Catherine A. Cooper[1*], Elisabeth Gasteiger[2*], Nicolle H. Packer[1,3]

[1]Proteome Systems Ltd, Sydney, Australia
[2]Swiss Institute of Bioinformatics, Geneve, Switzerland
[3]Macquarie University Center for Analytical Biotechnology, Sydney, Australia

GlycoMod – A software tool for determining glycosylation compositions from mass spectrometric data

GlycoMod (http://www.expasy.ch/tools/glycomod/) is a software tool designed to find all possible compositions of a glycan structure from its experimentally determined mass. The program can be used to predict the composition of any glycoprotein-derived oligosaccharide comprised of either underivatised, methylated or acetylated monosaccharides, or with a derivatised reducing terminus. The composition of a glycan attached to a peptide can be computed if the sequence or mass of the peptide is known. In addition, if the protein is known and is contained in the SWISS-PROT or TrEMBL databases, the program will match the experimentally determined masses against all the predicted protease-produced peptides (including any post-translational modifications annotated in these databases) which have the potential to be glycosylated with either *N*- or *O*-linked oligosaccharides. Since many possible glycan compositions can be generated from the same mass, the program can apply compositional constraints to the output if the user supplies either known or suspected monosaccharide constituents. Furthermore, known oligosaccharide structural constraints on monosaccharide composition are also incorporated into the program to limit the output.

Keywords: Oligosaccharide / *O*-Glycosylation / *N*-Glycosylation / Mass spectrometry / Monosaccharide composition PRO 0029

1 Introduction

The standard proteomic approach to the identification of proteins is to digest the separated proteins by a protease, such as trypsin, and to match the resulting masses with those generated by the theoretical digestion of the protein sequence or translated nucleotide sequence of the organism. Using this classic approach of peptide mass fingerprinting (PMF) it is clear that there are many masses generated which do not match the predicted mass of peptides generated by the theoretical protease digestion of an identified protein. Many of these masses may be accounted for by experimental artefacts, but many natural modifications occur to the read-out of the genome, both co- and post-translationally (*e. g.*, phosphorylation, glycosylation and deamidation). These modifications will usually produce peptides that have masses greater than that predicted. These can be overlooked if only the identity of the protein is required, as there are often sufficient correct peptide masses for a match to be made. If characterisation of the protein is required, however, then tools that can compute the difference between the observed and predicted masses, can give useful insight into the types of possible modifications that may have occurred to the protein. One such tool is FindMod [1] which considers 25 post-translational modifications, including acetylation, methylation, sulphation, and the addition of single *O*-GlcNAc residues. The complexity of protein glycosylation, however, precluded its inclusion in FindMod.

Glycosylation of a protein is one of the most common post-translational modifications in eukaryotes as well as prokaryotes [49]. It has recently been estimated that more than half of all proteins are glycosylated [2]. The two main types of glycosidic linkage in glycoproteins are an *N*-glycosidic link *via* the amide nitrogen of an asparagine residue; and an *O*-glycosidic link *via* the hydroxyl group of serine, threonine, tyrosine, hydroxylysine or hydroxyproline. Glycan units may also be *O*-linked to serine and threonine residues *via* a phosphate group, specifically known as a phosphodiester linkage [3].

N-linked glycans nearly always have an *N*-acetylglucosamine residue at the reducing terminus and are beta linked to the amide nitrogen of an asparagine residue in the consensus sequence -Asn-Xaa-Ser/Thr/Cys where Xaa $\neq$ Pro [4, 5]. Cysteine has been included in the motif

[*] C. Cooper and E. Gasteiger have contributed equally to the work shown.

Correspondence: Catherine Cooper, Proteome Systems Ltd, Locked Bag 2073, North Ryde, Sydney, NSW 1670, Australia
E-mail: catherine.cooper@proteomesystems.com
Fax: + 61-2-98891805

Abbreviations: GlcNAc, *N*-acetylgalactosamine; **Man**, mannose; **KDN**, 2-keto-3-deoxynonulosonic acid; **PGNase**, peptide-*N*-glycosidase; **M**, Peptide mass; **PA**, 2-aminopyridine; **Hex**, hexose; **Pent**, pentose; **HexNAc**, *N*-acetyl hexosamine; **NeuAc**, *N*-acetylneuraminic acid; **NeuGc**, *N*-glycolylneuraminic acid

© WILEY-VCH Verlag GmbH, 69451 Weinheim, 2001

because of its potential to be glycosylated *in vitro* [4], and the discovery of the Asn-Xaa-Cys sequence being glycosylated in murine fetal antigen 1 [5]. *N*-linked glycans are biologically synthesised by the transfer of $Glc_3Man_9GlcNAc_2$ from dolichol-pyrophosphate-$Glc_3Man_9GlcNAc_2$ to asparagine [6], followed by enzymatic processing to form the final *N*-linked glycan structures [7]. Due to this method of synthesis, the core pentasaccharide Man(α1–6)[Man(α1–3)]Man(β1–4)GlcNAc(β1–4)GlcNAc is nearly always present in *N*-linked glycans. This core is then further elongated, dividing the *N*-linked glycans into three distinct classes; high-mannose type, complex type and hybrid type (Fig. 1). If the core structure is predominantly substituted by mannose it is called a high-mannose type *N*-glycan. When the core structure is substituted by one or more of the sugars *N*-acetylglucosamine, galactose, fucose or sialic acid it is called a complex type. Hybrid type *N*-linked glycans have structural features of both the high-mannose and the complex type chains.

Unlike *N*-glycosylation there is no single motif for predicting *O*-glycosylation [8–11]. This is probably because of the number of monosaccharides and amino acids involved in *O*-glycan linkages, each different linkage probably requiring its own enzyme. *O*-linked glycans are characterised by the presence of fucose, galactose, *N*-acetylgalactosamine, *N*-acetylglucosamine, and may have acidic sugar residues such as *N*-acetylneuraminic acid, *N*-glycolylneuraminic acid, 2-keto-3-deoxynonulosonic acid (KDN) and glucuronic acid. *O*-linked glycans can also contain sulfate and phosphate residues. Mannose and glucose have been found as components of *O*-glycans [12], but they are not common. Unlike N-linked glycans, *O*-linked glycans are synthesised one residue at a time and at least nine core classes have been described [13–16]. There exists a wide diversity of structures within both the *N*- and *O*-linked glycans, the synthesis of which is a result of a finely controlled process dependent upon the availability and activity of the various glycosyltransferases, monosaccharides and precursors [17]. Consequently, one glycosylation site may have many glycan structures (microheterogeneity) [18], and one protein may have different structures at different sites (macroheterogeneity).

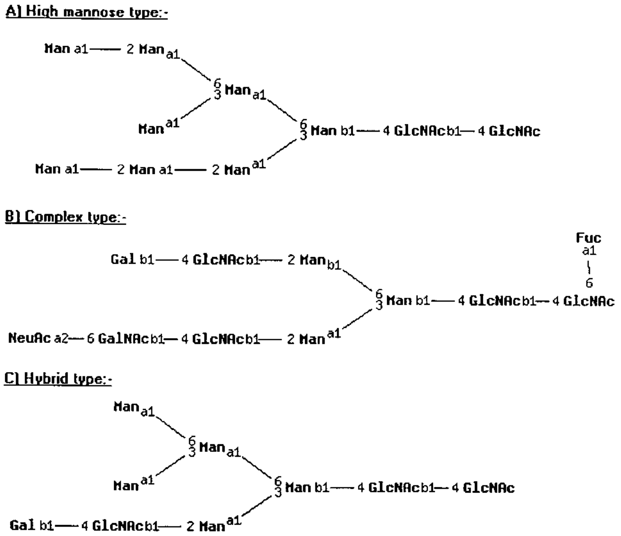
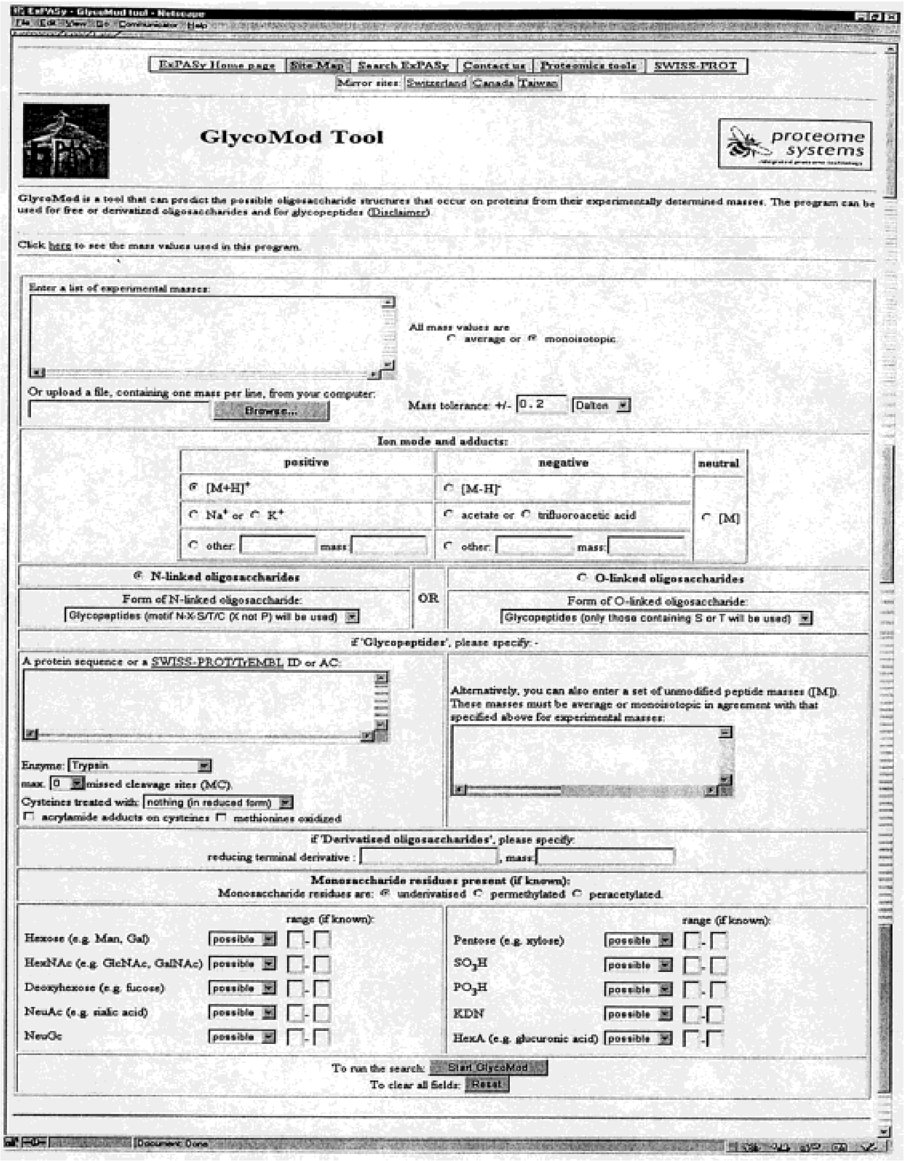

A) High mannose type:-
Man a1— 2 Man a1
6 3 Man a1
Man a1
6 3 Man b1— 4 GlcNAcb1— 4 GlcNAc
Man a1— 2 Man a1— 2 Man a1
B) Complex type:-
Fuc a1 6
Gal b1— 4 GlcNAcb1— 2 Man b1
6 3 Man b1— 4 GlcNAcb1— 4 GlcNAc
NeuAc a2— 6 GalNAcb1— 4 GlcNAcb1— 2 Man a1
C) Hybrid type:-
Man a1
6 3 Man a1
Man a1
6 3 Man b1— 4 GlcNAcb1— 4 GlcNAc
Gal b1— 4 GlcNAcb1— 2 Man a1

Figure 1. Classification of *N*-linked glycan structures.

There have been many structures described in the literature and their monosaccharide compositions have been found using a range of methods. These methods include high-performance anion-exchange chromatography with pulsed-amperometric detection (HPAEC-PAD) [19], thin-layer and paper chromatography [20], colorimetric detection [21–23], gas- or liquid-chromatography [20, 24], and capillary electrophoresis (CE) [25].

Colorimetric and thin-layer chromatographic methods require little instrumentation, but are time-consuming, with standard curves needing to be generated for each series of analyses, and they require large amounts of material, typically 10–100 μg. High-performance liquid chromatography (except HPAEC-PAD) and CE methods require pre- or post-column derivatisation of the carbohydrates to improve detection [25, 26]. Gas chromatographic methods are sensitive and rapid, but require the monosaccharides to be derivatised to form volatile compounds [27].

Mass spectrometry has developed into a method often used to study glycopeptides and oligosaccharides [28, 29, 50] and it is now becoming the method of choice for the rapid prediction of individual glycan compositions. The main advantages of MS are its sensitivity, accuracy and speed. It is possible to detect glycans in femtomole quantities [28, 30] and the mass accuracy is usually better than 50 ppm, with later model instruments and improved sample preparation [29]. The mass of the glycan on a protein can be determined either as the free oligosaccharides, after their release from the protein, or while still attached to the peptide backbone. MS cannot, however, distinguish between monosaccharides with the same mass, *e.g.* hexoses – glucose, mannose, galactose; hexosamines – glucosamine, galactosamine; *N*-acetylhexosamines – N-acetylglucosamine, *N*-acetylgalactosamine. This is further complicated by the fact that often different combinations of monosaccharides result in the same mass. For example, one hexose residue and one NeuAc residue combine together to give the same monoisotopic mass (453.1482 Da) as one deoxyhexose residue plus one NeuGc residue (453.1482 Da). To easily and quickly determine the possible glycan compositions correspond-


ExPASy - GlycoMod tool - Netscape
File Edit View Go Communicator Help
ExPASy Home page | Site Map | Search ExPASy | Contact us | Proteomics tools | SWISS-PROT
Mirror sites: Switzerland Canada Taiwan
proteome systems
Document: Done

GlycoMod Tool

GlycoMod is a tool that can predict the possible oligosaccharide structures that occur on proteins from their experimentally determined masses. The program can be used for free or derivatized oligosaccharides and for glycopeptides (Disclaimer).

Click here to see the mass values used in this program.

Enter a list of experimental masses:

All mass values are ○ average or ● monoisotopic

Or upload a file, containing one mass per line, from your computer: [Browse...]

Mass tolerance: +/- 0.2 Dalton

Ion mode and adducts:

positive	negative	neutral
● $[M+H]^+$	○ $[M-H]^-$	○ [M]
○ Na^+ or ○ K^+	○ acetate or ○ trifluoroacetic acid	
○ other: ___ mass: ___	○ other: ___ mass: ___	

● N-linked oligosaccharides OR ○ O-linked oligosaccharides

Form of N-linked oligosaccharide: Glycopeptides (motif N-X-S/T/C (X not P) will be used)

Form of O-linked oligosaccharide: Glycopeptides (only those containing S or T will be used)

if 'Glycopeptides', please specify: -

A protein sequence or a SWISS-PROT/TrEMBL ID or AC:

Enzyme: Trypsin

max. 0 missed cleavage sites (MC).

Cysteines treated with: nothing (in reduced form)

□ acrylamide adducts on cysteines □ methionines oxidized

Alternatively, you can also enter a set of unmodified peptide masses ([M]). These masses must be average or monoisotopic in agreement with that specified above for experimental masses:

if 'Derivatised oligosaccharides', please specify:

reducing terminal derivative : ___ , mass: ___

Monosaccharide residues present (if known):

Monosaccharide residues are: ● underivatised ○ permethylated ○ peracetylated.

		range (if known):			range (if known):
Hexose (e.g. Man, Gal)	possible	□ - □	Pentose (e.g. xylose)	possible	□ - □
HexNAc (e.g. GlcNAc, GalNAc)	possible	□ - □	SO_3H	possible	□ - □
Deoxyhexose (e.g. fucose)	possible	□ - □	PO_3H	possible	□ - □
NeuAc (e.g. sialic acid)	possible	□ - □	KDN	possible	□ - □
NeuGc	possible	□ - □	HexA (e.g. glucuronic acid)	possible	□ - □

To run the search: [Start GlycoMod]

To clear all fields: [Reset]

Figure 2. The input page of GlycoMod found at http://www.expasy.ch/tools/glycomod/.

ing to an experimentally determined mass of either a protease generated glycopeptide or a released and derivatised glycan, we found it necessary to have at our disposal a computational tool to predict the possible heterogeneous glycan compositions.

GlycoMod is a program designed to find all possible compositions of a glycan structure from its experimentally determined mass. This is done by comparing the mass of the glycan to a list of precomputed masses of glycan compositions. The program can be used with free or a range of derivatised glycans and for glycopeptides where the peptide mass or the identity of the protein is known. The latter can be entered as a SWISS-PROT/TrEMBL [31] accession number or a user-entered sequence. Monosaccharide compositional constraints entered from experimental data or known structural precedents can be applied to the output to limit the possible outcomes. GlycoMod is available on the internet as part of the ExPASy suite of proteomics tools at http://www.expasy.ch/tools/glycomod/.

2 Materials and methods

2.1 Input parameters (Fig. 2)

2.1.1 Experimental masses

Experimental masses (average or monoisotopic) to be analyzed may be manually entered separating them by spaces or new lines, or uploaded from a text file.

Table 1. Upper limits imposed on the number of residues of a particular monosaccharide (based on a survey of the literature) found in *O*-linked and *N*-linked oligosaccharides using GlycoMod

Monosaccharide residue	*O*-linked oligosaccharides	*N*-linked oligosaccharides
Hexose	0–14	0–20
HexNAc	0–14	0–20
Deoxyhexose	0–6	0–6
NeuAc	0–7	0–5
NeuGc	0–7	0–5
Pentose	0–3	0–3
Sulphate	0–6	0–3
Phosphate	0–6	0–2
KDN	0–2	0–0
HexA	0–2	0–0

A mass tolerance level can be selected in either Daltons or ppm.

2.1.2 Ion mode and adducts

The masses are entered as neutral ions, positive ions, or as negative ions. The possibility of sodium, potassium, TFA or user entered adducts can be accounted for.

2.1.3 Glycopeptides

GlycoMod can be used to calculate the possible compositions of the glycans attached to a peptide. The peptide data may be entered as a protein sequence, a SWISS-PROT/TrEMBL ID or AC, or as a set of unmodified peptide masses [M], where the masses are average or monoisotopic in agreement with that specified for the experimental masses of the data entered above. The user options for protein digestion, including cysteine adducts, methionine oxidation, protease selection, and missed cleavages, can be chosen as in PeptideMass [32] and FindMod [1].

1.1.4 Released glycans

GlycoMod can be used to calculate the possible compositions of free glycans. For example, *N*-linked glycans released using peptide-*N*-glycosidase (PNGase) F, PNGase A or anhydrous hydrazine; *N*-linked glycans released using endoglycosidase (Endo) H or Endo F; *O*-linked glycans released using *O*-glycanase, mild hydrazinolysis [33], or nonreductive beta-elimination [34]; and *O*-linked glycans released and reduced using reductive beta-elimination. Once released, free reducing oligosaccharides are often derivatised at the reducing terminus by a process of reductive amination, *i.e.* the reducing terminus of the glycan is reacted with an amine followed by reduction with a selective reducing agent. Common derivatives include 2-aminopyridine (PA) (Fig. 5), 2-aminobenzoic acid (ABA) or 8-aminonapthalene-1,3,6-trisulfonic acid (ANTS) [26]. GlycoMod allows the user to select “derivatised oligosaccharide” and to supply the mass of a derivative (M).

2.1.5 Monosaccharide residues

GlycoMod has been designed to calculate the masses of oligosaccharides using underivatised, permethylated or peracetylated monosaccharides since mass spectrometric data is often obtained from these derivatised oligosaccharides. To restrict the possible monosaccharide combinations it is possible to stipulate which monosaccharides and how many of each type, are (“yes”), are not (“no”), or may possibly (“possible”) be present in the glycan. The maximum numbers of each residue allowed (Table 1) has been set based on known structures reported in the text.

There are also some preprogrammed limits to the output of possible compositions allowed for *N*-linked glycans. These were implemented after careful investigation of the known *N*-linked glycan structures. (1). A composition may not contain both sulfate and phosphate; (2). The sum of the number of hexose plus HexNAc residues must be greater than or equal to the number of sulfate or phosphate residues; (3). The sum of the number of hexose plus HexNAc residues cannot be zero; (4). The number of fucose residues plus 1 must be less than or equal to the sum of the number of hexose plus HexNAc residues; (5).

If the number of HexNAc residues is less than or equal to 2 and the number of hexose residues is greater than 2, then the number of NeuAc and NeuGc residues must be zero.

The *N*-linked glycan types are classified by: (1). If the number of HexNAc residues equals 2 and the number of hexose residues is greater than or equal to 5, then the *N*-linked glycan is of the type "high mannose"; (2). If the number of HexNAc residues is greater than or equal to 3 and the number of hexose residues is also greater than or equal to 3, then the *N*-linked glycan is of the type "hybrid/complex".

All other combinations are not given a glycan type. There are no preprogrammed limits to the possible compositions allowed for *O*-linked glycans, except for the total number of any one particular monosaccharide residue (Table 1) which were set based on a survey of the literature. An upper limit on the total mass of the glycoform has also been set. This limit is 8000 Da for underivatised, 10 000 Da for permethylated and 13 000 Da for peracetylated *N*-linked glycans. For *O*-linked glycans the limit is 5000 Da for underivatised, 7000 Da for permetylated and 9500 Da for peracetylated oligosaccharides.

3 Results

The output for GlycoMod (Fig. 3) is divided into two main sections – a header and a table for each experimental mass entered. The header section lists the monosaccharide compositional data entered by the user and the calculated peptide masses of a protein sequence or SWISS-PROT/TrEMBL ID or AC if the glycopeptide option was chosen. The output tables report the monosaccharide compositions whose theoretical masses match the entered experimental user mass after any stated derivative or peptide modification has been subtracted. A separate table is generated for each entered mass. Each table shows the glycoform mass, Δmass in daltons or ppm (depending on the units entered by the user on the input form), and the possible matching monosaccharide compositions. If the glycan is *N*-linked the predicted glycan type, *i.e.*, hybrid/complex or high mannose (Fig. 1) is given.

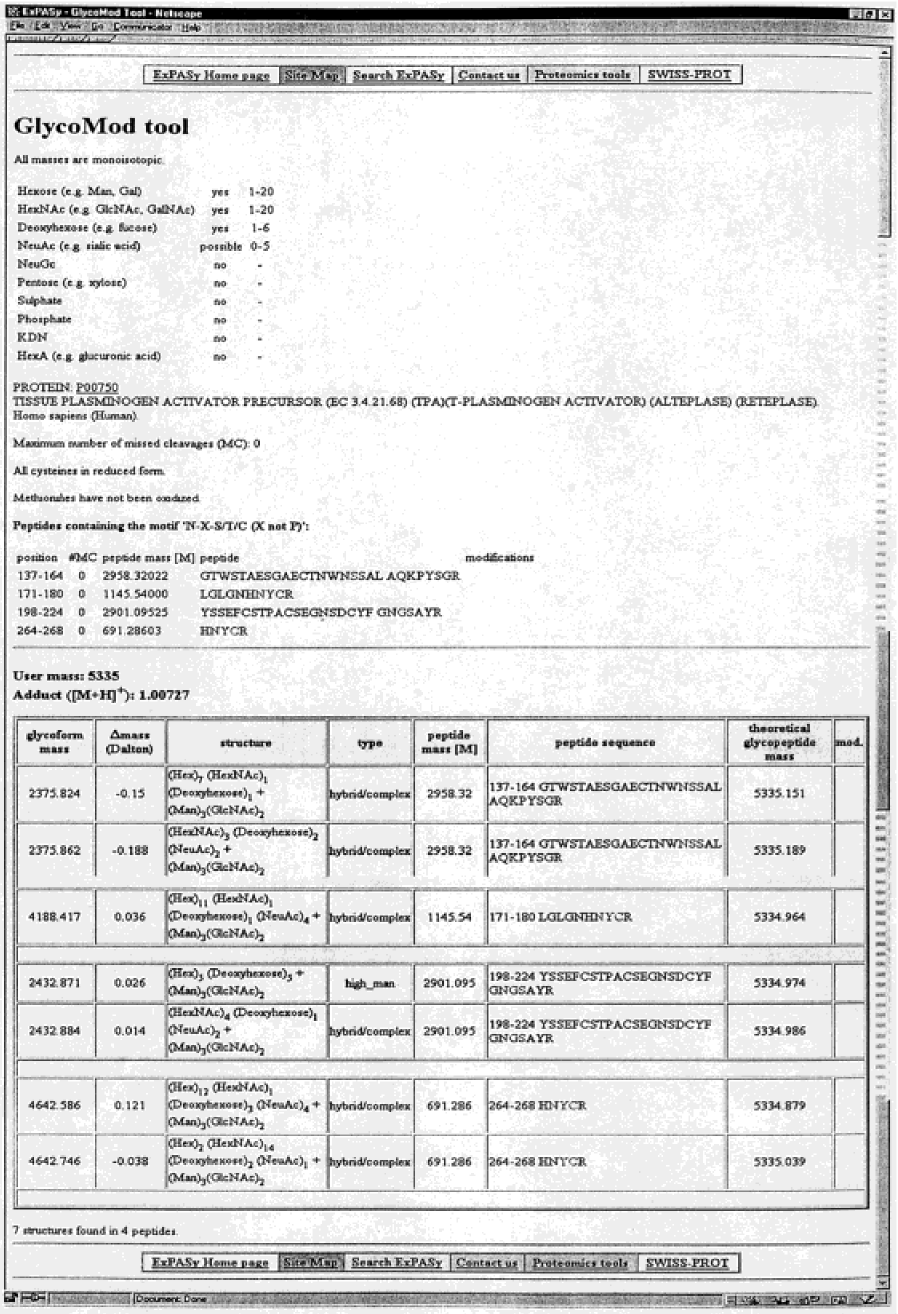
If a glycopeptide mass is entered together with a protein sequence or SWISS-PROT/TrEMBL ID or AC, then GlycoMod calculates the possible oligosaccharide compositions attached to the unmodified theoretical peptides formed after enzymatic or chemical digestion. GlycoMod also considers the peptides that may be biologically modified (as annotated in SWISS-PROT) and/or chemically modified (as specified by the user in the input form). If the entry has a SWISS-PROT/TrEMBL ID or AC the description line from the SWISS-PROT/TrEMBL entry detailing the protein name, synonym(s), contained proteins and the biological species from which the protein was derived is given, along with a hyperlink to the SWISS-PROT/TrEMBL entry.

For a glycopeptide output each table contains additional information on the peptide mass (M), peptide sequence or a SWISS-PROT/TrEMBL ID or AC (where entered by the user), the theoretical glycopeptide mass, and any modification noted in SWISS-PROT if a SWISS-PROT ID or AC was entered. To best show the utility of GlycoMod we present four case studies. In each case data was extracted from relevant journal articles and entered into GlycoMod. The output was compared with the results reported by the original authors.

3.1 Case 1: Permethylated *N*-linked glycan [35]

The *N*-linked glycans on glycodelin, a human glycoprotein, were released using PNGase F, permethylated and analysed by FAB-MS. The mass 2227.3 Da was assigned as the $[M+Na]^+$ mass of $NeuAcHex_4HexNAc_4$ by Dell and co-authors [35]. Using GlycoMod, the mass 2227.3 Da was entered as a monoisotopic mass, with a mass tolerance of 0.2 Da. Na^+ was selected as the ion mode and the form of the *N*-linked oligosaccharide was chosen as "Free/PNGase released oligosaccharides". The monosaccharide residues were selected as being permethylated and the search was run.

The output (Fig. 4) shows three different compositions that were found to match the input data. With some knowledge these results can be refined. For example, the composition containing NeuGc might be rejected because this sialic acid has not been found to date in normal healthy humans. Similarly, the composition with two pentose residues is not a typical composition for an *N*-linked oligosaccharide (although *N*-linked glycans containing arabinose and xylose have been characterised from carrots [36]). Therefore, the most likely composition for this *N*-linked permethylated glycan is $(Hex)_1(HexNAc)_2(NeuAc)_1$ + $(Man)_3(GlcNAc)_2$. The composition is written in this form, with the core monosaccharide residues written separately, when it contains at least two HexNAc residues and three Hexose residues. This is because the structures of *N*-linked glycans are generally well conserved with a core region consisting of two *N*-acetylglucosamine residues and three mannose residues, and branches containing a variety of hexose and HexNAc residues that may be further substituted with other residues such as sialic acid. To help the user to distinguish between those residues residing in the core of an *N*-linked glycan and those on the branches, the core monosac-



GlycoMod tool

All masses are monoisotopic.

Hexose (e.g. Man, Gal)	yes	1-20
HexNAc (e.g. GlcNAc, GalNAc)	yes	1-20
Deoxyhexose (e.g. fucose)	yes	1-6
NeuAc (e.g. sialic acid)	possible	0-5
NeuGc	no	-
Pentose (e.g. xylose)	no	-
Sulphate	no	-
Phosphate	no	-
KDN	no	-
HexA (e.g. glucuronic acid)	no	-

PROTEIN: P00750
TISSUE PLASMINOGEN ACTIVATOR PRECURSOR (EC 3.4.21.68) (TPA)(T-PLASMINOGEN ACTIVATOR) (ALTEPLASE) (RETEPLASE).
Homo sapiens (Human).

Maximum number of missed cleavages (MC): 0

All cysteines in reduced form.

Methionines have not been oxidized.

Peptides containing the motif 'N-X-S/T/C (X not P)':

```
position  #MC  peptide mass [M]  peptide                                 modifications
137-164   0    2958.32022        GTWSTAESGAECTNWNSSAL AQKPYSGR
171-180   0    1145.54000        LGLGNHNYCR
198-224   0    2901.09525        YSSEFCSTPACSEGNSDCYF GNGSAYR
264-268   0    691.28603         HNYCR
```

User mass: 5335
Adduct ([M+H]$^+$): 1.00727

glycoform mass	Δmass (Dalton)	structure	type	peptide mass [M]	peptide sequence	theoretical glycopeptide mass	mod.
2375.824	-0.15	$(Hex)_7$ $(HexNAc)_1$ $(Deoxyhexose)_1$ + $(Man)_3(GlcNAc)_2$	hybrid/complex	2958.32	137-164 GTWSTAESGAECTNWNSSAL AQKPYSGR	5335.151	
2375.862	-0.188	$(HexNAc)_3$ $(Deoxyhexose)_2$ $(NeuAc)_2$ + $(Man)_3(GlcNAc)_2$	hybrid/complex	2958.32	137-164 GTWSTAESGAECTNWNSSAL AQKPYSGR	5335.189	
4188.417	0.036	$(Hex)_{11}$ $(HexNAc)_1$ $(Deoxyhexose)_1$ $(NeuAc)_4$ + $(Man)_3(GlcNAc)_2$	hybrid/complex	1145.54	171-180 LGLGNHNYCR	5334.964	
2432.871	0.026	$(Hex)_3$ $(Deoxyhexose)_5$ + $(Man)_3(GlcNAc)_2$	high_man	2901.095	198-224 YSSEFCSTPACSEGNSDCYF GNGSAYR	5334.974	
2432.884	0.014	$(HexNAc)_4$ $(Deoxyhexose)_1$ $(NeuAc)_2$ + $(Man)_3(GlcNAc)_2$	hybrid/complex	2901.095	198-224 YSSEFCSTPACSEGNSDCYF GNGSAYR	5334.986	
4642.586	0.121	$(Hex)_{12}$ $(HexNAc)_1$ $(Deoxyhexose)_3$ $(NeuAc)_4$ + $(Man)_3(GlcNAc)_2$	hybrid/complex	691.286	264-268 HNYCR	5334.879	
4642.746	-0.038	$(Hex)_2$ $(HexNAc)_{14}$ $(Deoxyhexose)_2$ $(NeuAc)_1$ + $(Man)_3(GlcNAc)_2$	hybrid/complex	691.286	264-268 HNYCR	5335.039	

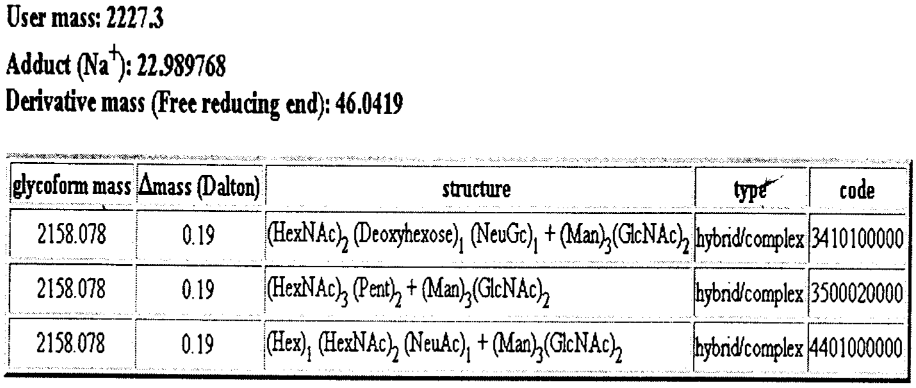
7 structures found in 4 peptides.

Figure 3. An example output page from GlycoMod.

charide residues are removed from the overall composition and written separately in the GlycoMod output.

3.2 Case 2: *N*-linked glycopeptide [37]

Human metalloproteinase inhibitor 1 (P01033) expressed in CHO cells, was digested using trypsin and the peptides analysed by MALDI-TOF-MS [37]. The authors assigned the mass 3726.7 Da to a glycopeptide corresponding to residues 23–37 (given as residues 46–60 including the signal sequence in SWISS-PROT) with an attached glycan of the composition $Gal_2Man_3GlcNAc_5Fuc_1$. Using GlycoMod, the mass 3726.7 Da was entered as an average mass, with a mass tolerance of 0.2 Da. [M+H]$^+$ was selected as the ion mode and "Glycopeptides (motif *N*-X-S/T/C (X not P) will be used)" was selected as the form of the *N*-linked oligosaccharide. The SWISS-PROT AC P01033 was entered under "if Glycopeptides" and "Trypsin" was chosen as the cleavage reagent. Monosaccharide residues were selected as being underivatised and



User mass: 2227.3

Adduct (Na^+): 22.989768

Derivative mass (Free reducing end): 46.0419

glycoform mass	Δmass (Dalton)	structure	type	code
2158.078	0.19	$(HexNAc)_2$ $(Deoxyhexose)_1$ $(NeuGc)_1$ + $(Man)_3(GlcNAc)_2$	hybrid/complex	3410100000
2158.078	0.19	$(HexNAc)_3$ $(Pent)_2$ + $(Man)_3(GlcNAc)_2$	hybrid/complex	3500020000
2158.078	0.19	$(Hex)_1$ $(HexNAc)_2$ $(NeuAc)_1$ + $(Man)_3(GlcNAc)_2$	hybrid/complex	4401000000

Figure 4. Table output from GlycoMod for a permethylated *N*-linked oligosaccharide released using PNGase F from glycodelin and analysed using FAB-MS [34] (Section 3.1).

the search was run. GlycoMod returned one possible peptide from amino acid position 46–60, corresponding to the sequence FVGTPEVNQTTLYQR, with a mass (M) of 1752.94 Da. Six different monosaccharide compositions were proposed as possible matches. These were reduced to two when pentose was disallowed. The resulting two compositions were $(HexNAc)_1(Deoxyhexose)_6$ + $(Man)_3$ $(GlcNAc)_2$ (a very unlikely composition for an *N*-linked glycan) and $(Hex)_2(HexNAc)_3(Deoxyhexose)_1$ + $(Man)_3$ $(GlcNAc)_2$, which is in agreement with that proposed by Sutton and co-authors [37].

3.3 Case 3: 2-Aminopyridine derivatised *N*-linked oligosaccharide [38]

Oligosaccharides *N*-linked to horseradish peroxidase were released using PNGase A, then derivatised by reductive amination with 2-aminopyridine (PA). The derivatised glycans were analysed by ESI-MS. The authors assigned the mass 1143.4 Da to $Hex_3GlcNAc_2$. Using GlycoMod the mass 1143.4 Da was entered as a monoisotopic mass corresponding to $[M+Na]^+$. A mass tolerance of 0.2 Da was chosen and "Derivatised oligosaccharide" was selected as the form of the *N*-linked oligosaccharide. The reducing terminal derivative was identified by 'PA' and 94.0531 was entered as its monoisotopic mass (M). The mass required is the monoisotopic or average mass of the nonreacted derivative, *e.g.* 94.053 for the monoisotopic mass of PA. GlycoMod automatically adds the mass of two hydrogen atoms which result from the reductive amination chemistry as shown in the example of the derivatisation of the reducing terminal *N*-acetylglucosamine with PA (Fig. 5). Monosaccharide residues were selected as being underivatised and the search was run.

Two possible compositions matched the input criteria: $(Hex)_1(NeuAc)_1(NeuGc)_1(Pent)_2$ (ΔMass = 0 018 Da) and $(Hex)_3(HexNAc)_2(Pent)_1$ (ΔMass = –0.018 Da). From these $(Hex)_3(HexNAc)_2(Pent)_1$ is the more obvious composition for a typical plant *N*-linked glycan since it contains the normal core $(Hex)_3(HexNAc)_2$ and since sialic acid has not been found on glycans released from plants to date. Many glycoproteins from plants have been analysed and have been found to contain the *N*-linked glycan [39] (Fig. 6).

3.4 Case 4: *O*-Linked oligosaccharide alditol [40]

O-linked oligosaccharides linked to bovine submaxillary mucin were released using reductive beta-elimination, in which the released oligosaccharides are reduced to alditols as they are released to prevent base degradation ("peeling"). The oligosaccharide alditols were analysed as $[M-H]^-$ ions by liquid secondary-ion mass spectrometry (LSI-MS). Using GlycoMod the mass 1040 Da was entered as a monoisotopic mass, with a mass tolerance of 0.5 Da. $[M-H]^-$ was chosen as the ion mode, "*O*-linked oligosaccharides" was highlighted and "Reduced oligosaccharide" was selected. Monosaccharide residues were selected as being underivatised and the search was run. Fifteen possible compositions matched the input parameters. In order to reduce this number, limits were placed on the types of monosaccharides allowed to be in the composition. The monosaccharide residues Hexose, HexNAc, Deoxyhexose, NeuAc and NeuGc were left as "possible", while Pentose, KDN and HexA were not allowed since these residues were not found during methylation compositional analysis. These limitations reduced the number of possible compositions to three:

$(Hex)_2(HexNAc)_2(NeuAc)_1$ (ΔMass = –0.378 Da); $(Hex)_1$ $(HexNAc)_2(Deoxyhexose)_1(NeuGc)_1$ (ΔMass = –0.378 Da); and $(Deoxyhexose)_5(NeuAc)_1$ (ΔMass = –0.403 Da). The last composition is highly unlikely, but further chemical analysis (such as methylation analysis or NMR) is required

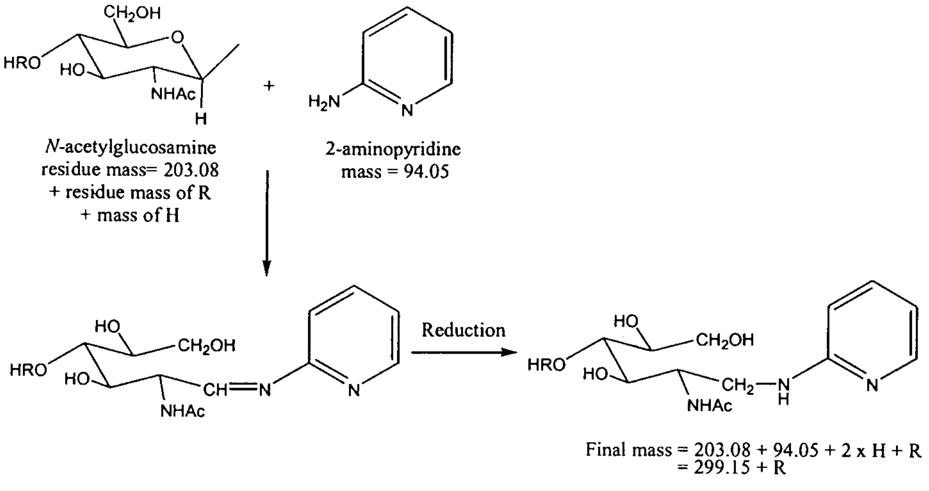

CH2OH
HRO
HO
O
NHAc
H
+
H2N
N
N-acetylglucosamine residue mass= 203.08 + residue mass of R + mass of H
2-aminopyridine mass = 94.05
HO
CH2OH
HRO
HO
CH=N
NHAc
N
Reduction
HO
CH2OH
HRO
HO
CH2—N
NHAc
H
N
Final mass = 203.08 + 94.05 + 2 x H + R = 299.15 + R

Figure 5. Derivatisation of an oligosaccharide with *N*-acetylglucosamine at the reducing terminus, with 2-aminopyridine (PA).

to determine which of the other two possible compositions (or both) correctly identify the oligosaccharide(s).

4 Discussion

GlycoMod provides the means to rapidly calculate the possible monosaccharide compositions of glycoprotein linked oligosaccharides from their masses obtained using a variety of analytical approaches. The main strengths of GlycoMod are its link to SWISS-PROT/TrEMBL, which enables glycopeptide possibilities to be easily computed, and the fact that monosaccharide constraints can be supplied so that all reasonably possible compositions are returned. This has the potential to highlight compositions that may not otherwise be considered. For example, 1396.4 Da corresponds to *N*-linked glycans which have been described previously with the compositions $(Hex)_3$ + $(Man)_3(GlcNAc)_2$ [41] as well as the composition $(HexNAc)_2$ $(Sulph)_1$ + $(Man)_3(GlcNAc)_2$ [42]. The monoisotopic mass 2093.74 Da also corresponds to two different compositions: $(Hex)_3(HexNAc)_2(NeuAc)_1$ + $(Man)_3(GlcNAc)_2$ and $(Hex)_2(HexNAc)_2(Deoxyhexose)_1(NeuGc)_1$ + $(Man)_3$ $(GlcNAc)_2$, both of which describe structures previously found [43, 44].

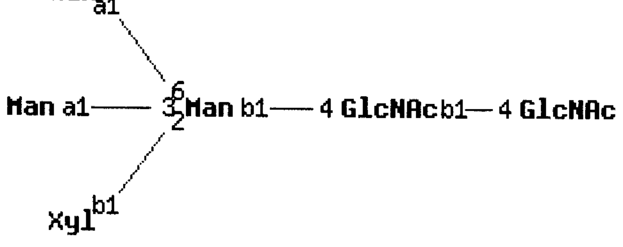


Figure 6. *N*-linked glycan commonly found in plant glycoproteins [39].

The multiplicity of possible compositions that are returned, especially if large errors are allowed, could be considered as a drawback of GlycoMod. This can be reduced by refining the output using monosaccharide compositional data. For example, if 1021.2 Da is entered as a monoisotopic mass (M) for a free *O*-linked oligosaccharide, 22 possible compositions are returned. When the input search is refined to disallow pentose, KDN and HexA residues, the number of possible compositions returned is reduced to six. On the other hand, increasingly unusual monosaccharide compositions and structures are being described. For example, $(Hex)_6(HexNAc)_6$ $(Deoxyhexose)_3(NeuAc)_3$ is the composition of an *N*-linked glycan found on human Alzheimer's disease amyloid A4 protein (P05067) expressed in the C6 cell line from a glial tumor of *Rattus norvegicus* [45], and $(Hex)_2$ $(NeuAc)_1(Pent)_1(HexA)_1(Sulph)_1$ is the composition of a glycan found on thrombomodulin isolated from human urine [46]. Thus, an output that has the flexibility to suggest unusual structures was considered to be an advantage in the construction of GlycoMod.

A limitation to the use of GlycoMod at present is the mass measurement of glycans, as reflected by the large errors given in the input for the presented case studies. Although MS and derivatisation chemistries are improving, the sensitivity for the analysis of carbohydrates is less than that of peptides [29]. This is also a reflection of the micro- and macroheterogeneity of glycan structures found on glycoproteins: the more different structures present on a protein, the more protein is needed to get enough of any one particular glycan structure for analysis. This heterogeneity, however, is what makes glycans important in the regulation of metabolic processes and therefore important in the field of proteomics. For example, the biological activity of erythropoietin is dependent on the types of glycans attached to the protein [47]. In particular, it has been shown that EPO-*bi*, a recombinant form of EPO that is enriched in biantennary *N*-linked glycans, has significantly less activity *in vivo* than a recombinant EPO enriched with tetra-antennary *N*-linked glycans [47]. This was shown to be due to rapid clearance of the EPO-*bi* protein from the systemic circulation by renal handling [48].

5 Concluding remarks

This paper describes the provision of a new computational tool, GlycoMod, included in the proteomics suite of tools available on ExPASy. It is the first step in recognising that the proteomic approaches of analysis require the means to rapidly identify possible protein glycosylation in the mass spectrometric data being accumulated from protease digests of known proteins. For this purpose GlycoMod is linked to the information available in the SWISS-PROT database and enables the prediction of *N*- and *O*-linked oligosaccharides on glycopeptides using experimentally obtained mass data.

GlycoMod goes further than just computing possible monosaccharide compositions corresponding to a mass, and allows a range of options of oligosaccharide release and derivatisation strategies to be included in the calculation. Furthermore, the program applies constraints on the output from both the user interface, in which specific known monosaccharide constituents can be selected for consideration, and from in-built criteria which account for known constraints on *N*- and *O*-linked oligosaccharide compositions.

Received March 29, 2000

6 References

[1] Wilkins, M. R., Gasteiger, E., Gooley, A. A., Herbert, B. R., Molloy, M. P., Binz, P. A., Ou, K., Sanchez, J. C., Bairoch, A., Williams, K. L., Hochstrasser, D. F., *J. Mol. Biol.* 1999, *289,* 645–657.

[2] Apweiler, R., Hermjakob, H., Sharon, N., *Biochim. Biophys. Acta* 1999, *1473,* 4–8.

[3] Haynes, P. A., *Glycobiology* 1998, *8,* 1–5.

[4] Bause, E., Legler, G., *Biochem. J.* 1981, *195,* 639–644.

[5] Krogh, T. N., Bachmann, E., Teisner, B., Skjodt, K., Hojrup, P., *Eur. J. Biochem.* 1997, *244,* 334–342.

[6] Hemming, F. W., in: Montreuil J., Vliegenthart J. F. G., Schachter H., (Eds.), *New Comprehensive Biochemistry,* Elsevier Science, Amsterdam 1995, pp. 127–143.

[7] Brockhausen, I., Schachter, H. in: Gabius, H.-J., Gabius, S. (Eds.), *Glycosciences: Status and Perspectives* 1997, Chapman & Hall, Weinheim, Germany 1997, pp. 79–114.

[8] Elhammer, A. P., Poorman, R. A., Brown, E., Maggiora, L. L., Hoogerheide, J. G., Kezdy, F. J., *J. Biol. Chem.* 1993, *268,* 10029–10038.

[9] Pisano, A., Redmond, J. W., Williams, K. L., Gooley, A. A., *Glycobiology* 1993, *3,* 429–435.

[10] Gooley, A. A., Williams, K. L., *Glycobiology* 1994, *4,* 413–417.

[11] Hansen, J. E., Lund, O., Engelbrecht, J., Bohr, H., Nielsen, J. O., Hansen, J-E. S., Brunak, S., *Biochem. J.* 1995, *308,* 801–813.

[12] Dutta, B., Rao, C. V. N., *Biochim. Biophys. Acta* 1982, *701,* 72–85.

[13] Schachter, H., Brockhausen, I. in: Allen H. J., Kisailus E. C., (Eds.), *Glycoconjugates: Composition, Structure and Function*, Marcel Dekker, New York 1992, pp. 263–332.

[14] Scharfman, A., Lamblin, G., Roussel, P., *Biochem. Soc. Trans.* 1995, *23,* 836–839.

[15] Hounsell, E. F., Davies, M. J., Renouf, D. V., *Glycoconj. J.* 1996, *13,* 19–26.

[16] Cooper, C. A., Wilkins, M. R., Williams, K. L., Packer, N. H., *Electrophoresis* 1999, *20,* 3589–3598.

[17] Dwek, R. A., Edge, C. J., Harvey, D. J., Wormald, M. R., Parekh R. B., *Ann. Rev. Biochem.* 1993, *62,* 65–100.

[18] Sasaki, H., Ochi, N., Dell, A., Fukuda, M., *Biochemistry* 1988, *27,* 8618–8626.

[19] Hardy, M. R., Townsend, R. R., Lee, Y. C., *Anal. Biochem.* 1988, *170,* 54–62.

[20] Robards, K. L., Whitelaw, J., *J. Chromatogr.* 1986, *373,* 81–170.

[21] Strominger, V., Park, J. T., Thompson, R. E., *J. Biol. Chem.* 1959, *243,* 3263–3268.

[22] Bitter, T., Muir, H. M., *Anal. Biochem.* 1962, *4,* 330–334.

[23] Skoza, L., Mohos, S., *Biochem. J.* 1976, *159,* 457–462.

[24] Honda, S., *Anal. Biochem.* 1984, *140,* 1–47.

[25] El Rassi, Z., *Electrophoresis* 1997, *18,* 2400–2407.

[26] Hase, S., *J. Chromatogr. A.* 1996, *720,* 173–182.

[27] Merkle, R. K., Poppe, I., *Methods Enzymol.* 1994, *230,* 1–15.

[28] Harvey, D. J., *Glycoconj. J.* 1992, *9,* 1–12.

[29] Packer, N. H., Harrison, M. J., *Electrophoresis* 1998, *19,* 1872–1882.

[30] Whittal, R. M., Palcic, M. M., Hindsgaul, O., Li, L., *Anal. Chem.* 1995, *67,* 3509–3514.

[31] Bairoch, A., Apweiler, R., *Nucleic Acids Res.* 2000, *28,* 45–48.

[32] Wilkins, M. R., Lindskog, I., Gasteiger, E., Bairoch, A., Sanchez, J. C., Hochstrasser, D. F., Appel, R. D., *Electrophoresis* 1997, *18,* 403–408.

[33] Patel, T., Bruce, J., Merry, A., Bigge, C., Wormald, M., Jaques, A., Parekh, R., *Biochemistry* 1993, *32,* 679–693.

[34] Cooper, C. A., Packer, N. H., Redmond, J. W., *Glycoconj. J.* 1994, *11,* 163–167.

[35] Dell, A., Morris, H. R., Easton, R. L., Panico, M., Patankar, M., Oehniger, S., Koistinen, R., Koistinen, H., Seppala, M., Clark, G. F., *J. Biol. Chem.* 1995, *270,* 24116–24126.

[36] Sturm, A., *Eur. J. Biochem.* 1991, *199,* 169–179.

[37] Sutton, C. W., O'Neill, J. A., Cottrell, J. S., *Anal. Biochem.* 1994, *218,* 34–46.

[38] Takahashi, N., Lee, K. B., Nakagawa, H., Tsukamoto, Y., Masuda, K., Lee, Y. C., *Anal. Biochem.* 1998, *255,* 183–187.

[39] Lerouge, P., Cabanes-Macheteau, M., Rayon, C., Fischette-Laine, A. C., Gomord, V., Faye, L., *Plant Mol. Biol.* 1998, *38,* 31–48.

[40] Chai, W., Hounsell, E. F., Cashmore G. C., Rosankiewicz, J. R., Feeney, J., Lawson, A. M., *Eur. J. Biochem.* 1992, *207,* 973–980.

[41] Kubelka, V., Altmann, F., Kornfeld, G., Marz, L., 1994, *Arch. Biochem. Biophys. 308,* 148–157.

[42] Weisshaar, G., Hiyama, J., Renwick, A. G., Nimtz, M., *Eur. J. Biochem.* 1991, *195,* 257–268.

[43] Geyer, R., Dabrowski, J., Dabrowski, U., Linder, D., Schluter, M., Schott, H. H., Stirm, S., *Eur. J. Biochem.* 1990, *187,* 95–110.

[44] Yoneda, A., Ogawa, H., Matsumoto, I., Ishizuka, I., Hase, S., Seno, N., *Eur. J. Biochem.* 1993, *218,* 797–806.

[45] Saito, F., Tani, A., Miyatake, T., Yanagisawa, K., *Biochem. Biophys. Res. Commun.* 1995, *210,* 703–710.

[46] Wakabayashi, H., Natsuka, S., Mega, T., Otsuki, N., Isaji, M., Naotsuka, M., Koyama, S., Kanamori, T., Saki, K., Hase, S., *J. Biol. Chem.* 1995, *274,* 5436–5442.

[47] Takeuchi, M., Inoue, N., Strickland, T. W., Kubota, M., Wada, M., Shimizu, R., Hoshi, S., Kozutsumi, H., Takasaki, S., Kobata, A., *Proc. Natl. Acad. Sci. USA* 1989, *86,* 7819–9822.

[48] Misaizu, T., Matsuki, S., Strickland, T. W., Takeuchi, M., Kobata, A., Takasaki, S., *Blood* 1995, *86,* 4097–4104.

[49] Schäffner, C., Graninger, M., Messner, P., *Proteomics* 2000, *1,* 248–261.

[50] Harvey, D. J., *Proteomics* 2000, *1,* 311–328.