher charged species, which are typical for glycopeptide analysis conducted by HPLC-ESI-MS. Consequently, the mass tolerance needs to be high for good matches (high matched intensity), but an unsymmetric tolerance window should be used with that kind of data. Setting the mass tolerance to a relative value and assuming an unsymmetric mass deviation, favoring higher masses, is a huge advantage of *FindGlycoPeptides* compared to *GlycoPeptideSearch* which only takes absolute values in form of a symmetric window. Because of this problem with *GPS* the mass tolerance window had to be set to values as high as 1 Da to obtain reasonable numbers of assignments, leading to very high false discovery rates. With *FGP* a reasonable value for the given mass accuracy was 750 ppm.

FGP does not analyze the MS^1 scans at the moment, so it must rely on the precursor mass value determined by the instrument software or other external programs, which often does not correspond to the monoisotopic peak, particularly in case of the large analytes such as glycopeptides ($\mathrm{M}_r > 2000$). The wrong mass values have a drastic impact on the assignment rate if not corrected, especially concerning species with a high mass, where the higher isotopic peaks are often reported as precursor masses. In the investigated data sets there were also a few cases in which the reported masses were 1 or 2 Da lower than the right isotopic peak. Therefore glycan masses were searched up to 2 Da in both directions, although in most cases correction towards lower masses should be sufficient. This correction can complicate the problem of assigning the correct peptide if there are multiple peptides with masses differing by nearly 1 or 2 Da or one or more saccharide masses plus 1 or 2 Da. In future versions, the MS^1 data may be used to assign the correct monoisotopic peak by evaluation of the isotope cluster. As mentioned, however, with Orbitrap-MS the observed distribution can differ greatly from the theoretical one for low intensity signals.

An example occured with the samples containing Fetuin and Asialofetuin. The Asialofetuin peptide RPTGEVYDIEIDTLETTCHVLDPTPLANCSVR has a mass of 3670.7656 Da and the Fetuin peptide PSSLLSLDCNSSYVLDIANDILQDINRDR 3305.6247 Da, the difference of 365.1409 Da is very close to the mass of HN with 365.1322, the absolute difference being only 0.0087 Da. If the smaller peptide has the glycan H5N4 and the larger one H4N3, the relative difference is only 1.77 ppm, which is too small to exclude one of them by the mass deviation even with the mass accuracy of the Orbitrap analyzer. A few assignments for the Fetuin peptide had a score slightly higher than 20 and were therefore reported but could not be confirmed. This could have been avoided by a higher score threshold, with the trade-off of missing correct hits. A routine that identifies such cases could be implemented. A warning that the peptide identification is ambiguous might be reported, in combination with rules for selecting the most likely peptide.

The problem of ambiguous peptide masses increases drastically with more complex samples, effectively prohibiting confident assignments in the case of the MCF-7 secretome, where 20 or more protein sequences were supplied. More sophisticated decision routines can be implemented, but the requirement of a tightly restricted search space remains fundamental. Analysis of such samples is only feasible if the expected glycopeptides are known and their number is small. This means that the strength of this program lies in targeted glyoproteomics type experiments. When dealing with very complex samples, additional information on the peptides which are glycosylated has to be incorporated. This can be achieved by PNGase digestion in $H_2{}^{18}O$ followed by peptide analysis, e.g. by *X!Tandem*. Better yet would be the acquisition of ETD spectra, from which the peptide sequence can be determined.

Challenges in Achieving Complete Glycopeptide Coverage Another issue hindering the complete coverage of all relevant glycopeptides is the size of tryptic peptides, which should generally have a mass of less than 2000 Da, even though some larger peptides produced good spectra, such as the bovine AGP peptide QNGTLSKVESDREHFVDL-LLSK, with 2 missed cleavages and a mass of 2514.3 Da. Lysine and arginine are usually common, however, in many glycoproteins the smallest tryptic peptide bearing a particular glycosylation site has more than 30 residues. Furthermore, glycans near the cutting site can inhibit cleavage, although in other cases the peptide bond is efficiently hydrolyzed even if the next residue is glycosylated. Multiple glycans on the same peptide are another problem and *FindGlycoPeptides* is not capable of dealing with such species. Enzymes with a broader specificity or mixtures of proteases could be employed for such cases. Such possibilities are not implemented yet in *FGP*, but can easily be introduced. Unspecific proteases like Proteinase K are another option, allowing to some extent to control the average size of the peptides, but requiring a more sophisticated approach for automatic analysis and a drastic increase of the search space, making it applicable to pure proteins only or in combination with ETD fragmentation.

Finally, the somewhat random nature of isolating ions for fragmentation complicates the exhaustive identification of all glycopeptides present. Assuming that in bovine AGP each combination of Neu5Ac and Neu5Gc, that is four different compositions for a glycan with 3 sialic acids, is present, many of the low abundant glycans were not found, although about 14500 spectra of glycopeptides were acquired in three runs of the digest. The coverage could be increased by a higher number of replicates, or better yet, by supervised mass spectrometric methods, e.g. by blacklisting assigned ions in subsequent experiments. An alternative would be the inspection of the MS^1 spectra in an effort to find the signals of expected analytes, but this approach is rather complicated, relying on data processing methods like denoising and centroidation, and would be prone to errors due to salt clusters, peptide modifications and distorted isotope clusters, particularly affecting weak signals near the detection limit. Therefore, the interpretation of the parent mass spectra can give only clues of the remaining glycopeptides, whereas for certain assignments fragment spectra are essential.

Despite these limitations making an exhaustive assignment of all glycan-peptide pairs species difficult, *FindGlycoPeptides* can be a useful tool for targeted glycopeptide analysis of large HPLC-MS data sets, providing information at nearly zero cost. All that is needed is the protein sequence(s), a list of possible glycan compositions and the experimental data, conversed to mzXML format. Alternative tools for high-throughput evaluation of such data sets are scarce, with *GlycoPeptideSearch* being the only other freely available multiplatform software for this task. With the experimental setup used, *FGP* performed considerably better, both in terms of sensitivity (number of assigned species) and specificity (false discovery ratio) at comparable computation times. In addition it provides an informative output, including a summary of the found species, both easily readable and easy to parse. Glycopeptide discovery can benefit from utilizing both programs, as some species are found only by one of them, therefore increasing the number of species found. Conversely, assignments made by both programs consistently are usually correct, allowing for highly confident results.

7. Outlook

The program code will be published after some modifications and improvements. It was developed for analyzing data sets of tryptic glycoprotein digests gathered by an LTQ Orbitrap Velos with fragment scans made by the ion trap, therefore using low resolution and low accuracy spectra. For general usability, options for other proteolytic enzymes will be included as well as as options for high-resolution MS^2 data. Thorough testing will be needed for optimizing the parameters for data from different sources. In any case, the algorithm was designed for centroided peak data, so profile data must be transformed by external programs. For usability some kind of user interface may be implemented, such as configuration files or a graphical user interface.

The program is still in an early stage of development and various features may be added in the future. Matching calculated glycan compositions, similar to the *GlycoMod* tool, was used experimentally but eventually discarded. It may be implemented again for the version that will be published, together with more sophisticated glycan selection routines. The scoring, which was developed empirically, can discriminate between good and bad matches but is far from being fully developed. A more sophisticated model of glycopeptide fragmentation could be employed by valuating the presence of specific, more indicative, fragments differently. The loss of a single sialic acid or lactosamine unit for example is frequently observed and rewards can be given, if the respective fragments were found with high intensity. A score reflecting the statistical significance is desirable and an approach similar to *SEQUEST* or *X!Tandem* may be introduced.

The utilization of high-resolution CID fragment spectra, having high mass accuracies, could drastically decrease FDR. Evaluation of ETD spectra for peptide sequence confirmation would greatly increase the confidence of the assignments, so this possibility should be implemented as well. B-ions other than the NH^+ fragment could be used as indicators for for glycopeptide fragment scans to increase the number of assignments, depending on the mass range covered. Analysis of MS^1 scans has also great potential. Precursor mass correction could be employed by examination of the isotopic distribution. The deviation of the theoretical isotope cluster could be incorporated in the score.

Predicted glycopeptides, which were not isolated for fragmentation, could be searched, although the confidence of such assignments would be lower.

In the long run, porting of the program to the more powerful *C++* programming language and integration into existing frameworks like *Open-MS* would be worthwile.

Supplemental Data

bovine AGP

List of all confirmed bovine AGP glycopeptides found by *FindGlycoPeptides.* Numbers in the “count” columns represent the number of assigned spectra in the three technical repeats. The “height” columns show the maximum peak heights of the M+1 isotopic peaks and the “area” columns show the peak integrals, as determinded by *mzMine* .

				count			**height**			**area**		
composition	**m/z exp**	**z**	**avg. RT**	**I**	**II**	**III**	**I**	**II**	**III**	**I**	**II**	**III**
						WFYIGSAFRNPEYNK, m = 1890.910, site 2						
H5N4A1	952.15	4	32.26	2		2	6.54E+005	7.00E+005	1.13E+006	4.40E+006	9.80E+006	8.00E+006
H5N4A1G1	1028.93	4	37.45	51	58	45	1.67E+007	2.29E+007	2.28E+007	2.50E+008	3.40E+008	3.60E+008
H5N4A1G2	1105.7	4	43.98	42	41	27	4.79E+006	5.96E+006	6.59E+006	8.70E+007	2.30E+008	1.30E+008
H5N4A2	1024.93	4	37.57	39	40	41	8.32E+006	1.02E+007	1.31E+007	9.10E+007	1.20E+008	1.40E+008
H5N4A2G1	1101.701	4	44.10	45	49	34	6.69E+006	8.20E+006	8.53E+006	1.50E+008	1.50E+008	2.70E+008
H5N4A3	1097.703	4	44.26	25	28	36	3.63E+006	4.83E+006	5.15E+006	3.30E+007	9.70E+007	4.80E+007
H5N4G1	956.15	4	32.12	3	1	2	8.65E+005	1.25E+006	1.51E+006	1.72E+007	1.50E+007	1.10E+007
H5N4G2	1032.92	4	37.30	53	55	49	1.59E+007	1.94E+007	2.23E+007	2.20E+008	3.10E+008	3.30E+008
H5N4G3	1109.698	4	43.80	3	3	8	2.23E+006	2.76E+006	2.97E+006	5.20E+007	3.10E+007	7.10E+007
H6N5A1	1043.44	4	31.86	1	2	1	2.72E+005	4.53E+005	4.78E+005	3.10E+006	6.80E+006	7.40E+006
H6N5A1G1	1120.21	4	37.22	3	2	3	1.29E+006	1.48E+006	2.38E+006	8.30E+006	4.10E+007	1.20E+007
H6N5A1G2	1196.982	4	44.04	5	7	16	1.14E+006	1.58E+006	1.57E+006	4.30E+007	6.10E+007	1.24E+008
H6N5A2	1116.21	4	37.38	6	4	3	6.93E+005	1.09E+006	1.29E+006	5.40E+006	1.60E+007	1.90E+007
H6N5A2G1	1192.984	4	44.10	8	7	23	1.19E+006	2.30E+006	2.42E+006	3.20E+007	1.20E+008	2.90E+007
H6N5A2G2	1269.755	4	47.85			4	3.29E+005	1.63E+005	2.76E+005	n.a.	n.a.	6.80E+006
H6N5A3	1188.986	4	44.33	1	1	1	1.04E+006	1.69E+006	1.71E+006	3.18E+007	1.10E+007	4.50E+007
H6N5G2	1124.205	4	37.17	3		4	7.59E+005	3.06E+005	1.14E+006	5.30E+007	n.a.	1.10E+007
H6N5G3	1200.98	4	43.58	3	6	3	5.15E+005	5.94E+005	6.72E+005	6.60E+006	6.90E+006	2.40E+007
H7N6A1G1	1211.491	4	36.38		2	1	7.12E+004	1.59E+005	1.21E+005	n.a.	n.a.	n.a.
H7N6A1G2	1288.267	4	43.29	1	1	1	1.23E+005	2.18E+005	1.81E+005	8.40E+005	3.80E+006	1.80E+006
H7N6A2G1	1284.267	4	43.52		3	2	1.87E+005	3.23E+005	2.60E+005	1.30E+006	2.30E+006	2.30E+006
						NPEYNKSAR, m = 1077.525, site 2						

H5N4A1G1	825.579	4	6.01	3		1	5.65E+004	5.26E+004	1.22E+005	n.a.	n.a.	1.20E+006
H5N4A1G1F1	862.092	4	6.17	1	1	2	2.12E+004	5.20E+004	3.79E+004	n.a.	n.a.	n.a.
H5N4A1G2	902.351	4	9.65	1	2	3	2.62E+005	8.77E+004	1.74E+005	n.a.	3.00E+006	2.50E+006
H5N4A2	821.582	4	6.29	4	4	5	3.31E+005	2.44E+005	5.93E+005	2.50E+006	4.00E+006	5.70E+006
H5N4A2G1	898.353	4	9.99	5	7	4	7.33E+005	4.46E+005	5.94E+005	5.10E+006	1.20E+007	1.10E+007
H5N4A3	894.353	4	10.32	3	2	5	6.71E+005	4.21E+005	6.93E+005	9.40E+006	5.80E+007	1.00E+007
H6N5A1G2	993.637	4	9.51	1		1	1.35E+005	8.85E+003	1.26E+005	1.10E+006	1.30E+006	1.50E+006
H6N5A2G1	989.636	4	9.94	3	2	2	3.59E+005	1.99E+005	3.26E+005	2.10E+006	9.40E+006	9.50E+006
H6N5A2G2	1066.409	4	14.94	1		2	7.35E+004	6.14E+004	1.13E+005	n.a.	n.a.	2.20E+006
H6N5A3	985.637	4	10.15	4	3	5	4.61E+005	1.24E+005	4.33E+005	2.90E+006	4.20E+006	4.70E+006
H6N5G3	997.634	4	9.18			1	1.95E+004	1.06E+004	1.75E+004	n.a.	n.a.	n.a.
H7N6A1G2	1084.919	4	9.11	3	2	4	3.38E+004	2.96E+004	4.79E+004	n.a.	6.70E+005	8.10E+005
H7N6A2G1	1080.918	4	9.41	4	2	3	4.40E+004	4.39E+004	6.76E+004	1.20E+006	1.10E+006	1.70E+006
H7N6A3	1076.919	4	9.27			1	9.48E+003	7.07E+003	3.72E+004	n.a.	n.a.	n.a.

EYQTIEDKCVYNCSFIK, m = 2195.992, site 3

H4N3A1	937.14	4	25.87			4	1.51E+005	1.29E+005	9.44E+004	n.a.	n.a.	n.a.
H4N3G1	941.138	4	25.76	2	2	1	1.73E+005	1.78E+005	1.17E+005	n.a.	n.a.	n.a.
H4N4A1	987.911	4	25.81	2	2	3	2.43E+005	1.87E+005	1.47E+005	n.a.	2.60E+006	2.00E+006
H4N4G1	991.911	4	25.67	2	2	2	2.41E+005	2.88E+005	2.07E+005	2.00E+006	2.20E+006	n.a.
H5N4	955.652	4	22.13			2	3.12E+004	7.52E+004	9.44E+004	n.a.	n.a.	n.a.
H5N4A1	1028.426	4	25.62	2	4	4	2.12E+006	4.02E+006	3.84E+006	9.30E+007	4.60E+007	1.35E+008
H5N4A1G1	1105.197	4	29.82	69	63	54	2.34E+007	3.36E+007	3.70E+007	n.a.	1.30E+008	n.a.
H5N4A1G2	1181.971	4	35.14	2	3	1	8.50E+006	1.49E+007	1.35E+007	n.a.	3.80E+006	n.a.
H5N4A2	1101.199	4	29.94	38	47	38	1.35E+007	2.05E+007	2.20E+007	2.40E+008	1.50E+008	2.70E+008
H5N4A2G1	1177.973	4	35.29	23	13	15	1.13E+007	1.95E+007	1.77E+007	4.70E+008	3.17E+008	1.30E+008
H5N4A3	1173.974	4	35.47	17	25	22	8.77E+006	1.22E+007	1.18E+007	6.90E+007	7.50E+007	1.20E+008
H5N4G1	1032.424	4	25.53	4	5	3	2.82E+006	4.69E+006	4.84E+006	8.00E+007	8.40E+007	1.50E+008
H5N4G1F1	1068.937	4	25.46	1			8.37E+004	1.51E+005	1.29E+005	n.a.	n.a.	n.a.
H5N4G2	1109.195	4	29.74	37	46	53	2.18E+007	1.52E+007	3.05E+007	n.a.	2.10E+007	n.a.
H5N4G3	1185.968	4	34.97	1	3	2	3.87E+006	7.73E+006	6.89E+006	n.a.	n.a.	n.a.
H6N5A1G2	1273.256	4	35.00	2	2	1	8.98E+004	1.42E+005	1.45E+005	2.20E+006	1.20E+006	1.50E+006
H6N5A2G1	1269.255	4	35.10	4	1	3	1.13E+005	2.24E+005	2.34E+005	2.80E+006	3.30E+006	8.00E+006

H6N5A3	1265.255	4	34.78			1	1.16E+005	2.25E+005	1.93E+005	3.10E+007	1.60E+005	4.30E+006
CVYNCSFIK, m = 1189.531, site 3												
H5N4A1	1035.409	3	24.10	2	2	2	4.39E+005	1.23E+006	6.67E+005	7.90E+006	2.00E+007	1.20E+007
H5N4G1	1040.742	3	23.96	2	1	2	8.61E+005	1.89E+006	1.06E+006	1.00E+007	5.10E+006	9.00E+006
H5N4A2	1132.442	3	30.69	7	11	7	4.37E+006	1.16E+007	9.12E+006	4.60E+007	3.60E+008	1.00E+008
	849.583	4					4.50E+005	1.23E+006	9.18E+005	5.70E+006	9.50E+006	6.70E+006
H5N4A1G1	1137.775	3	30.50	13	29	20	9.78E+006	2.41E+007	1.46E+007	1.80E+008	3.50E+008	2.70E+008
	853.582	3					9.86E+005	2.14E+006	1.38E+006	6.00E+006	9.30E+007	1.30E+007
H5N4G2	1143.108	3	30.28	6	8	6	7.94E+006	1.97E+007	1.04E+007	4.80E+007	1.80E+008	n.a.
	857.582	4					7.25E+005	1.47E+006	1.01E+006	6.10E+006	1.80E+007	7.60E+006
H5N4A3	1229.472	3	40.39	5	8	8	1.16E+006	2.63E+006	1.81E+006	1.80E+007	3.80E+007	2.50E+007
	922.356	4					2.78E+005	4.33E+005	3.84E+005	4.60E+006	9.00E+006	3.10E+006
H5N4A2G1	1234.806	3	40.05	3	4	7	2.24E+006	4.93E+006	3.51E+006	8.20E+006	1.70E+007	5.20E+007
	926.36	4					3.74E+005	8.58E+005	6.29E+005	5.50E+006	1.10E+007	1.30E+007
H5N4A1G2	1240.139	3	39.69	3	10	7	1.71E+006	4.37E+006	3.05E+006	1.60E+007	1.10E+008	7.70E+007
	930.356	4					2.57E+005	5.94E+005	5.30E+005	4.70E+006	1.10E+007	4.80E+006
H5N4G3	1245.469	3	39.33	3	4	7	8.41E+005	2.18E+006	1.14E+006	1.50E+007	3.10E+007	1.60E+007
	934.355	4					1.03E+005	2.85E+005	1.94E+005	1.70E+006	4.10E+006	3.50E+006
QNGTLSKVESDREHFVDLLLSK, m = 2514.313, site 4												
H4N3A1	813.578	5	30.97	1			9.18E+004	8.60E+004	9.57E+004	n.a.	n.a.	n.a.
H4N3G1	816.777	5	30.91	2		1	1.25E+005	1.13E+005	1.55E+005	n.a.	n.a.	n.a.
H4N4G1	857.393	5	30.81	1			2.25E+005	1.38E+005	1.36E+005	n.a.	n.a.	n.a.
H5N4A1	886.605	5	30.77	1	3	2	1.04E+006	1.24E+006	1.76E+006	6.30E+006	8.90E+006	1.00E+007
H5N4A1G1	948.023	5	33.97	39	37	38	1.79E+007	1.90E+007	2.57E+007	2.00E+008	2.80E+008	2.90E+008
H5N4A1G2	1009.445	5	37.86	28	21	35	4.61E+006	5.44E+006	6.81E+006	n.a.	6.60E+007	n.a.
H5N4A2	944.825	5	34.16	39	31	33	9.15E+006	1.05E+007	1.34E+007	1.20E+008	1.70E+008	1.80E+008
H5N4A2G1	1006.244	5	38.00	28	25	34	6.18E+006	7.70E+006	9.70E+006	5.50E+007	7.20E+007	8.60E+007
H5N4A3	1003.047	5	38.09	9	9	19	3.22E+006	4.07E+006	5.12E+006	3.60E+007	4.30E+007	5.70E+007
H5N4A4	1061.263	5	42.61	1	1	0	4.71E+005	5.40E+005	8.02E+005	6.20E+006	5.00E+007	7.30E+006
H5N4G1	889.804	5	30.70	1	2	3	1.26E+006	1.48E+006	2.89E+006	7.90E+006	2.80E+007	5.60E+007
H5N4G2	951.222	5	33.91	37	24	35	1.52E+007	1.40E+007	2.05E+007	n.a.	n.a.	n.a.
H5N4G3	1012.641	5	37.80	3	1	8	1.84E+006	2.43E+006	2.65E+006	1.50E+007	1.90E+007	n.a.

H6N5A1	959.63	5	30.53	1	2	3	1.95E+005	3.08E+005	5.06E+005	n.a.	n.a.	n.a.
H6N5A1G1	1021.047	5	33.74	1	5	6	1.21E+006	1.66E+006	2.32E+006	6.20E+006	2.20E+007	1.20E+007
H6N5A1G2	1082.468	5	37.77	4	4	6	1.11E+006	1.72E+006	1.60E+006	1.80E+007	9.60E+006	1.30E+007
H6N5A1G3	1143.888	5	42.11	1	1	2	1.46E+005	2.10E+005	2.76E+005	3.70E+006	3.10E+006	1.30E+006
H6N5A2	1017.85	5	33.93	3	3	3	8.18E+005	1.28E+006	1.16E+006	4.00E+006	6.50E+006	7.60E+006
H6N5A2G1	1079.267	5	37.91	2	6	9	1.74E+006	2.36E+006	2.35E+006	1.70E+007	3.70E+007	2.50E+007
H6N5A2G2	1140.688	5	42.27	4	2	6	3.01E+005	4.75E+005	5.31E+005	6.10E+006	1.90E+006	3.10E+006
H6N5A3	1076.068	5	38.01	3	3	3	1.00E+006	1.45E+006	1.48E+006	1.30E+007	5.50E+007	7.90E+007
H6N5A4	1134.289	5	42.52	1	2	1	2.24E+005	2.69E+005	4.20E+005	3.60E+006	3.40E+006	1.60E+006
H6N5G1	962.83	5	30.45	4	2	1	2.25E+005	3.37E+005	5.82E+005	2.60E+006	n.a.	n.a.
H6N5G2	1024.25	5	33.64	2	3	2	6.11E+005	9.06E+005	1.18E+006	3.10E+006	3.60E+007	5.60E+006
H6N5G3	1085.667	5	37.64	4	2	4	4.60E+005	7.30E+005	9.39E+005	3.10E+007	1.00E+007	4.70E+006
H6N5G4	1147.093	5	41.62			1	too weak	6.67E+004	1.19E+005	n.a.	2.00E+006	1.90E+006
H7N6A1G1	1094.075	5	33.40	1	1	3	9.40E+004	3.37E+005	2.34E+005	n.a.	3.30E+006	2.10E+006
H7N6A1G2	1155.495	5	37.33		1	1	1.72E+005	3.01E+005	2.78E+005	n.a.	4.40E+006	n.a.
H7N6A2G1	1152.291	5	37.43	1	2	3	1.62E+005	2.61E+005	3.31E+005	3.10E+006	n.a.	5.80E+006
H7N6A2G2	1213.714	5	42.10	1		2	7.87E+004	8.78E+004	1.10E+005	1.50E+006	2.40E+006	1.50E+006
H7N6A3G1	1210.51	5	42.14	2	2	4	7.45E+004	1.19E+005	1.62E+005	2.00E+006	1.60E+006	2.70E+006
H7N6A4	1207.317	5	42.23			1	2.81E+004	5.28E+004	9.00E+004	n.a.	1.90E+006	n.a.
H8N7A2G2	1286.757	5	41.95			1	4.82E+004	7.18E+004	6.91E+004	n.a.	n.a.	n.a.
IYRQNGTLSK, m = 1178.646, site 4												
H4N3A1	910.07	3	6.94	2	1	2	5.90E+005	5.27E+005	4.16E+005	1.10E+007	8.40E+006	7.20E+006
H4N3A1G1	1012.434	3	10.48	2	2	1	5.75E+004	5.86E+004	4.98E+004	n.a.	n.a.	n.a.
H4N3G1	915.401	3	6.64	7	5	5	7.54E+005	6.81E+005	5.47E+005	1.60E+007	1.40E+007	7.50E+006
H4N4A1	733.574	4	6.77	1			3.40E+004	3.06E+004	1.68E+004	n.a.	n.a.	n.a.
H4N4G1	737.5714	4	6.53	1		1	5.87E+004	4.02E+004	3.11E+004	n.a.	n.a.	n.a.
H5N4A1	1031.78	3	6.63	5	1	1	3.74E+006	7.10E+006	5.32E+006	8.60E+007	1.40E+008	1.10E+008
	774.087	4					4.45E+005	7.18E+005	5.16E+005	6.20E+006	5.30E+006	6.10E+006
H5N4A1G1	850.86	4	9.84	20	19	24	1.96E+007	2.40E+007	2.30E+007	2.50E+008	3.20E+008	3.00E+008
H5N4A1G2	927.634	4	14.48	4	5	3	4.33E+006	4.05E+006	5.78E+006	n.a.	n.a.	n.a.
H5N4A1G2F1	964.148	4	14.11		1		9.94E+004	1.13E+005	1.08E+005	n.a.	n.a.	2.20E+006
H5N4A2	846.861	4	10.09	10	12	5	1.01E+007	1.34E+007	1.21E+007	1.49E+008	1.80E+008	1.60E+008
H5N4A2G1	923.634	4	14.66	4	6	7	9.40E+006	5.90E+006	6.80E+006	1.20E+008	1.50E+008	1.60E+008

H5N4A3	919.635	4	15.13			2	4.65E+006	4.57E+006	4.93E+006	6.00E+007	6.80E+007	7.60E+007
H5N4G1	1037.113	3	6.47	9	7	5	5.06E+006	8.08E+006	7.47E+006	9.10E+007	1.50E+008	1.30E+008
	778.087	4					4.42E+005	7.03E+005	6.90E+005	4.90E+006	8.00E+006	7.10E+006
H5N4G1F1	814.6	4	6.50		3	1	6.29E+004	1.17E+005	1.11E+005	n.a.	1.70E+006	n.a.
H5N4G2	854.859	4	9.58	26	32	21	1.68E+007	2.08E+007	1.98E+007	1.90E+008	2.50E+008	2.40E+008
H5N4G2F1	1188.159	3	9.61			1	1.68E+006	2.80E+006	2.21E+006	1.60E+007	2.60E+007	9.80E+007
H5N4G2F1	891.372	4	9.62				8.28E+005	1.38E+006	1.23E+006	7.60E+006	2.20E+007	1.00E+007
H5N4G3	931.632	4	14.41		1	1	3.32E+006	3.52E+006	3.43E+006	n.a.	8.10E+007	n.a.
H6N5	1056.457	3	6.24			1	1.24E+004	3.93E+004	3.20E+004	n.a.	n.a.	n.a.
H6N5A1	865.37	4	6.33	3	4	4	2.76E+005	8.77E+005	7.51E+005	3.00E+006	8.30E+006	3.30E+007
H6N5A1G1	942.143	4	9.69	4	3	4	2.00E+006	3.53E+006	2.80E+006	5.80E+007	4.60E+007	4.00E+007
H6N5A1G2	1018.92	4	14.44	7	4	8	3.63E+006	2.46E+006	2.80E+006	4.20E+007	5.50E+007	3.80E+006
H6N5A1G2F1	1055.428	4	14.19		1	1	2.65E+004	3.87E+004	4.64E+004	n.a.	1.10E+006	n.a.
H6N5A1G3	1095.688	4	19.57	2	5	3	1.17E+005	1.81E+005	1.51E+005	5.10E+006	2.10E+007	1.00E+007
H6N5A2	938.144	4	9.93	3	3	2	1.53E+006	2.51E+006	2.08E+006	2.10E+007	6.20E+007	2.90E+007
H6N5A2G1	1014.916	4	14.81	3	3	5	3.84E+006	3.65E+006	4.29E+006	5.70E+007	6.80E+007	9.60E+007
H6N5A2G2	1091.689	4	20.10	5	4	5	3.91E+005	3.21E+005	3.60E+005	4.10E+006	1.00E+007	7.70E+006
H6N5A3	1010.918	4	15.09	2	3	3	2.31E+006	3.27E+006	3.36E+006	3.20E+007	4.00E+007	4.80E+007
H6N5G1	869.367	4	6.16		6	4	2.44E+005	9.64E+005	7.73E+005	2.10E+006	7.80E+006	1.13E+007
H6N5G2	946.142	4	9.40	4	3	6	1.37E+006	2.13E+006	1.75E+006	4.60E+007	4.30E+007	2.00E+007
H6N5G2F1	982.656	4	9.50	1	2	1	4.64E+004	7.17E+004	5.86E+004	n.a.	1.30E+006	n.a.
H6N5G3	1022.914	4	14.05	3	5	6	1.10E+006	1.46E+006	1.30E+006	4.20E+006	n.a.	9.30E+006
H6N5G4	1099.688	4	19.96	2	4	1	7.09E+004	9.55E+004	5.18E+004	n.a.	n.a.	n.a.
H7N6A1G1	1033.427	4	9.30	2	1	1	1.80E+005	3.86E+005	2.83E+005	n.a.	n.a.	n.a.
H7N6A1G2	1110.199	4	13.67	3	4	4	2.51E+005	4.42E+005	3.49E+005	1.40E+007	n.a.	5.10E+006
H7N6A2	1029.425	4	9.51	1		1	1.22E+005	2.89E+005	2.23E+005	2.70E+006	5.80E+006	2.70E+006
H7N6A2G1	1106.2	4	13.99	4	3	4	3.64E+005	6.48E+005	4.79E+005	n.a.	n.a.	n.a.
H7N6A2G2	1182.971	4	20.07		1		3.04E+004	5.11E+004	8.26E+004	2.20E+006	n.a.	n.a.
H7N6A3	1102.202	4	14.26	3	4	4	2.39E+005	4.16E+005	3.58E+005	5.50E+006	9.10E+006	4.80E+007
H7N6A3G1	1178.98	4	20.21		3	2	4.25E+004	5.61E+004	9.05E+004	n.a.	n.a.	1.20E+006
H7N6A4	1174.979	4	20.95	1			3.17E+004	6.73E+004	3.37E+004	n.a.	n.a.	n.a.
H7N6G2	1037.445	4	8.97	1	1	1	7.22E+004	1.39E+005	1.07E+005	1.30E+006	2.90E+006	5.70E+007
H7N6G3	1114.198	4	13.29	2	3	3	9.64E+004	1.92E+005	1.46E+005	2.10E+006	3.90E+006	3.20E+006
H8N7A1G2	1201.482	4	13.17			1	4.49E+004	9.55E+004	7.21E+004	n.a.	n.a.	n.a.

H8N7A2G1	1197.484	4	13.42	3	2	3	7.24E+004	1.68E+005	1.12E+005	5.70E+005	3.90E+006	2.60E+006
H8N7A3	1193.485	4	13.67		1	4	4.58E+004	1.01E+005	7.59E+004	9.70E+005	n.a.	1.30E+006
TFMLAASWNGTK, m = 1325.649, site 5												
H4N3A1G1	1061.436	3	41.06			2	6.19E+004	5.13E+004	7.72E+004	n.a.	n.a.	n.a.
H4N3A2	1056.102	3	41.35	2		2	8.56E+004	1.54E+005	7.96E+004	n.a.	n.a.	n.a.
H4N3G1	964.401	3	33.09	1	1	1	3.46E+005	4.12E+005	2.95E+005	1.50E+007	n.a.	4.30E+006
H4N4A1	1026.764	3	33.11	1		1	3.45E+005	4.09E+005	2.34E+005	3.30E+006	5.00E+006	2.40E+006
H4N4A1G1	1129.128	3	40.84	1			7.03E+004	9.24E+004	6.63E+004	n.a.	n.a.	n.a.
H4N4G1	1032.096	3	32.90	1			4.34E+005	6.85E+005	3.89E+005	5.70E+006	n.a.	n.a.
H5N4A1	1080.778	3	32.80	3	8	6	2.00E+006	5.66E+006	3.28E+006	2.50E+007	1.20E+008	4.10E+007
	810.835	4					9.88E+004	2.66E+005	1.62E+005	n.a.	n.a.	2.00E+006
H5N4A1G1	1183.147	3	40.17	64	59	53	2.28E+007	3.79E+007	3.19E+007	n.a.	1.10E+009	8.00E+008
	887.611	4					4.79E+006	7.11E+006	6.79E+006	1.60E+008	2.50E+008	1.90E+008
H5N4A1G1F1	1231.833	3	39.60	6	7	11	2.66E+006	4.21E+006	3.40E+006	6.70E+007	1.00E+008	3.90E+007
	924.126	4					8.32E+005	9.24E+005	8.15E+005	3.40E+006	1.50E+007	1.00E+007
H5N4A1G2	1285.509	3	47.82	22	30	23	8.95E+006	8.87E+006	7.42E+006	1.80E+008	3.10E+008	2.70E+008
	964.383	4					1.91E+006	1.51E+006	1.79E+006	4.40E+007	8.00E+007	5.80E+007
H5N4A2	1177.814	3	40.41	30	55	38	1.28E+007	2.21E+007	1.61E+007	5.60E+008	5.40E+008	4.10E+008
	883.612	4					2.64E+006	4.24E+006	3.78E+006	1.40E+008	1.20E+008	2.00E+008
H5N4A2G1	1280.178	3	47.93	16	22	22	3.61E+006	6.24E+006	7.02E+006	2.50E+008	4.40E+008	3.10E+008
	960.385	4					7.64E+005	1.50E+006	1.94E+006	5.60E+007	9.80E+007	6.60E+007
H5N4A3	1274.846	3	48.23	22	17	17	1.49E+006	2.60E+006	2.37E+006	1.10E+008	3.20E+008	1.50E+008
	956.386	4					2.87E+005	8.60E+005	5.28E+005	6.00E+007	6.00E+007	1.10E+008
H5N4G1	1086.114	3	32.64	5	9	5	2.51E+006	3.68E+006	4.40E+006	7.30E+007	8.20E+007	1.50E+008
	814.837	4					1.08E+005	1.01E+005	1.59E+005	1.30E+006	n.a.	n.a.
H5N4G1F1	1134.797	3	41.35	4	3	2	1.66E+005	3.70E+005	2.34E+005	3.40E+006	n.a.	8.90E+006
H5N4G2	1188.473	3	39.95	36	43	41	2.11E+007	3.64E+007	3.50E+007	n.a.	n.a.	7.60E+008
	891.61	4					4.67E+006	7.06E+006	6.73E+006	9.50E+007	1.70E+008	2.80E+008
H5N4G2F1	1237.163	3	39.39	3	2	4	1.77E+006	3.60E+006	3.11E+006	3.50E+007	1.50E+008	8.80E+007
	928.111	4					4.69E+005	8.53E+005	1.60E+006	2.90E+006	1.30E+007	1.50E+007
H5N4G3	1290.841	3	47.81	14	18	11	4.48E+006	3.91E+006	3.52E+006	1.00E+008	1.80E+008	1.80E+008
	968.382	4					9.26E+005	8.47E+005	8.13E+005	4.40E+007	4.30E+007	3.10E+007
H6N5A1G1	1304.865	3	39.77	5	3	3	9.21E+005	2.15E+006	1.80E+006	2.40E+007	1.40E+008	6.90E+007

	978.894	4					1.15E+005	2.32E+005	2.05E+005	n.a.	n.a.	n.a.
H6N5A1G2	1055.667	4	47.80	35	29	26	6.15E+005	5.26E+005	5.14E+005	2.00E+007	7.70E+007	5.30E+007
H6N5A2	1299.524	3	39.95	1	3	3	5.02E+005	1.06E+006	9.43E+005	3.50E+007	n.a.	1.90E+007
	974.894	4					1.21E+005	1.85E+005	1.39E+005	n.a.	n.a.	n.a.
H6N5A2G1	1051.668	4	47.81	28	19	18	7.65E+005	8.32E+005	7.46E+005	4.50E+007	7.30E+007	4.00E+007
H6N5A3	1396.556	3	47.82	6	20	16	3.24E+005	5.36E+005	4.92E+005	1.30E+007	6.50E+007	5.50E+007
	1047.669	4					2.72E+005	4.90E+005	4.56E+005	6.20E+007	4.80E+007	7.70E+006
H6N5G1	1207.824	3	32.04	1	1	1	9.23E+004	2.74E+005	1.66E+005	n.a.	n.a.	1.10E+007
	906.12	4					too weak	4.48E+004	2.35E+004	n.a.	n.a.	n.a.
H6N5G2	1310.189	3	39.55		1	2	4.10E+005	7.93E+005	7.28E+005	1.40E+007	1.20E+007	3.00E+007
	982.89	4					4.95E+004	2.81E+005	8.04E+004	n.a.	n.a.	2.20E+006
H6N5G3	1059.665	4	47.82		4	2	2.35E+005	3.28E+005	3.28E+005	1.40E+007	7.40E+006	8.50E+006
H7N6A1G1	1070.176	4	38.91		1		too weak	too weak	3.87E+004	n.a.	n.a.	n.a.
H7N6A1G2	1146.961	4	47.52	1	2	4	5.54E+004	1.15E+005	7.58E+004	9.60E+006	2.30E+006	9.70E+005
H7N6A2G1	1142.953	4	47.68		6	3	7.90E+004	2.36E+005	1.85E+005	3.20E+007	3.90E+006	9.90E+006
TFmLAASWNGTK, m = 1341.644, site 5												
H5N4A1	1086.113	3	29.58	2	3	1	6.95E+005	4.78E+005	5.87E+005			
H5N4A1G1	1188.473	3	32.92	13	12	8	3.43E+005	9.08E+005	5.66E+005			
	891.604	4					1.14E+005	1.36E+005	1.62E+005			
H5N4A1G1F1	1237.163	3	35.59	6	3	4	3.73E+005	7.06E+005	7.17E+005			
H5N4A1G2	1290.842	3	45.33	21	21	31	2.71E+006	3.55E+006	3.50E+006			
	968.383	4					8.46E+005	1.13E+006	9.89E+005			
H5N4A2	1183.144	3	33.17	10	10	8	1.70E+005	3.48E+005	2.66E+005			
	887.611	4					6.30E+004	7.68E+004	7.79E+004			
H5N4A2G1	1285.51	3	45.59	6	11	7	2.79E+006	5.28E+006	4.26E+006			
	964.384	4					9.03E+005	1.75E+006	1.34E+006			
H5N4A3	1280.178	3	45.82	6	4	6	1.39E+006	2.29E+006	2.45E+006			
	960.385	4			.		4.99E+005	1.08E+006	8.32E+005			
H5N4G1	1091.445	3	29.46	1	4	1	5.50E+005	6.93E+005	8.82E+005			
H5N4G2	1193.808	3	32.70	14	10	10	2.88E+005	4.86E+005	5.00E+005			
	895.606	4					9.44E+004	1.11E+005	1.47E+005			
H5N4G2F1	1242.494	3	35.37	1	2	1	3.05E+005	8.07E+005	5.21E+005			
H5N4G3	1296.172	3	45.12	4	2	2	9.59E+005	1.82E+006	1.57E+006			

	972.381	4					3.23E+005	5.44E+005	4.46E+005
H6N5A1G2	1059.667	4	45.58	4	15	17	1.17E+005	2.40E+005	1.94E+005
H6N5A2G1	1055.667	4	45.91	7	5	11	1.69E+005	2.73E+005	2.74E+005
H6N5A2G2	1132.437	4	46.65			3	too weak	too weak	4.85E+004
H6N5A3	1051.668	4	46.15	2	4	2	9.75E+004	2.62E+005	1.78E+005
H6N5G3	1063.662	4	45.28			1	4.45E+004	9.05E+004	6.64E+004

hAGP

List of all confirmed human AGP glycopeptides and the numbers of assigned spectra in each sample.

composition	0.125 µg	0.25 µg	1 µg	5 µg	sum
		site 2 - both			
H5N4A1		1		1	2
H5N4A2			2	1	3
H5N4A2F1			1	1	2
H5N4A3	2	1	1	2	6
H5N5A2	2	2	1	2	7
H5N5A2F1		1	1		2
H6N5A1	2	1	1	2	6
H6N5A1F1		1	1		2
H6N5A2	6	4	6	10	26
H6N5A2F1	6	2	4	3	15
H6N5A2F2	1	1		1	3
H6N5A3	2	3	6	4	15
H6N5A3F1	1	2	5	3	11
H7N6A3F1	1				1
		site 3 - both			
H4N3A1			1		1
H4N3A1F2			1		1
H4N3A2				1	1
H4N4A1			2		2
H4N4A2			1	3	4
H4N4A2F1			2	1	3
H5N4A1			6	6	12
H5N4A1F1			3	1	4
H5N4A2		3	9	9	21
H5N5A1			2		2
H5N5A1F1			1		1
H5N5A2F1			1		1
H5N5A2F2	1				1
H6N5A1			1	2	3
H6N5A1F2				1	1
H6N5A2			5	4	9
H6N5A2F1			3	3	6
H6N5A2F2			1	1	2
H6N5A3		4	10	14	28
H6N5A3F1		2	4	5	11
H6N5A3F2			1	2	3
H6N5A4				2	2
H6N5F2			1		1
H6N6A1F2				1	1
H6N6A2				1	1
H6N6A2F1			2	2	4
H6N6A3				2	2
H6N6A3F1				1	1

H6N6A4F1				1	1
H7N6A1			1	1	2
H7N6A1F2				1	1
H7N6A2			1	1	2
H7N6A2F1				2	2
H7N6A3			1	10	11
H7N6A3F1			1	1	2
H7N6A3F2		1	1	1	3
H7N6A4		2	2	2	6
H7N6A4F1			1	2	3
H7N7A1F2			1		1
site 4 - A1					
H4N4A1		1		1	2
H4N4A2		2	1	1	4
H4N4A2F1			1	1	2
H5N4A2		1		1	2
H5N5A1				1	1
H5N5A2		1	3	2	6
H5N5A2F1			1		1
H5N5A3F2		1			1
H6N5A1	1	1		3	5
H6N5A1F1			1		1
H6N5A2	1		1	2	4
H6N5A2F1	1		1	2	4
H6N5A2F2		1	2		3
H6N5A3	1	4	2	1	8
H6N5A3F1	1	1	1	3	6
H6N5A3F2			1		1
H6N6A1			2	1	3
H6N6A1F2			1	2	3
H6N6A2	1				1
H6N6A2F1			3	2	5
H6N6A3F1		1	1		2
H7N6A1	1	1	3	3	8
H7N6A1F1	1	1	4	4	10
H7N6A2		1	1	2	4
H7N6A2F1		1			1
H7N6A3		1		1	2
H7N6A4		1	3	1	5
H7N6A4F1		1	2		3
H7N7			1		1
H7N7A2F1			1		1
site 5 - A1					
H4N4A1			2		2
H4N4A1F2				1	1
H4N4A2	1	2	1	1	5
H4N4A2F1	1	1	3	1	6
H5N4A1				1	1
H5N4A1F2				1	1
H5N4A2	1	1	7	3	12
H5N4A2F1	1		1	1	3
H5N5A1				1	1

H5N5A2	1	3	3	2	9
H5N5A2F1			1		1
H6N5A1	1	1	2	3	7
H6N5A1F1	2	1	1	1	5
H6N5A1F2		2	1		3
H6N5A2	3	3	8	7	21
H6N5A2F1	2	2	3	5	12
H6N5A2F2	1	1	2	2	6
H6N5A3	6	5	9	11	31
H6N5A3F1		3	6	4	13
H6N6A2F1			1		1
H7N6A2	2	1	1	1	5
H7N6A3	1	3	3	2	9
H7N6A3F1	1	2	3	1	7
H7N6A3F2	1		2	1	4
H7N6A4		1			1
H7N6A4F1			3	3	6
H7N6A4F2		1		2	3
H7N7A3F2			1		1
H7N7A4F2		1			1
site 4 - A2					
H4N4A1			1		1
H6N5A1	1				1
H6N5A2	2	2	1	2	7
H6N5A2F1	1	1	3	1	6
H6N5A3		1	2	2	5
H6N5A3F1			1	4	5
H6N6A2				1	1
H6N6A2F1			3		3
H6N6A2F2			1		1
H6N6A3F1			1		1
H7N6A1	1				1
H7N6A1F1	1	1			2
H7N6A2			2	3	5
H7N6A3				1	1
H7N6A4			1		1
H7N6A4F1			1		1
H7N7A1F2			1		1
site 5 - A2					
H5N4A1			1	1	2
H5N4A2	1	1	1	3	6
H5N4A2F1		1			1
H5N5A2		1	2	1	4
H5N5A2F1		1	1	1	3
H6N5A2	4	2	3	5	14
H6N5A2F1	2	1	2	4	9
H6N5A2F2	1	1	1	1	4
H6N5A3	5	4	12	5	26
H6N5A3F1	6	3	3	4	16
H7N6A2			1		1
H7N6A3F1	1	1	1	2	5
H7N6A3F2		2	1	1	4

H7N6A4F1		2	1	4	7
H7N6A4F2	1			2	3

MCF-7

List of all confirmed glycopeptides found in the samples derived from MCF-7 supernatants.

Thrombospondin-1 (cA1, iA1)		
VVNSTTGPGEHLR (site 4)		
H4N5F2	H5N4F2	H6N4G1
H5N4	H5N4F3	H7N4
H5N4A1	H5N4F3	
H5N4F1	H5N5	
Desmoplakin (cA1)		
SLNESKIEIERLQSLTENLTK (site 9)		
H4N4A2		
ANSSATETINKLK (site 19)		
H7N6A3		
Glucose-6-phosphate isomerase (cA2)		
HFVALSTNTTK (site 3)		
H7N6A3		
Heat shock-related 70 kDa protein 2 (cA2, iA2)		
VHSAVITVPAYFNDSQR (site 2)		
H5N5A2	H5N5A3	H5N5A3G1
Heat shock cognate 71 kDa protein (iA2)		
NQTAEKEEFEHQQK (site 6)		
H5N5A3		
Vitamin D-binding protein (iA2)		
ELPEHTVKLCDNLSTK (site 1)		
H6N5A3		
Alpha-1-antichymotrypsin (iA2, iA3, iA4)		
YTGNASALFILPDQDK (site 6)		
H6N5A3	H6N5A4	
Heat shock 70 kDa protein 1-like (iA2)		
LLQDYFNGRDLNK (site 3)		
H6N5A3F1	H6N6A4	H6N5A3F1
POTE ankyrin domain family member F (iA3, iA4)		
KEKDILHENSTLR (site 6)		
H6N5A4	H6N5A3	
Alpha-enolase (iA3)		

IDKLmIEMDGTENKSK (site 3)
H6N5A2F2
Complement factor B (iAl)
GSANRTCQVNGR (site 3)
H5N4
Elongation factor 2 (iAl)
AYLPVNESFGFTADLR (site 4)
H6N5A4

Bibliography

[1] Varki A, Cummings RD, Esko JD, et al. Essentials in Glycobiology, Chapter 36 - Glycans in Glycoprotein Quality Control. Cold Spring Harbor Laboratory Press (2009)

[2] Varki A, Cummings RD, Esko JD, et al. Essentials in Glycobiology, Chapter 6 - Biological Roles of Glycans. Cold Spring Harbor Laboratory Press (2009)

[3] Varki A, Cummings RD, Esko JD, et al. Essentials in Glycobiology, Chapter 8 N-Glycans. Cold Spring Harbor Laboratory Press (2009)

[4] Inoue S, Sato C and Kitajima K. Extensive enrichment of N-glycolylneuraminic acid in extracellular sialoglycoproteins abundantly synthesized and secreted by human cancer cells . Glycobiology (2010) 20: p. 752-762.

[5] Valmu L, Alfthan H, Hotakainen K, Birken S and Stenman UH. Site-specific glycan analysis of human chorionic gonadotropin $\beta-$subunit from malignancies and pregnancy by liquid chromatography-electrospray mass spectrometry. Glycobiology (2006) 16: p. 1207–1218.

[6] Zaia J. Mass Spectrometry and Glycomics. OMICS A Journal of Integrative Biology (2010) 14: p. 401-418.

[7] Hua S, Hu CY, Kim BJ, Totten SM, Oh MJ, Yun N, Nwosu CC, Yoo JS, Lebrilla CB and An HJ. Glyco-Analytical Multispecific Proteolysis (Glyco-AMP): A Simple Method for Detailed and Quantitative Glycoproteomic Characterization. Journal of Proteome Research (2013) 12: p. 4414—4423.

[8] Sparbier K, Koch S, Kessler I, Wenzel T and Kostrzewa M. Selective Isolation of Glycoproteins and Glycopeptides for MALDI- TOF MS Detection Supported by Magnetic Particles. Journal of Biomolecular Techniques (2005) 16: p. 405-411.

[9] Drake RR, Schwegler EE, Malik G, Diaz J, Block T, Mehta A and Semmes OJ. Lectin Capture Strategies Combined with Mass Spectrometry for the Discovery of Serum Glycoprotein Biomarkers. Molecular and Cellular Proteomics (2006) 5: p. 1957-1967.

[10] Ongay S, Boichenko A, Govorukhina N and Bischoff R. Glycopeptide enrichment and separation for protein glycosylation analysis. J. Sep. Sci (2012) 35: p. 2341–2372.

[11] Wohlgemuth J, Karas M, Eichhorn T, Hendriks R and Andrecht S. Quantitative site-specific analysis of protein glycosylation by LC-MS using different glycopeptide-enrichment strategies. Analytical Biochemistry (2009) 395: p. 178-188.

[12] Liu J, Wang F, Zhu J, Mao J, Liu Z, Cheng K, Qin H and Zou H. Highly efficient N-glycoproteomic sample preparation by combining C18 and graphitized carbon adsorbents. Analytical and Bioanalytical Chemistry (2014) 406: p. 3103-3109.

[13] Snovida SI, Bodnar ED, Viner R, Saba J and Perreault H. A simple cellulose column procedure for selective enrichment of glycopeptides and characterization by nano LC coupled with electron-transfer and high-energy collisional-dissociation tandem mass spectrometry. Carbohydrate Research (2010) 345: p. 792-801.

[14] Lewandrowski U, Zahedi RP, Moebius J, Walter U and Sickmann A. Enhanced N-Glycosylation Site Analysis of Sialoglycopeptides by Strong Cation Exchange Prefractionation Applied to Platelet Plasma Membranes. Molecular and Cellular Proteomics (2007) 6: p. 1933-1941.

[15] Morelle W and Michalski JC. Analysis of protein glycosylation by mass spectrometry. Nature Protocolls (2007) 2: p. 1585-1602.

[16] Toyama A, Nakagawa H, Matsuda K, Sato TA, Nakamura Y and Ueda K. Quantitative Structural Characterization of Local N-Glycan Microheterogeneity in Therapeutic Antibodies by Energy-Resolved Oxonium Ion Monitoring. Analytical Chemistry (2012), 84: p. 9655—9662.

[17] Alley WR, Mechref Y and Novotny MV. Characterization of glycopeptides by combining collision-induced dissociation and electron-transfer dissociation mass spectrometry data. Rapid Communications in Mass Spectrometry (2009) 23: p. 161-170.

[18] Aebersold R and Mann M. Mass spectrometry-based proteomics. Nature (2003) 422: p. 198-207.

[19] Kapp EA, Schütz F, Connolly LM, Chakel JA , Meza JE, Miller CA, Fenyo D, Eng JK, Adkins JN, Omenn GS and Simpson RJ. An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: Sensitivity and specificity analysis. Proteomics (2005) 5: p. 3475-3490.

[20] Craig R and Beavis RC. A method for reducing the time required to match protein sequences with tandem mass spectra. Rapid Commun. Mass Spectrom. (2003), 17: p. 2310–2316.

[21] Fenyo D and Beavis RC. A Method for Assessing the Statistical Significance of Mass Spectrometry-Based Protein Identifications Using General Scoring Schemes. Analytical Chemistry (2003) 75: p. 768-774.

[22] Eng JK, McCormack AL Yates RJ. An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. J. Am. Sot. Mass Spectrom. (1994), 5: p. 976-989.

[23] Perkins DN, Pappin DJC, Creasy DM and Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis (1999) 20: p. 3551-3567.

[24] Dallas DC, Martin WF, Hua S and German JB. Automated glycopeptide analysis - review of current state and future directions. Briefings in Bioinformatics (2012).

[25] Cooper CA, Gasteiger E and Packer N. GlycoMod - A software Tool for Determining Glycosylation Compositions from Mass Spectrometric Data. Proteomics (2001) 1: p. 340-349.

[26] Deshpande N, Jensen PH, Packer NH and Kolarich D. GlycoSpectrumScan: Fishing Glycopeptides from MS Spectra of Protease Digests of Human Colostrum sIgA . Journal of Proteome Research (2010) 9: p. 1063-1075.

[27] Chandler KB, Pompach P, Goldman R & Edwards N. Exploring Site-Specific N-Glycosylation Microheterogeneity of Haptoglobin Using Glycopeptide CID Tandem Mass Spectra and Glycan Database Search . Journal of Proteome Research (2013) 12: p. 3652—3666.

[28] Ozohanics O, Krenyacz J, Ludányi K, Pollreisz F, Vékey K and Drahos L. GlycoMiner: a new software tool to elucidate glycopeptide composition. Rapid Commun. Mass Spectrom. (2008) 22: p. 3245–3254.

[29] Gornik O and Lauc G. Glycosylation of serum proteins in inflammatory diseases. Disease Markers (2008) 25: p. 267–278.

[30] Varki A, Cummings RD, Esko JD, et al. Essentials in Glycobiology, Chapter 44 - Glycosylation Changes in Cancer. Cold Spring Harbor Laboratory Press (2009)

[31] Kessner D, Chambers M, Burke R, Agus D and Mallick P. ProteoWizard: open source software for rapid proteomics tools development. Bioinformatics (2008) 21: p. 2534–2536.

[32] Sturm M, Bertsch A, Gröpl C, Hildebrandt A, Hussong R, Lange E, Pfeifer N, Schulz-Trieglaff O, Zerck A, Reinert K and Kohlbacher O. OpenMS – An open-source software framework for mass spectrometry. BMC Bioinformatics (2008) 9: p. 163-174.

[33] Pluskal T, Castillo S, Villar-Briones A and Orešič M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinformatics (2010) 11: p. 395-406.

[34] Wuhrer M, Koeleman CAM, Hokke CH and Deelder AM. Mass spectrometry of proton adducts of fucosylated N-glycans: fucose transfer between antennae gives rise to misleading fragments. Rapid Communications in Mass Spectrometry (2006) 20: p. 1747–1754.

[35] Rodriguez J, Gupta N, Smith RD and Pevzner PA. Does Trypsin Cut Before Proline? . Journal of Proteome Research (2008) 7: p. 300–305.

Lebenslauf

Name	Nikolaus Voulgaris BSc
Staatsbürgerschaft	Österreich
Telefon	+43 680 1255 190
Email	n.voulgaris@gmx.net

Schulausbildung

1995 – 1999	VS Felixdorf
1999 – 2007	BG Babenbergering in Wiener Neustadt
2003 – 2007	Oberstufe mit Informatikzweig
2007	Matura mit ausgezeichnetem Erfolg

Studienverlauf

Oktober 2008 - März 2012	Bachelorstudium Chemie an der Universität Wien
seit März 2012	Masterstudium Chemie an der Universität Wien

Studienschwerpunkte:

- Analytische Chemie
- Theoretische Chemie

Titel der Masterarbeit:

- „FindGlycoPeptides – an Open Source Program for High-Throughput N-Glycopeptide Identification in Large LC-MS/MS Data Sets“

Besondere Auszeichnungen

2009, 2010, 2013	Leistungstipendium der Universität Wien

Berufserfahrung

2004 – 2006	jeweils 1-monatige Ferialpraktika (Essilor, Erste Bank Gruppe)
Aug. 2007 – Apr. 2008	Zivildienst bei der Lebenshilfe NÖ
Juli 2008	Ferialpraktikum bei Brokerjet
Okt. 2011 – Jän. 2014	geringfügige Beschäftigung als Nachtportier/ Rezeptionist bei Meininger Hotels

Kenntnise

Sprachen

Deutsch	Muttersprache
Englisch	fließend

Computer

HTML, C#, C++	Grundkenntnisse
R	Grundkenntnisse
Unix	fortgeschritten
Perl	fortgeschritten