

Design Studies Related to the Development of Distributed Web-based European Carbohydrate Databases

Design Study DS2: Creating a peer-to-peer network of distributed databases for applications related to Glycosciences.

Task Titles:

DS2-SUB2: Recommendation for a concept of a digital cell-by-cell catalogue of glycan structures and glycosyltransferases

Deliverable

SUB-D2: Report: "Cell-by-cell catalogue of glycan structures and glycosyltransferases".

Dissemination: PU; Partners: 1, 4

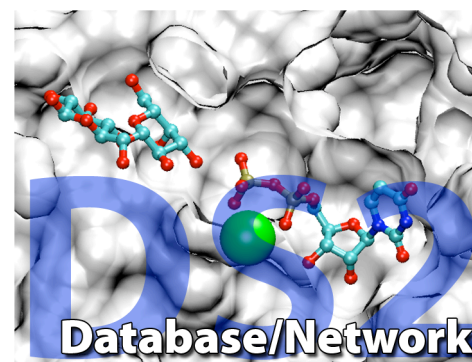
Due date of deliverable: 31.03.2007; Actual submission date: 31.03.2007

Start date of project: 01.04.2005; Duration: 48 months

Organisation name of lead contractor for this deliverable:
Deutsches Krebsforschungszentrum, Heidelberg, Germany

Authors

Tony Merry



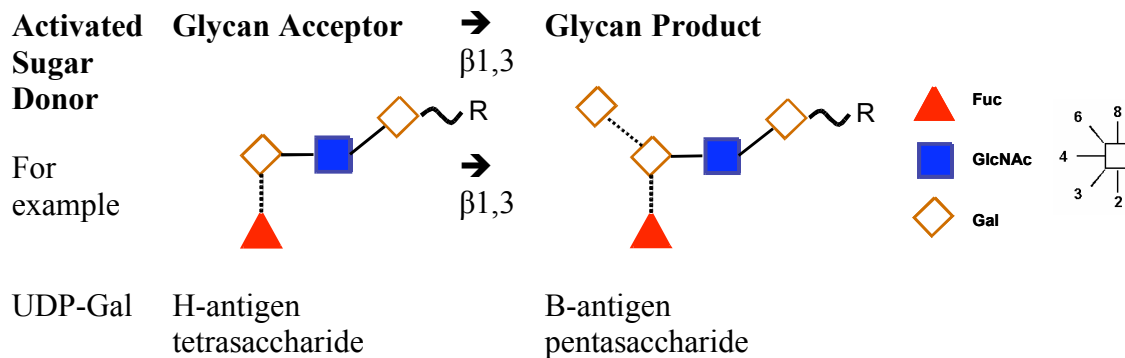
DS1-DSU2: Report: “Cell-by-cell catalogue of glycan structures and glycosyltransferases”

Tony Merry
Subcontractor to EBI

1.	Introduction	3
2.	Activity of Glycosyltransferases.....	4
3.	Assembly of complex glycans.....	5
4.	Diversity of glycosyltransferases.....	8
5.	Biosynthetic Pathways for Glycan Assembly	9
6.	Techniques for investigating glycosyltransferase gene expression	10
7.	Techniques for investigating glycan structures	10
8.	Tissue/cell Distribution of Glycosyl transferases	10
9.	Tissue/cell Distribution of glycans	11
10.	Currently available Databases	11
11.	Recommendations for a digital cell-by-cell catalogue of glycan structure and glycosyltransferases	12
12.	Development of the EUROCarbDB GT database	12
	References	15

1. Introduction

Glycosyltransferases are key enzymes in glycan synthesis that build up glycan chains. They follow the general reaction pattern as follows



These enzymes generally show a high degree of specificity for the nucleotide sugar donor and in the anomericity and position of the linkage formed to the glycan chain. However since the acceptor specificity is generally limited to at most three of the monosaccharides in the glycan they may act on a wide range of glycan acceptors and thus produce a wide range of glycan products.

It is therefore important that any study of the distribution of the glycosyltransferases is also linked to the glycan structures that are actually found. It is generally possible to exclude certain structures on the basis of the transferases present but quite difficult to be sure which particular glycan structures are present. This is particularly the case when the range of glycans present at any particular glycosylation site on a glycoprotein are considered

It is also important to consider that in the context of biological activity or disease markers it is frequently only part of the glycan structure that is important. Thus in many cases the association is with a particular epitope or antigen which may form part of a large number of glycan structure and these may even be present on different glycoconjugates e.g. glycoproteins or glycolipids.

In considering the tissue or cell distribution of glycosyltransferases and glycan all these factors need to be taken into account and should ideally be borne in mind in designing a database

The glycan structures found on the glycoconjugates of any given cell are determined to a greater or lesser extent by the glycosyltransferases which are expressed in that particular cell, and the glycans and the transferases responsible for their biosynthesis may therefore be conveniently considered together, as will be done in this report.

The glycosyltransferases are a remarkable class of enzymes in many ways. They work with the most complex and wide variety of acceptor molecules and yet carry out reactions with a high degree of specificity. The activity of the enzymes may be judged from the resulting glycoconjugate products and the consistency and

stereospecificity of the glycosyltransferases is evident from the precision by which the glycoconjugates are assembled.

They may be generally expressed throughout life in most tissues or may be transiently expressed in a single cell type. They may act on a wide variety of glycoconjugates or be specific for a very limited subset. The degree of specificity of the reactions performed has made them valuable tools for carbohydrate chemists where chemo-enzymatic synthesis allows only the required isomer to be made which is often virtually impossible or only in extremely low yields by conventional chemistry.

Their importance may be judged from studies of individuals with congenital disorders of glycosylation (CDG) or in knockout mouse mutants (Raman, Venkataraman et al. 2006) with severe or lethal phenotypes are found. However such studies have also shown there are several alternative 'back-up' systems which can take over in cases where the usual pathway cannot operate. This may make studies on the biological roles of glycosylation more difficult, but such redundancy is often employed in critical systems, which highlights the importance of glycosylation.

Although the identification and cloning of glycosyltransferase genes has only taken place in the last 20 years, there are now over 200 glycosyltransferases which have been identified. These may perform a number of reactions in the different biosynthetic pathways for glycoconjugates but in many cases catalyse the transfer of a single monosaccharide in a given linkage stereochemistry to the growing glycan chain. In addition to the glycosyltransferases there are other enzymes involved in glycoconjugate biosynthesis including isomerases, sulpho- and phospho-transferases but these will not be studied in this report.

2. Activity of Glycosyltransferases

Glycosyltransferase may be involved in the biosynthesis of a wide variety of glycan structures ranging from a single monosaccharide to a polymer with several thousands of residues. However it is a general principle that the product of one glycosyltransferase reaction becomes the acceptor for the next glycosyltransferase reaction. The major details of the pathway for N-glycan biosynthesis were initially elaborated in the 1980's by the groups of Spiro (Spiro and Spiro 1982) and Kornfeld (Kornfeld and Kornfeld 1985) The reactions are temporally and spatially separated within the cell and any consideration of the overall synthesis of a glycan must take account of this.

An exception to this general rule in the first stages in the biosynthesis of the N-glycan core structure where a lipid linked intermediate is transferred as a single unit. It should be noted that when branching (Reitman, Varki et al. 1981) of the glycan chain is present then the same monosaccharide may be transferred to the non-reducing termini of more than one branch.

In the previous DS1_SUB1 report the very wide diversity of glycan structure found was presented. In the present report only those transferases involved in the biosynthesis of N- and O-linked glycans on glycoproteins and those involved in the biosynthesis of glycolipids will be considered. This study will also be restricted to two species, mouse and human as the majority of studies on glycosyltransferases have been carried out on these. It should be noted however that some glycosyltransferases may have a very limited species distribution as is often the case in bacteria or may be

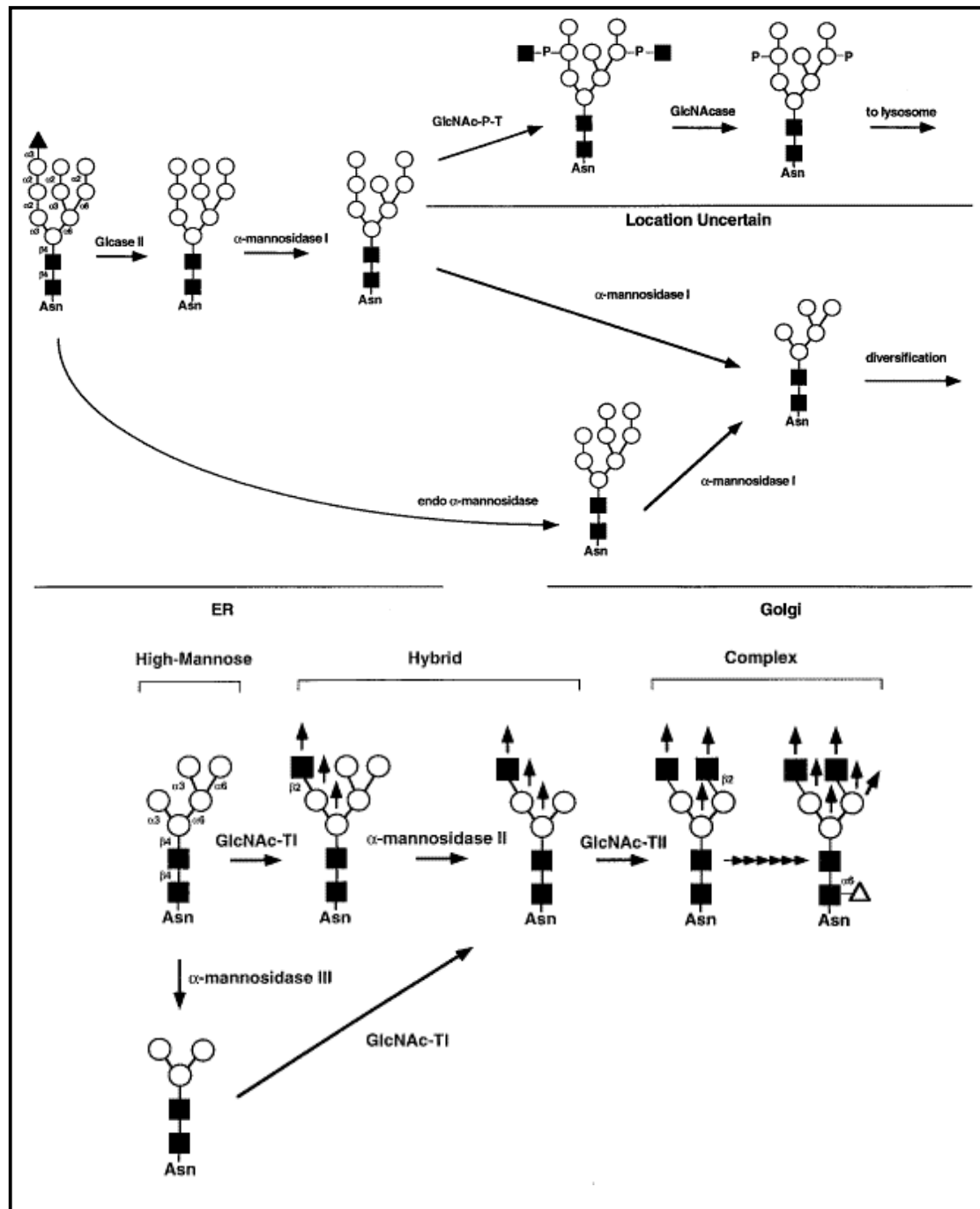
exclusively involved in the biosynthesis of polysaccharides. There are also large number of glycosyltransferases or other glycan modifying enzymes, such as epimerases, that are involved in glycosaminoglycan biosynthesis. Finally in addition to glycosyltransferases there are a number of *glycosidases* which are actually involved in biosynthesis primarily through ‘trimming’ of intermediate structures, although they also play an important part in recycling and degradative pathways. ((Reitman, Varki et al. 1981))

3. Assembly of complex glycans

The pathways for the biosynthesis of glycan are highly conserved and will now be outlined for each of the major class of glycans under consideration here.

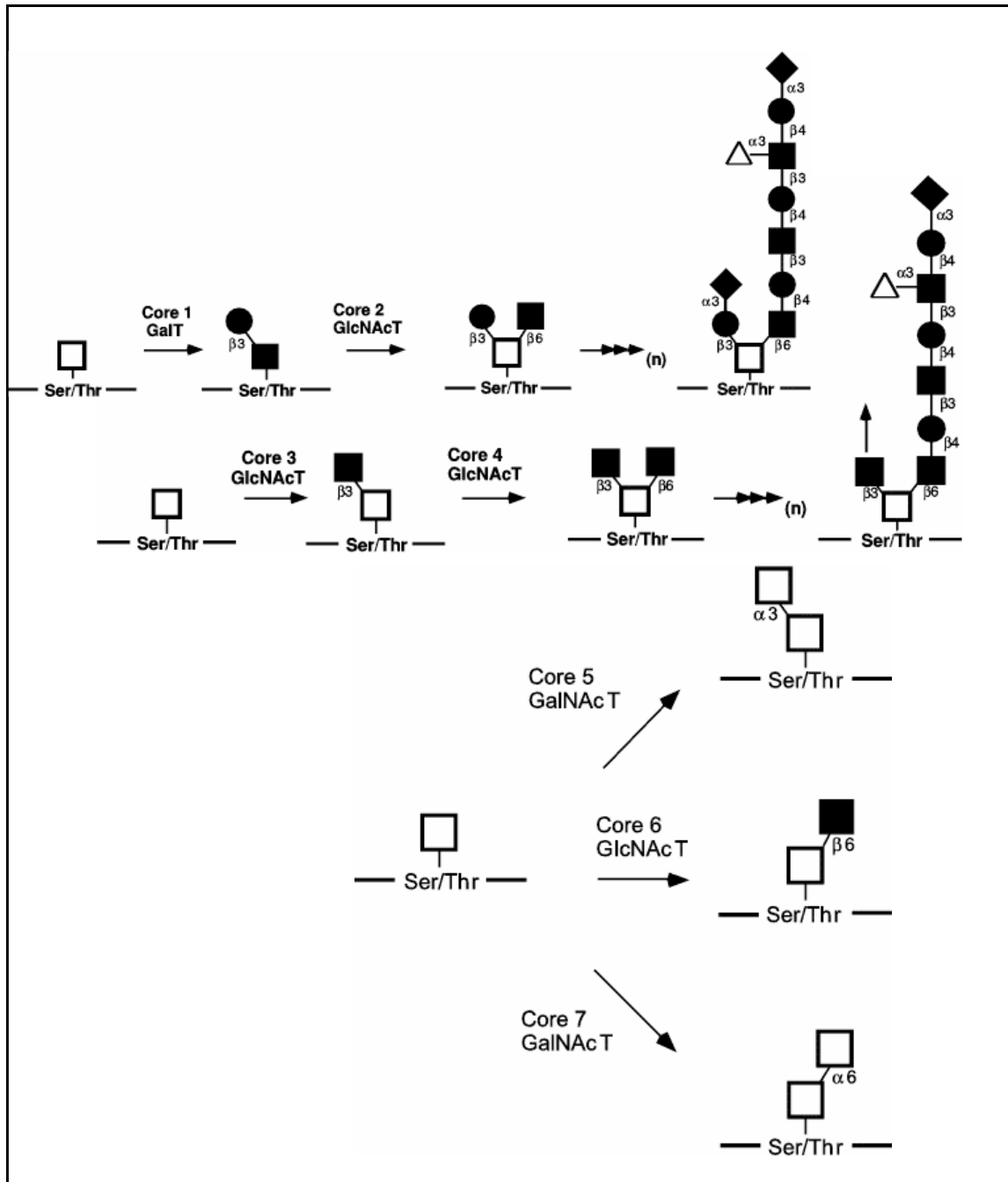
The pathway for the assembly of N-linked glycans is highly conserved through evolution. A common feature of all N-linked glycans described to date is the so called ‘tri-mannosyl core’ which consists of two N-acetylglucosamine residues in a β 1,4 linkage to an asparagine residue with a β 1,4 linked mannose residue to which are attached two further mannose residues in α 1,6 and α 1,3 linkages. This is a direct result of the glycosyltransferases and involved in forming the precursor of this structure as a dolichol-lipid linked donor and the subsequent action of trimming glycosidases. This basic core structure may then be elaborated by glycosyltransferase action as shown in Fig 1. The action of these glycosyl transferases result in the final N-glycan structures present on glycoprotein produced by any cell type.

Figure 1 N-Glycan Assembly (reproduced from *Essentials of Glycobiology* by Ajit Varki (Editor), Richard O. Cummins (Editor), Jeffrey Esko (Editor), Richard Cummings (Editor), Hudson Freeze (Editor), Gerald Hart (Editor), Jamey Marth (Editor) Cold Spring Harbor Laboratory Press, U.S. (31 Dec 1999))



In contrast the O-linked glycans have at least eight different core structures in the case of those linked to serine and threonine residues in addition to less common types of linkage. Also the O-glycans are not built upon a common precursor but are constructed stepwise by the addition of monosaccharides by glycosyltransferases as shown in Fig 2. Following the construction of the various types of core the extension may be carried out by enzymes which are common to many pathways.

Table 2 Protein O-glycan Biosynthesis (reproduced from *Essentials of Glycobiology* by Ajit Varki (Editor), Richard O. Cummins (Editor), Jeffrey Esko (Editor), Richard Cummings (Editor), Hudson Freeze (Editor), Gerald Hart (Editor), Jamey Marth (Editor) Cold Spring Harbor Laboratory Press, U.S. (31 Dec 1999))



The biosynthesis of glycolipids starts with the addition of a single glucose residue to the lipid. This is generally a derivative of glycerol with lipids attached which are often tissue specific. As far as is known there is only one glycan chain attached to the glycerol derivative with two lipid chain which may be the same or different, Following glucose addition the extension of the chain is again performed by glycosyl transferases common to several pathways as shown in Fig 3.

Thus although the core structures in each of these classes of glycoconjugate are very different the terminal structures may be quite similar. It is not uncommon to find the

same terminal structures at the non-reducing end of glycan chains of N- O-linked proteins and from glycolipids from a given cell type.

The biosynthesis of glycosaminoglycans will not be considered in detail here but generally they are quite different from glycoprotein or glycolipid glycans and a completely different set of glycosyltransferases, epimerases and sulphotransferases are involved in the construction of the long glycosaminoglycan chains found in proteoglycans.

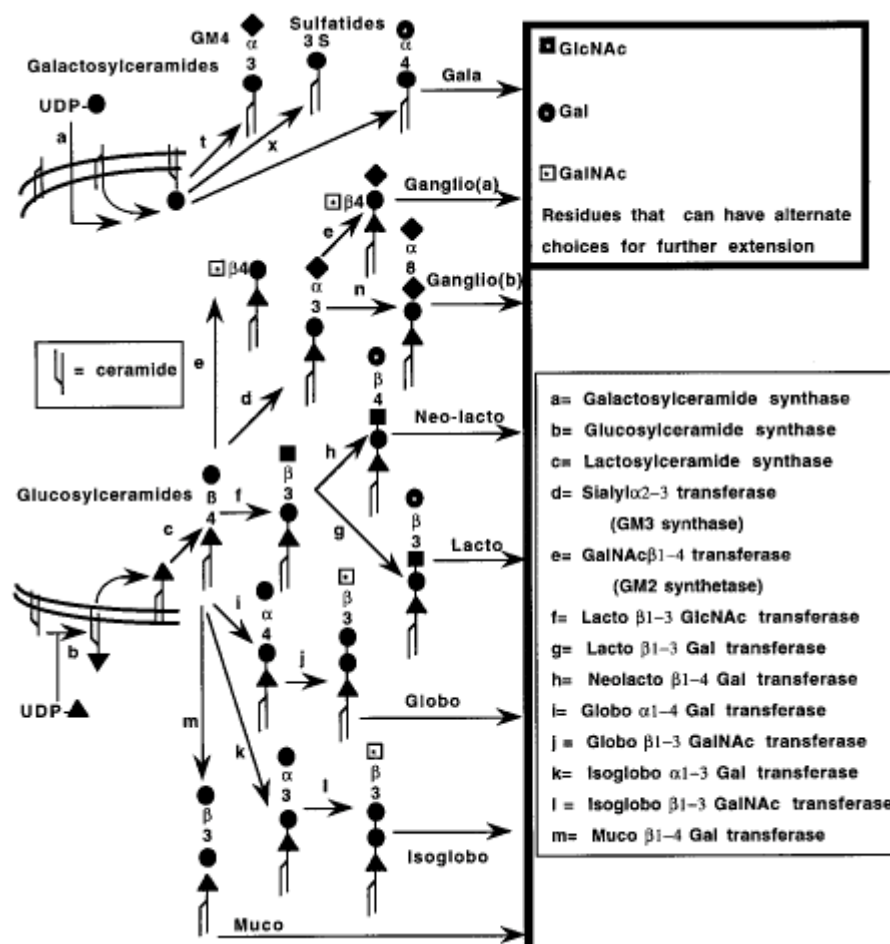
4. Diversity of glycosyltransferases

It has been estimated that there are over 200 different types of glycan linkage that are found in nature. As there may be more than one glycosyltransferase responsible for the same glycosidic linkage the total number of reported glycosyl transferases is over 200 (Narimatsu 2004). The specificity of the enzymes means that in contrast to the theoretical combination of several thousand glycan structures even in a pentasaccharide the total number of distinct glycan structures is much less

Table 1 Currently described glycosyltransferases involved in N- O- and glycolipid glycan structures. (CFG Mouse Phenotype Core)

Given the widespread occurrence and diversity of glycosyltransferase genes it is surprising that the first glycosyl transferase gene was not characterised until 1986 (the blood group B β 1,galactosylglycosyltransferase(Elices, Blake et al. 1986)). This was in part due to the difficulties in isolation of the gene products, the glycosyltransferases, which are generally only found at relatively low amounts within the cell. Many common motifs have been identified and the detailed molecular structure of most glycosyltransferases is known as is the mechanism of action of these

Figure 3 Glycolipid Biosynthesis (reproduced from *Essentials of Glycobiology* by Ajit Varki (Editor), Richard O. Cummins (Editor), Jeffrey Esko (Editor), Richard Cummings (Editor), Hudson Freeze (Editor), Gerald Hart (Editor), Jamey Marth (Editor) Cold Spring Harbor Laboratory Press,U.S. (31 Dec 1999)



5. Biosynthetic Pathways for Glycan Assembly

The biosynthetic pathways for N- and O-linked glycans of glycoproteins and O-linked glycans on glycolipids are well characterised and are shown in Figures 1 and 2. The stepwise addition of individual monosaccharides can clearly be seen indicating the interdependency of glycosyltransferases which is highlighted in the changes in glycosylation patterns due to absence of certain glycan structures shown when enzyme defects occur in cases of congenital diseases of glycosylation.

The pathway for biosynthesis of N-linked protein glycans is highly conserved and involves many steps which can be divided into several stages. Firstly a lipid (dolichol)-linked glycan is assembled through the action of mannosyl transferases and glycosyltransferases. This lipid-linked glycan is then transferred to selected asparagine residues in the protein back bone by the action of the oligosaccharyl transferase, enzyme .

There are then a series of glycosidases which trim down the initial precursor to form the 'core' region found in all N-linked glycans. These processes are closely linked to interaction with the protein folding chaperones Calnexin and Calreticulin and this provides an explanation for this seemingly complex process of biosynthesis (Helenius and Aebi 2004)

In the next stage the final glycan is assembled by sequentially adding monosaccharides by means of specific glycosyltransferases and it is the activity of these which determine whether the glycans are of the oligomannose, hybrid or complex type and also whether they are modified by monosaccharides such as sialic acids or fucose which are often found in terminal positions. Many of these glycosyl transferases will also act on O-linked glycans on proteins or glycolipids provided the appropriate acceptor glycans are present. For this reason the terminal sections of glycans may be common to all these glycan types produced in a particular cell.

6. Techniques for investigating glycosyltransferase gene expression

There are a number of different approaches that may be used to find if a glycosyltransferase gene is actually expressed in a particular cell type. This may lead to some confusion when considering the distribution of the transferase genes as some techniques may be more reliable than others. The investigations are more often carried out on material extracted from a certain tissue rather than a particular cell type although there are some studies on individual cells mostly when these have been grown in tissue culture

7. Techniques for investigating glycan structures

Techniques for glycan analysis until recently were complex and time consuming and therefore only a relatively few glycoconjugates had been completely characterised. However the requirements of such programs as that of the Functional Glycomics Consortium in the USA and large scale projects in Japan have lead to development of much more rapid screening techniques to show all the glycans structures present in a particular sample.

In these techniques after an initial analysis to determine the structures present it is frequently possible to profile the glycans more rapidly be it by mass spectrometry or lectin array techniques. This may be done at the level of individual proteins extracted or secreted by the cell type under study or by an analysis of all the glycan present in a particular tissue.

8. Tissue/cell Distribution of Glycosyl transferases

It is important to remember that this list of glycosyltransferase distribution shows the *capability* of any given cell-type or groups of cells with a tissue to perform a certain reaction in synthesis of a complex glycan. As already mentioned the synthesis of any given linkage will depend on a number of factors which may or may not be known.

It has been found that many transferases have a wide cell distribution which is not surprising given the common biosynthetic pathways especially for the N-glycan core.

The most varied expression (in species considered here) is seen in the fucosyl and sialyl transferases which are not present in the expression libraries of some cell types. The N-acetylgalactosaminyl transferases responsible for initiation of different O-glycan cores also may have very restricted cellular distribution . (Gerken, Tep et al. 2004). After these the next transferases showing limited distribution are the set of N-acetylglucosaminyl transferases involved in extending the N-glycan core and the branching of complex glycans.

Although the vast majority of the glycosyltransferases found in mouse cell types are also found in humans there are two notable exceptions : the α -galactosyl-N-acetylglucosaminyl transferase (Vanhove, Goret et al. 1997) and the N-glycolylneuraminic acid transferases (Odaka, Yuki et al. 1998) .

9. Tissue/cell Distribution of glycans

Given the complexities of glycan analysis it is not until relatively recently that anything like a widespread evaluation of glycan expressed in particular cell types has been available or even feasible. However there has been intense activity in recent years along with advancements in technology so for a growing number of cell types a representative profile of glycans is now available. However there is also a caveat to this. Generally the glycans are analysed as a pool from all or a substantial sub-set of the glycoproteins or glycolipids synthesised by that cell. This by no means indicates that the glycans found on any particular glycoconjugate from that cell will have all of these glycans present indeed it is unlikely that it will do so. Factors such as accessibility of enzymes to sites of different stereospecificity or other conditions such as the biosynthetic organelles through which they have passed will limit the variety found on any given glycoconjugate from a variety. These profiles are useful nonetheless for comparison with the distribution of the glycosyltransferases shown above.

10. Currently available Databases

The large collection of data and its diversity highlights the need for a comprehensive, constantly updated database to make this accessible to the community Existing databases (Table 2) such as (<http://www.cazy.org/>) are comprehensive but to increase their utility their integration into glycan and other databases need to be considered

Table 2 Current Databases of Glycosyltransferases

CAZY	http://www.cazy.org/
KEGG	http://www.genome.ad.jp/kegg/pathway/map/map01170.html
CFG GT database	http://www.functionalglycomics.org/static/gt/gtdb.shtml
BRENDA	http://www.brenda.uni-koeln.de/

11. Recommendations for a digital cell-by-cell catalogue of glycan structure and glycosyltransferases

The catalogue of glycan structures should be compiled according to the standard taxonomy generally used in other databases but it should be appreciated that much of data currently available (with the exception of recombinant glycoproteins or studies in cell culture) is performed on whole tissues or in the case of secreted material may have come from a variety of cell types. This is not a problem but the structures should be recorded as coming from a tissue rather than a cell type. It is also important that way in which the sample was processed for example if it is a 'total' glycan extract or if it was taken from a purified glycoconjugate

The catalogue of transferases should take into account the importance of recording the *expression* of the enzymes as highlighted above and not just the presence of the gene. For current data it is also true that it is generally available on an organ or tissue basis rather than by the cell. The database should record all the known reactions which have been characterised in terms of the nucleotide sugar donor, the position and anomericity of the linkage, the types of acceptor identified, the sub-cellular localisation, and any activators or inhibitors. Where kinetic data is available this should also be included showing the actual reaction studied.

12. Development of the EUROCarbDB GT database

A EUROCarbCB database has been set up to test the ability to input information regarding glycosyltransferase expression in cells/tissues and of the corresponding glycans that have been found.

This database was designed with the considerations mentioned above taken into account and data from some selected publications representative of glycosyltransferase distribution in the mouse as collected by the CFG Mouse phenotyping core facility (Comelli, Head et al. 2006) and also on transferase and antigen distribution for ABH, Lewis, P and Ii Blood group systems in Man (Ravn and Dabelsteen 2000). (Martins, de Oliveira Corvelo et al. 2006). These data sets were selected on the basis of different types of data that are likely to be encountered in such a database.

Data has been entered on the glycan donors, known glycan epitope acceptors, Cell or tissue distribution, typical glycans found in that tissue/cell, glycan antigens or epitopes found, and examples of glycoproteins or glycolipids carrying the gene products on their glycans where known.

The data entered into the prototype database was in the following categories

Term	Unique ID
Glycosyltransferase	GENEID
Species	MeSH (NCBI)
Tissue	MeSH (NCBI)
Cell	MeSH (NCBI)
Antigen	MeSH (NCBI)
Glycan epitope	LINCUS and MeSH (Carbohydrate Antigens)
Acceptor Glycan	LINCUS (Bohne-Lang, Lang et al. 2001)
Product Glycan	LINCUS (Bohne-Lang, Lang et al. 2001)
Donor Nucleotide	MeSH (Nucleotides) (NCBI)
Glycoprotein	UniProt
Glycolipid	LMSD LIPID MAPS structure database.
Publication Reference	PubMedID

The input of glycan structure through LINCUS at present is to allow the input of sub-structures.

This data will enable the tissue/cell distribution of transferases to be correlated with glycan structures, antigens, epitopes (e.g. of monoclonal antibodies), specificity of transferase, and glycoconjugates expressing the gene products in that tissue.

The utility of the GT transferase database will be assessed and the feasibility of automating data input and cross linking to other databases examined.

Refinement and optimisation of the database will be continued and will form a key part in the linking of the experimental data available from the project with the biological roles of glycans and to gene and proteomics databases.

It is expected that this database will play an important role in the following recommendation documented in the Frontiers in Glycomics white paper published by NIH

- **Recommendation 4: Support the development of open source software for automated analysis of analytical data and data mining in the Glycomics domain**

The focus group assigns a high priority to supporting open source software projects that will provide robust solutions for often-required functions in glycomics research. These include software for the automated interpretation of high-throughput analytical data (such as mass spectral data) and data mining tools that facilitate, for example, the discovery of correlations between glycan structure and function .

White Paper Report from Focus Groups at the NIH Workshop on Frontiers in Glycomics and Glycobiology (Combined Draft 4: May 9, 2007)

Data from over 200 publications was used in compilation of the database.

References

- Bohne-Lang, A., E. Lang, et al. (2001). "LINUCS: linear notation for unique description of carbohydrate sequences." *Carbohydr Res* **336**(1): 1-11.
- Comelli, E. M., S. R. Head, et al. (2006). "A focused microarray approach to functional glycomics: transcriptional regulation of the glycome." *Glycobiology* **16**(2): 117-31.
- Elices, M. J., D. A. Blake, et al. (1986). "Purification and characterization of a UDP-Gal:beta-D-Gal(1,4)-D-GlcNAc alpha(1,3)-galactosyltransferase from Ehrlich ascites tumor cells." *J Biol Chem* **261**(13): 6064-72.
- Gerken, T. A., C. Tep, et al. (2004). "Role of peptide sequence and neighboring residue glycosylation on the substrate specificity of the uridine 5'-diphosphate-alpha-N-acetylgalactosamine:polypeptide N-acetylgalactosaminyl transferases T1 and T2: kinetic modeling of the porcine and canine submaxillary gland mucin tandem repeats." *Biochemistry* **43**(30): 9888-900.
- Helenius, A. and M. Aebi (2004). "Roles of N-linked glycans in the endoplasmic reticulum." *Annu Rev Biochem* **73**: 1019-49.
- Kornfeld, R. and S. Kornfeld (1985). "Assembly of asparagine-linked oligosaccharides." *Annu Rev Biochem* **54**: 631-64.
- Martins, L. C., T. C. de Oliveira Corvelo, et al. (2006). "ABH and Lewis antigen distributions in blood, saliva and gastric mucosa and H pylori infection in gastric ulcer patients." *World J Gastroenterol* **12**(7): 1120-4.
- Narimatsu, H. (2004). "Construction of a human glycogene library and comprehensive functional analysis." *Glycoconj J* **21**(1-2): 17-24.
- Odaka, M., N. Yuki, et al. (1998). "N-glycolylneuraminic acid-containing GM1 is a new molecule for serum antibody in Guillain-Barre syndrome." *Ann Neurol* **43**(6): 829-34.
- Raman, R., M. Venkataraman, et al. (2006). "Advancing glycomics: implementation strategies at the consortium for functional glycomics." *Glycobiology* **16**(5): 82R-90R.
- Ravn, V. and E. Dabelsteen (2000). "Tissue distribution of histo-blood group antigens." *Apmis* **108**(1): 1-28.
- Reitman, M. L., A. Varki, et al. (1981). "Fibroblasts from patients with I-cell disease and pseudo-Hurler polydystrophy are deficient in uridine 5'-diphosphate-N-acetylglucosamine: glycoprotein N-acetylglucosaminylphosphotransferase activity." *J Clin Invest* **67**(5): 1574-9.
- Spiro, R. G. and M. J. Spiro (1982). "Studies on the synthesis and processing of the asparagine-linked carbohydrate units of glycoproteins." *Philos Trans R Soc Lond B Biol Sci* **300**(1099): 117-27.
- Vanhove, B., F. Goret, et al. (1997). "Porcine alpha1,3-galactosyltransferase: tissue-specific and regulated expression of splicing isoforms." *Biochim Biophys Acta* **1356**(1): 1-11.