GlycoSuiteDB: a new curated relational database of glycoprotein glycan structures and their biological sources

Catherine A. Cooper*, Mathew J. Harrison, Marc R. Wilkins and Nicolle H. Packer

Proteome Systems Ltd, 1/35–41 Waterloo Road, North Ryde, Sydney, NSW 2113, Australia

Received September 5, 2000; Revised and Accepted October 25, 2000

ABSTRACT

GlycoSuiteDB is a relational database that curates information from the scientific literature on glycoprotein derived glycan structures, their biological sources, the references in which the glycan was described and the methods used to determine the glycan structure. To date, the database includes most published O-linked oligosaccharides from the last 50 years and most N-linked oligosaccharides that were published in the 1990s. For each structure, information is available concerning the glycan type, linkage and anomeric configuration, mass and composition. Detailed information is also provided on native and recombinant sources, including tissue and/or cell type, cell line, strain and disease state. Where known, the proteins to which the glycan structures are attached are reported, and cross-references to the SWISS-PROT/TrEMBL protein sequence databases are given if applicable. The GlycoSuiteDB annotations include literature references which are linked to PubMed, and detailed information on the methods used to determine each glycan structure are noted to help the user assess the quality of the structural assignment. GlycoSuiteDB has a user-friendly web interface which allows the researcher to query the database using monoisotopic or average mass, monosaccharide composition, glycosylation linkages (e.g. N- or O-linked), reducing terminal sugar, attached protein, taxonomy, tissue or cell type and GlycoSuiteDB accession number. Advanced queries using combinations of these parameters are also possible. GlycoSuiteDB can be accessed on the web at http://www.glycosuite.com.

INTRODUCTION

GlycoSuiteDB is a curated and annotated database of glycan structures. It was initiated in April 1999 and first made available in September 2000. It aims to simplify the study and understanding of glycosylation and glycobiology through storing glycan information in logical, integrated and easily queried ways. Glycan structures are not presented in isolation, but instead are viewed in the context of the protein they are associated with (where appropriate), the cell, tissue and/or developmental stage of the source organism, and other factors such as environmental conditions or disease state of the source organism. We believe this database will be an essential resource for the glycobiologist as well as the biochemist, biotechnologist and those studying human diseases.

In many respects, a database of glycosylation faces a different set of complexities from those of nucleic acid or protein sequence databases. One fundamental difference is that nucleic acid and protein sequences are linear, and users search sequence databases to find homology between sequences. By comparison, glycan structures are (in most cases) branched, and branching occurs through different linkages and anomeric configurations. Thus, a glycan structure database must contain branching information as well as the monosaccharides comprising the structure. A second difference is that whilst a certain nucleic acid or protein sequence is generally unique for a given species, one glycan structure may occur on many different proteins. Also, the glycan structures attached to any particular glycoprotein may change depending on the tissue in which a protein is expressed, the conditions under which an organism is grown or, in the case of recombinant and viral glycoproteins, the host organism in which a protein is expressed. To help address these complexities, the GlycoSuiteDB has been made as a relational, rather than flat file, database.

On a more technical level, the main features of the database are:

- Curation: all data for GlycoSuiteDB is taken from scientific literature and curation is done by trained glycobiologists. Direct entry by researchers is not allowed. This ensures consistency and integrity in the data.
- Relational format: GlycoSuiteDB is stored in a relational format. This helps to ensure data consistency and allows very flexible querying of the database.
- Minimal redundancy: GlycoSuiteDB strives to have minimal redundancy. For example, if a glycan structure has been described from a protein in two or more articles, the structure and source information will only appear once, with the two or more references noted.
- Integration with other online databases: GlycoSuiteDB currently cross-references MEDLINE/PubMed and SWISS-PROT/TrEMBL (1), with more links planned.

*To whom correspondence should be addressed at: Proteome Systems Ltd, Locked Bag 2073, North Ryde, Sydney, NSW 1670, Australia. Tel: +61 2 9889 1830; Fax: +61 2 9889 1805; Email: catherine.cooper@proteomesystems.com
ORGANISATION OF THE GLYCOSUITE DB DATABASE

The fundamental data type in GlycoSuiteDB is the glycan structure. This is analogous to the amino acid sequence in a protein sequence database. Each unique structure in the database is numbered according to the order in which it was entered. However, as a single structure can be found in many different sources (both within and between species), we have combined structure numbers and source numbers to create a structure-source id (GlycoSuiteDB accession number). GlycoSuiteDB accession numbers are analogous to SWISS-PROT or GenBank accession numbers, insofar as a unique GlycoSuiteDB accession number allows the retrieval of a specific glycan structure from a specific biological source.

DATA FORMATS

Figure 1 shows an example ‘entry’ from GlycoSuiteDB. Each entry follows a similar profile, however, some fields are optional. Each major field is explained in detail below.

GlycoSuiteDB number

The GlycoSuiteDB number is constructed from the structure id and the source id, separated by a hyphen. In Figure 1, therefore, the GlycoSuiteDB number 1619-195 refers to structure number 1619 in source number 195.

Glycan structure image

The glycan structures are entered in a condensed linear form (see glycan structure format description below) and visualised within the GlycoSuiteDB web interface as a full structural image, such as that shown in Figure 1.

Species

Species names and preferred common names, if applicable, are used. These are cross-checked against the National Center for Biotechnology Information (NCBI) taxonomy database, http://www.ncbi.nlm.nih.gov/Taxonomy.

Class

Information on the taxonomic class details is also taken from the NCBI taxonomy database. Where no class is given, the closest taxonomic group is used and its classification noted in brackets, e.g. Halobacteriales (order) is the order to which the species Haloferax volcanii belongs.

Source

This field specifies the tissue or cell type from which the glycan was isolated. A challenge is that tissue or cell type information can be described in a number of ways, e.g. mucin samples from the lung are sometimes described as respiratory, bronchial or tracheobronchial. This inconsistency can make searching difficult as structures listed from different sources may ultimately be the same, creating data redundancy and
diluting data consistency. To overcome this, the anatomy categories of the National Library of Medicine’s medical subject headings (MeSH) (http://www.nlm.nih.gov/mesh/meshhome.html) were adopted, with minor changes, to reflect the entries in GlycoSuiteDB. A current list of the biological systems and divisions used to describe tissue and/or cell type is documented at http://www.glycosuite.com/docs/systems.html. Where applicable, the disease, strain and/or developmental stage of the source organism is given, as well as the cell line name where relevant. For example, source 19 reads ‘cell line: Zajdela hepatoma, strain: Sprague-Dawley, disease: hepatocarcinoma, life stage: 7–9 week old’.

Where known and appropriate, the blood group(s) is also given. For example, A, Se, Le(A-B+) and St(a).

Source notes
This is a field that contains any extra information deemed relevant to the biological source. For example, in Figure 1 this field reads ‘Components X1 and X2 contain same structures’. Other examples include, ‘Sample collected 4 days post partum’ or ‘Protein is a mutant in which Asn-251 has been mutated to Ser-251’.

Attached to
If the protein to which a glycan is attached is known, the protein name is stored in GlycoSuiteDB. When this protein is found in SWISS-PROT/TrEMBL, the name used is that from SWISS-PROT/TrEMBL and the accession number is also stored. For example, the preferred name for the protein ‘Stuart Factor’ is Coagulation Factor X, SWISS-PROT number P00743.

Where known, the amino acids to which an individual structure is linked are also given in this field, e.g. in Figure 1, structure 1619 has been found linked to Asn-218 of Bovine Coagulation Factor X. Where the protein is known, but the glycans have not been assigned to individual sites, the term ‘unmapped’ appears. It is important to note that the localisation of individual glycan structures on any particular amino acid is experimentally difficult and is therefore not widely documented. It is also common for one particular glycosylation site to have more than one glycan structure (known as microheterogeniety; 3) and for a particular glycan structure to be present on more than one glycosylation site in a protein.

Glycosylation sites
Where known, all the confirmed sites of glycosylation on a protein are given, along with the reference for the article detailing the discovery of the glycosylation sites. For example, in Figure 1, the sites of glycosylation that have been identified for the protein are ‘T-208 [Inoue and Monta (1993) Eur. J. Biochem. 218:156–163], N-218 & T-485 [Titani et al. (1975) PNAS 72:3082–3086]’.

Where applicable, the numbering of the glycosylated amino acids follows the sequence given in SWISS-PROT. If the sequence is not in SWISS-PROT, numbering will follow the sequence numbering given in the relevant literature article, until such time as the sequence becomes available in SWISS-PROT.

Identified by methods
Every article is carefully read and the methods used to determine a particular glycan structure in an individual biological source are recorded in this field of the database. By listing the methods used, users of GlycoSuiteDB can determine their level of confidence in the assignment of the structure without needing to refer to the original journal article.

Additional notes
This field contains any additional notes regarding the current entry, especially comments concerning any assumptions made or other issues regarding the structural assignment. For example, ‘Structure not confirmed—no data given’ and ‘Mixture analysed. Presence of individual structure not confirmed’.

Glycan structure
Each structure entered at present into the database in a condensed linear form that conforms to the nomenclature given by International Union of Pure and Applied Chemistry (IUPAC) (4, 5). When using the condensed format, the IUPAC convention is that the reducing terminal monosaccharide is at the right-hand end and that a branched glycan is represented in one line form by placing the branch inside square brackets. The guidelines available for deciding which chain is the parent and which is the branch are, however, limited. To overcome these problems, rules previously developed were adopted for depicting branched glycan structures in GlycoSuiteDB (6).

Reference
All references used to date in the construction of GlycoSuiteDB are journal research articles, since only the principal reference to the elucidation of a glycan structure describes the biological source, methods used to analyse the glycan and gives a full explanation as to how the glycan structure was assigned. GlycoSuiteDB therefore stores the first author, year, journal name, journal volume and the page numbers of the article. The PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed) unique identifier (MUID) of the reference entry is also stored where applicable. In the future, full author lists and article titles will be made available.

Mass
For every glycan structure the corresponding glycan mass, both monoisotopic and average, is stored in the database to four decimal places. The masses are calculated automatically from the composition of the glycan structure using the corresponding monosaccharide masses. A table of these monosaccharide masses is available from http://www.glycosuite.com/docs/mass.html.

Composition
The monosaccharide composition of each glycan structure is given as monosaccharide types (e.g. Fig. 1, Hex:5 HexNAc:4 NeuAc:4).

RECOMBINANT AND VIRAL PROTEINS
Where a protein has been expressed in a recombinant system, the species name is given 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, if human erythropoietin is recombinantly expressed in CHO cells, the species field would be
Homo sapiens and the recombinant field would be Cricetus griseus.

Viral proteins are like recombinant proteins in that the glycosylation of the viral protein is dependent on the glycosylation machinery of the host species. Therefore, like recombinant proteins, the species name of a viral glycoprotein is 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 in which the viral protein has been made.

DATABASE ACCESS

The web version of GlycoSuiteDB (http://www.glycosuite.com) can be used to perform simple or complex queries on the database. Simple queries by composition, mass, protein name, biological tissue/cell type or taxonomy return a simple summary page with the results of the query indexed by fields appropriate to the type of query. Advanced queries allow the user to query the database by combining several of the query types described above. From the index of results returned by a query, the user may select one, several or all of the entries that matched their query criteria, thereby allowing the user to interactively narrow their results set. Following this selection, the full GlycoSuiteDB entries for the selected result set are displayed.

Database usage

There are currently no restrictions on the use of GlycoSuiteDB by non-profit organisations as long as its content is not modified in any way. After an initial free access period, use by and for commercial entities will require a licence. Full conditions of use will be made available on the web site and through GeneBio (www.genebio.com), the exclusive worldwide distributor of GlycoSuiteDB.

CURRENT STATUS

This release (no. 1.0) of GlycoSuiteDB has been constructed from data in 578 references. Currently, the database contains most O-linked glycans published since 1950, and N-linked glycans in the literature from the years 1990–2000.

GlycoSuiteDB at present contains 5849 structures from different biological sources. Of these, 2153 are unique structures (1281 N-linked, 834 O-linked and 38 other), with 692 different monosaccharide compositions.

To date there are 742 different biological sources represented in GlycoSuiteDB, from 143 different species. There are ~446 different proteins (where known), with 213 different SWISS-PROT/TrEMBL AC numbers.

SUBMITTING DATA, UPDATES AND CORRECTIONS

We welcome all comments on GlycoSuiteDB. If you would like to make a comment, submit data for possible inclusion in GlycoSuiteDB or if you have updates or corrections for GlycoSuiteDB, please contact the author of this paper or email: glycosuitedb@proteomesystems.com.

FUTURE DEVELOPMENTS

GlycoSuiteDB will continue to grow in both content and functionality. There will be at least two release updates each year, comprised of all recently published glycan structures and, over time, all the N-linked glycan structures in literature published before 1990.

A series of developments are planned for the near- and long-term future. Links to other relevant online databases will be made. For example, it is anticipated that, where possible, disease information will be linked to the Online Mendelian Inheritance in Man (OMIM) database situated at the NCBI (http://www.ncbi.nlm.nih.gov/omim). We are also currently developing a series of tools to use with GlycoSuiteDB. The first of these tools is GlycoMod (http://www.expasy.ch/tools/glycomod/), a software tool designed to find all possible compositions of a glycan structure from the experimentally determined mass of a glycopeptide or released glycan (7). This will soon be linked directly to the GlycoSuiteDB to further increase the power of this approach. Other tools under development focus on glycosidase treatment, mass spectrometry analysis and improved means of rapidly analysing analytical data to better understand glycan structure and function.

REFERENCES