Pacific Symposium on Biocomputing 7:297-309 (2002)

DATA STANDARDISATION IN GLYCOSUITEDB

C.A. COOPER, M.J. HARRISON, J.M. WEBSTER, M.R. WILKINS, N.H. PACKER
Proteome Systems Ltd, Locked Bag 2073, North Ryde, NSW 1670, Australia

GlycoSuiteDB, a database of glycan structures, has been constructed with an emphasis on quality, consistency and data integrity. Importance has been placed on making the database a reliable and useful resource for all researchers. This database can help researchers to identify what glycan structures are known to be attached to certain glycoproteins, as well as more generally identifying what types of glycan structures are associated with different states, for example, different species, tissues and diseases. To achieve this, a major effort has gone into data standardisation. Many rules and standards have been adopted, especially for representing glycan structure and biological source information. This paper describes some of the challenges faced during the continuous development of GlycoSuiteDB.

1 Introduction

GlycoSuiteDB[1] is a curated and annotated database of glycan structures, available on the Internet at www.glycosuite.com. It was initiated in April 1999 and first made available in September 2000. There are currently more than 6000 entries in GlycoSuiteDB, extracted from approximately 700 references and covering 250 distinct proteins from about 160 different species.

The glycan structures are presented with the biological source from which they were obtained, the literature references in which the glycan structure was described, and the methods used by the researchers to determine the structure. An example entry from GlycoSuiteDB is given in Figure 1. The main aim of GlycoSuiteDB is to store and disseminate information on protein glycosylation in a logical, integrated and searchable way, in order to simplify the study and understanding of glycobiology.

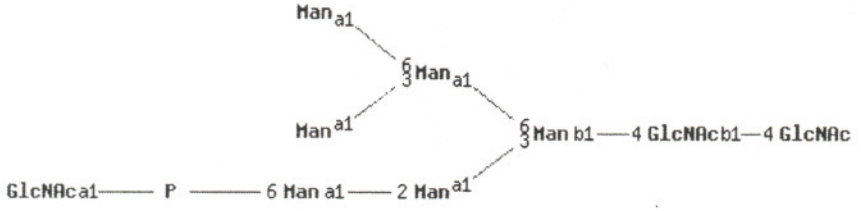
A database of glycosylation faces a different set of complexities from those of nucleic acid or protein sequence databases. An obvious difference is that unlike nucleic and protein sequences, where the individual bases or amino acids are linked together in a linear fashion, glycan structures are branched. Parameters such as anomeric configuration and position of linkage between monomers, contribute to five to six orders of magnitude higher structural diversity than found in proteins. For example, the number of possible structures from six known amino acids is $6! = 720$. The number of possible linear structures from six D-hexose molecules is $6! \times 2^6 \times 4^5 = 47\ 185\ 920$, where the first term is the number of permutations of six linear molecules; the second, the number of possible anomeric configurations; and the third, the position of the linkage. When the ring size (pyranose or furanose) and L-sugars are considered this number increases by 4096. If branching of the chains is

considered the number of possibilities increases by more than 100 fold, and naturally occurring substituents such as sulfates and phosphates increase it again. Having said this, the conservative nature of biology dictates that nowhere near this number of possibilities actually occurs, but the discovery of more and more unusual structures emphasizes the need for a consistent, curated catalogue of known glycans.

A glycan database is also complicated by the fact that glycosylation is a finely controlled process dependent upon the availability and activity of the various glycosyltransferases, glycosidases, monosaccharides and precursors[2]. As a result of these factors, one glycosylation site may have many glycan structures, and the same protein expressed at different times of development or from different tissues, can possess different glycan chains. Similarly, in the case of recombinant or viral proteins, the glycosylation machinery of the host organism is the main influence on glycosylation of the protein. More indirectly, levels of various hormones also affect protein glycosylation through a variety of cell type-dependent changes and the production of differentiated phenotypes[3].

GlycoSuiteDB / source / protein / top

entry: 3053-1328



Species	*Homo sapiens* (HUMAN); sample isolated from species *Mesocricetus auratus*
Class	MAMMALIA
Source	UROGENITAL SYSTEM, KIDNEY (cell line BHK-21)
Source notes	NONE
Attached to	ERYTHROPOIETIN (swiss-prot entry P01588); amino-acid ASN-51
Linkage	N-LINKED
Glycosylation sites	N-51, N-65 AND N-110 [NIMTZ ET AL. (1993) EUR. J. BIOCHEM. 213:39-56].
Identified by methods	MALDI-TOF MS, METHYLATION ANALYSIS, MONOSACCHARIDE ANALYSIS, PROTON NMR
References	Nimtz (1995) *Febs Lett.* **365**: 203-208
Glycan structure	GlcNAc(a1-P-6)Man(a1-2)Man(a1-3)[Man(a1-3)[Man(a1-6)]Man (a1-6)]Man(b1-4)GlcNAc(b1-4)GlcNAc
Mass	1679.5319 Da (*monoiso*), 1680.4346 Da (*avg*), total residues: 9
Composition	$Hex_6HexNAc_3P_1$
Release date	04-NOV-00 (*last updated* 10-JAN-01)

Figure 1: An example entry from GlycoSuiteDB

Due to these complexities, all the information in GlycoSuiteDB is manually extracted from the original scientific literature by trained glycobiologists. Whilst doing this, a number of inconsistencies in the way data is represented in these publications became apparent. This paper describes the way in which we have addressed these variations in order to achieve a standardised format in which the entries are presented in GlycoSuiteDB.

2 Glycan Structure Representation

As the glycan structures are the most important data type in the database, it was important to standardise their representation in such a way as to provide consistency and to enable different searching criteria.

2.1 Linear Representation

Since glycans are often branched structures, this poses special challenges for electronic storage. To enable storage and searching of the glycan structures in a textual form, a linear representation or sequence was formulated. The monosaccharide abbreviations[4] and condensed linear form[5] recommended by IUPAC were adopted with additional rules[6] designed to ensure consistent representation of all glycan structures. For example, the IUPAC recommendation states that a branched glycan is represented as a string by placing branches inside square brackets. However, the guidelines for deciding which chain is the parent and which is a branch were limited and not comprehensive enough to ensure constant and precise representation of all glycan structures. In particular, structures are often described in the literature without full assignment of all linkages and anomeric configurations. This complicates converting branched structures to linear form, as the branch linkages are not known.

To address these issues rules were designed[6] and adopted for converting branched glycan structures into linear form:

- Alpha and beta are represented by 'a' and 'b', respectively
- Where the anomeric configuration or linkage point is not known, a question mark is used
- The parent, or primary, chain is defined as the longest chain, all other chains are branches
- If two chains are of the same length, the more branched chain is the primary chain
- If two chains have the same length and degree of branching, then the chain with the lowest alphabetical terminal residue is considered primary,

working from the most terminal residue towards the branch point until a difference is found

- If two chains are still indistinguishable, then the chain with the lowest terminal linkage, working in towards the branch point until a difference is observed, is considered primary

These rules work equally well with *N*-,*O*- and *C*-linked glycan structures. They also cope with multiply branched structures. For example, structure 1555 (Figure 2) is represented in linear form as: GalNAc(b1-4)[NeuAc(a2-3)]Gal(b1-4)GlcNAc(b1-2)[GalNAc(b1-4)[NeuAc(a2-3)]Gal(b1-4)GlcNAc(b1-4)]Man(a1-3)[GalNAc(b1-4)[NeuAc(a2-3)]Gal(b1-4)GlcNAc(b1-2)[GalNAc(b1-4)[NeuAc(a2-3)]Gal(b1-4)GlcNAc(b1-6)]Man(a1-6)]Man(b1-4)GlcNAc(b1-4)[Fuc(a1-6)]GlcNAc.

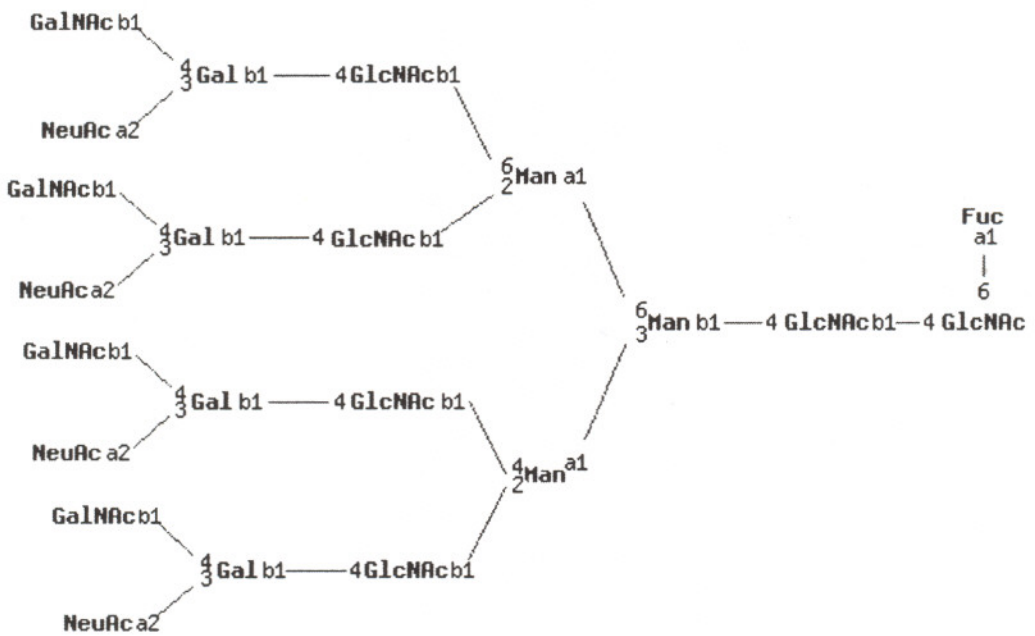


Figure 2: Structure 1555 from GlycoSuiteDB

We have not tried to simplify the terminology of the linear form, as we believe that the 2-D graphical representation that is presented in GlycoSuiteDB (see section 2.3) gives the user a more realistic image of the glycan.

Currently only full glycan structures are entered into GlycoSuiteDB. Fragments of structures, such as lectin recognition patterns, and monosaccharide compositions without any linkage order are not recorded.

2.2 *Searching*

Using the linear format it is possible to search the glycan structure field for structures containing specific residues and linkages. For example, it is possible to perform a search on GlycoSuiteDB for all structures containing the Lewis X motif (Figure 3A).

Figure 3: A) Lewis X motif; and B) Sialyl Lewis X motif

Since the motif is terminal on one branch, the order of the residues in the linear form, following the rules created, is Fuc(a1-3)[Gal(b1-4)]GlcNAc. However, the overall order of this branch within a full linear glycan structure may not be terminal, therefore it is necessary to look in the database for the text 'Fuc(a1-3)[Gal(b1-4)]GlcNAc%' or '%[Fuc(a1-3)[Gal(b1-4)]GlcNAc%' where % is a wild card.

Using this query it was found that there are more than 100 entries in GlycoSuiteDB containing the Lewis X motif. Structures were found where this motif was not terminal in the linear form. For example, structures 1227 and 3803, shown in Figure 4, are represented in linear form as

i)NeuAc(a2-6)Gal(b1-4)GlcNAc(b1-2)Man(a1-3)***[Fuc(a1-3)[Gal(b1-4)]GlcNAc***(b1-2)Man(a1-6)]Man(b1-4)GlcNAc(b1-4)[Fuc(a1-6)]GlcNAc, and

ii)Gal(a1-3)Gal(b1-4)GlcNAc(b1-2)Man(a1-6)***[Fuc(a1-3)[Gal(b1-4)]GlcNAc***(b1-2)Man(a1-3)]Man(b1-4)GlcNAc(b1-4)[Fuc(a1-6)]GlcNAc, respectively.

A similar search was performed looking for Sialyl Lewis X motif (Figure 3B). 32 entries were found to contain this structural feature. Interestingly, Sialyl Lewis X and Lewis X motifs were found on both *N*- and *O*-linked glycans. In addition the search revealed that Lewis X containing structures were found only on native human proteins, whereas Sialyl Lewis X containing structures were reported on glycoproteins isolated from humans, pigs, cattle, mice, rabbit, spotted salamander, tiger salamander and Iberian ribbed newt.

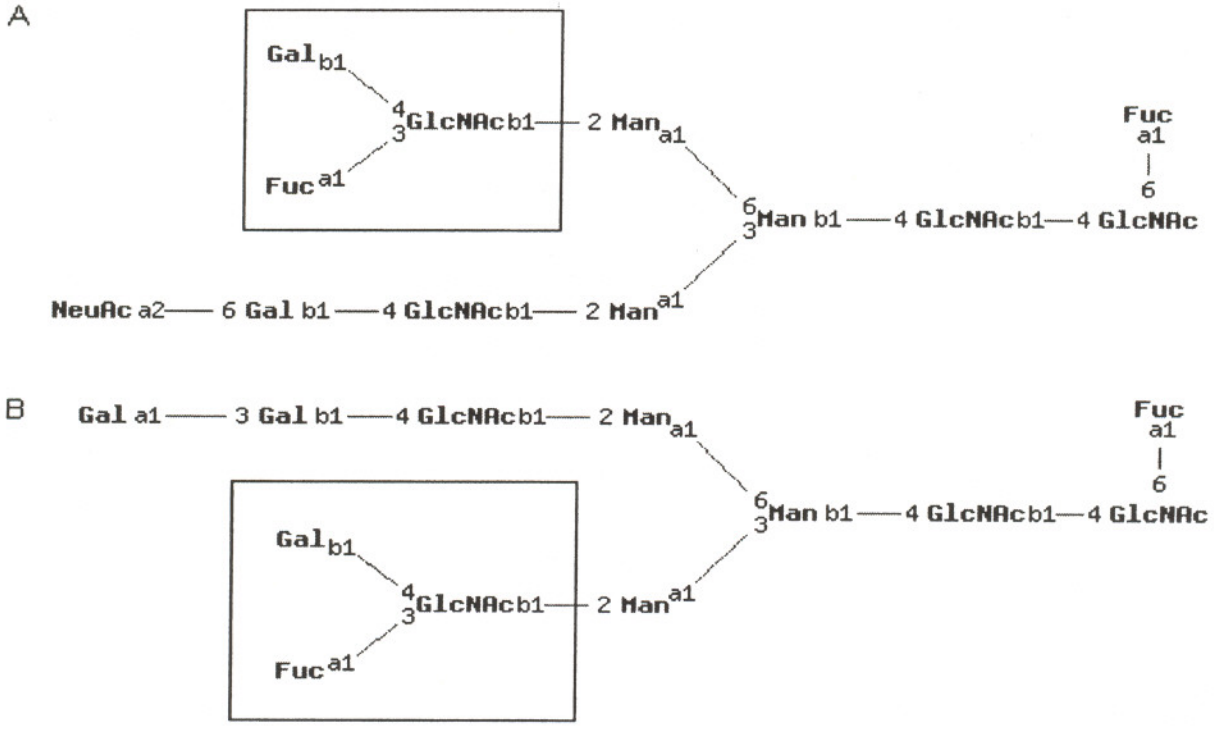


Figure 4: A) Structure 1277 and B) Structure 3803, both containing Lewis X motif (boxed)

2.3 Two-Dimensional Graphical Representation

Standardised linear structures, whilst essential for a searchable database, are not user-friendly because of their visual complexity. GlycoSuiteDB converts the linear string to a more acceptable 2-D graphical representation (as shown in Figures 2, 3 and 4). The use of such software also ensures that the glycan structures are free of syntactical errors and that all glycan structures are valid in terms of linkage. It ensures that monosaccharide residues do not have linkages to positions occupied by other residues, such as other monosaccharides or acetamido groups on GlcNAc.

In addition, based on the linear glycan structure rules given above, we have developed a 'sugarbuilder' tool that allows for the user construction of complex glycans through the incremental addition of monosaccharides. This tool can be used to generate the syntactically correct glycan structure for any given glycan structure. This tool also forms the basis of a user interface that will enable the user to query the glycan structure for full glycan structures, or substructures (such as epitopes), without having to know the linear code rules.

Basic substructure searches, such as those performed in section 2.2, use a pattern-matching technique on the linear code. However, these searches may be

Pacific Symposium on Biocomputing 7:297-309 (2002)

limited in that not all structures with a given substructure may be found. Other branches that originate from, or include, the search substructure may be hidden by the presence of nested branch sequences that interrupt the continuous sequence of the search substructure. In the case of GlycoSuiteDB, substructure searches are implemented using mathematical tree-matching algorithms. This type of search algorithm ensures that all glycans that contain the given substructure will be found.

3 Biological Source

Because the glycan structures are very dependent on their biological origin in terms of species, tissue and cell type, each entry in GlycoSuiteDB records these details.

3.1 Taxonomy

Species names are checked against the NCBI taxonomy database (http://www.ncbi.nlm.nih.gov/Taxonomy). Taxonomy class information is also taken from this database. Where the class classification is not given, the closest division is used and the division definition is noted in brackets. For example, Rosidae is the subclass to which *Brassica rapa* (field mustard), *Glycine max* (soybean) and many other plants belong. It is recorded in the database as 'Rosidae (subclass)'.

3.2 Recombinant Proteins

Many glycoproteins under study have been expressed in a recombinant system. This is a common situation if the native protein is not able to be isolated in sufficient quantities to enable the study of the glycan structures. Recombinant glycoproteins are also common due to the increasing demand for glycoprotein drug products. The expression of a protein in different recombinant systems results in different glycosylation patterns compared to the non-recombinant protein and between the different expression systems. In these cases GlycoSuiteDB lists the species name as that from which the DNA encoding the protein originates. The recombinant field then contains the name of the species in which the protein has been expressed. For example, human Interferon omega-1 expressed in *Spodoptera frugipeda*, would have *Homo sapiens* as the species and *Spodoptera frugipeda* in the recombinant field.

Like recombinant proteins, the glycosylation of viral proteins is dependent on the glycosylation machinery of the host species. Thus the species name of a viral glycoprotein is also given as that from which the DNA encoding the protein originates, i.e. the name of the virus, and the recombinant field contains the species from which the viral protein was isolated.

Pacific Symposium on Biocomputing 7:297-309 (2002)

3.3 Tissue and Cell Type

A major source of inconsistency in the literature is that the tissue and cell type source of a glycoprotein is often described in a variety of ways with differing degrees of specificity. For example, mucin samples from the lung are sometimes described as respiratory, bronchial, or tracheobronchial mucins. Since glycosylation is tissue-dependent as well as protein-dependent it is often desirable to be able to search for all types of glycan structures coming from the same tissue source and to have as much information on the biological source as possible. To enable this we developed a standard format for describing the tissue or cell type based on the anatomy categories of the National Library of Medicine's medical subject headings (MeSH).

For each entry in GlycoSuiteDB the tissue or cell type is described in a maximum of four columns: system, division1, division2, division3 and division4 where each is a nested subset of the previous column. For example, Table 1 shows the various subdivisions of the respiratory and hemic systems adopted for GlycoSuiteDB. It is a common problem that tissues or cell types can exist under multiple systems or divisions. Our current position is to standardise our classification so that most of the tissues or cell types are found in one system only. For example, the nose can be a subdivision of the respiratory system or the sensory system. Currently this tissue is classified under the sensory system since the sub-tissue from which the glycoprotein was isolated was the vomeronasal organ, the primary function of which appears to be in sensing pheromones.

Using this information it is possible to search for which tissue/cell types express particular monosaccharides, thus reflecting the activity of glycosyltransferases. For example, a search for where Gal(b1-4) residues are found in *N*-linked glycans derived from human non-recombinant proteins, showed that this particluar residue and linkage has only been isolated to date from *N*-linked glycans in milk, a secretion subdivision of the exocrine system, and urine, an excretion subdivision of the urogenital system. A similar search for Gal(b1-3) shows that this residue is found in *N*-linked glycans from glycoproteins from nearly all human systems.

Table 1: Example of the organisation of tissue and cell type information in GlycoSuiteDB, adapted from the anatomy categories of the National Library of Medicine's medical subject headings (MeSH).

System	Division1	Division2	Division3	Division4
Hemic system	Blood cell	Erythrocyte	Erythrocyte membrane	
		Leukocyte	Mononuclear leukocyte	Lymphocyte
Respiratory system	Lung	Mucosa		
	Pleura	Fluid		

3.4 Protein Name, Amino Acid Numbering and Glycosylation Sites

When the protein from which a particular glycan structure has been characterised is known, this information is stored in GlycoSuiteDB. To minimize inconsistencies the protein is searched for in the SWISS-PROT/TrEMBL protein databases and the name used is that preferred in these databases. For example, alpha-1-protease inhibitor and alpha-1-antiproteinase both describe the same protein, for which the preferred name from the SWISS-PROT database is alpha-1-antitrypsin. GlycoSuiteDB entries are cross-linked to the corresponding SWISS-PROT protein entry, and SWISS-PROT links directly to GlycoSuiteDB, where appropriate.

Where known, the amino acids to which an individual glycan structure is linked are also entered, with the numbering of the glycosylated amino acids following the sequence given in SWISS-PROT. If the sequence is not in SWISS-PROT, the numbering follows the sequence numbering given in the relevant literature article.

Not all glycan structures are linked to a particular protein however. Whilst many researchers separate a protein to purity before analyzing its glycans, it is also common for researchers to look at the total glycans present in mixtures of proteins, for example, from a particular tissue or cell line. There are many (greater than 55%) of structures in GlycoSuiteDB therefore which do not have a link to a particular SWISS-PROT protein entry.

3.5 Disease Names and Cell Lines

Glycosylation is known to be altered in disease states. For example, alpha fucose linked to the 6 position of reducing terminal GlcNAc in the *N*-linked glycans of alpha-1-antitrypsin is only found when the protein is isolated from patients with hepatocellular carcinoma[7]. In addition different cell lines can result in different

Pacific Symposium on Biocomputing 7:297-309 (2002)

glycosylation of the same recombinantly-produced protein, e.g., there is a wide difference in the *N*-linked glycan structures of human interferon-gamma expressed in Chinese hamster ovary cells and Sf9 cells[8].

Disease names have been standardised using the names and definitions National Library of Medicine's medical subject headings (MeSH) and CancerWEB's online medical dictionary (http://www.graylab.ac.uk/omd/). Cell line names have been adopted from American Type Culture Collection (ATCC) (www.atcc.org) and HyperCLDB, the hypertext on cell culture availability extracted from the Cell Line Data Base of the Interlab Project (http://www.biotech.ist.unige.it/cldb/indexes.html).

4 Methods

The quality of GlycoSuiteDB relies on information published in the literature. For glycan structures, the confidence we have that a published structure is correct is dependent on the method, or methods, used in its determination. In our in-house analytical work we use a confidence rating to distinguish data quality based on these methods. Each method used to determine glycan structures has been critically assessed and has been given a confidence value based on the reliability of the method and the value and the extent of the information obtainable from the method. For example, proton NMR is given a value of 10 because it can be used to obtain information on the complete structure of the glycan, e.g., the type and ring form of monosaccharides, relative number of each sugar residue, the linkage positions, anomeric configurations and the sugar sequence. Mass spectrometric methods have been given a value of 5 if they only give the total mass of the glycan from which the possible composition can be predicted[9]. However, if a single structure was fragmented by tandem mass spectrometry, the sugar sequence and linkage position of the glycan can also be deduced. In this case, a method called "fragmentation", with a confidence value of 3, is noted in addition to the mass spectrometry method, to reflect the extra information obtained. Mass spectrometry can, however, only give generic monosaccharide information as many monosaccharides, e.g., glucose and mannose, have the same mass.

The relative values for each method were carefully chosen so that the sum of the confidence values for certain combinations of methods were comparable. For example, a structure determined by proton NMR and methylation analysis is very reliable and has a confidence value of 18. Likewise structures characterized by: i) methylation analysis, monosaccharide analysis and glycosidase treatment; or ii) monosaccharide analysis, glycosidase treatment, mass spectrometry and periodate oxidation; are quite dependable and would have confidence values of 18. A structure that had been determined by its monosaccharide composition and chromatographic

elution position compared to a standard would not be considered to be as reliable as this and would have a resultant confidence value of 5 assigned to reflect this.

Although these confidence values are not included in the public version of GlycoSuiteDB because of the controversial nature of the ratings, the user can make their own judgements from the methods field provided.

5 Conclusion

In this article, we have described some of the features of GlycoSuiteDB. There are only 2 other web-based glycan structural databases, CarbBank and Glycominds. Funding for CarbBank was discontinued and the web site now runs unattended, with no new data entries. In its construction, because of its dependence on user entry, there was no standardisation of data procedures implemented, which lead to variability in data formatting and inconsistencies in search output[6]. Glycominds is a new glycan structural database launched on the web in November 2000. Unlike GlycoSuiteDB however, Glycominds advocates the representation of glycan structures in a linear format in order to enable searching for specific epitopes. The data in GlycoSuiteDB is also based on a linear format entry and can be searched in the same way. At present this functionality in both databases is limited, as discussed in Section 2.3 of this paper, and we are addressing this.

GlycoSuiteDB is available on the web (www.glycosuite.com) and there has been considerable focus on data standardisation, which means that it is easily searchable and accurate. Queries can be performed using monosaccharide composition, glycan mass, species, biological tissue/cell type, protein name or any combination of these. GlycoSuiteDB is already extensively linked with the SWISS-PROT protein database and PubMed. Further links to other online databases, such as the Online Mendelian Inheritance in Man (OMIM) database, are planned.

GlycoSuiteDB has been designed to allow researchers to search for precedence and thus to have more confidence in making assumptions on glycan structure. However, in the development of GlycoSuiteDB it has become obvious that there are many variations in glycan structure and that not all assumptions are valid. For example, there are 16 monosaccharide compositions that correspond to more than 10 unique structures each. Moreover, at least 18 unique glycan structures have been isolated with the composition hexose=3, hexNAc=3 and deoxyhexose=1. More than 290 unique *N*-linked and 270 unique *O*-linked glycan structures have been characterised from humans.

An example of a glycan structure that does not conform to precedence is an *N*-linked glycan isolated from chicken ovalbumin[10] (Figure 5). This structure was characterised by FAB-MS and proton NMR, and has three GlcNAc residues all individually linked to the Man6 branch. Using the FAB-MS results only, precedence

would probably have predicted that the structure would have only two branches per mannose arm with GlcNAc-GlcNAc units added linearly.

As the above example indicates, we would make the point that precedence does not necessarily mean that a mass equates with a certain structure, and that researchers should be careful as to what structure is assigned. Initially journal articles describing glycan structures focused solely on defining the analytical methods used and on the characterization of major glycans on a particular protein. As advances have been made in the field the focus has shifted to trying to see all the glycan structures present by using more sensitive approaches, such as mass spectrometry, and to try to determine the function of the glycans. This has led to scientists making more assumptions about the glycan structures rather than systematically determining all linkages and anomeric configurations. This is particularly true with N-linked glycans. However, GlycoSuiteDB can assist researchers as a resource to see what is already known about what glycan structures are attached to certain glycoproteins.

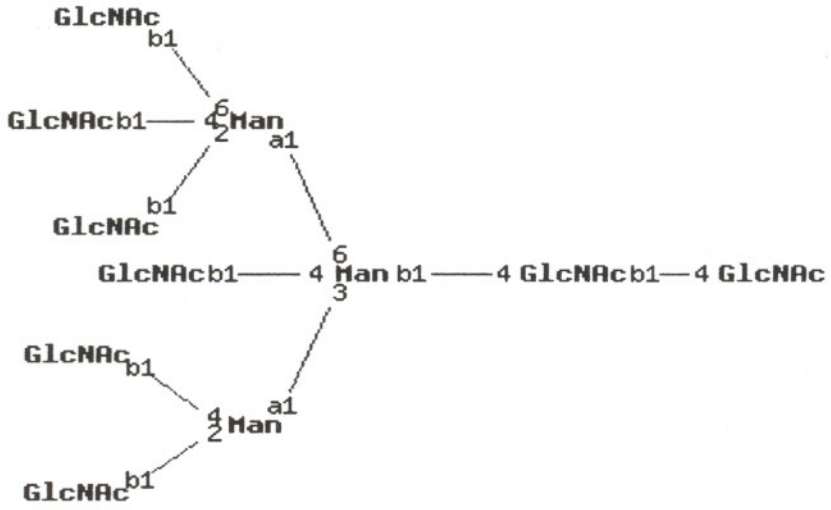


Figure 5: *N*-linked glycan structure isolated from chicken ovalbumin[10]

References

1. C.A. Cooper et al, *Nucleic Acids Res.* **29**, 332 (2001)
2. R. A. Dwek et al, *Annu. Rev. Biochem.* **62**, 65 (1993)
3. C. F. Goochee and T. Monica, *Biotechnology (N. Y.)* **8**, 421 (1990)
4. A. D. McNaught, *Pure Appl. Chem.* **68**, 1919 (1996)
5. N. Sharon, *Eur. J. Biochem.* **159**, 1 (1986)
6. C. A. Cooper et al, *Electrophoresis* **20**, 3589 (1999)
7. A. Saitoh et al, *Arch. Biochem. Biophys.* **303**, 281-287 (1993)

8. D. C. James et al, *Biotechnology (N. Y.)* **13**, 592 (1995)
9. C.A. Cooper et al, *Proteomics* **1**, 340 (2001)
10. M. L. Corradi Da Silva et al, *Arch. Biochem. Biophys.* **318**, 465 (1995)