Development and Analysis of Web Resources using Glycoinformatics

September 2016

Yukie Akune

Contents

Chapter 1

Introduction

Glycans are macromolecular substances that are known to be extremely vital in regards to recognition signals for biological phenomena such as cell-cell communication, biological development and cancer metastases (Fig. 1.1) [1, 2, 3]. For example, selectins, such as L-, E- and P-selectin, are molecules expressed on the cell surface of leukocytes. L-selectin and sulfated glycans play an important role for the infiltration of lymphocytes [4, 5, 6, 7]. First, lymphocytes, which through blood vessel with high speed, start rolling on vascular endothelial cell surface by the specific interaction of L-selectin of lymphocytes and glycans endothelial cell surface. Then, activated integrins of leukocytes strongly adhere with endothelial cells and promote the infiltration of leukocytes. Furthermore, studies about a tumor-associated antigen using glycan structures have been reported in the last decades. Hamid *et al.* have suggested that tri-sialilated N-glycans can be applied as a stage diagnostic marker of breast cancer [8].

Despite their importance, it has been difficult to study glycans and their synthesis. Because they are molecules that are synthesized by enzymes, in contrast to template-based

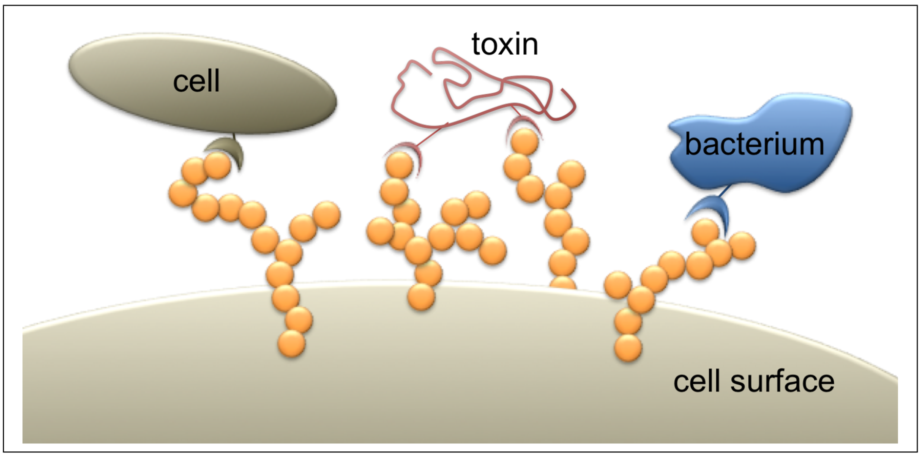


Figure 1.1: An example of glycan biological role. Glycans have complicated structures such as the variety of monosaccharides and multi-branched form. They are, in general, bound to proteins or lipids on the cell surface to act as a key role for the adhesion of other cells, toxins and bacteria.

synthesis of nucleic acids and proteins. Therefore, it is not possible to predict glycan structures directly from genetic information. Hence, the study of the biological functions of glycans is still at early stage.

Glycans are classified according to their core structure. The *N*-glycan class is on of the major classes that covalently bind to asparagine (Asn) residues found in the consensus-peptide sequence Asn-X-Ser/Thr (X can be any amino acid residue). Although many enzymes are involved in the biosynthesis of glycans, and the glycome of tissues and species is highly complex, the prediction of *N*- and *O*-glycan structures that can be formed is becoming feasible with our expanding knowledge of the glyco-machinery.

Glycan structures and related information that are experimentally described are stored in several databases such as Kyoto Encyclopedia of Genes and Genomes (KEGG) GLYCAN database [9, 10, 11], the Consortium for Functional Glycomics (CFG) [12], and GlycomeDB [13]. These databases use different formats to represent glycan structures which were developed for their respective purposes. However, there is no standard universal format for all databases, and so users are required to convert the format manually when they want to search or compare between databases. Another issue lies in the fact that the majority of the algorithms for glycoscience research has not been developed as tools, and thus biologists in practice are not able to apply them to their original data. Hence, we initially focused on the development of a web-based resource for glycan analysis named RINGS (Resource for informatics of glycomes as Soka) [14].

It is evident, based on the network size proposed by Krambeck and Betenbaugh and others, that existing glycan structure and experimental databases severely underrepresent the complexity of the glycome. For example, UniCarbKB [15, 16] is restricted to approximately 3,000 N-linked glycan structures compared to an estimated 10,000-20,000 generated by Krambeck and Betenbaugh [17, 18]. As such it is likely that we are missing critical information that could be of biological importance. Previously, we developed the Glycan Pathway Predictor (GPP) tool as part of RINGS, which takes a single N-glycan structure and displays all possible glycans that can be synthesized by a defined set of glycosyltransferases. It implements the mathematical model of N-glycosylation described by Krambeck *et al.* that characterizes substrate-specificity, and the resulting glycans are displayed as a pathway map - albeit the current list of glycosyltransferases is currently limited. Therefore, we focused on a systematic method to defined the properties and substrate specificity rules of mammalian glycosyltransferases involved in the biosyntheisis of *N*-linked glycans as a part of the UniCorn project. The proposed mode is designed to streamline: i) the automated construction of single glycosylation reactions using enzyme definitions; ii) the simulation of an entire *N*-glycosylation reaction network; and iii) mass spectrometry glycan analysis by providing a library of structures for data mining/matching.

The complexity of glycan structures

A glycan is constructed by linking monosaccharides with various glycosidic linkages. When a monosaccharide forms a stereoisomer by cyclization reaction, depending on the direction of an anomeric carbon and an anomeric reference atom, the anomer conformation of a monosaccharide is defined as alpha or beta (Fig. 1.2) [19, 20]. A disaccharide is formed by glycosidic linkage between an anomeric carbon with another monosaccharide. The monosaccharide, which anomeric carbon used for the glycosidic linkage, is defined as non-reducing end, and which has a free anomeric carbon is defined as reducing end (Fig. 1.3) [21]. Theoretically, it is possible that a glycosidic linkage can be at position C2, C3, C4, C6, and C8 of a reducing end monosaccharide. Therefore, a glycan structure is able to form a wide variety structure [22].

Glycans can be classified based on their structure patterns or connection manner with other molecules. *N*-glycans binds to the nitrogen atom (N) of an asparagine residue. Furthermore, *N*-glycans have a core structure that are constructed with three mannoses and two N-acetylglucosamines. *O*-glycans bind to the oxygen atom (O) of a serine or threonine residue. On the other hand, glycosphingolipids and GPI-anchor

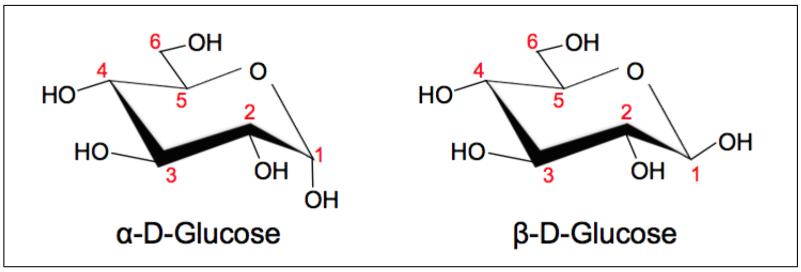


Figure 1.2: An example of anomer conformation. An anomer conformation of a monosaccharide is determined based on the configuration of the anomeric carbon and the anomeric reference atom. For example, a glucose, if the anomeric carbon (C1) and the anomeric reference atom (C5) are the same direction, the anomer conformation is defined as "α". If they are different, it is defined as "β".

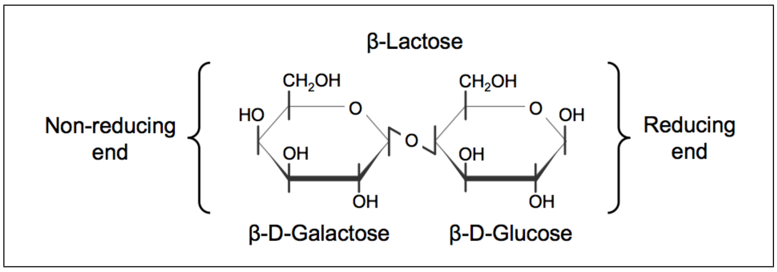


Figure 1.3: An example of a disaccharide. Anomeric information of the non-reducing end monosaccharide and a carbon number of the reducing end monosaccharide is used to represent a glycosidic linkage. In this example, a galactose and a glucose is linked "β1-4", and this disaccharide is named Lactose.

(glycosylphosphatidylinositol-anchor) are major classes that binds to lipids [20]. Figure 1.4 shows an example structures of each class. Each structure has an identifier named G-id which is composed of G and five numbers. These identifiers are given by KEGG so that users are able to see the detailed information referring to the structures from glycan-related databases.

N-glycans can be classified into three sub-groups based on the pattern of substructures that extend from the core structure (Fig.1.5). A high mannose type contains only mannose in the substructure extending from the core structure. If an N-acetylglucosamine residue is connected to the core structure, the glycan is grouped in the complex type. A hybrid typed structure has high mannose and complex type characteristics: one of the mannoses of the core structure, Man(α1-3), has an N-acetyl glucosamine extension, and the other core mannose, Man(α1-6), has mannose extension.

A definition of the tree representation in glycome informatics

A glycan structure can be treated as a "tree" in glycome informatics field. Figure 1.6 explains how a glycan is treated as a tree structure. A node represents a monosaccharide, an edge represents a glycosidic linkage and a link represents the combination of a node and an edge. A reducing end is named root and non-reducing ends are named leaves because a glycan structure looks like a tree structure. Moreover, a glycan substructure is represented as subtree. A reducing end node of disaccharide is described as parent, and non-reducing end node is described as child. When a parent has two carbohydrate, they

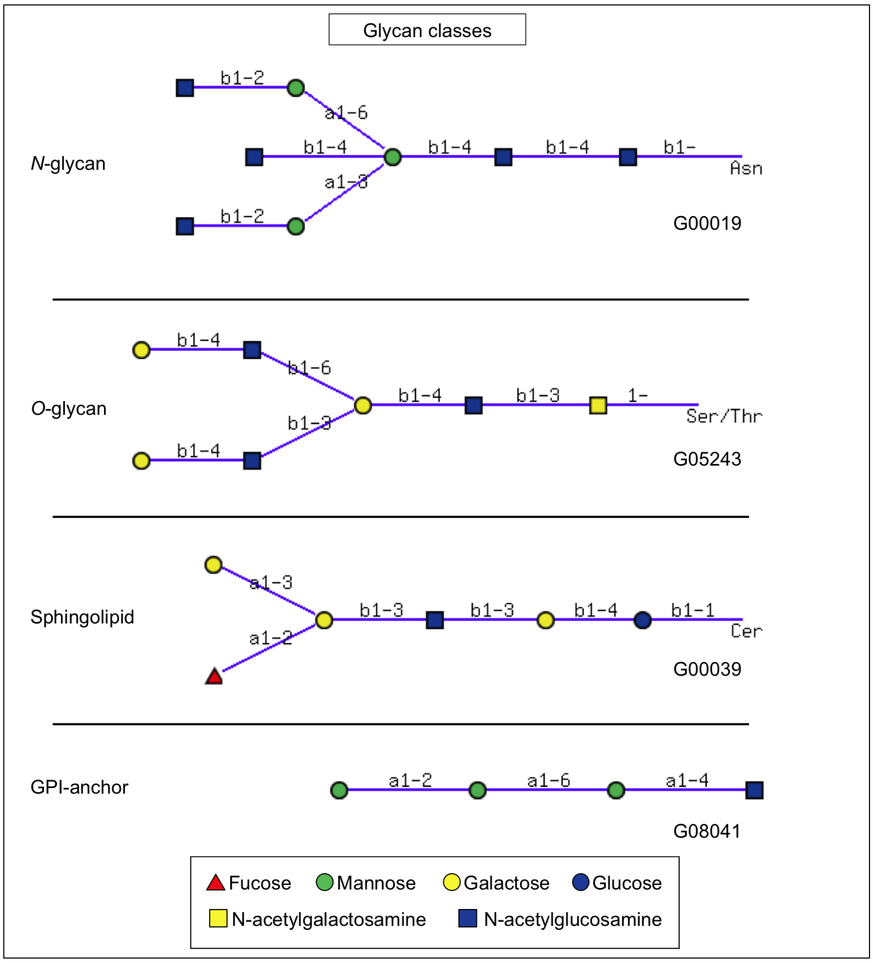


Figure 1.4: An example of representative glycan classes. Glycans are classified into different classes depending on their structure patters and binding to proteins or lipids. As shown in this in this example, a glycan connected to the asparagine residue of a consensus sequence "Asn-X-Ser/Thr (where X is any amino acid)", is defined as an N-linked glycan. If a glycan connected to a serine or threonine residue is defined as an O-linked glycan. Glycosphingolipid and GPI-anchor classes are the major that bind to lipids.

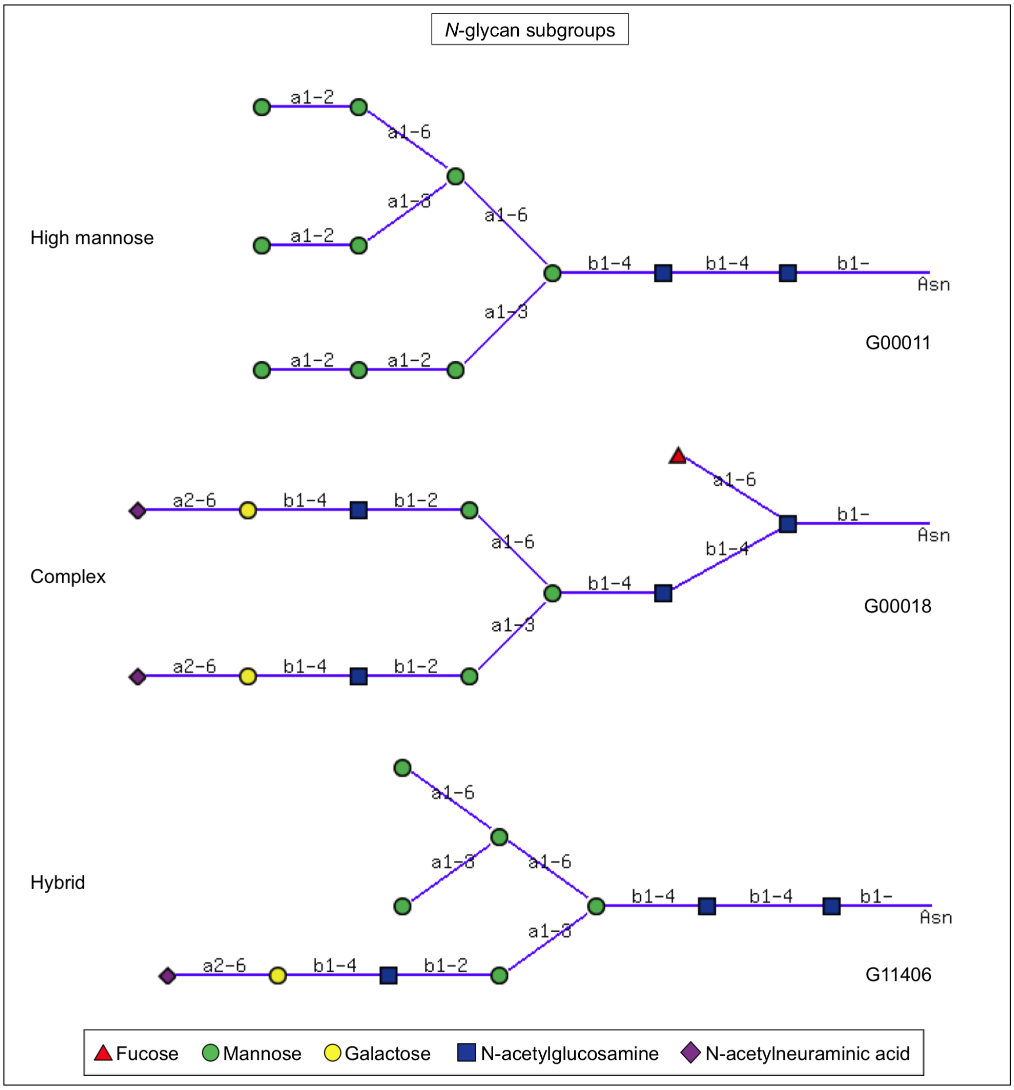


Figure 1.5: An example of *N*-glycan subclasses. N-glycans are classified based on the extension types from the core structure which consists of two N-acetylglucosamine and three mannose residues. When only mannose residues are extended, the structure is grouped as a high mannose type. When other residues, such as N-acetylglucosamine and galactose are extended, the structure is complex type. A glycan having both properties is classified as hybrid.

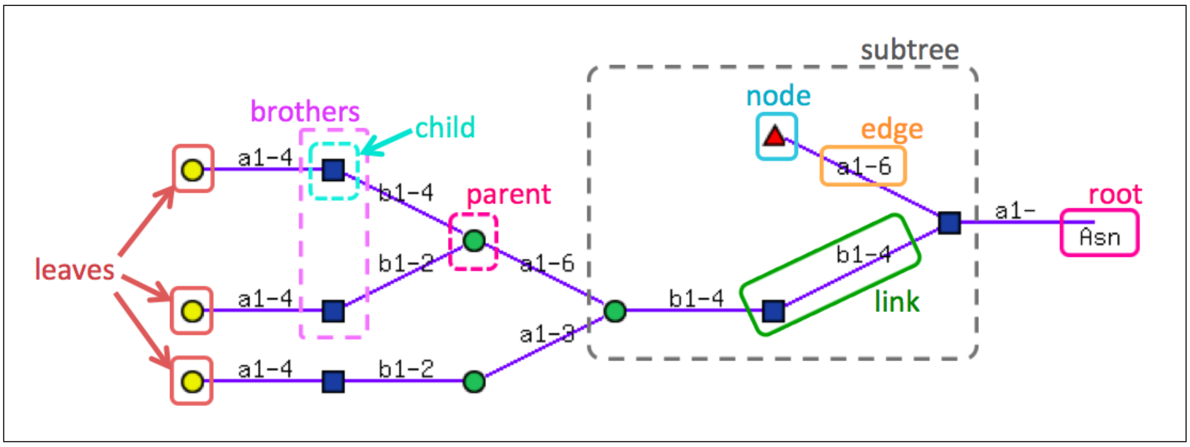


Figure 1.6: An example of lexicon used in glycome informatics. A glycan structure is represented as tree, and a substructure is described as subtree. A node stands for a monosaccharide and an edge stands for a glycosidic link. A link represents a pair of a node and an edge. A reducing end of disaccharide is defined as a parent and its non-reducing end is a child. When a parent has two or more nodes, they are defined as brothers.

are represented as brothers or children.

In the next chapter, we describe RINGS, which is the first web resource related to glycan structure analysis, especially focused on the development of tools and utilities followed by a description of the algorithm and the usage. We describe the development of theoretical glycan database in chapter 3. We also describe an overview of future work and concluding remarks.

Chapter 2

Development of RINGS: Resource for INformatics of Glycomes at Soka

2.1 Introduction

2.1.1 Glycome informatics

Storage of experimentally identified glycan structures and other experiment-related data, annotating the analysis results and development of glycan analysis algorithms and tools are promoted in the glycome informatics area. For example, Agravat *et al.* have developed a web resource named GlycoPattern [23]. This is used for helping users to analyze glycan array data which may promote understanding of the relationship between glycans and glycan binding proteins. Moreover, Li *et al.* have reviewed about the glycoinformatics tools for analysis of mass spectrometry data [24]. However, the analysis of not only glycan structures but also their functions may increase the difficulty of the development of glycoinformatics. For instance, the mechanisms of recognition of glycan binding proteins

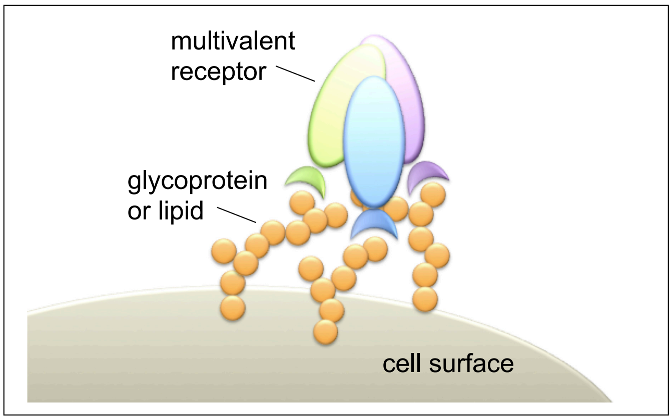


Figure 2.1: An example of recognition of multivalent glycan binding proteins. A single glycan is not able to strongly interact with glycan binding proteins (GBPs). However, glycans may take to cover it. Furthermore, some GPBs have multiple glycan recognition sites, such as Concanavalin A, which recognizes mannoses via four glycan-binding sites. Thus, it is possible to exert a strong interaction with a larger number of recognition sites of GPBs and glycans.

(GBPs) are depends on species and organ tissues specificity. Furthermore, GBPs may recognize multiple glycans at the same time (Fig.2.1). On the other hand, it is not easy to synthesize or predict glycans because they are synthesized by glycosyltransferases. Therefore, it is difficult to identify the expression of glycan structures and their functions in biological samples [25].

By combining identification of experimental data and glycoinformatics technology, the accuracy of predicting glycan structures and functions may be increased, just as BLAST [26] and FASTA [27] is essential in protein analysis to accurately identify and predict the function of a given amino acid sequence. When RINGS was developed, there was no web-based glycan analysis tool that biologists could use freely. One of the reasons for this was that there was no standard description format of glycan structures defined yet.

Therefore, glycobiologists were forced to understand the characteristics of each glycan related databases and their glycan structure formats when they search for particular glycans across different databases. Furthermore, since most of the algorithms for glycan analysis reported in the past decades had not been developed as a web tool, biologists could not apply them to their experimental data. Hence, as the first purpose of this study, we have developed a web resources named RINGS (Resource for informatics of glycomes at Soka), which incorporates glycan analysis tools and a database, for breakthrough research in the glycoinformatics field. We are developing glycan analysis tools based on reported algorithms and continually update various glycan related information, which are freely available via web.

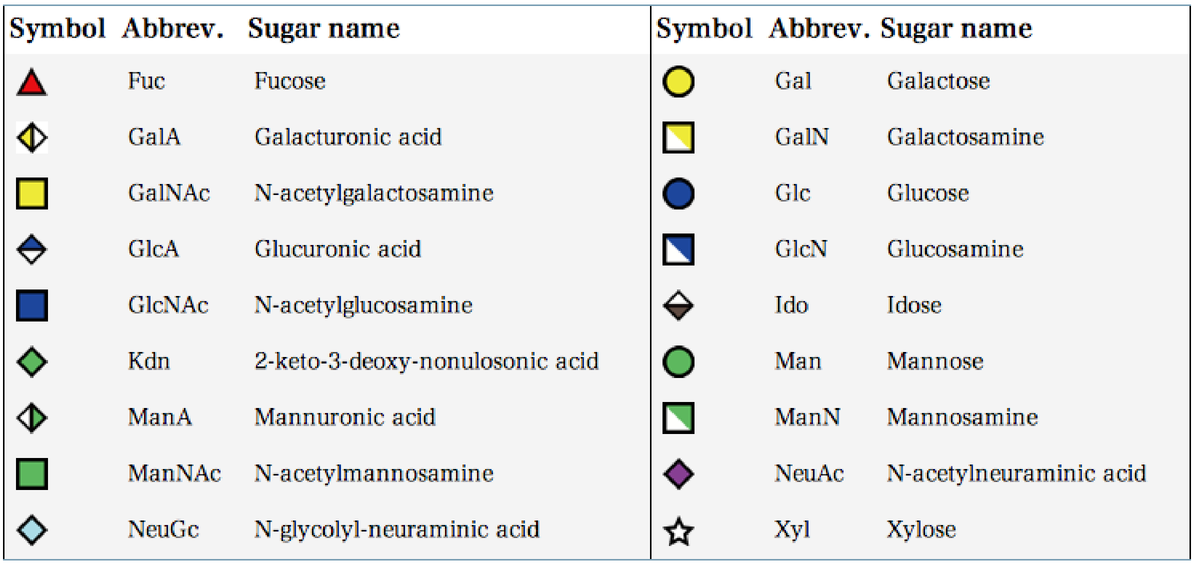
2.1.2 Glycan related Databases associated with RINGS

One of the roles for glycoinformatics is the management of glycan related data. A variety of glycan related databases have been developed over the years. Not only glycan structure information, but annotation information, including experimental protocols have been collected in the individual databases from unique perspectives (Table.2.1). For example, KEGG (Kyoto Encyclopedia of Genes and Genomes) has been developed by the Institute for Chemical Research of Kyoto University since 1995 [9, 11]. KEGG stores a wide variety of data related to molecular interaction networks of systems biology such as the cell, the organism and the ecosystem from molecular level information based on large-scale molecular datasets obtained by genome sequencing and other experimental technologies. CFG (Consortium for Functional Glycomics) developed a function glycomics gateway as a resource for experimental data for providing glycan-protein interactions [12]. Users can search the data based on species, experimental conditions, cell lines and other experimetal backgrounds. GlycomeDB has collected structural information from several glycan-related databases, and stored glycan structures in their own format allowing users to cross search glycan related information between different databases at once [13].

Table 2.1: A list of glycan related databases associated with RINGS.

Name	Major features	Year released
KEGG	a wide variety of information about pathways, genes and chemical compounds	1995
CFG	experimental data related to glycan-protein interaction	2001
GlycomeDB	glycan structure data integrated from several databases	2008

Each individual database developed their own format to represent a glycan structure that is suitable for their objective. KEGG Chemical Function (KCF) is a standard format used in KEGG [28]. It gives X and Y coordinate data on monosaccharides so that developers and users are able to treat a glycan structure as a graph on a canvas. Another format named GlycoCT is specialized for glycan structures [29]. This format is mainly used in GlycomeDB, UniCarbKB and UniCarb-DB. GlycoCT is able to describe more chemical information of monosaccharides compared to KCF. When drawing a glycan structure, CFG symbols are often used for representing monosaccharides so that users are able to understand a structure more intuitively. Figure 2.2 lists symbols especially found in mammalian glycan structures. Glycans are also represented as linear strings, and major formats include LinearCode ® [30] and IUPAC nomenclature for monosaccharides [31].



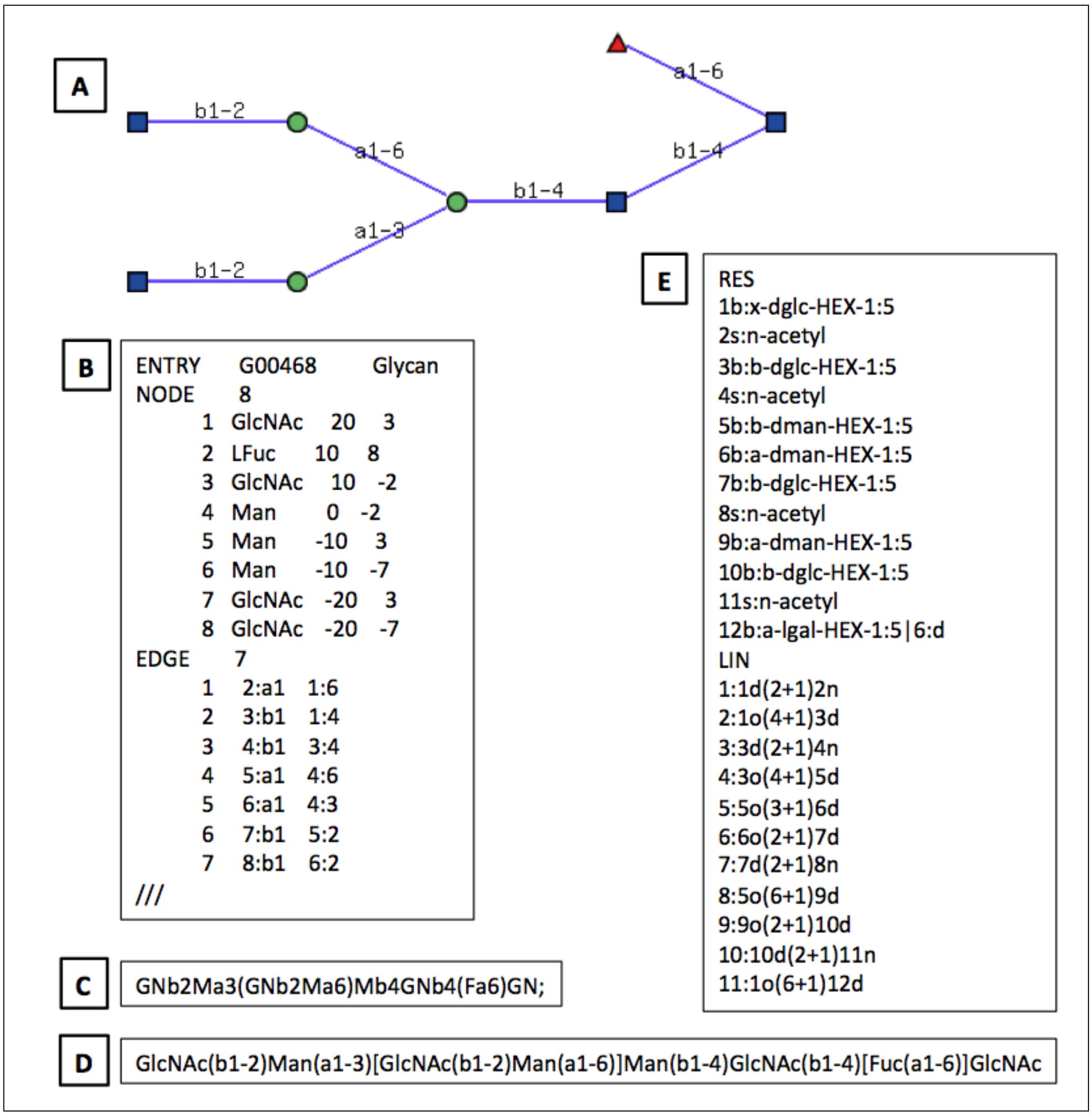
Symbol	Abbrev.	Sugar name	Symbol	Abbrev.	Sugar name
🔺	Fuc	Fucose	🟡	Gal	Galactose
◈	GalA	Galacturonic acid	◪	GalN	Galactosamine
🟨	GalNAc	N-acetylgalactosamine	🔵	Glc	Glucose
⬘	GlcA	Glucuronic acid	◩	GlcN	Glucosamine
🟦	GlcNAc	N-acetylglucosamine	⬘	Ido	Idose
◆	Kdn	2-keto-3-deoxy-nonulosonic acid	🟢	Man	Mannose
◈	ManA	Mannuronic acid	◪	ManN	Mannosamine
🟩	ManNAc	N-acetylmannosamine	◆	NeuAc	N-acetylneuraminic acid
◇	NeuGc	N-glycolyl-neuraminic acid	☆	Xyl	Xylose

Figure 2.2: A list of CFG symbols. These monosaccharides are especially seemed in mammalian.

A monosaccharide is represented with one or two alphabets in the LinearCode ® format, so that a glycan can be written in the shortest manner. The IUPAC foramt represents a glycan structure without abbreviation, so that users can understand a structure easily. Figure 2.3 shows an example of a glycan structure represented in each format respectively.

2.1.3 Algorithms used in RINGS

Various algorithms for glycan structure analysis have been reported in the last decades. Aoki *et al.* have reported the glycan alignment algorithm, named KCaM (KEGG Carbohydrate Matcher) [28]. This algorithm is used for calculating the similarity between two glycan structures, and implemented in query searching systems of several databases including KEGG and RINGS. In RINGS, this algorithm is installed in database searching system in DrawRINGS and similarity calculations of the algorithm for Glycan Score



```
ENTRY     G00468        Glycan
NODE      8
          1  GlcNAc   20    3
          2  LFuc     10    8
          3  GlcNAc   10   -2
          4  Man       0   -2
          5  Man     -10    3
          6  Man     -10   -7
          7  GlcNAc  -20    3
          8  GlcNAc  -20   -7
EDGE      7
          1   2:a1   1:6
          2   3:b1   1:4
          3   4:b1   3:4
          4   5:a1   4:6
          5   6:a1   4:3
          6   7:b1   5:2
          7   8:b1   6:2
///
```

```
GNb2Ma3(GNb2Ma6)Mb4GNb4(Fa6)GN;
```

```
GlcNAc(b1-2)Man(a1-3)[GlcNAc(b1-2)Man(a1-6)]Man(b1-4)GlcNAc(b1-4)[Fuc(a1-6)]GlcNAc
```

```
RES
1b:x-dglc-HEX-1:5
2s:n-acetyl
3b:b-dglc-HEX-1:5
4s:n-acetyl
5b:b-dman-HEX-1:5
6b:a-dman-HEX-1:5
7b:b-dglc-HEX-1:5
8s:n-acetyl
9b:a-dman-HEX-1:5
10b:b-dglc-HEX-1:5
11s:n-acetyl
12b:a-lgal-HEX-1:5|6:d
LIN
1:1d(2+1)2n
2:1o(4+1)3d
3:3d(2+1)4n
4:3o(4+1)5d
5:5o(3+1)6d
6:6o(2+1)7d
7:7d(2+1)8n
8:5o(6+1)9d
9:9o(2+1)10d
10:10d(2+1)11n
11:1o(6+1)12d
```

Figure 2.3: An example of glycan structure (A) represented in different formats; (B) KCF, (C) LinearCode, (D) IUPAC and (E) GlycoCT{condensed}.

Matrix tool.

Another algorithm used in RINGS is used for calculating glycan score matrices [32]. Amino acid scoring matrices, such as BLOSUM (BLOcks SUbstitution Matrix) [33], reflects the physicochemical similarity between a pair of amino acids. Therefore, high accuracy alignments of the amino acid sequences are possible by using the appropriate amino acid scoring matrix. Although glycans do not necessarily have evolutionary relationships, their recognition processes may be due to the similarity of monosaccharides. Therefore, Kinoshita *et al.* have reported that glycan score matrices are possible to improve the accuracy of glycan alignments [32]. Hence, glycan score matrices may reflect physicochemical properties between a pair of monosaccharides and the glycosidic bonds.

Glycan structures are synthesized by carbohydrate precursors and the activities of glycosyltransferases and glycosidases. Krambeck *et al.* have reported an algorithm for simulating glycan synthesis in 2005 [17]. Two kinds of glycosidases and nine glycosyltransferases are used in this model. They then extended the model to use two types of glycosidases and seventeen types of glycosyltransferases in 2009 [18]. These enzymes used in these algorithms are the basic enzymes involved in the synthesis of N-glycans. Krambeck *et al.* constructed the algorithm based these enzymatic reactions. In the 2005 model, they were able to calculate 7,500 glycans as a result of simulation. Moreover, in the 2009 model, they simulated 20,000 types of N-glycans even though they limited the mass $< 4,000$. We have developed Glycan Pathway Predictor (GPP) in RINGS based on these models.

Algorithms for extracting featured substructure(s) by comparing two types of glycan datasets have been developed since 2005. Hizukuri *et al.* [34] and Kuboyama *et al.* [35]

have reported the learning model to classify large amounts of data based on Support Vector Machines (SVMs) [36]. In this algorithm, input glycans are first decomposed into substructures, which contains the distance information from the reducing end. Hizukuri *et al.* have focused on tri-saccharides substructures, while Kuboyama *et al.* have decomposed glycans into substructures ranging from one to nine monosaccharides. The substructures are, then, vectorized for kernel classification to extract the most featured substructure. Hao *et al.* [37] have developed a new model by introducing similarities of monosaccharides and glycosidic bonds based on SIMCOMP [38]. The Glycan Kernel Tool in RINGS was developed based on this newest model.

Hosoda *et al.* have developed a novel algorithm for multiple glycan alignment named MCAW (Multiple Carbohydrate Alignment with Weight) [39]. This algorithm is based on BLAST [26], which is used for multiple alignment of protein sequences. Users are able to calculate configuration properties of a glycan data set. The MCAW calculation is promoted based on the dynamic programming algorithm. This novel model is used for calculating "blocks" in the Glycan Score Matrix algorithm.

2.1.4 RINGS architecture

RINGS is a resource that provides glycan related data as well as tools that have been developed for glycan analysis and data mining. Glycan structure information that is stored in RINGS have been extracted mainly from KEGG GLYCAN and GlycomeDB. In RINGS, glycans are represented using CFG symbols (Fig.2.2) so that users are able to understand the structure intuitively.

RINGS was developed using Perl, Java, Ruby and HTML language for coding tools

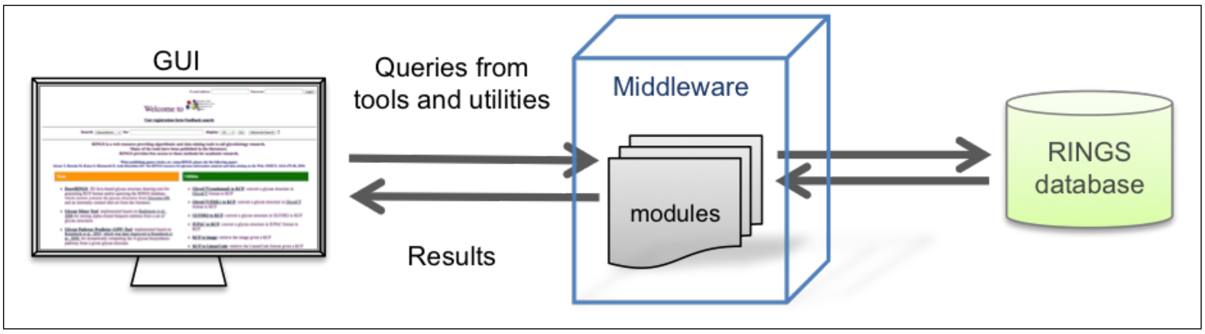


Figure 2.4: An example of RINGS architecture. Queries from RINGS tool and utilities are computed via the middleware, and input and results are displayed on the GUI (Graphical User Interface).

and utilities, and MySQL was used for constructing database. Figure 2.4 is an example of the framework of RINGS, which is composed of GUI, middleware and RINGS database. Users are able to run queries and tools, and view the results via GUI (Graphical User Interface), which are basically coded by HTML and Java. Middleware contains software modules which are groups of functions written in Perl. Modules are used to enhance the efficient operation of databases.

2.2 Materials and Methods

To develop RINGS as a web-resource, we mainly used R and Perl languages for coding Glycan Score Matrix, Glycan Kernel Tool, IUPACtoKCF and GlycoCT{condensed}toKCF programs. Moreover, we used HTML language for developing each web page, and MySQL language for modifying the RINGS database. Furthermore, we used glycan data stored in RINGS database for glycan structure analysis.

2.2.1 Glycan Score Matrix development

DrawRINGS allows users to draw a glycan structure on the canvas by mouse operation or using KCF format. The drawn glycan is able to be used as a query for searching similar structures from RINGS database and/or GlycomeDB. Furthermore, users are able to choose the scoring methods from “Matched” or “Similarity”. These scoring methods have been developed based on the KCaM algorithm. The matched scoring method, which developed for as a local alignment algorithm, calculates a score by adding the score of matched nodes and edges (Fig. 2.5(a)). Meanwhile, a score calculated by the similarity method represents a percentage of the similarity between a pair of glycan structures (Fig. 2.5(b)). This calculation was developed for as a global alignment algorithm. These scoring algorithms give constant values when a pair of monosaccharides or linkages are completely matched. For example, it gives 70 for matched monosaccharides, ten for each matched linkage information, and 0 for not matched.

During the calculation of a glycan score matrix, a clustering tree is used as a guide tree for glycan multiple alignment. Distance scores are required to generating a clustering tree. A distance score is given by subtracting an alignment score from the maximum score. We computed clusters based on the Fitch-Margoliash algorithm [40]. First, we generated a distance table composed of glycans and their distances. Second, we searched for the pair of glycans that has the nearest (smallest) distance value. Third, we calculated the average distance from glycans contained in the new cluster to all of the other glycans and clusters. Fourth, we computed the branch distances using the equation 2.1, 2.2 and 2.3. When a new cluster contains two glycans, such as A and B, equation 2.1 was calculated for the distance ($D(a)$) (Fig.2.6(i)). When, a cluster contains two or more glycans, equation 2.2

and 2.3 are used for calculating each branch distance (Fig.2.6(ii)) Based on these model, we calculated a distance table for input data set.

$$D(a) = (D(a+b) - D(b+x) + D(a+x))/2 \tag{2.1}$$

$$D(a) = (D(bc+a+c) + D(a+x) - D(c+x) - D(b))/2 \tag{2.2}$$

$$D(bc) = (D(a+bc+c) + D(c+x) - D(a+x) - D(b))/2 \tag{2.3}$$

KCaM algorithms are able to calculate the concordance percentage between a pair of glycans, although it is difficult to measure the similarity with taking biological background into the consideration. To improve the accuracy of the glycan searching algorithm, we have added a concept of "blocks" into the scoring system called Glycan Score Matrix [32], and then we introduced the new Glycan Score Matrix into DrawRINGS. Blocks are one of the calculation processes of BLOSUM, which are non-gapped amino acid sequences of the result of multiple alignment (Fig. 2.7). Moreover, blocks are considered to be conserved sequences of the query, such as species and protein families.

We have used MCAW for glycan multiple alignment to calculate the blocks of glycan structures. First, in this study, a hierarchical clustering of the input glycan structures are dynamically calculated. We have developed a code for determining the order of the multiple alignments by using this clustering as a guide tree. Furthermore, in order to obtain the largest blocks, we calculated blocks after dividing glycans based on each cluster. Figure 2.8 is an example of the calculation. A clustering tree is divided based on the threshold, which is decided by a parameter. Then, the alignment is progressed from the closest pairs. First, a glycan pair, G12723 and 12722, is aligned. They are very similar structures, so that both structures are treated as a block (orange line). In the other

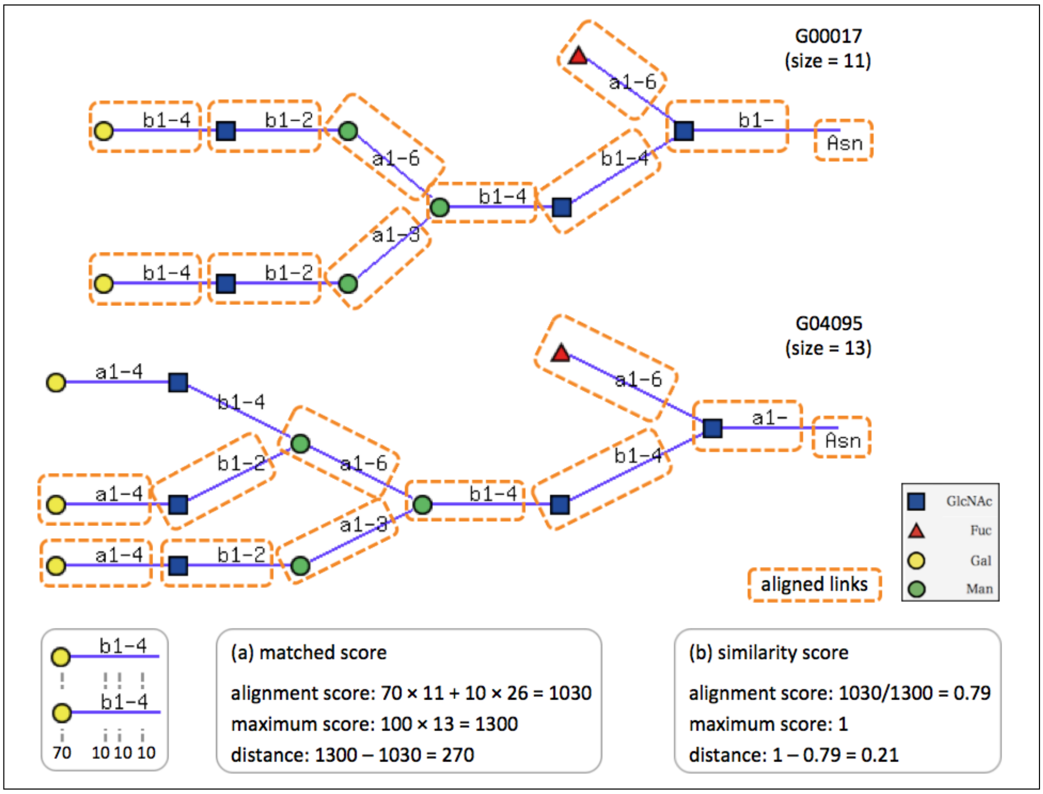


Figure 2.5: An example of scoring for matched and similarity. The substructures framed with dash lines are aligned links. When 70 points are given to aligned monosacchrides, and ten points are given to each linkage information, the matched alignment score is 1030 and similarity is 0.79. The distances between two glycans are required to calculate a clustering tree. The distance is calculated by subtracting an alignment score from the maximum score.

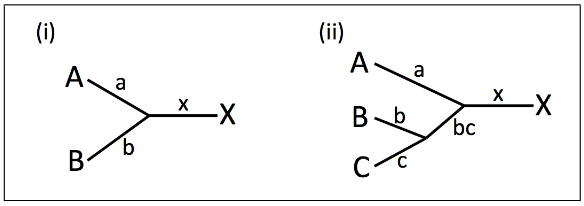
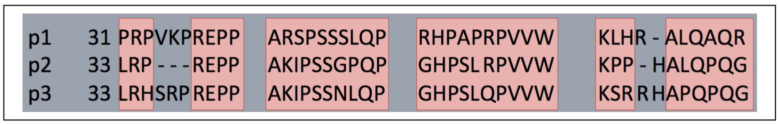


Figure 2.6: An example of the branch distance calculation. (i) When a new cluster contains two glycans (A and B), distance (D) a is calculated by equation 2.1. (ii) If a new cluster contains a glycan and a group of two glycans (B, C and A), equation 2.2 is used for calculating $D(a)$. Equation 2.3 is used for calculating $D(bc)$. In this graph, X refers to any glycan or cluster.



```
p1   31 PRPVKPREPP  ARSPSSSLQP  RHPAPRPVVW   KLHR - ALQAQR
p2   33 LRP - - -REPP  AKIPSSGPQP  GHPSL RPVVW   KPP - HALQPQG
p3   33 LRHSRP REPP  AKIPSSNLQP  GHPSLQPVVW   KSR R HAPQPQG
```

Figure 2.7: An example of blocks of amino acid sequences. Blocks are the results of multiple alignments which are non-gaped amino acid sequences. They are considered to be conserved sequences.

hand, another cluster, G00374 and G00272 are first aligned, and then G00232 is aligned. Blocks do not contain gaps, therefore, blue dash lined substructures are defined as the block of this cluster. Then, pairs of links, monosaccharide and its glycosidic linkage, were extracted from blocks, and glycan score matrices are calculated based on the frequency of the links. Unlike proteins, it has not been confirmed genetic conservation in glycan structures. However, we suggest that glycan blocks may considered as domain or motif structures. We have introduced the new glycan score matrix that contains this blocks concept. $F_{i,j}$ represents the frequency of aligned link i and j, and F_{total} represents the total number of all alignments. Then, the probability of alignment $G_{i,j}$ is obtained by equation 2.4. The probability of alignment of a link i (p_i) is calculated by equation 2.5. The expected probability of alignment of link i and j ($E_{i,j}$) is obtained based on equation 2.6. We are, then, able to calculate the score of link i and j by equation 2.7.

$$G_{i,j} = F_{i,j} \,/\, F_{total} \tag{2.4}$$

$$p_i = G_{i,i} + \sum_{i \neq j} G_{i,j}/2 \tag{2.5}$$

$$E_{i,j} = \begin{cases} p_i p_j & for\, i = j \\ 2p_i p_j & for\, i \neq j \end{cases} \tag{2.6}$$

$$S_{i,j} = log_2\,(G_{i,j} \,/\, E_{i,j}) \tag{2.7}$$

2.2.2 Glycan Kernel Tool

We have developed Glycan Kernel Tool based on Jiang [37] model. Input glycan data sets, such as target (X) and control (Y) groups, are decomposed into substructures. When

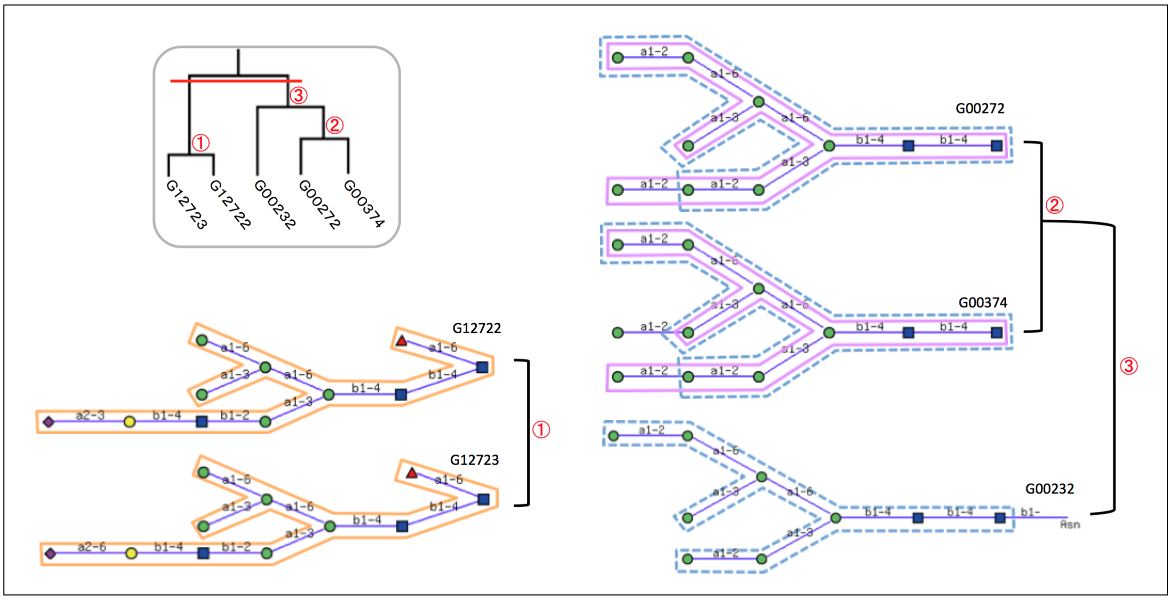


Figure 2.8: An example of glycan blocks calculation. A clustering tree is used as a guide tree for multiple alignments. To obtain larger blocks, the The clustering tree is divided into smaller class to extract the largest block(s). The threshold to divide the clusters is decided by the parameter. In this example, the closest pair, G12723 and 12722, are aligned (1). This alignment does not have any gap, so that whole structures are treated as a block (orage lined). G00272 and G00374 are aligned based on another cluster (2), and then G00232 is aligned (3). As the result of this multiple alignment, blue dash lined substructure is defined as another block.

N kinds of substructures are generated in total, the decomposed glycans are transformed into vectors based on the occurrences of each substructures (X and Y). Then, the vectors are positioned into a certain space. Generically, it is not able to classify the target (black) and control (plane) in this space. Therefore, these positions are transferred by kernel computing for extracting the featured structure(s) (Fig.2.9).

This calculation is coded in Matlab and Perl which is specialized for mathamatical calculation. However, because of the large amount of memory and calculation time, we have divided these process into four steps. Figure 2.10 shows an example of the process of Glycan Kernel Tool. First, a calculation id is generated when a user run a query to save the input, parameter and result data into RINGS database. Then, glycan kernel calculation promoted, and its progress situation is saved as well. Therefore, the user is able to check the calculation progress and the result using calculation id.

2.2.3 RINGS Utilities

I have developed two utilities, IUPACtoKCF and GlycoCT{condensed}toKCF, using Perl and HTML languages in this study. Both utilities uses RINGS modules individually. Figure 2.11 shows the workflow of each utilities. When users input glycan structure which form is IUPAC or GlycoCT{condensed}, the query is passed into each modules. IUPAC and GlycoCT{condensed} are totally different formats, however, the conversion process into KCF is similar. The input glycan is decomposed into monosaccharides and glycosidic bonds, at the same time the relationship between each monosaccharide, such as which monosaccharide is connected to the other, are saved. Then, monosaccharides are converted into Node section of KCF format, and glycosidic bond are Edge section. The relationships between each monosaccharide are referred to the X and Y coordinates

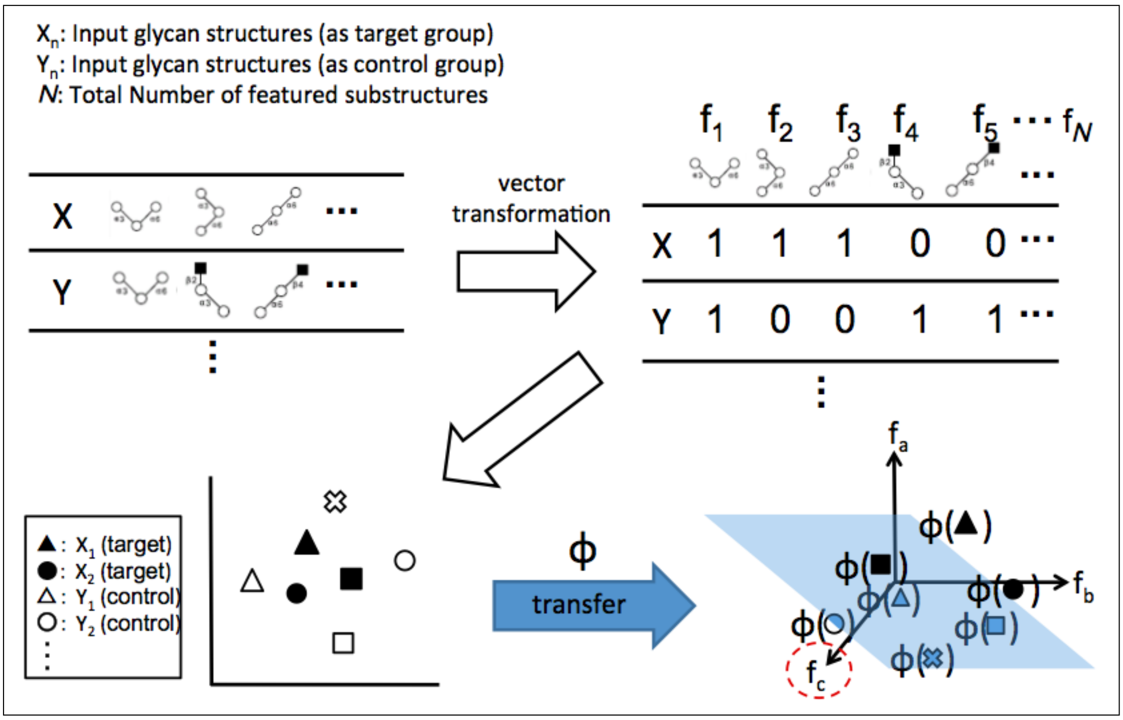


Figure 2.9: An example of kernel classification. When users input two types of glycan data, such as target (X) and control (Y). They are decomposed into substructures (X_n,Y_n), and the total number of substructures is N. The substructures are transformed into vectors by the occurrence of each structure. The positioned vectors into a space is not able to classify into target (black) and control (plane). Hence, the space is further transferred by kernel computing for extracting the featured structure(s)

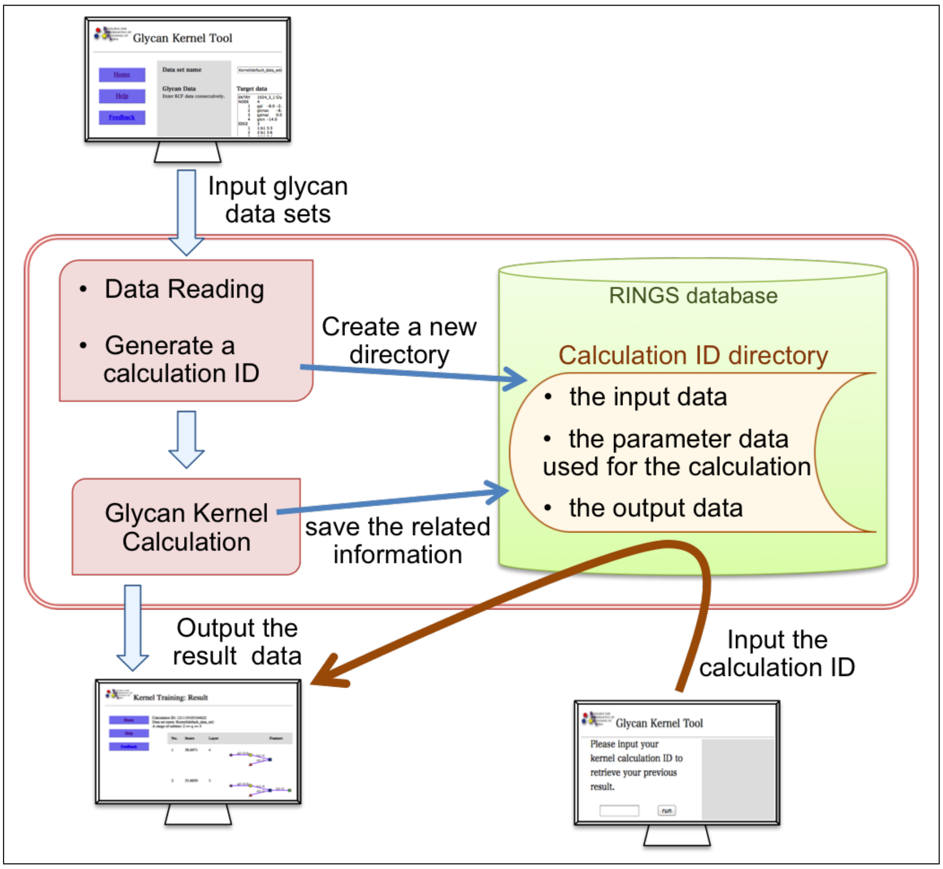


Figure 2.10: An example of the process of Glycan Kernel Tool. Input glycan structures and parameters are saved in RINGS database with unique calculation ID. Users are able to see the progress and the result data by using the calculation ID.

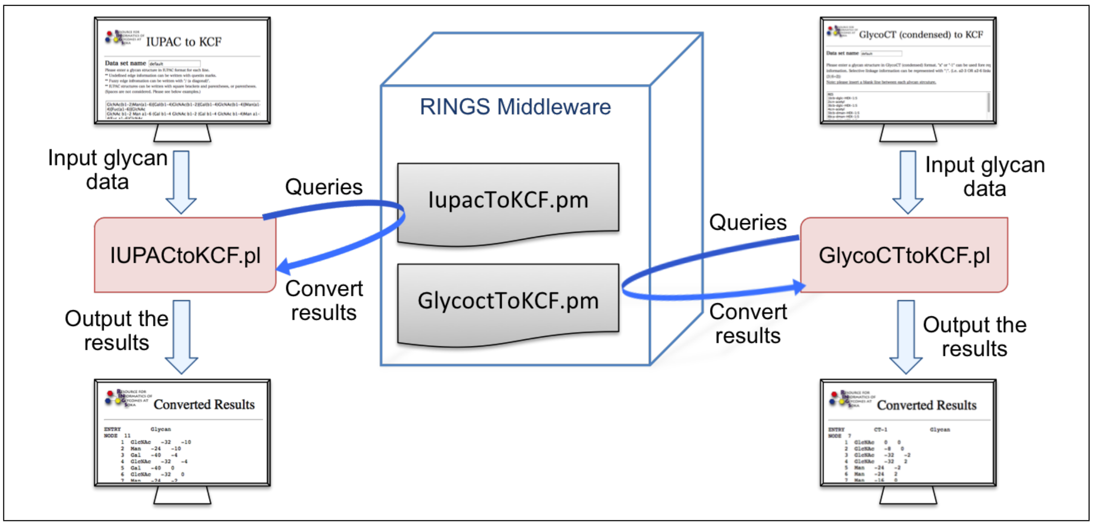


Figure 2.11: An example of the progress of utilities. When a user input a glycan structure in IUPAC or GlycoCT{condensed}, it is passed in RINGS module (IupacToKCF.pm or GycoctToKCF.pm). Then, the input structure is converted into KCF format, and returned as the output data.

of Node section and the order of the index of Edge section. Finally, the converted KCF format is viewed on the screen via IUPACtoKCF and GlycoCT{condensed}toKCF perl program respectively.

2.3 Results

2.3.1 Glycan Score Matrices

We were able to construct the calculation algorithm of distance matrices and phylogenetic analysis. Figure 2.12 shows the result of clustering using test data based on Fitch-Margoliash method. We were successfully able to classify the test data into five classes based on each feature. High-mannose group were divided by the number of mannoses,

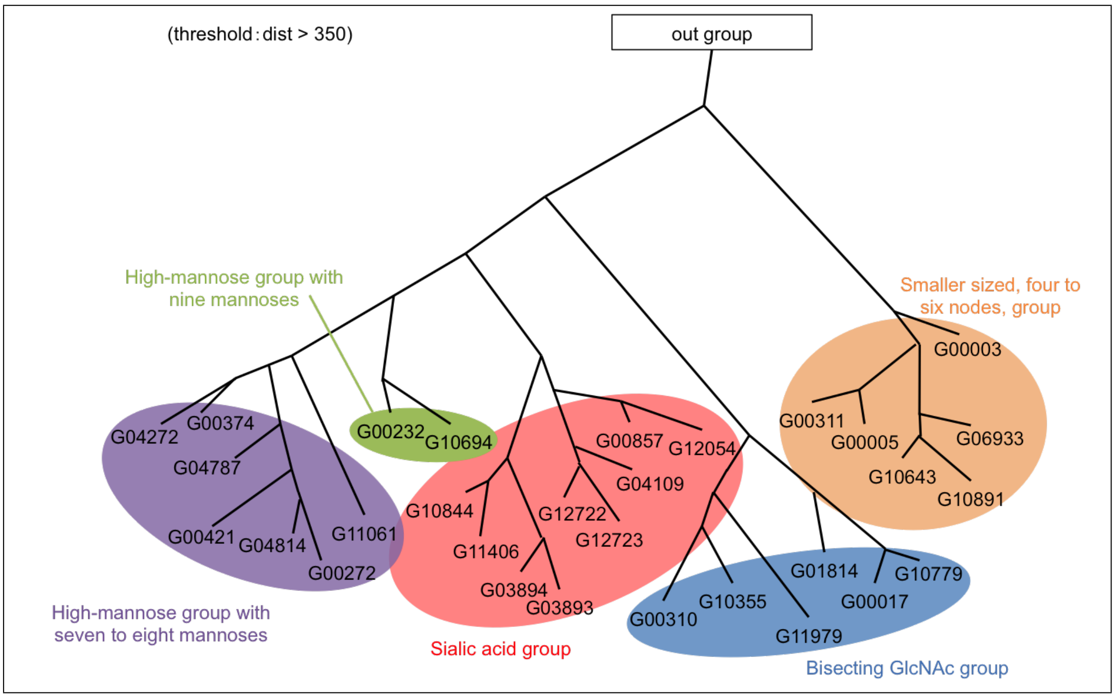
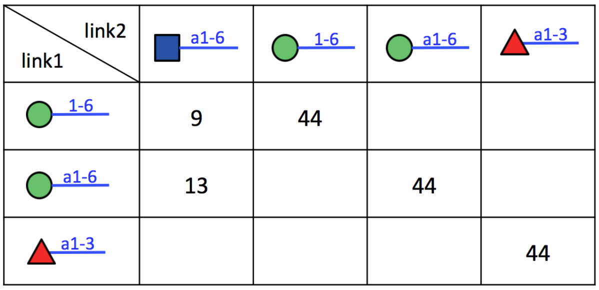


Figure 2.12: An example to glycan clustering based on Fitch-Margoliash method. Each cluster contains different features. The high-mannose groups, which contains seven or eight mannoses, and the nine mannoses group are the closest classes. Glycans, which have sialic acid(s) at the non-reducing end, and which have a bisecting GlcNAc, are separated individually. The furthest group contains small structures which nodes are from four to six.

however position were closest. Glycans with sialic acid(s), and a bisecting GlcNAc were clearly divided. Moreover, glycans which contains from four to six monosaccharides, were classified as the furthest group.

We have calculated a glycan score matrix using *N*-glycans that stored in RINGS. Figure 2.13 is a part of the score matrix. The highest score was 44 point were given to pairs of identical link paired. The pairs of GlcNAc(a1-6) - Man(1-6) and a pair GlcNAc(a1-6) - Man(a1-6) were given lower scores because the pairs were aligned lower times than



link2 / link1	🟦 a1-6	🟢 1-6	🟢 a1-6	🔺 a1-3
🟢 1-6	9	44		
🟢 a1-6	13		44	
🔺 a1-3				44

Figure 2.13: An example of glycan score matrix. This is a part of the result of N-glycan score matrix. The highest score was 44, which the same link was paired. The scores of a pair GlcNAc(a1-6) and Man(1-6), and a pair GlcNAc(a1-6) and Man(a1-6) were lower than others. It is because that the pairs were aligned but the frequency was less than the others. The blanks means that the pair was not aligned.

the others. Blanks represents that the pairs were not aligned within the blocks. We calculated glycan score matrices using N- and O-glycans and Sphingolipids, and saved them into RINGS database. We also programmed these matrices into DrawRINGS, so that the score matrix is able to be reflected to the query search (Fig. 2.14).

2.3.2 Glycan Kernel Tool

We have developed Glycan Kernel Tool based Jian *et al.* algorithm [37]. Figure 2.15 is a snapshot of the tool. Users need to input a data set name to save the input data and the calculation results into RINGS database. Furthermore, both control and target glycan data sets are required. The glycan structures can be input or load a file in KCF format. As a parameter, users are able to input the number of monosaccharides in the substructures to extract. The range of substructures should be at least one and at most nine. Once the program starts, a calculation ID is given to the user (Fig 2.16 (a)). Users

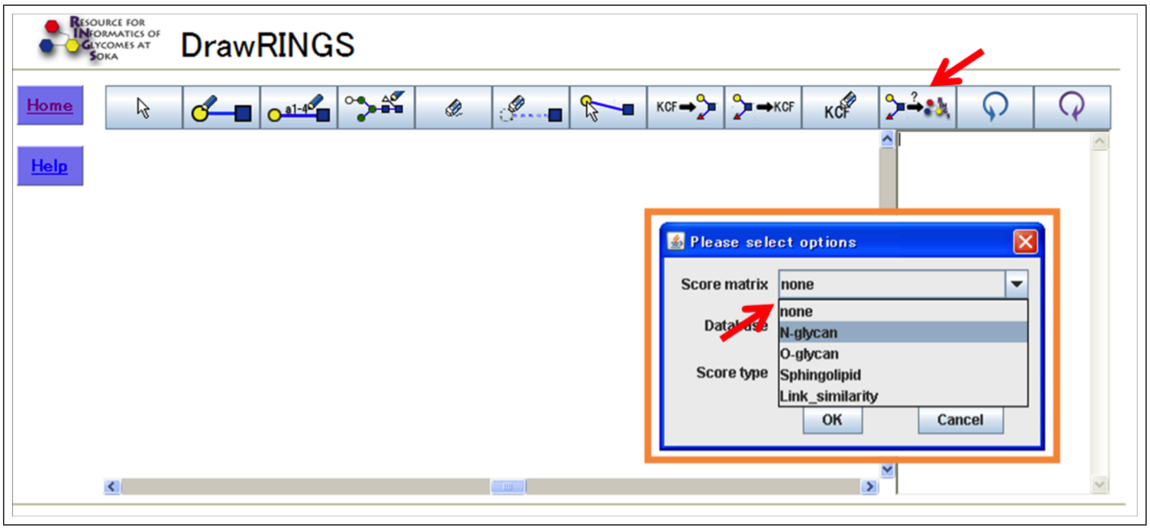
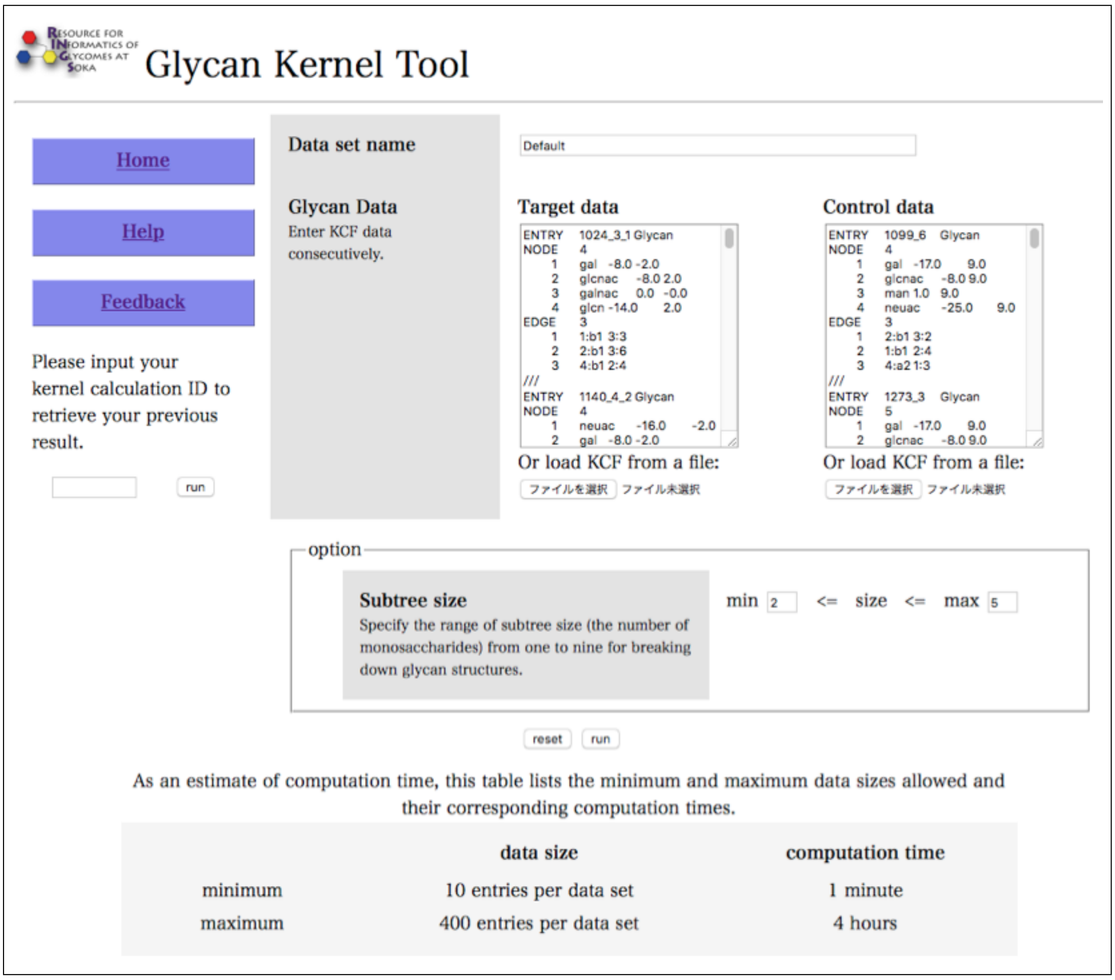


Figure 2.14: The interface of query search of DrawRINGS. Users are able to draw a glycan structure by using DrawRINGS. The drawn structure is used as a query for searching the same or similar structures in RINGS database. Searching is executed by "Run query" button, and then a pop-up dialog opens. Users are able to choose a score matrix from the option in the dialog.

are able to browse the results by input the calculation ID. Figure 2.17 is an example of the result page. The scores represent the degree of the differences of the substructure (Feature) from the control data. The layer represents the depth of the substructure from the reducing end of the input glycan.

2.3.3 RINGS Utilities

We have developed utilities, IUPACtoKCF and GlycoCT(condensed)toKCF (Fig. 2.18 and 2.19). Users are able to use a bracket or square bracket for a branching site in IUPAC format. Both utilities are available to input multi glycans, however, a blank line is required for GlycoCT(condensed)toKCF between glycans. Input glycans are converted by clicking "Get KCF".



Glycan Kernel Tool

Please input your kernel calculation ID to retrieve your previous result.

Data set name: Default

Glycan Data

Enter KCF data consecutively.

Target data

```
ENTRY     1024_3_1 Glycan
NODE      4
     1    gal   -8.0 -2.0
     2    glcnac     -8.0 2.0
     3    galnac     0.0   -0.0
     4    glcn -14.0       2.0
EDGE      3
     1    1:b1  3:3
     2    2:b1 3:6
     3    4:b1 2:4
///
ENTRY     1140_4_2 Glycan
NODE      4
     1    neuac      -16.0       -2.0
     2    gal   -8.0 -2.0
```

Or load KCF from a file:

ファイルを選択 ファイル未選択

Control data

```
ENTRY     1099_6    Glycan
NODE      4
     1    gal   -17.0       9.0
     2    glcnac     -8.0 9.0
     3    man 1.0   9.0
     4    neuac      -25.0       9.0
EDGE      3
     1    2:b1 3:2
     2    1:b1  2:4
     3    4:a2 1:3
///
ENTRY     1273_3    Glycan
NODE      5
     1    gal   -17.0       9.0
     2    glcnac     -8.0 9.0
```

Or load KCF from a file:

ファイルを選択 ファイル未選択

option

Subtree size

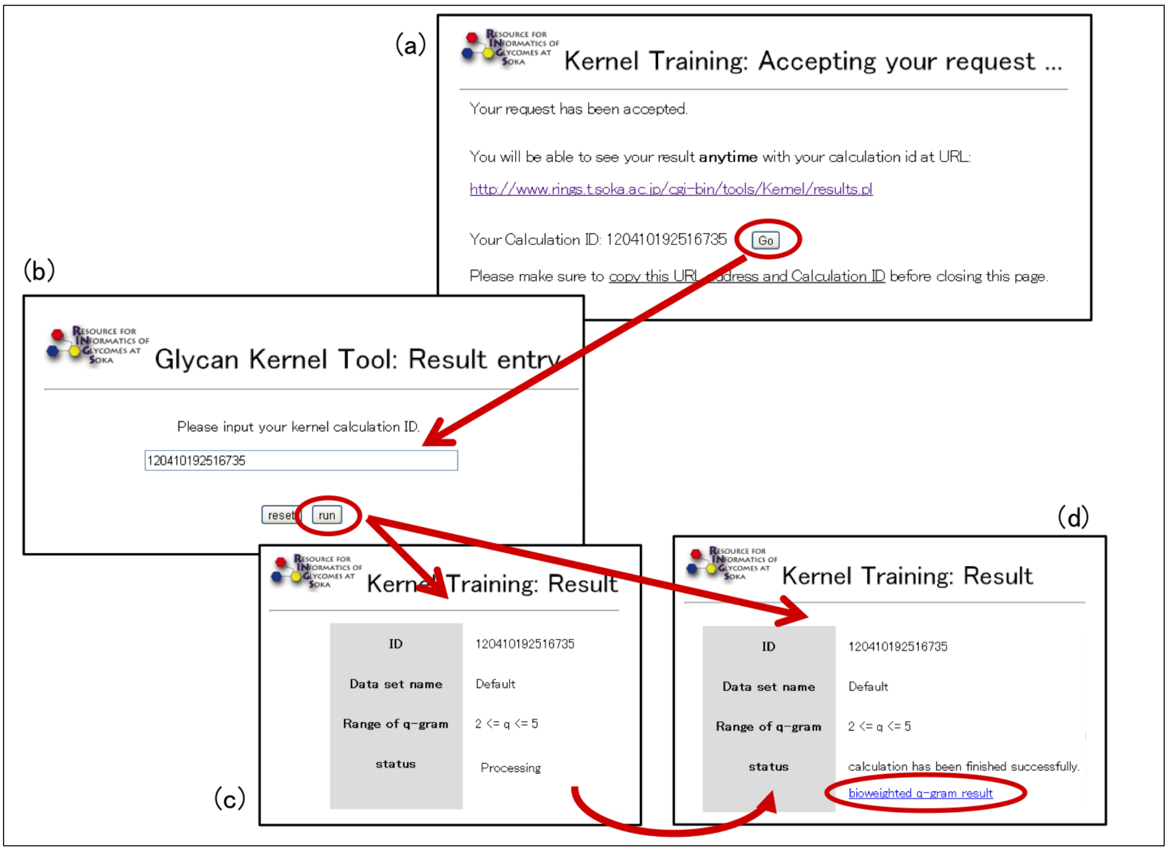
Specify the range of subtree size (the number of monosaccharides) from one to nine for breaking down glycan structures.

min 2 <= size <= max 5

As an estimate of computation time, this table lists the minimum and maximum data sizes allowed and their corresponding computation times.

	data size	computation time
minimum	10 entries per data set	1 minute
maximum	400 entries per data set	4 hours

Figure 2.15: A snapshot of Glycan Kernel Tool. Users are required to input data set name and target and control glycan data sets. The data set name is used for saving the input data and the calculation results into RINGS database. The glycans are required to written in KCF format. As a parameter, users can input the number of monosaccharides (size) in the substructures to extract. The range of the size should be at least one and at most nine. If users have used this tool previously, users are able to view the result by typing the calculation ID.



Kernel Training: Accepting your request ...

Your request has been accepted.

You will be able to see your result **anytime** with your calculation id at URL:

http://www.rings.t.soka.ac.jp/cgi-bin/tools/Kernel/results.pl

Your Calculation ID: 120410192516735

Please make sure to copy this URL address and Calculation ID before closing this page.

Glycan Kernel Tool: Result entry

Please input your kernel calculation ID.

120410192516735

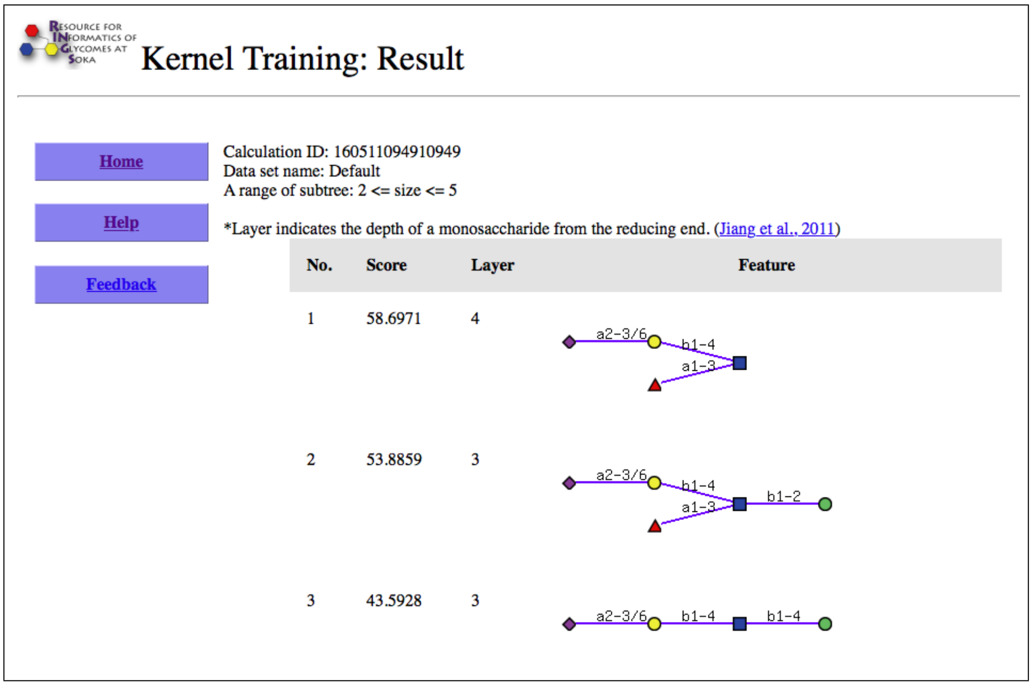
Kernel Training: Result

ID	120410192516735
Data set name	Default
Range of q-gram	2 <= q <= 5
status	Processing

Kernel Training: Result

ID	120410192516735
Data set name	Default
Range of q-gram	2 <= q <= 5
status	calculation has been finished successfully. bioweighted q-gram result

Figure 2.16: An example of usability of calculation ID. Glycan Kernel Tool requires some time for the calculation. So that, users are given a calculation ID to retrieve the result. When the calculation process is finished, a new link to the result page is generated (d).



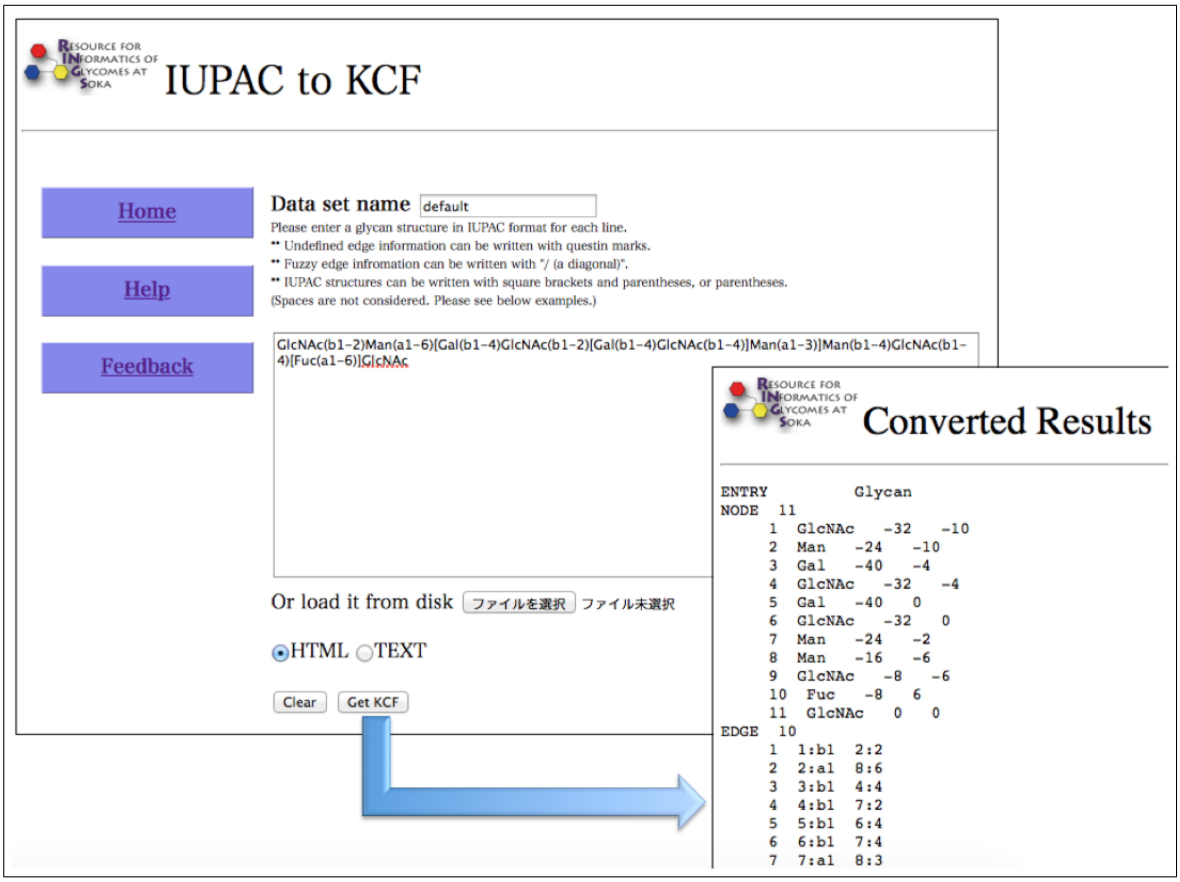
Kernel Training: Result

Calculation ID: 160511094910949
Data set name: Default
A range of subtree: 2 <= size <= 5

*Layer indicates the depth of a monosaccharide from the reducing end. (Jiang et al., 2011)

No.	Score	Layer	Feature
1	58.6971	4	a2-3/6 b1-4 a1-3
2	53.8859	3	a2-3/6 b1-4 a1-3 b1-2
3	43.5928	3	a2-3/6 b1-4 b1-4

Figure 2.17: An example of the result page of Glycan Kernel Tool. The scores represent the degree of the differences of the substructure (Feature) from the control data. The layer represents the depth of the substructure from the reducing end of the input glycan.



IUPAC to KCF

Data set name default

Please enter a glycan structure in IUPAC format for each line.
** Undefined edge information can be written with questin marks.
** Fuzzy edge infromation can be written with "/ (a diagonal)".
** IUPAC structures can be written with square brackets and parentheses, or parentheses.
(Spaces are not considered. Please see below examples.)

```
GlcNAc(b1-2)Man(a1-6)[Gal(b1-4)GlcNAc(b1-2)[Gal(b1-4)GlcNAc(b1-4)]Man(a1-3)]Man(b1-4)GlcNAc(b1-
4)[Fuc(a1-6)]GlcNAc
```

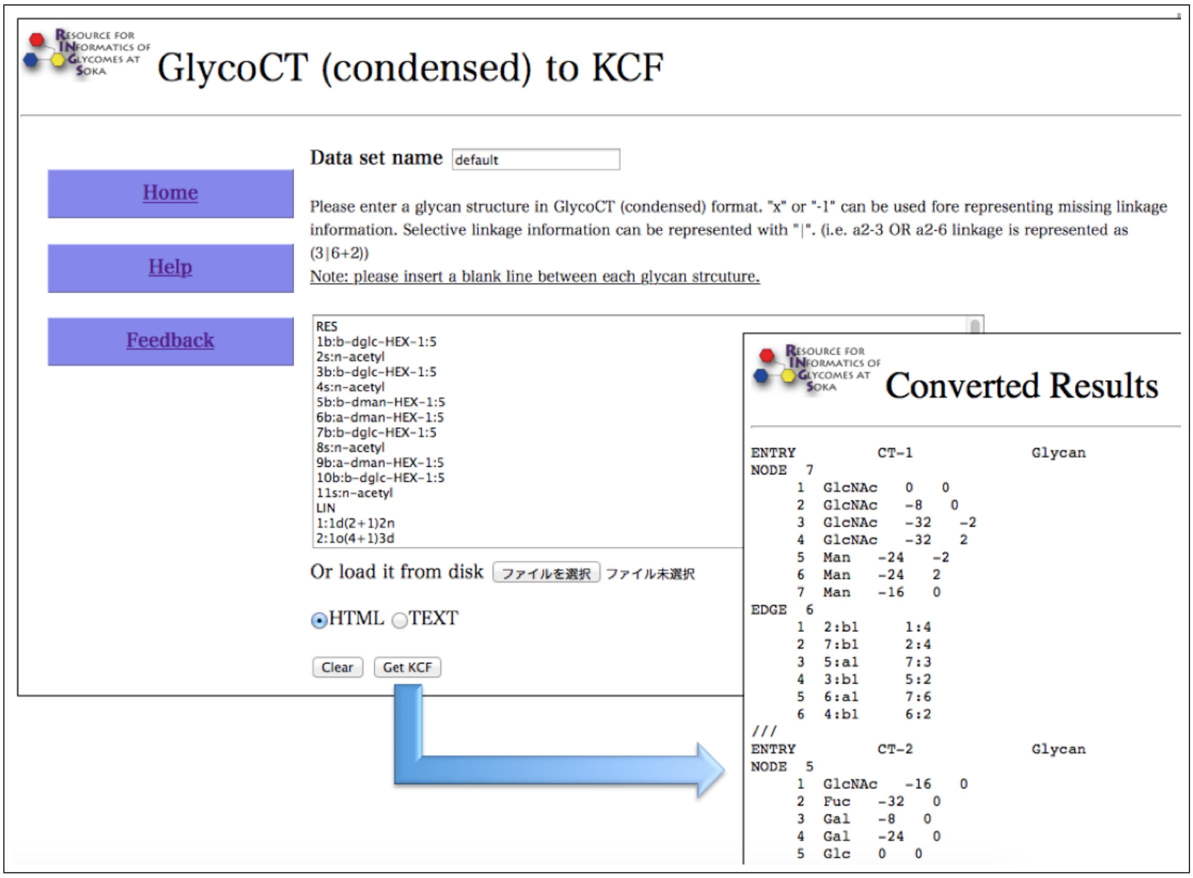
Or load it from disk ファイルを選択 ファイル未選択

● HTML ○ TEXT

Converted Results

```
ENTRY          Glycan
NODE  11
     1  GlcNAc   -32   -10
     2  Man   -24   -10
     3  Gal   -40   -4
     4  GlcNAc   -32   -4
     5  Gal   -40   0
     6  GlcNAc   -32   0
     7  Man   -24   -2
     8  Man   -16   -6
     9  GlcNAc   -8   -6
     10  Fuc   -8   6
     11  GlcNAc   0   0
EDGE  10
     1  1:bl  2:2
     2  2:al  8:6
     3  3:bl  4:4
     4  4:bl  7:2
     5  5:bl  6:4
     6  6:bl  7:4
     7  7:al  8:3
```

Figure 2.18: A snapshot of IUPAC to KCF utility. To represent a branching position, users are able to use bracket or square bracket. Users are also able to input multiple glycans in each line. The input structure(s) is converted by clicking "Get KCF".



RESOURCE FOR INFORMATICS OF GLYCOMES AT SOKA
Home
Help
Feedback
ファイルを選択
Clear
Get KCF
RESOURCE FOR INFORMATICS OF GLYCOMES AT SOKA

GlycoCT (condensed) to KCF

Data set name: default

Please enter a glycan structure in GlycoCT (condensed) format. "x" or "-1" can be used fore representing missing linkage information. Selective linkage information can be represented with "|". (i.e. a2-3 OR a2-6 linkage is represented as (3|6+2))

Note: please insert a blank line between each glycan strcuture.

```
RES
1b:b-dglc-HEX-1:5
2s:n-acetyl
3b:b-dglc-HEX-1:5
4s:n-acetyl
5b:b-dman-HEX-1:5
6b:a-dman-HEX-1:5
7b:b-dglc-HEX-1:5
8s:n-acetyl
9b:a-dman-HEX-1:5
10b:b-dglc-HEX-1:5
11s:n-acetyl
LIN
1:1d(2+1)2n
2:1o(4+1)3d
```

Or load it from disk ファイル未選択

● HTML ○ TEXT

Converted Results

```
ENTRY         CT-1              Glycan
NODE   7
       1   GlcNAc   0    0
       2   GlcNAc   -8    0
       3   GlcNAc   -32    -2
       4   GlcNAc   -32    2
       5   Man    -24    -2
       6   Man    -24    2
       7   Man    -16    0
EDGE   6
       1   2:b1      1:4
       2   7:b1      2:4
       3   5:a1      7:3
       4   3:b1      5:2
       5   6:a1      7:6
       6   4:b1      6:2
///
ENTRY         CT-2              Glycan
NODE   5
       1   GlcNAc   -16    0
       2   Fuc    -32    0
       3   Gal    -8    0
       4   Gal    -24    0
       5   Glc    0    0
```

Figure 2.19: A snapshot of GlycoCT(condensed) to KCF utility. Users are also able to input multiple glycans by inserting a blank line between glycans. The input structure(s) is converted by clicking "Get KCF".

2.4 Discussions

In this study,I have developed calculating system for glycan score matrices, Glycan Kernel tool and conversion utilities. I have developed and introduced glycan score matrices into the DrawRINGS query search tool to improve the accuracy of the database searching including physico-chemical meanings. However, current score matrices give higher score to the link, a monosaccharide and its glycosidic linkage, pairs which alignment frequency is low. For example, the score given to a link pair "Glc(?1-3)" and "Gal(?1-3)" is 44 points which is one of the second highest score among the N-glycan score matrix. Furthermore, the highest score in this matrix was given to a link pair "Gal(β1-4)" and "GlcNAc(α1-6)". This may be caused by the scoring calculation during pair-wised alignments and/or multiple alignments. Unlike amino acids neither genome sequences, genetic conservation of glycan sequences in process of evolution has not been proven. Therefore, it is difficult to compute the alignment scores with biological weights. However, this study is still in the first stage, and there is a necessity for variety of efforts, such as the investigation of the input data and the development of a new alignment and clustering algorithms, to improve this study.

I also have developed a Glycan Kernel tool. In order to assess the generality of the ability of this tool to extract meaningful substructures, I utilized a data set obtained from the CFG to evaluate the feature selection performance. I obtained O- and N-glycan profile data extracted from the brain of mouse train C57BL/6 [41],which consisted of 47 structures in wildtype and 50 structures in FucTIV+VII knockout mice[42]. I also chose the size to extract substructures as at least one to five to avoid substructures of large sizes from the learning algorithm. As a result, the feature with the top score extracted by this tool was "NeuAc(2-3/6)Gal(β1-4)[Fuc(1-3)]GlcNAc" at layer four, which represents a depth (the number of monosaccharide from the reducing end) of the input glycan structure. This featured structures is sialyl-Lewis X, which was previously confirmed for this sample [43]. Thus, this tool is shown to be able to extract unique and accurate glycan structures from the target data set compared to the control. Moreover, I have allowed biologists to take advantage of a powerful tool for glycan feature extraction with this new tool.

The implementation of utilities IUPACtoKCF and GlycoCT{condensed}toKCF are essential to the users of RINGS because these are the most commonly used format in glycoscience research. Before these utilities were developed, many researchers had difficulty using the RINGS tools and were forced to start drawing all their input glycans using DrawRINGS. However, because of these utilities, users can directly convert their structures and quickly use the RINGS tools.

Chapter 3

Construction of theoretical *N*-glycan database

3.1 Introduction

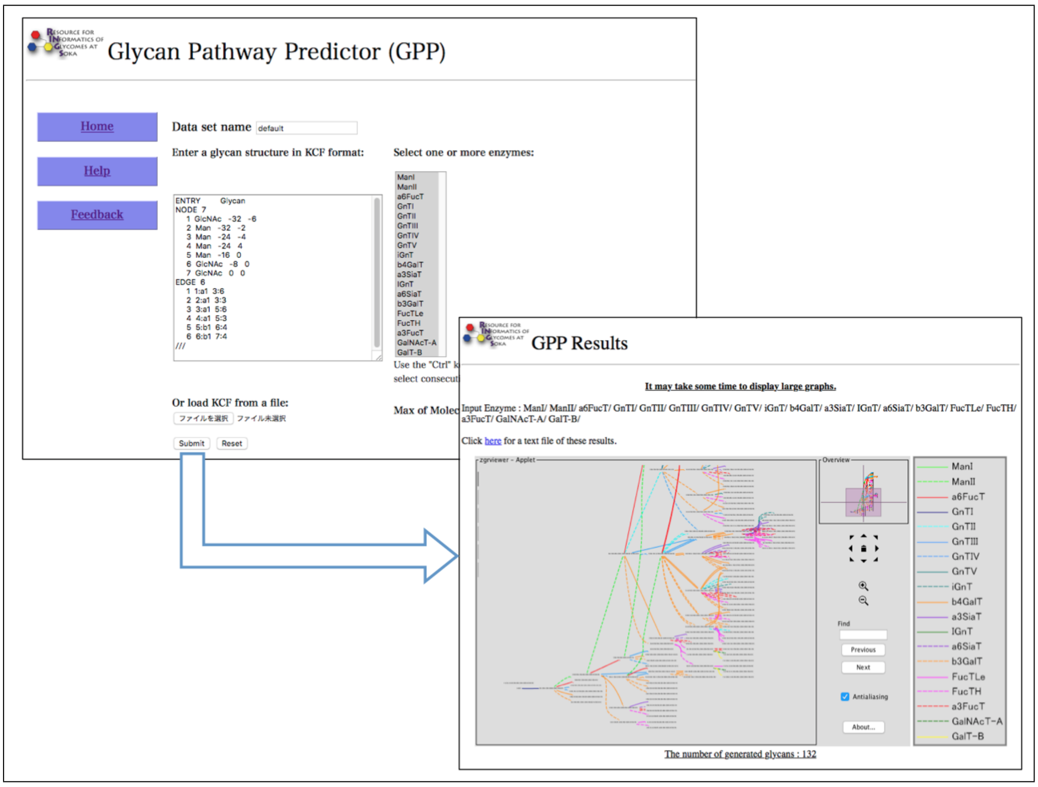
3.1.1 Mathematical models for predicting glycan synthetic pathway

Because various computational models have been reported in the last decades, glycome informatics including mathematical models has progressed. Umana and Bailey has reported an algorithm to predict small glycan synthesis computationally [44]. Their theory has demonstrated that the alterations of glyco-patterns (or glycan profiles), which may occur by mislocalization or by a gene knockout of a specific glycosyltransferase, can be predicted. This model is one of the earliest computational models related to glycan synthesis, moreover, it is the first report that an *in silico* approach of glyco-engineering is feasible. Similar to the Umana and Bailey approach, in 2009, Krambeck *et al.* reported a mathematical model using reaction conditions of seventeen glycosyltransferases and two glycosidases associated with N-glycans [18, 17]. 10,000-20,000 theoretical N-glycans were calculated from this model. One feature of this model is that concentrations of the enzymes are adjusted based on a numerical optimization method to help correlate experimentally observed glycans with calculated glycans.

Several algorithms to analyze glycan functions and structures have been reported. However, these algorithms are constructed based on biological backgrounds, such as enzymatic conditions and reaction properties along their aims. Therefore, it may be necessary to limit the target sample or the parameters. For the glycan synthetic systems, it depends on a number of factors, such as the effectiveness of a donor, the adjustment of the paradigm and/or bio-processing, a culture kinetics of a sample and the design of a metabolic networks.

3.1.2 A tool for predicting glycan synthetic pathway

In RINGS, we have developed several web tools, that users are able to freely use, based on published algorithms. One of the tools, we have developed is Glycan Pathway Predictor (GPP) (Fig. 3.1). This tool allows users to predict glycan synthesis using a selection of N-glycan glycosyltransferases and glycosidases. GPP implements the Krambeck 2009 mathematical models, and outputs the dynamically calculated result as synthetic pathway maps.



Glycan Pathway Predictor (GPP)

Home

Help

Feedback

Data set name: default

Enter a glycan structure in KCF format:

```
ENTRY      Glycan
NODE  7
   1  GlcNAc  -32  -6
   2  Man  -32  -2
   3  Man  -24  -4
   4  Man  -24   4
   5  Man  -16   0
   6  GlcNAc  -8   0
   7  GlcNAc   0   0
EDGE  6
   1  1:a1  3:6
   2  2:a1  3:3
   3  3:a1  5:6
   4  4:a1  5:3
   5  5:b1  6:4
   6  6:b1  7:4
///
```

Select one or more enzymes:

ManI
ManII
a6FucT
GnTI
GnTII
GnTIII
GnTIV
GnTV
iGnT
b4GalT
a3SiaT
IGnT
a6SiaT
b3GalT
FucTLe
FucTH
a3FucT
GalNAcT-A
GalT-B

Use the "Ctrl" k
select consecut

Max of Molec

Or load KCF from a file:
ファイルを選択 ファイル未選択

Submit Reset

GPP Results

It may take some time to display large graphs.

Input Enzyme : ManI/ ManII/ a6FucT/ GnTI/ GnTII/ GnTIII/ GnTIV/ GnTV/ iGnT/ b4GalT/ a3SiaT/ IGnT/ a6SiaT/ b3GalT/ FucTLe/ FucTH/ a3FucT/ GalNAcT-A/ GalT-B/

Click here for a text file of these results.

The number of generated glycans : 132

Figure 3.1: A snapshot of Glycan Pathway Predictor (GPP). Users can use an N-glycan structure written in KCF format as input. Two glycosidases and seventeen glycosyltransferases are listed. The chosen enzymes are used for calculating the potential biosynthetic pathways, and the calculation continues until the maximum of the molecular mass is reached. The calculated pathways can be browsed as a map. The detail information of each reaction is displayed by clicking each structure in the map.

3.1.3 glycan related databases associated with theoretical *N*-glycan database

Activity and the gene expression levels of glycosyltransferases fluctuate by pathophysiological conditions such as nucleotide sugars and receptor molecules, nucleotide sugar transporters, and expression levels of endogenous lectins in tumor cells. For example, the epithelial-mesenchymal transition (EMT) refers to the phenomenon where epithelial cells obtain the ability of migration or invasion [45]. The EMT processes is one of the key roles for cancer metastasis. The decreasing of GnT-III concentration, which synthesizes a bisecting GlcNAc, during EMT processes has been reported by Xu *et al.* [46].

Several databases are storing glycan related information. For instance, KEGG GLYCAN has been developed for collecting glycan related data such as KEGG Pathway Maps for Glycans, Glycans in Cancer Pathways, Glycosyltransferases, Glycan Binding Proteins and KEGG GLYCAN Structures [47]. Figure 3.2 shows an example of *N*-glycan biosynthesis pathways in KEGG Pathway Maps. *N*-glycan precursors are synthesized in the cytosol and endoplasmic reticulum (ER). Then, it is transferred to an asparagine residue by oligosaccharyltransferase (OST) and passed to golgi membranes after the trimming by glucosidases and mannosidases. Glycosyltransferases related to *N*-glycans are thought to be expressed in the golgi membrane, however, it is remained to be elucidated. KEGG GLYCAN also has developed Composite Structure Map (CSM), which is a static representation of all carbohydrate structures in KEGG GLYCAN overlapped with one another and displayed in a global tree format (Fig. 3.3) [48]. CSM contains links to corresponding glycosyltransferase information, if the data is available in the KEGG database.

One of the other databases related to glycosyltransferases is Carbohydrate-Active en-

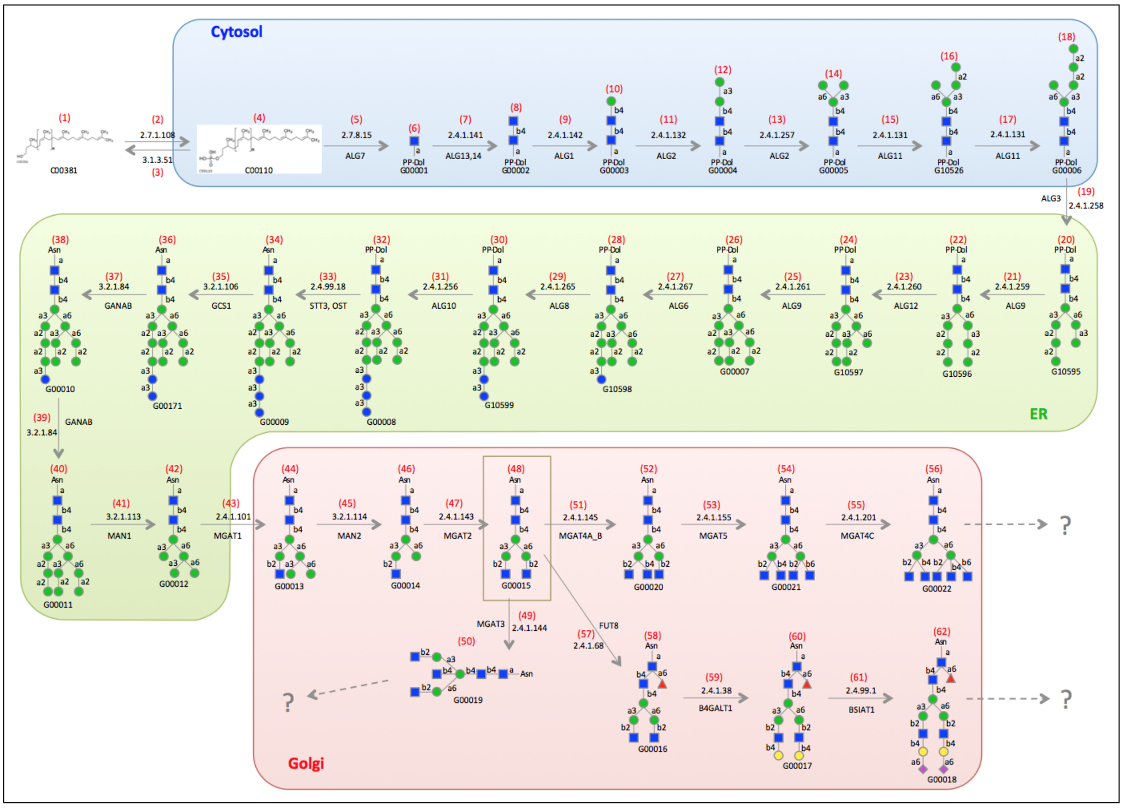
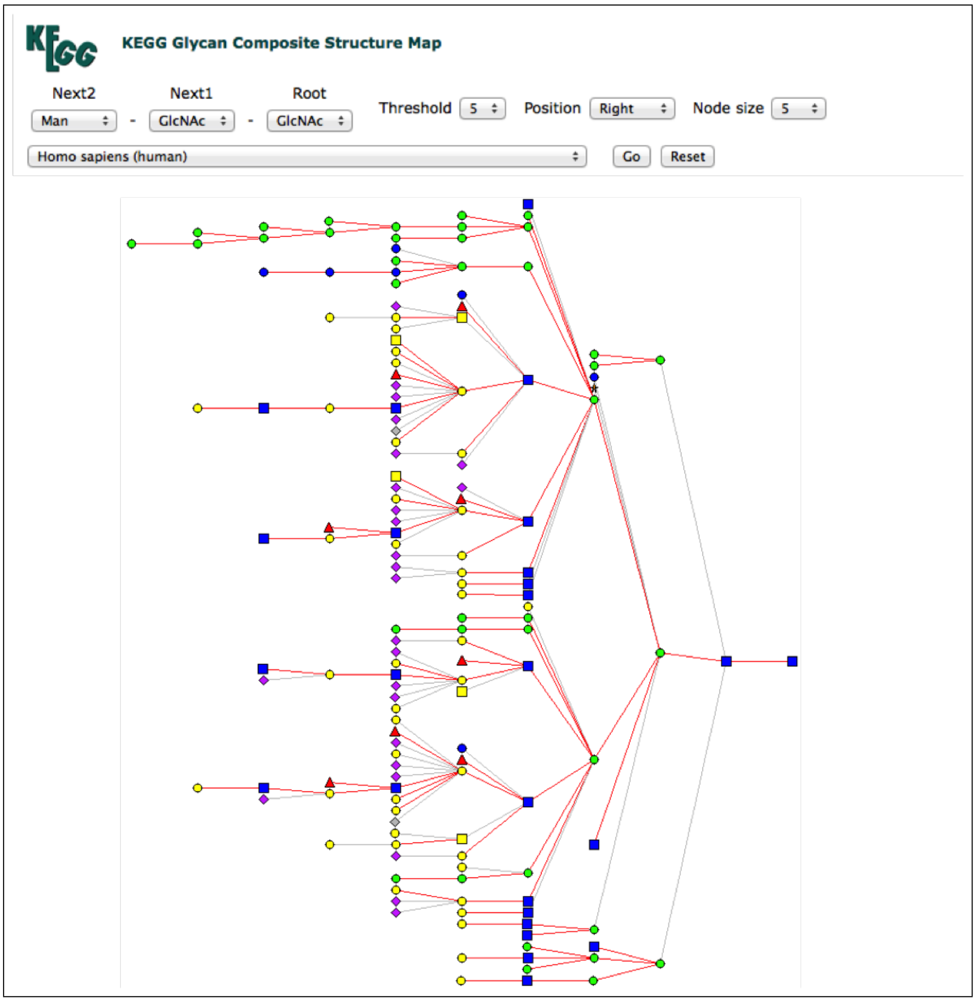


Figure 3.2: An example of *N*-glycan synthetic pathway map. Each synthetic reaction and structure are save in KEGG GLYCAN database. However, glycan synthesis may be depends on species, organs and pathological conditions. So that it is difficult to elucidate the whole pathway maps.



KEGG Glycan Composite Structure Map

Next2 Man - Next1 GlcNAc - Root GlcNAc

Threshold 5

Position Right

Node size 5

Homo sapiens (human)

Figure 3.3: A snapshot of KEGG Glycan Composite Structure Map. Users are able to limit for overlapping structures by obtain a map by choosing non-reducing end monosaccharide(s) and species.

ZYmes (CAZy) database, which have been developed since 1998 in France [49]. In CAZy, Glycan related proteins are classified into five families, such as glycosyltransferases, glycoside hydrolase, polysaccharide lyase, Carbohydrate esterase and carbohydrate-binding module family. Moreover, BRaunschweig ENzyme DAtabase (BRENDA) have been developed as enzyme information repository since 1987 in Germany [50]. The enzymes are classified by the Enzyme Nomenclature and contains background data related to molecular biology, biochemistry, medical research, and biotechnology. GlycoGene DataBase (GGDB) [51] is one of the project included in Japan Consortium for Glycobiology and Glycotechnology DataBase (JCGGDB). GGDB collected data based on glycan related gene, for instance, gene names, enzyme names and types of substrate specificity. The Universal Protein Resource (UniPort) has been developed as a resource for protein sequence and annotation data. Users are able to find accurate, consistent and rich annotation information via the UniProt Knowledgebase (UniProtKB) [52]. Furthermore, UniCarbKB has stored experimentally described glycan structures reported in the literature [53, 16].

3.1.4 Development of the novel database for glycan synthetic pathway prediction

In this manuscript we describe a systematic method to use the properties and substrate specificity rules of all the described human glycosyltransferases involved in the biosynthesis of *N*-glycans to produce a theoretically possible collection of *N*-glycan structures based on current pathway knowledge. The proposed model is designed to facilitate: i) the automated construction of single glycosylation reactions using enzyme-substrate relationships; ii) the simulation of an entire *N*-glycosylation synthetic network; and iii) mass spectrometry glycan analysis by providing a library of theoretically possible glycan structures for data mining/matching. The proposed theoretical N-glycan structures that result from this computation are available in the UniCorn section of UniCarbKB. UniCorn has been introduced in response to concerns that existing structural and experimental databases under-represent the N-glycome, that current experimental glyco-analysis protocols are not observing all structures and that the traditional time- and labor-intensive manual curation of the published literature may be complemented by such a database. UniCorn also aims to provide an effective curation and quality-checking tool of newly discovered curated glycan structures populating UniCarbKB. Such an application will assist researchers in validating previously unrecognized glycan structures, provide information on biosynthetic glycan pathways and enable the generation of mass spectral fragments from biosynthetically putative structures.

3.2 Materials and Methods

3.2.1 Glycosyltransferase catalog

We have collected the human glycosyltransferase information related to N- and precursor N-glycans from previous studies [17, 18] and existing databases including KEGG, GGDB, CAZy, CFG, BRENDA and UniProt. In particular we included in our computation any described additional residues, substrate structures, intercellular localization, genes and synthetic condition information. The collected information is used for calculating a theoretical glycan database and glycosyltransferase (GT) candidates analysis (Fig. 3.4). In this work, glycosyltransferases are represented based on the name of UniProt database due to the usages of different protein name or identifiers among databases. A glycan is

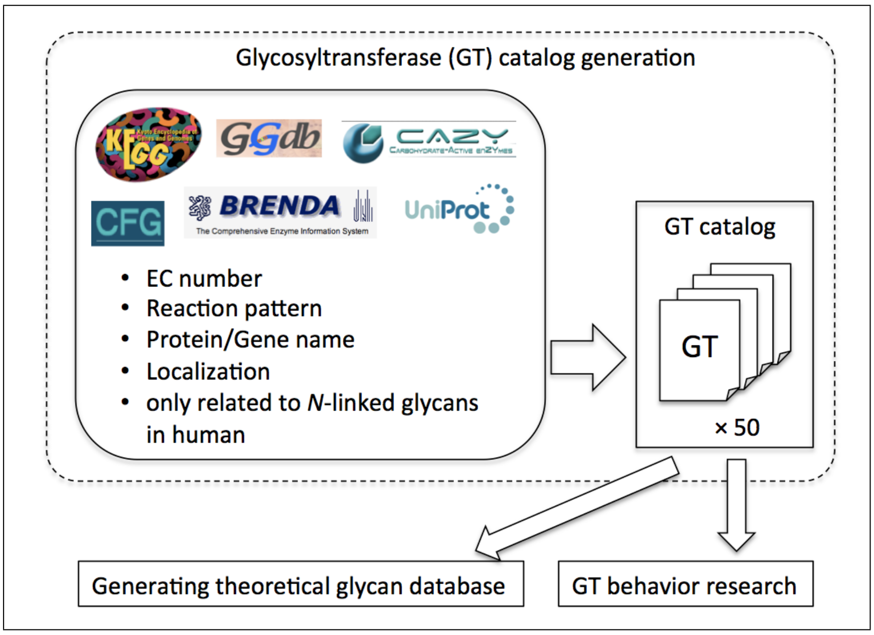


Figure 3.4: A flowchart of generating a glycosyltransferase catalog. We have collected human glycosyltransferase (GT) information related to N-glycans including precursor N-glycans from previous reports and databases. We finally gathered 50 types of GTs, and they were used for calculating theoretical glycan database and GT candidates research.

represented in IUPAC format in our catalog.

3.2.2 Development for a tool for glycosyltransferase candidates prediction

A tool for glycosyltransferase candidates prediction have been developed for predicting glycosyltransferases activities may worked on a glycan structure. For example, users are able to obtain a list of glycosyltransferases that may associated to synthesize an input glycan structure by comparing each glycosidic linkage of input glycan and glycosyltransferase catalog which contains a donor and an acceptor information (Fig. 3.5).

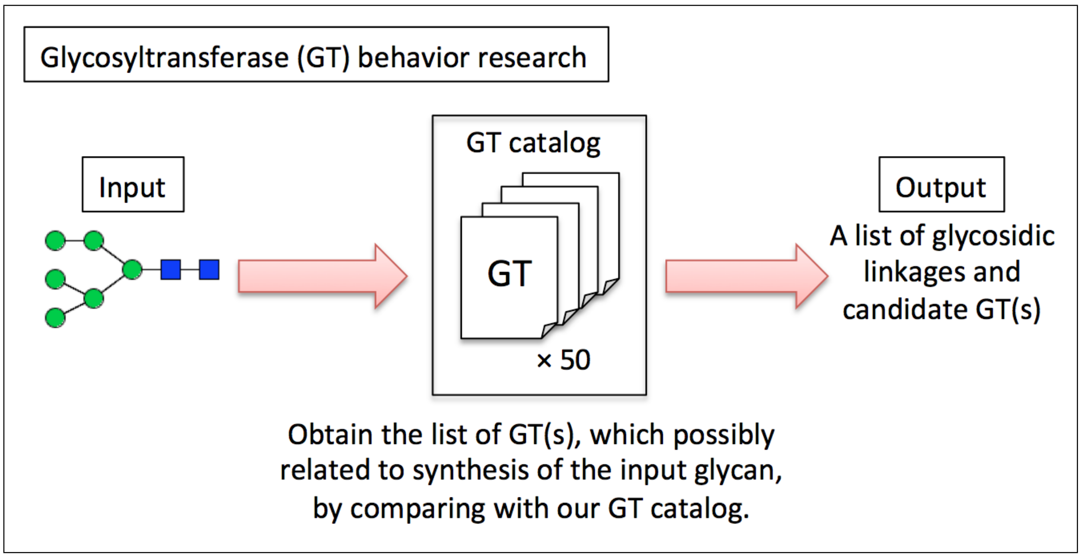

Glycosyltransferase (GT) behavior research
Input
GT catalog
GT
× 50
Output
A list of glycosidic linkages and candidate GT(s)
Obtain the list of GT(s), which possibly related to synthesis of the input glycan, by comparing with our GT catalog.

Figure 3.5: A flowchart of a tool for glycosyltransferase candidates prediction. When a glycan structure was input, each glycosidic linkage is compared with our glycosyltransferase (GT) catalog that contains a donor and an acceptor information. A list of glycosyltrnasferases are output if the synthetic pattern is matched.

3.2.3 Theoretical *N*-glycan calculation algorithm

We have developed a code for generating theoretical *N*-glycans by using glycosyltransferase catalog of human and database to save the generated structures. We used Perl language for the code and PostgreSQL for the database. The database contains two main tables, glycan_structure and glycan_tree. Glycan_structure table is used for saving the generated structures without duplication. A structure is written in LinearCode format which is able to represent a glycan in the shortest manner. Each structure is given an id as an identifier and a level as the number of monosaccharide contained in the structure. Glycan_tree table is used for saving the synthetic pathway relationships. A structure prior to the synthesis is labeled as "parent" and its id is stored as parent_id. On the other hand, the structure after the reaction is stored as child_id. This reaction pattern is saved with

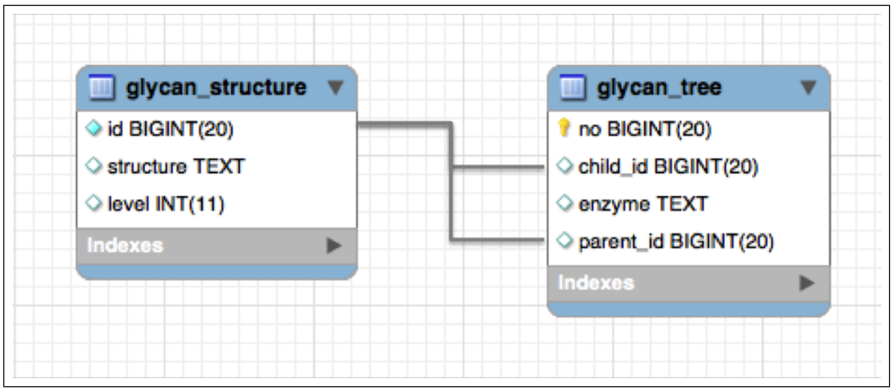


Figure 3.6: An entity relationship diagram of theoretical glycan database. Glycan_structure table stores "id" for the identifier, "structure" for generated glycan structure and "level" for the number of monosaccharides in the glycan. Glycan_tree table stores "no" for the identifier, "child_id", "enzyme" and "parent_id". A parent_id and a child_id represents the prior and after the reaction structure ids. The glycosyltransferase is stored as "enzyme".

the identifier "no" and its glycosyltransferase "enzyme".

Figure 3.7 shows an image of this program. When a glycan structure is given to the program, the structure is compared with the reaction pattern of the GT catalog. If a reaction pattern is matched, a theoretically synthesized glycan is generated. This calculation continues the size, the number of monosaccharide in a glycan, is less than fifteen.

In total, 46 unique reaction patterns are specified in our model that involve the transfer of a monosaccharide from a nucleotide-sugar donor to an acceptor from 50 human glycosyltransferases including which related to precursor *N*-glycans. For example, MGAT5 adds GlcNAc(β1-6) onto Man(α1-6). However, it may also have the potential to add GlcNAc(β1-6) on Man(α1-3) [54]. Moreover, additional possible (previously described)

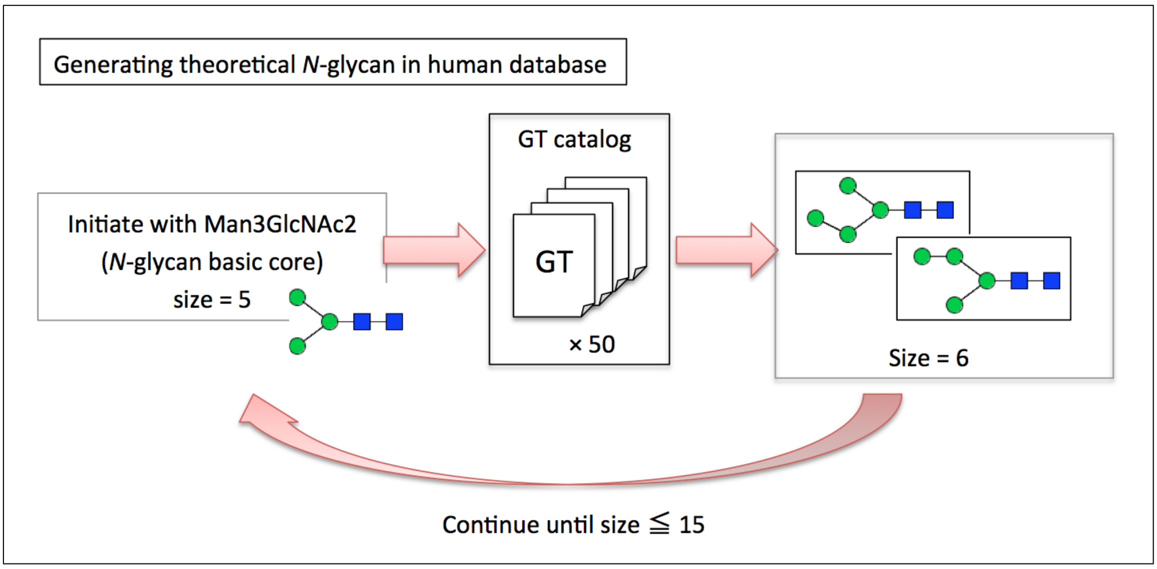


Figure 3.7: An example of generating theoretical N-glycans in human. The calculation starts with Man3GlcNAc2 which is a core structure of N-glycan. The input structure is compared with the reaction patterns of the list of glycosyltransferases of N-glycans in human (GT catalog). When a substrate the reaction pattern is matched with the input glycan, a theoretical structure is generated. This calculation continues until the size, the number of monosaccharide in a glycan, $\leq$ 15.

substrates are included in the catalogue, such as "GlcNAc(β1-2) [ˆ] Man(α1-3)" and "GlcNAc(β1-2) [GlcNAc(β1-4)] [ˆ] Man(α1-3)", where a caret represents the insertion position and square brackets represent a branching position. For simplicity, the stereochemical information (α/β) is inferred from the known specificities of the enzymes. For instance, all fucosyltransferases and sialyltransferases produce α-linked structures, the galactosyltransferases and N-acetylglucosaminyltransferases are assumed to form β-glycosidic linkages while N-acetylgalactosaminyltransferases is assumed to form α products.

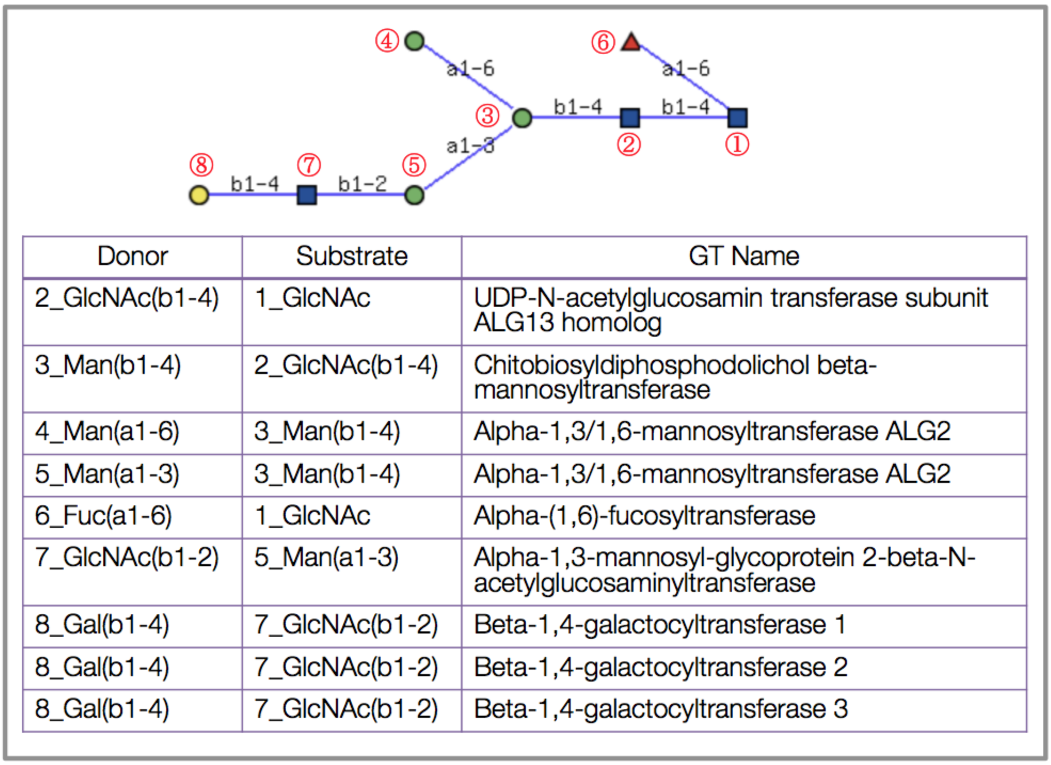
3.3 Results

3.3.1 A tool for glycosyltransferase candidates prediction

We have developed a tool for predicting glycosyltransferase candidates. Figure 3.8 shows an example of the result. Each monosaccharide and its glycosidic linkage in the input glycan is given a number as a position. So that, it is able to make difference when the same monosaccharide and glycosidic linkage appears more than twice in a glycan. This program output a list of two residues, donor and substrate, and possible glycosyltransferase which may synthesize the residue. The number before each residue represents the position in the input glycan structure. This tool is available in the structure page of UniCarbKB (Fig. 3.9)

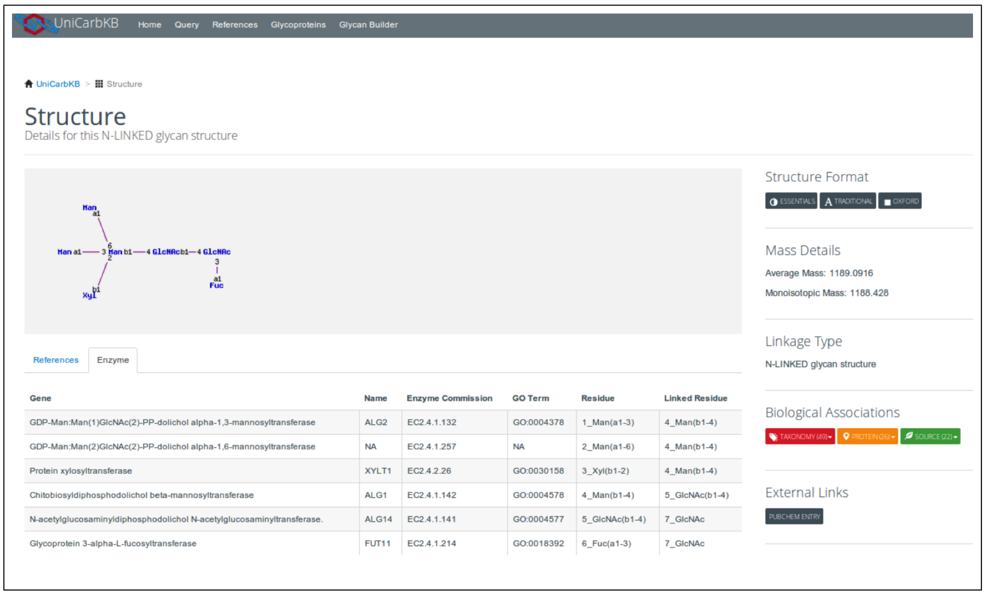
3.3.2 Glycosyltransferase catalog

We collected glycosyltransferase information based on N-glycosylation pathways in humans as currently known and described by databases, such as KEGG, CFG, CAZy, GGDB and BRENDA. Table 3.1 lists the protein and gene names used in our catalog, and Table 3.2 lists the gene names and their reaction pattern used for the calculation of theoretical glycan database. Our catalog includes enzymes that related to the synthesis of precursor N-glycans in human. Moreover, in Table 3.2, glycans are represented in LinearCode format that a monosaccharide is represented in one or two letters of the alphabet. Furthermore, we used a caret for representing the insertion position and square brackets for a branching position. We were able to collect 45 reaction patterns to calculate the theoretical N-glycans from 50 glycosyltransferases.



Donor	Substrate	GT Name
2_GlcNAc(b1-4)	1_GlcNAc	UDP-N-acetylglucosamin transferase subunit ALG13 homolog
3_Man(b1-4)	2_GlcNAc(b1-4)	Chitobiosyldiphosphodolichol beta-mannosyltransferase
4_Man(a1-6)	3_Man(b1-4)	Alpha-1,3/1,6-mannosyltransferase ALG2
5_Man(a1-3)	3_Man(b1-4)	Alpha-1,3/1,6-mannosyltransferase ALG2
6_Fuc(a1-6)	1_GlcNAc	Alpha-(1,6)-fucosyltransferase
7_GlcNAc(b1-2)	5_Man(a1-3)	Alpha-1,3-mannosyl-glycoprotein 2-beta-N-acetylglucosaminyltransferase
8_Gal(b1-4)	7_GlcNAc(b1-2)	Beta-1,4-galactocyltransferase 1
8_Gal(b1-4)	7_GlcNAc(b1-2)	Beta-1,4-galactocyltransferase 2
8_Gal(b1-4)	7_GlcNAc(b1-2)	Beta-1,4-galactocyltransferase 3

Figure 3.8: An example of the result of glycosyltransferase candidates prediction. Each monosaccharide and its glycosidic linkage is given a number as a position, so that users are able to see the specific pair of monosaccharides and its possible glycosyltransferase.



Structure

Details for this N-LINKED glycan structure

Gene	Name	Enzyme Commission	GO Term	Residue	Linked Residue
GDP-Man:Man(1)GlcNAc(2)-PP-dolichol alpha-1,3-mannosyltransferase	ALG2	EC2.4.1.132	GO:0004378	1_Man(a1-3)	4_Man(b1-4)
GDP-Man:Man(2)GlcNAc(2)-PP-dolichol alpha-1,6-mannosyltransferase	NA	EC2.4.1.257	NA	2_Man(a1-6)	4_Man(b1-4)
Protein xylosyltransferase	XYLT1	EC2.4.2.26	GO:0030158	3_Xyl(b1-2)	4_Man(b1-4)
Chitobiosyldiphosphodolichol beta-mannosyltransferase	ALG1	EC2.4.1.142	GO:0004578	4_Man(b1-4)	5_GlcNAc(b1-4)
N-acetylglucosaminyldiphosphodolichol N-acetylglucosaminyltransferase.	ALG14	EC2.4.1.141	GO:0004577	5_GlcNAc(b1-4)	7_GlcNAc
Glycoprotein 3-alpha-L-fucosyltransferase	FUT11	EC2.4.1.214	GO:0018392	6_Fuc(a1-3)	7_GlcNAc

Structure Format

Mass Details

Average Mass: 1189.0916

Monoisotopic Mass: 1188.428

Linkage Type

N-LINKED glycan structure

Biological Associations

External Links

Figure 3.9: An example of the interface of glycosyltransferase candidates prediction. This tool is available in the structure page of UniCarbKB.

Table 3.1: A list for the proteins and genes used in our glycosyltransferase catalog.

No.	Protein Name	Gene Name
1	Histo-blood group ABO system tranferase	ABO
2	Dol-P-Man:Man(5)GlcNAc(2)-PP-Dol alpha-1,3-mannosyltransferase	ALG3
3	Dolichyl pyrophosphate Man9GlcNAc2 alpha-1,3-glucosyltransferase	ALG6
4	Probable dolichyl pyrophosphate Glc1Man9GlcNAc2 alpha-1,3-glucosyltransferase	ALG8
5	Alpha-1,2-mannosyltransferase ALG9	ALG9
6	Dol-P-Glc:Glc(2)Man(9)GlcNAc(2)-PP-Dol alpha-1,2-glucosyltransferase	ALG10
7	GDP-Man:Man(3)GlcNAc(2)-PP-Dol alpha-1,2-mannosyltransferase	ALG11
8	Dol-P-Man:Man(7)GlcNAc(2)-PP-Dol alpha-1,6-mannosyltransferase	ALG12
9	Galactosylgalactosylxylosylprotein 3-beta-glucronosyltransferase 1	B3GAT1
10	Galactosylgalactosylxylosylprotein 3-beta-glucronosyltransferase 2	B3GAT1
11	Beta-1,3-galactosyltransferase 1	B3GALT1
12	Beta-1,3-galactosyltransferase 2	B3GALT2
13	N-acetyllactosaminide beta-1,3-N-acetylgulcosaminyltransferase	B3GNT1
14	UDP-GlcNAc: beta Gal beta-1,3-N-acetylglucosaminlytransferase 7	B3GNT7
15	Beta-1,4-N-acetylgalactosaminyltransferase 3	B4GALNT3
16	N-acetyl-beta-glucosaminyl-glycoprotein 4-beta-N-acetylgalactosaminyltransferase 1	B4GALNT4
17	Beta-1,4-galactosyltransferase 1	B4GALT1
18	Beta-1,4-galactosyltransferase 2	B4GALT2
19	Beta-1,4-galactosyltransferase 3	B4GALT3
20	Beta-1,4-galactosyltransferase 4	B4GALT4
21	Galactoside 2-alpha-L-fucosyltransferase 1	FUT1
22	Galactoside 2-alpha-L-fucosyltransferase 2	FUT2
23	Galactoside 3(4)-L-fucosyltransferase	FUT3
24	Alpha-(1,3)-fucosyltransferase 5	FUT5
25	Alpha-(1,3)-fucosyltransferase 6	FUT6
26	Alpha-(1,3)-fucosyltransferase 7	FUT7
27	Alpha-(1,6)-fucosyltransferase	FUT8
28	Alpha-(1,3)-fucosyltransferase 9	FUT9
29	Alpha-(1,3)-fucosyltransferase 11	FUT11
30	Beta-1,3-galactosyl-O-glycosyl-glycoprotein beta-1,6-N-acetylglucosaminyltransferase3	GCNT3
31	Alpha-1,3-mannosyl-glycoprotein 2-beta-N-acetylglucosaminyltransferase	MGAT1
32	Alpha-1,6-mannosyl-glycoprotein 2-beta-N-acetylglucosaminyltransferase	MGAT2
33	Beta-1,4-mannosyl-glycoprotein 4-beta-N-acetylglucosaminyltransferase	MGAT3
34	Alpha-1,3-mannosyl-glycoprotein 4-beta-N-acetylglucosaminyltransferase A	MGAT4A

Table 3.1: A list for the proteins and genes used in our glycosyltransferase catalog.

No.	Protein Name	Gene Name
35	Alpha-1,3-mannosyl-glycoprotein 4-beta-N-acetylglucosaminyltransferase B	MGAT4B
36	Alpha-1,3-mannosyl-glycoprotein 4-beta-N-acetylglucosaminyltransferase C	MGAT4C
37	Alpha-1,6-mannosyl-glycoprotein 6-beta-N-acetylglucosaminyltransferase A	MGAT5
38	CMP-N-acetylneuraminate-beta-1,4-galactoside alpha-2,3-sialyltranferase 1	ST3GAL1
39	CMP-N-acetylneuraminate-beta-1,4-galactoside alpha-2,3-sialyltranferase 2	ST3GAL2
40	CMP-N-acetylneuraminate-beta-1,4-galactoside alpha-2,3-sialyltranferase	ST3GAL3
41	CMP-N-acetylneuraminate-beta-1,4-galactoside alpha-2,3-sialyltranferase 4	ST3GAL4
42	Type 2 lactosamine alpha-2,3-sialyltransferase	ST3GAL6
43	Beta-galactoside alpha-2,6-sialyltransferase 1	ST6GAL1
44	Beta-galactoside alpha-2,6-sialyltransferase 2	ST6GAL2
45	Alpha-N-acetylneuraminide alpha-2,8-sialyltransferase	ST8SIA1
46	Alpha-2,8-sialyltransferase 8B	ST8SIA2
47	Sia-alpha-2,3-Gal-beta-1,4-GlcNAc-R:alpha 2,8-sialyltransferase	ST8SIA3
48	CMP-N-acetylneuraminate-poly-alpha-2,8-sialyltransferase	ST8SIA4
49	Alpha-2,8-sialyltransferase 8E	ST8SIA5
50	Alpha-2,8-sialyltransferase 8F	ST8SIA6

Table 3.2: A list of genes and their synthetic activities used in our catalog. A caret represents the insertion position and square brackets represent a branching position.

No.	Gene Name	Substrate	Donor residue	Condition rule
1	ABO	Fa2[^]Ab	ANa3	-
2	ABO	Fa2[^]Ab	Aa3	-
3	ALG3	^Ma6	Ma3	On the "Ma6" arm of the *N*-glycan core. No A,F,NN in the glycan
4	ALG6	^Ma2Ma2Ma3	Ga3	On the "Ma3" arm of the *N*-glycan core. No A,F,NN in the glycan
5	ALG8	^Ga3Ma2Ma2Ma3	Ga3	On the "Ma3" arm of the *N*-glycan core. No A,F,NN in the glycan
6	ALG9	^Ma3Ma6	Ma2	On the "Ma6" arm of the *N*-glycan core. No A,F,NN in the glycan
7	ALG9	Ma3[^Ma6]Ma6	Ma2	On the "Ma6" arm of the *N*-glycan core. No A,F,NN in the glycan
8	ALG10	^Ga3Ga3Ma2Ma2	Ga2	On the "Ma3" arm of the *N*-glycan core. No A,F,NN in the glycan
9	ALG11	^Ma3	Ma2	On the "Ma3" arm of the *N*-glycan core. No A,F,NN in the glycan
10	ALG11	^Ma2Ma3	Ma2	On the "Ma3" arm of the *N*-glycan core. No A,F,NN in the glycan
11	ALG12	Ma3[^]Ma6	Ma6	On the "Ma6" arm of the *N*-glycan core. No A,F,NN in the glycan
12	B3GALT1, B3GALT2	^GNb	Ab3	Not on the bisecting GlcNAc
13	B3GAT1, B3GAT2	^Ab4GN	Ub3	-
14	B3GNT1	^Ab4GN	GNb3	-
15	B3GNT7	^AN	GNb3	-
16	B4GALNT3, B4GALNT4	^GN	ANb4	Not on the bisecting GlcNAc

Table 3.2: A list of genes and their synthetic activities used in our catalog. A caret represents the insertion position and square brackets represent a branching position.

No.	Gene Name	Substrate	Donor residue	Condition rule
17	B4GALT1, B4GALT2	^Ga	Ab4	-
18	B4GALT1, B4GALT2, B4GALT3, B4GALT4	^GN	Ab4	Not on the bisecting GlcNAc
19	FUT1, FUT2	^Ab3GNb	Fa2	-
20	FUT1, FUT2	^Ab4GNb	Fa2	-
21	FUT3	Ab3[^]GNb	Fa4	Fa2 or NNa3 can connect on Ab3
22	FUT3, FUT4, FUT5, FUT6, FUT9	Ab4[^]GNb	Fa3	Fa2 or NNa3 is not connected on Ab4
23	FUT3, FUT4, FUT5, FUT6, FUT7	NNa3Ab4[^]GNb	Fa3	Fa2 is not connected on Ab4
24	FUT3, FUT4, FUT5, FUT6, FUT9	Fa2Ab4[^]GNb	Fa3	NNa3 is not connected on Ab4
25	FUT8	GNb4[^]GN	Fa6	On the *N*-glycan core (root). No Ab in the glycan
26	FUT10, FUT11	GNb4[^]GN	Fa3	On the *N*-glycan core (root). No Ab in the glycan
27	GCNT3	^Ab4GN	GNb6	-
28	GCNT3	Ab4GNb3[^]Ab	GNb6	-
29	GCNT3	Ab4GNb3[^]Ab	GNb6	-
30	MGAT1	^Ma3	GNb2	On the "Ma3" arm of the *N*-glycan core
31	MGAT2	^Ma6	GNb2	On the "Ma6" arm of the *N*-glycan core
32	MGAT3	Ma3[Ma6][^]Mb4	GNb4	On the *N*-glycan core. No bisecting GlcNAc and Ab in the glycan.

Table 3.2: A list of genes and their synthetic activities used in our catalog. A caret represents the insertion position and square brackets represent a branching position.

No.	Gene Name	Substrate	Donor residue	Condition rule
33	MGAT4A, MGAT4B, MGAT4C	GNb2[ˆ]Ma3	GNb4	On the “Ma3” arm of the *N*-glycan core. No bisecting GlcNAc in the glycan.
34	MGAT4A, MGAT4B, MGAT4C	GNb2[GNb6][ˆ]Ma3	GNb4	On the “Ma3” arm of the *N*-glycan core. No bisecting GlcNAc in the glycan.
35	MGAT5	GNb2[ˆ]Ma6	GNb6	On the “Ma6” arm of the *N*-glycan core. No bisecting GlcNAc in the glycan.
36	MGAT5	GNb2[GNb4][ˆ]Ma6	GNb6	On the “Ma6” arm of the *N*-glycan core. No bisecting GlcNAc in the glycan.
37	MGAT5	GNb2[ˆ]Ma3	GNb6	On the “Ma3” arm of the *N*-glycan core. No bisecting GlcNAc in the glycan.
38	MGAT5	GNb2[GNb4][ˆ]Ma3	GNb6	On the “Ma3” arm of the *N*-glycan core. No bisecting GlcNAc in the glycan.
39	ST3GAL1, ST3GAL2, ST3GAL4	ˆAb3	NNa3	-
40	ST3GAL3, ST3GAL4, ST3GAL6	ˆAb4GN	NNa3	-
41	ST6GAL1, ST3GAL2	ˆAb4GN	NNa6	-
42	ST8SIA1, ST8SIA3, ST8SIA4, ST8SIA5, ST8SIA6	ˆNNa3Ab	NNa8	-

Table 3.2: A list of genes and their synthetic activities used in our catalog. A caret represents the insertion position and square brackets represent a branching position.

No.	Gene Name	Substrate	Donor residue	Condition rule
43	ST8SIA2, ST8SIA3, ST8SIA4, ST8SIA5	ˆNNa8NNa3	NNa8	-
44	ST8SIA2, ST8SIA3, ST8SIA4, ST8SIA5, ST8SIA6	ˆNNa6Ab	NNa8	-
45	ST8SIA2, ST8SIA3, ST8SIA4, ST8SIA5	ˆNNa8NNa6	NNa8	-

3.3.3 Theoretical *N*-glycan database

Table 3.3 lists the number of *N*-glycan structures generated by our deductive method, based on the sugar-transition catalogue described in Table 3.2. In this table, we list the number of structures generated based on the number of monosaccharides in a glycan (size) as well as the number of reactions in which the glycan was used as a substrate for the next monosaccharide addition. In total, by using the constraints defined in Table 3.2 and by restricting the size of the *N*-glycan to fifteen monosaccharide residues (corresponding to approximately 90% of glycans in UniCarbKB), we were able to generate almost 1.1 million potential N-glycan structures based on our model. These were generated using over 4.7 million enzyme reactions, and these are stored in the theoretical N-glcyan database. Figure 3.10 shows an example of the result of our theoretical pathway model. These structures are stored in our database with identifier and its size. Moreover, precursor and produced structure and related enzyme name are also stored. Those generated structures are used for calculating the subsequent products.

Next, we compared the theoretically generated N-glycan structures with the literature described structures listed in UniCarbKB. The structure listings were filtered based on the following conditions; i) structure must contain the chitobiose core, ii) the compositions are restricted to common human N-glycan monosaccharide residues (not containing sulfates, phosphates, methyl groups or non-human monosaccharides), iii) the structure size is between five and fifteen monosaccharide residues, and iv) only defined structures (whereby glycosidic linkages are not fuzzy) are compared. Interestingly, only 310 unique structures (582 in total), ranging from fully defined elongated biantennary to fully sialylated penta-antennary structures, stored in UniCarbKB were matched against our database - demonstrating the huge gulf between the observed structures reported in the literature and the theoretically possible structures.

3.4 Discussions

In this manuscript, we have described a computational model to generate over 1.1 million theoretically possible N-glycan structures based on a strict set of biosynthetic rules, which have been obtained from that described in the established databases KEGG Gly-

Table 3.3: A summary of the number of glycan structures and reactions determined by the described model and classified by monosaccharide size 5>x<15 residues. The 'reactions' value corresponds to the number of enzymatic reactions between each composition size, for example, 134 putative substrate-glycosyltransferase reactions were recorded between glycan composition size 7 and 8.

Monosaccharide composition size	Number of generated structures	Reactions
5	1	7
6	7	37
7	29	152
8	115	646
9	461	2,966
10	1,822	13,351
11	7,094	57,388
12	26,964	236,385
13	99,762	939,249
14	360,793	3,623,307
15	1,280,472	-
Total	1,777,520	4,873,488

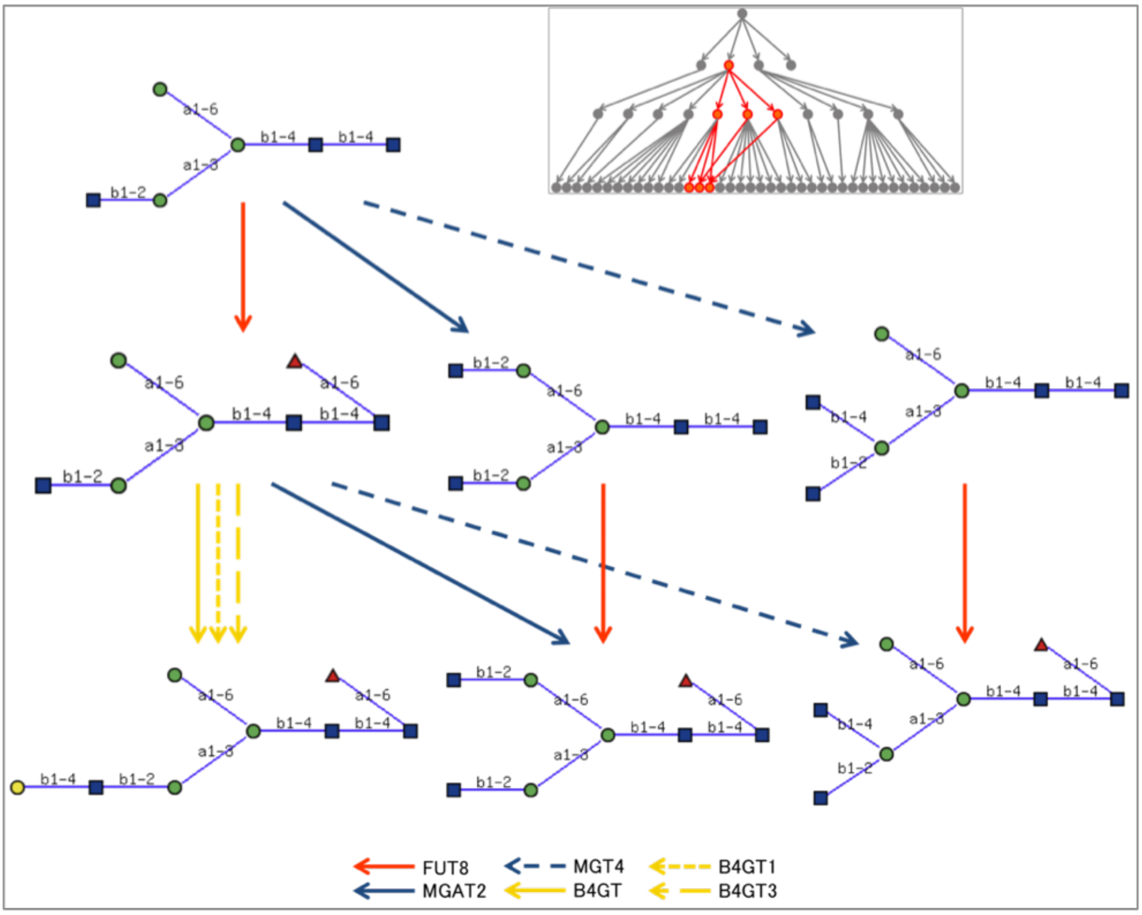


Figure 3.10: These structures are stored in our database with identifier and its size. Moreover, precursor and produced structure and related enzyme name are also stored.

can, CFG, CAZy, GGDB and BRENDA. To constrain the output, a maximum cutoff of 15 monosaccharides was imposed, which is representative of the largest composition typically reported in the majority of structure databases including UniCarbKB. Even by constraining the maximum size, the predicted library has more than 500 times the number of glycan structures that have been fully characterized experimentally (including all linkages) on eukaryotic glycoproteins, as there are only approximately 2,000 fully-defined N-glycans described in existing glycan databases.

To explore this difference in number, we compared the structures in the theoretically generated glycan database with the experimentally determined structures reported in UniCarbKB. Surprisingly, approximately half of all the experimentally determined N-glycan structures in UniCarbKB were not contained in our theoretically generated human pathway based data. This may be explained by the associated biological source, with many literature based reported structures derived from non-human glycoproteins or biological fluids. For example, a high-mannose glycan (bearing more than nine mannose residues) with a terminal α1,2 galactose, which was found on yeast proteins [55] and a bisecting GlcNAc with a terminal α1,3 galactose, which was reported in non-primate mammalian and new world monkey glycoproteins [56], are not found in the theoretical database. In addition, some glycans reported in UniCarbKB that are not predicted in our database are not a product of known glycan biosynthetic pathway rules. For example, the experimentally determined GlcNAc(β1-6)Man(α1-6)[Man(α1-3)]Man(β1-4)GlcNAc(β1-4)[Fuc(α1-6)]GlcNAc in UniCarbKB (ID=911) may be the result of degradation during sample preparation, in source fragmentation, incorrect structure assignment, glycosidase action or indeed the product of a previously unreported pathway. Based on our library, GlcNAc(β1-6) may be transferred by MGAT5 or GCNT3 but both are reported to require substrates that are part of this structure. Therefore, our theoretical N-glycan database does not contain this structure because the substrate does not satisfy any of our reaction rules.

Glycosylation reactions are dependent on many variables such as the concentrations of enzymes and availability of precursor-sugars, as well as reaction kinetics and enzyme location constraints. Furthermore, some synthetic reactions may occur so rapidly that the intermediate products may not be detected by current technologies. Figure 3.11 shows a snapshot of the glycan biosynthetic pathway using structure entries from our theoretical database and experimentally reported UniCarbKB databases. In this example UniCarbKB has no information associated with Man(α1-6) [Gal(β1-3)GlcNAc(β1-2)Man(α1-3)] Man(β1-4)GlcNAc(β1-4) [Fuc(α1-6)] GlcNAc (structure B) or GlcNAc(β1-2)Man(α1-6) [Gal(β1-3)GlcNAc(β1-2)Man(α1-3)] Man(β1-4)GlcNAc(β1-4) [Fuc(α1-6)] GlcNAc (structure E), but the products (C) and (F) are listed. However information is available for two fuzzy structures such like Gal(β1-?)GlcNAc(β1-?)Man(α1-?) [Man(α1-?)] Man(β1-4)GlcNAc(β1-4) [Fuc(??-?)] GlcNAc (UniCarbKB ID: 1534) and Gal(??-?)GlcNAc(??-?)Man(α1-?) [GlcNAc(??-?)Man(α1-?)] Man(β1-4)GlcNAc(β1-4) [Fuc(??-6)] GlcNAc (UniCarbKB ID: 1797), that may be suitable descriptors for these fully defined structures.

Unlike genomics and proteomics, glycomics has no template strategies to completely determine a glycan structure with all sequence and linkage information defined. Furthermore, when we considered enzyme activities and substrate specificities, it was not possible to assign kinetic rates of activity because of the lack of information in the literature and the variation imposed by the sample conditions and cellular concentrations of donor and acceptors. We know from the BRENDA database the enzyme activity and

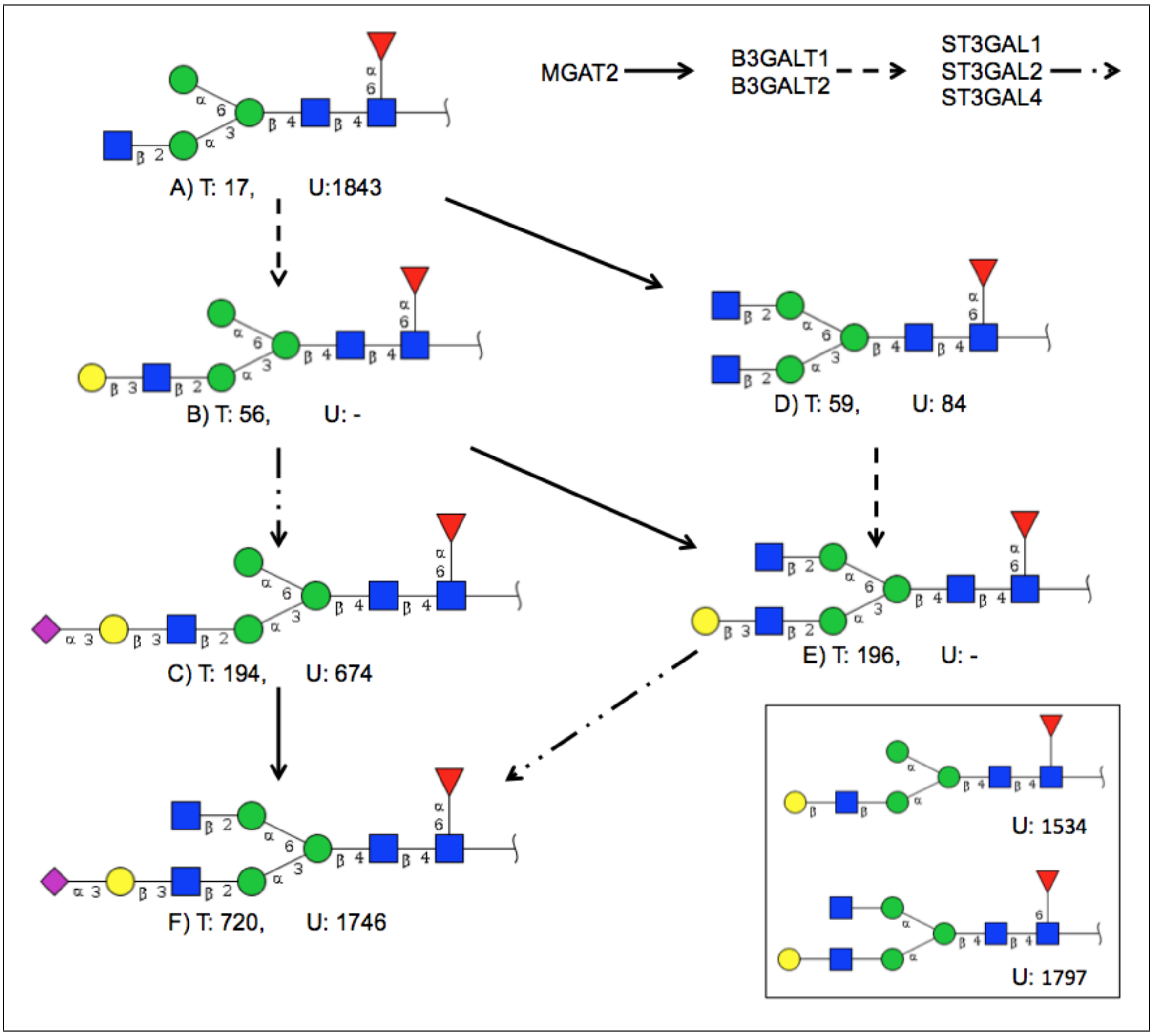


Figure 3.11: An example of the glycan synthetic pathway model used to generate the theoretical *N*-glycan database. A glycan is represented with the accession numbers from the theoretical database (T) alongside the matching UniCarbKB (U) entry. Structures "B" and "E" are not registered in UniCarbKB, however, they can be represented by two UniCarbKB entries (1534 and 1797) that are missing some glycosidic bond linkage information.

reaction rates vary widely depending on the environmental and experimental conditions for the same enzyme [57]. For example, the specific activity of MGAT3 in normal liver is $1.3 * 10^{-7}[\mu mol/min/mg]$ compared to $1.24 * 10^{-5}[\mu mol/min/mg]$ in hepatic carcinoma cells [58]. Thus, the fact that many glycan structures that are the substrates of these enzymes are not found in experimentally derived databases may be due to the low concentrations and high turnovers that cannot be detected by current technologies.

We also analysed the usage frequency of glycosyltransferases used in our model (Fig. 3.12). MGAT3, which adds a bisecting GlcNAc, and FUT8, which adds a fucose to the core structure, have smaller usage frequencies compared to the other enzymes because of the characteristic feature that they only add single monosaccharides to existing oligosaccharide structures. Furthermore, glycosyltransferases responsible for attaching terminal residues (e.g sialyltransferases) are more frequent in the synthesis of the larger glycans. Therefore, these results correspond well with known facts about the N-glycan biosynthesis pathway.

We further compared glycan structure stored in UniCarb KB and UniCorn database. The number of N-glycan in UniCarb KB is 3,224 (2014 data). Within these structures, 659 structures are fully determined N-glycans, and we were able to find the half of these glycans in UniCorn database. Ten percent of the unmatched glycans were formed by more than 15 monosaccharides. 50 percent of the unmatched glycans were experimentally determined in non-human species. 40 percent of the unmatched glycans contained a specific glycan moieties which were not covered by our glycosyltransferase catalog. For example, glycan substructures "Fuc(α1-6)Gal(β1-4)" and "GlcNAc(β1-4)Man(α1-6)" are not able to be generated using our catalog. These moieties may become a key role for

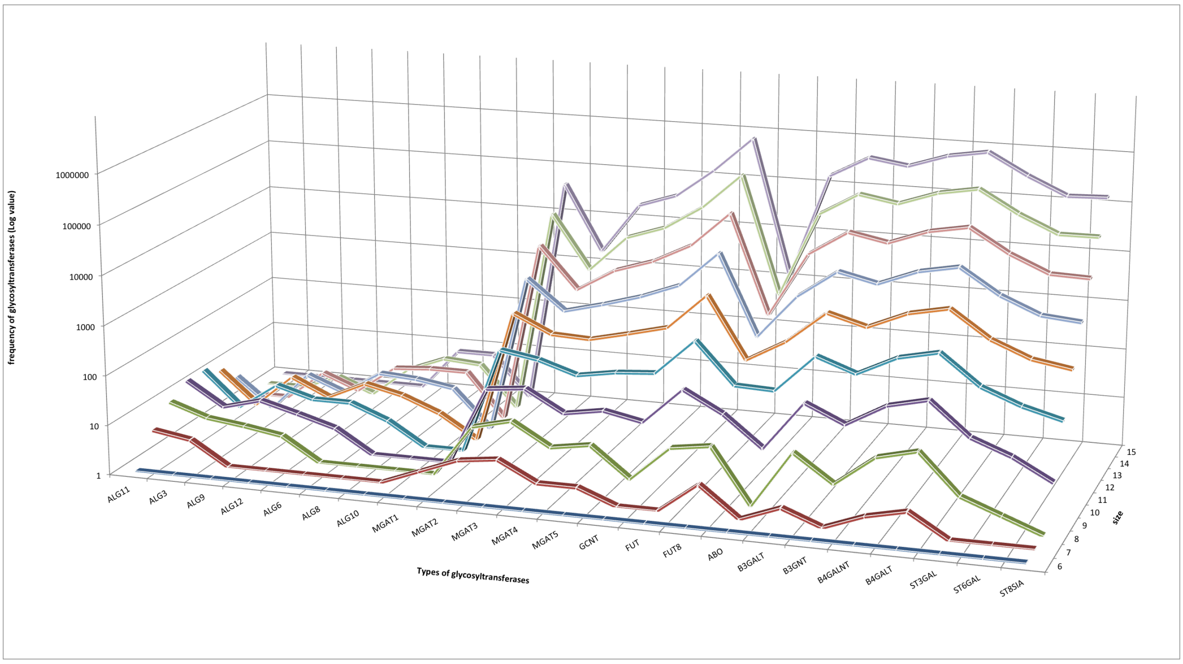


Figure 3.12: A plot of the frequency of glycosyltransferase usage. Glycosyltransferases are grouped based on their gene types (x-axis). For example, the label "MGAT4" represents the sum of the frequency of MGAT4A, B and C. Note that FUT represents the fucosyltransferases other than FUT8. The size (z-axis) represents the size of the glycan structure generated.

understanding unknown regulatory systems.

It is clear that there are many unknown regulatory systems in biological cells that strictly control protein glycosylation patterns. Our comprehensive glycan pathway analysis highlights the gaps between glycobiological analysis *in vivo* and *in silico*. The availability of a predicted collection of such glycan structural data may facilitate increased glycomics knowledge and enable easier adoption by other disciplines. Moreover, the prediction of these "missing" glycans may also provide some light into the enzyme kinetics of the glycosyltransferases involved. Clearly our theoretical N-glycan database is not fully supported by the experimentally determined structures at this time. It is not known how many N-glycan structures actually exist in nature, and we acknowledge that the structures predicted and stored in our database are more than two orders of magnitude greater than currently reported structures. As glyco-analysis technologies improve, our database can be used to validate newly assigned structures with the knowledge that they are biosynthetically plausible.

Chapter 4

Conclusion

Calculation models have been developed to aid reducing costs and saving time for the processes of predicting and synthesizing glycans. In the last decade, these models are used to predict the industrial relevance of glycan structures. Computing models for glycan synthesis will be further extended and sophisticated to affect the detailed analysis of glycan-omics research. Moreover, it will be one of the valuable resources for the glycan engineering. Spahn and Lewis was discussed that the development of small glycan synthesis model which is specialized in the synthesis of glycans on the specific recombinant protein would be of the one of the most interest in a number of approaches [59]. For instance, Liu *et al.* have focused on the O-glycan modification of the glycoprotein ligand I of P-selectin, and have developed the optimization technique to reconstruct the reaction pathway as the most consistent with the observed glycan abundance [60]. We suggest that these small-scale models are significant to reduce the complexity of glycan structure researches, and it is able to develop the lager-scale models for the glyan-omics anlaysis.

This study is a first step to fill the large gap of the glycan science based on development of the first web resource that contains data mining tools and algorithms focusing on the sugar chain structure. Several useful tools have become available for glycobiology analysis. Furthermore, we have developed a comprehensive human-related N-glycan synthesis pathway based on utilizing the algorithm of Glycan Pathway Predictor tool. We were able to generate 1.1 million theoretical N-glycans that are mostly not reported. Hence, we were able to highlight the large gaps between the glycan science *in vivo* and *in silico*, and considered to be an important role studies to fill these gaps.

Acknowledgements

My work was supported by KAKENHI (20016025) Grant-in-Aid for Young Scientists (A), Japan Society for the Promotion of Science (JSPS), and the Ministry of Education, Culture, Sports, Science and Technology (MEXT) and the Macquarie University Research Excellence Scheme Postgraduate Co-tutelle Scholarship.

I would like to thank my adviser in Soka University, Prof. Kiyoko F. Kinoshita, and in Macquarie University, Prof. Nicolle H. Packer and Dr. Matthew P. Campbell. I could not finish my degree without their continual encouragements and advices. I thank the members of Kinoshita laboratory and Nicki's group for helping my research. I was able to spend my days with enjoying and laugh. Masae helped my research about Glycan Score Matrices by developing MCAW algorithm. Jodie gave me advices about mass spectrometry. Chi-Hung gave me advices about glycosyltransferases.

I would like to give my appreciation to my mother for all the support she have given to me.

I would like to thank the founder and his wife, Dr. Daisaku Ikeda and Mrs. Ikeda, of Soka University for providing supports in a number of ways throughout my days in university.

Bibliography

[1] N. Taniguchi, T. Endo, W.G. Hart, H.P. Seeberger, and C.H. Wong. *Glycoscience: Biology and Medicine*, volume 1, book 1. Springer, 2015.

[2] K. Ohtsubo and D.J. Marth. Glycosylation in cellular mechanisms of health and disease. *Cell*, 126(5):855–867, 2006.

[3] E.B. Collins and C.J. Paulson. Cell surface biology mediated by low affinity multivalent protein–glycan interactions. *Current Opinion in Chemical Biology*, 8(6):617–625, 2004.

[4] Y. Imai, M.S. Singer, C. Fennie, L.A. Lasky, and S.D. Rosen. Identification of a carbohydrate-based endothelial ligand for a lymphocyte homing receptor. *J Cell Biol*, 113(5):1213–21, 1991.

[5] B. Wang and G.J. Boons. *Carbohydrate Recognition: Biological Problems, Methods, and Applications*, book 1. Wiley, 2011.

[6] S.D. Rosen. Endothelial ligands for l-selectin - from lymphocyte recirculation to allograft. *Am. J. Pathol*, 155:1013–1020, 1999.

[7] R. Kannagi and A. Kanamori. Glycobiology of sialyl 6-sulfo lewis x, a new carbohydrate ligand for selectins. *Trends Glycosci. Glycotechnol*, 11:329–344, 1999.

[8] U.M. Abd Hamid, L. Royle, R. Saldova, C.M. Radcliffe, D.J. Harvey, S.J. Storr, M. Pardo, R. Antrobus, C.J. Chapman, N. Zitzmann, J.F. Robertson, R.A. Dwek, and P.M. Rudd. A strategy to reveal potential glycan markers from serum glycoproteins associated with breast cancer progression. *Glycobiology*, 18(12):1105–1118, 2008.

[9] M. Kanehisa and S. Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27–30, 2000.

[10] Kanehisa M., Goto S., Sato Y., Furumichi M., and Tanabe M. Kegg for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res*, 40(Database issue):D109–114, 2012.

[11] M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe. Data, information, knowledge and principle: back to metabolism in kegg. *Nucleic Acids Res*, 42(Database issue):D199–205, 2014.

[12] R. Raman, M. Venkataraman, S. Ramakrishnan, W. Lang, S. Raguram, and R. Sasisekharan. Advancing glycomics: implementation strategies at the consortium for functional glycomics. *Glycobiology*, 16(5):82R–90R, 2006.

[13] René Ranzinger, Stephan Herget, Claus-Wilhelm von der Lieth, and Martin Frank. Glycomedb - a unified database for carbohydrate structures. *Nucleic Acids Research*, 39:D373–D376, 2011.

[14] Y. Akune, M. Hosoda, S. Kaiya, D. Shinmachi, and K. F. Aoki-Kinoshita. The rings resource for glycome informatics analysis and data mining on the web. *OMICS*, 14(4):475–86, 2010.

[15] M. P. Campbell, C. A. Hayes, W. B. Struwe, M. R. Wilkins, K. F. Aoki-Kinoshita, D. J. Harvey, P. M. Rudd, D. Kolarich, F. Lisacek, N. G. Karlsson, and N. H. Packer. Unicarbkb: putting the pieces together for glycomics research. *Proteomics*, 11(21):4117–21, 2011.

[16] M. P. Campbell, R. Peterson, J. Mariethoz, E. Gasteiger, Y. Akune, K. F. Aoki-Kinoshita, F. Lisacek, and N. H. Packer. Unicarbkb: building a knowledge platform for glycoproteomics. *Nucleic Acids Res*, 42(Database issue):D215–21, 2014.

[17] F. J. Krambeck and M. J. Betenbaugh. A mathematical model of n-linked glycosylation. *Biotechnol Bioeng*, 92(6):711–28, 2005.

[18] F. J. Krambeck, S. V. Bennun, S. Narang, S. Choi, K. J. Yarema, and M. J. Betenbaugh. A mathematical model to derive n-glycan structures and cellular enzyme activities from mass spectrometric data. *Glycobiology*, 19(11):1163–75, 2009.

[19] A.D. McNaught and A. Wilkinson. *Compendium of chemical terminology*, page 1. Wiley, 1997.

[20] Ajit Varki, Richard D. Cummings, Jeffrey D. Esko, Hudson H. Freeze, Pamela Stanley, Carolyn R. Bertozzi, Gerald W. Hart, and Marilynn E. Etzler, editors. *Essentials of Glycobiology*, chapter 2. Cold Spring Harbor Laboratory Press, second edition, 2009.

[21] Brooks S., Dwek M., and Schumacher U. *Functional and Molecular Glycobiology*, book 1. Bios Scientific Pub Ltd, 2002.

[22] von der Lieth C.W., Luetteke T., and Frank M. *Bioinformatics for Glycobiology and Glycomics: An Introduction*, book 2. Wiley, 2009.

[23] S.B. Agravat, J.H. Saltz, R.D. Cummings, and D.F. Smith. Glycopattern: a web platform for glycan array mining. *Bioinformatics*, 30(23):3417–3418, 2014.

[24] F. Li, O. V. Glinskii, and V. V. Glinsky. Glycobioinformatics: current strategies and tools for data mining in ms-based glycoproteomics. *Proteomics*, 13(2):341–54, 2013.

[25] H. Lis and N. Sharon. Lectins: Carbohydrate-specific proteins that mediate cellular recognition. *Chem Rev*, 98(2):637–674, 1998.

[26] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–402, 1997.

[27] W.R. Pearson. Finding protein and nucleotide similarities with fasta. *Curr Protoc Bioinformatics*, 53:3.9.1–3.9.25, 2016.

[28] K. F. Aoki, A. Yamaguchi, N. Ueda, T. Akutsu, H. Mamitsuka, S. Goto, and M. Kanehisa. Kcam (kegg carbohydrate matcher): a software tool for analyzing the structures of carbohydrate sugar chains. *Nucleic Acids Res*, 32(Web Server issue):W267–72, 2004.

[29] S. Herget, R. Ranzinger, K. Maass, and C. W. Lieth. Glycoct-a unifying sequence format for carbohydrates. *Carbohydr Res*, 343(12):2162–71, 2008.

[30] Ehud Banin, Yael Neuberger, Yaniv Altshuler, Asaf Halevi, On Inbar, Dotan Nir, Avinoam Dukler, and Ken-ichi Kasai. A novel linear code((r)) nomenclature for complex carbohydrates. *Trends in Glycoscience and Glycotechnology*, 14(77):127–137, 2002.

[31] A. D. McNaught. Nomenclature of carbohydrates (recommendations 1996). *Adv Carbohydr Chem Biochem*, 52:43–177, 1997.

[32] K. F. Aoki, H. Mamitsuka, T. Akutsu, and M. Kanehisa. A score matrix to reveal the hidden links in glycans. *Bioinformatics*, 21(8):1457–63, 2005.

[33] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, 89(22):10915–9, 1992.

[34] Y. Hizukuri, Y. Yamanishi, O. Nakamura, F. Yagi, S. Goto, and M. Kanehisa. Extraction of leukemia specific glycan motifs in humans by computational glycomics. *Carbohydr Res*, 340(14):2270–2278, 2005.

[35] T. Kuboyama, K. Hirata, K. F. Aoki-Kinoshita, H. Kashima, and H. Yasuda. A gram distribution kernel applied to glycan classification and motif extraction. *Genome Inform*, 17(2):25–34, 2006.

[36] A. Zien, G. Rätsch, S. Mika, B. Schölkopf, T. Lengauer, and K.R. Müller. Engineering support vector machine kernels that recognize translation initiation sites. *Bioinformatics*, 16(9):799–807, 2000.

[37] H. Jiang, K.F. Aoki-Kinoshita, and W.K. Ching. Extracting glycan motifs using a biochemicallyweighted kernel. *Bioinformation*, 7(8):405–412, 2011.

[38] M. Hattori, N. Tanaka, M. Kanehisa, and S. Goto. Simcomp/subcomp: chemical structure search servers for network analyses. *Nucleic Acids Res.*, 38:(Web Server issue)W652–656, 2010.

[39] Masae Hosoda, Yukie Akune, and Flora Kiyoko Aoki-Kinoshita. Multiple tree alignment with weights applied to carbohydrates to extract binding recognition patterns. *Pattern Recognition in Bioinformatics*, pages 49–58, 2012.

[40] W.M. Fitch and E. Margoliash. Construction of phylogenetic trees. *Science*, 155(3760):279–284, 1967.

[41] `http://www.functionalglycomics.org/glycomics/SampleServlet?operationType=view&key=12`.

[42] S.L. Orr, D. Le, J.M. Long, P. Sobieszczuk, B. Ma, H. Tian, X. Fang, J.C. Paulson, J.D. Marth, and N. Varki. A phenotype survey of 36 mutant mouse strains with gene-targeted defects in glycosyltransferases or glycan-binding proteins. *Glycobiology*, 23(3):363–380, 2013.

[43] S. Parry, V. Ledger, B. Tissot, S.M. Haslam, J. Scott, Morris H. R., and A. Dell. Integrated mass spectrometric strategy for characterizing the glycans from glycosphingolipids and glycoproteins: direct identification of sialyl le(x) in mice. *Glycobiology*, 17:645–654, 2007.

[44] P. Umaña and J. E. Bailey. A mathematical model of n-linked glycoform biosynthesis. *Biotechnol Bioeng*, 55(6):890–908, 1997.

[45] X. Li, X. Wang, Z. Tan, S. Chen, and F. Guan. Role of glycans in cancer cells undergoing epithelial-mesenchymal transition. *Front Oncol*, 6(33):1–5, 2016.

[46] Q. Xu, T. Isaji, Y. Lu, W. Gu, M. Kondo, T. Fukuda, Y. Du, and J. Gu. Roles of n-acetylglucosaminyltransferase iii in epithelial-to-mesenchymal transition induced by transforming growth factor β1 (tgf-β1) in epithelial cell lines. *J Biol Chem*, 287(20):16563–16574, 2012.

[47] K. Hashimoto, S. Kawano, S. Goto, K. F. Aoki-Kinoshita, M. Kawashima, and M. Kanehisa. A global representation of the carbohydrate structures: a tool for the analysis of glycan. *Genome Inform*, 16(1):214–22, 2005.

[48] K. Hashimoto, S. Goto, S. Kawano, K. F. Aoki-Kinoshita, N. Ueda, M. Hamajima, T. Kawasaki, and M. Kanehisa. Kegg as a glycome informatics resource. *Glycobiology*, 16(5):63R–70R, 2006.

[49] V. Lombard, H. Golaconda Ramulu, E. Drula, P. M. Coutinho, and B. Henrissat. The carbohydrate-active enzymes database (cazy) in 2013. *Nucleic Acids Res*, 42(Database issue):D490–5, 2014.

[50] I. Schomburg, A. Chang, C. Ebeling, M. Gremse, C. Heldt, G. Huhn, and D. Schomburg. Brenda, the enzyme database: updates and major new developments. *Nucleic Acids Res*, 32(Database issue):D431–3, 2004.

[51] Akira Togayachi, Kwon-Yeon Dae, Toshihide Shikanai, and Hisashi Narimatsu. *A Database System for Glycogenes (GGDB)*, pages 423–425. Springer Japan, 2008.

[52] UniProt Consortium. Uniprot: a hub for protein information. *Nucleic Acids Res*, 43(Database issue):D204–12, 2015.

[53] M. P. Campbell and N. H. Packer. Unicarbkb: New database features for integrating glycan structure abundance, compositional glycoproteomics data, and disease associations. volume 16.

[54] K. Inamori, T. Endo, Y. Ide, S. Fujii, J. Gu, K. Honke, and N. Taniguchi. Molecular cloning and characterization of human GnT-IX, a novel beta1,6-N-acetylglucosaminyltransferase that is specifically expressed in the brain. *J Biol Chem*, 278(44):43102–9, 2003.

[55] A. Flores-Carreón, S. H. Hixson, A. Gómez, M. C. Shao, G. Krudy, P. R. Rosevear, and F. Wold. The processing of n-linked glycans in yeast. mutually exclusive steps in the processing of a man6 derivative by yeast membrane preparations. *J Biol Chem*, 265(2):754–9, 1990.

[56] U. Galili. The alpha-gal epitope and the anti-gal antibody in xenotransplantation and in cancer immunotherapy. *Immunol Cell Biol*, 83(6):674–86, 2005.

[57] H.B. Guo, A.L. Jiang, T.Z. Ju, and H.L. Chen. Opposing changes in n-acetylglucosaminyltransferase-v and -iii during the cell cycle and all-trans retinoic acid treatment of hepatocarcinoma cell line. *Biochim Biophys Acta*, 1495(3):297–307, 2000.

[58] E.Y. Song, S.K. Kang, Y.C. Lee, Y.G. Park, T.H. Chung, D.H. Kwon, S.M. Byun, and C.H. Kim. Expression of bisecting n-acetylglucosaminyltransferase-iii in human hepatocarcinoma tissues, fetal liver tissues, and hepatoma cell lines of hep3b and hepg2. *Cancer Invest*, 19(8):799–807, 2001.

[59] P. N. Spahn and N. E. Lewis. Systems glycobiology for glycoengineering. *Curr Opin Biotechnol*, 30:218–24, 2014.

[60] G. Liu, D. D. Marathe, K. L. Matta, and S. Neelamegham. Systems-level modeling of cellular glycosylation reaction networks: O-linked glycan formation on natural selectin ligands. *Bioinformatics*, 24(23):2740–7, 2008.

List of Tables

List of Figures

Appendix A

Source codes

Source codes for each algorithm in this thesis is uploaded in Bitbucket. In the following sections, each source codes are saved in the listed URL(s) of Bitbucket.

Glycan Score Matrix

Source codes for calculating a glycan score matrix are stored in:

`https://yukie_akune@bitbucket.org/appendix_ya/score-matrix-generator.git`

"scorematrix_index.pl" is used for displaying the input. "score_matrix.pl" is a main program for calculating glycan score matrices, and used for output as well.

Glycan Kernel Tool

Source codes for Glycan Kernel Tool are stored in: `https://yukie_akune@bitbucket.org/appendix_ya/glycankerneltool.git`

"kernel_Input_viewer.pl" is used for displaying the input. "kernel-tool.pl" is used for generating a calculation id, and the calculation process is displayed via “kernel-request.pl”. “bioweightedq.pl” is a main code, and it uses “run_Kernel_Comp.sh”, “getqgrams.sh”, “getqgram_file.rb” and “Linkagesimi.mat” for kernel calculation. The result page is displayed using “show-results.pl”.

Our kernel calculation is coded in Matlab. Source codes of Matlab are stored in: `https://yukie_akune@bitbucket.org/appendix_ya/matlab-codes.git`

IUPACtoKCF

Source codes for IUPACtoKCF are stored in:

`https://yukie_akune@bitbucket.org/appendix_ya/iupactokcf.git`

“iupactokcf_index_au.pl” is used for displaying the input. “iupactokcf_au.pl” is a main program for converting input data, and used for output as well.

GlycoCT{condensed}toKCF

Source codes for GlycoCT{condensed}toKCF are stored in:

`https://yukie_akune@bitbucket.org/appendix_ya/glycocttokcf.git`

“glycoct_index_au.pl” is used for displaying the input. “glycoct_to_kcf_au.pl” is a main program for converting input data, and used for output as well. “transNODEs_uniq.txt” is a list of monosaccharides written in GlycoCT format and KCF formats.

Generating glycosyltransferase candidates

Source codes for generating glycosyltransferase candidates are stored in:

`https://yukie_akune@bitbucket.org/appendix_ya/get_enzyme_candidates.git`

"get_Nenz_JSON.pl" is a main program for generating glycosyltransferase candidates. "transNODEs_uniq.txt" is a list of monosaccharides written in GlycoCT format and KCF formats.

Theoretical *N*-glycan database

Source codes for generating glycosyltransferase candidates are stored in:

`https://yukie_akune@bitbucket.org/appendix_ya/theoretical-db.git`

"gp.pl" is a main program for calculating theretical *N*-glycans using human glycosyltransferases which data is stored in "Nenzyme_local.txt". "sqlCodes.txt" saves codes for generating database using MySQL language.

Appendix B

Glycosyltransferase Catalog

Table B.1: Glycosyltransferase catalog (reaction patterns)

No.	Substrate	Donor residue	Condition rule
1	ˆMan(a1-6)	Man(a1-3)	On the core M̈an(a1-6)ärm. No Gal/Fuc/Neu5Ac in the parent glycan
2	ˆMan(a1-2)Man(a1-2)Man(a1-3)	Glc(a1-3)	On the core M̈an(a1-3)ärm. No Gal/Fuc/Neu5Ac in the parent glycan
3	ˆGlc(a1-3)Man(a1-2)Man(a1-2)Man(a1-3)	Glc(a1-3)	On the core M̈an(a1-3)ärm. No Gal/Fuc/Neu5Ac in the parent glycan
4	ˆMan(a1-3)Man(a1-6)	Man(a1-2)	On the core M̈an(a1-6)ärm. No Gal/Fuc/Neu5Ac in the parent glycan
5	Man(a1-3)[ˆMan(a1-6)]Man(a1-6)	Man(a1-2)	On the core M̈an(a1-6)ärm. No Gal/Fuc/Neu5Ac in the parent glycan
6	ˆGlc(a1-3)Glc(a1-3)Man(a1-2)Man(a1-3)	Glc(a1-2)	On the core M̈an(a1-3)ärm. No Gal/Fuc/Neu5Ac in the parent glycan
7	ˆMan(a1-3)	Man(a1-2)	On the core M̈an(a1-3)ärm. No Gal/Fuc/Neu5Ac in the parent glycan
8	ˆMan(a1-2)Man(a1-3)	Man(a1-2)	On the core M̈an(a1-3)ärm. No Gal/Fuc/Neu5Ac in the parent glycan
9	Man(a1-3)[ˆ]Man(a1-6)	Man(a1-6)	On the core M̈an(a1-6)ärm. No Gal/Fuc/Neu5Ac in the parent glycan
10	ˆGal(b1-4)GlcNAc	GlcNAc(b1-3)	-
11	ˆGalNAc	GlcNAc(b1-3)	-

Table B.1: Glycosyltransferase catalog (reaction patterns)

No.	Substrate	Donor residue	Condition rule
12	^GlcNAc(b1-	Gal(b1-3)	Not on the bisecting GlcNAc
13	^Glc(a1-	Gal(b1-4)	-
14	^GlcNAc	Gal(b1-4)	Not on the bisecting GlcNAc
15	^Gal(b1-4)	GalNAc(b1-4)	-
16	^GlcNAc	GalNAc(b1-4)	Not on the bisecting GlcNAc
17	Fuc(a1-2)[^]Gal(b1-	GalNAc(a1-3)	-
18	Fuc(a1-2)[^]Gal(b1-	Gal(b1-3)	-
19	^Gal(b1-4)GlcNAc(b1-	Fuc(a1-2)	-
20	Gal(b1-3)[^]GlcNAc(b1-	Fuc(a1-4)	Fuc(a1-2) or Neu5Ac(a2-3) can connect on Gal(b1-3)
21	Gal(b1-4)[^]GlcNAc(b1-	Fuc(a1-3)	Fuc(a1-2) or Neu5Ac(a2-3) can connect on Gal(b1-3)
22	GlcNAc(b1-4)[^]GlcNAc	Fuc(a1-6)	On the core GlcNAc (root). No Gal (b1-) in the parent glycan
23	Gal(b1-4)[^]GlcNAc(b1-	Fuc(a1-3)	stop the extension
24	Gal(b1-3)[^]GlcNAc(b1-	Fuc(a1-3)	Fuc(a1-2) can connect on Gal(b1-3)
25	^Gal(b1-4)GlcNAc	GlcNAc(b1-6)	-
26	Gal(b1-4)GlcNAc(b1-3)[^]Gal(b1-	GlcNAc(b1-6)	-
27	^Man(a1-3)	GlcNAc(b1-2)	On the core Man(a1-3)
28	^Man(a1-6)	GlcNAc(b1-2)	On the core Man(a1-6)
29	Man(a1-3)[Man(a1-6)][^]Man(b1-4)	GlcNAc(b1-4)	On the core Man(b1-4). No bisecting GlcNAc and Gal (b1-) in the parent glycan
30	GlcNAc(b1-2)[^]Man(a1-3)	GlcNAc(b1-4)	On the core Man(a1-3). No bisecting GlcNAc in the parent glycan
31	GlcNAc(b1-2)[GlcNAc(b1-6)][^]Man(a1-3)	GlcNAc(b1-4)	On the core Man(a1-3). No bisecting GlcNAc in the parent glycan
32	GlcNAc(b1-2)[GlcNAc(b1-6)][^]Man(a1-6)	GlcNAc(b1-4)	On the core Man(a1-6)
33	GlcNAc(b1-2)[^]Man(a1-6)	GlcNAc(b1-4)	On the core Man(a1-6)
34	GlcNAc(b1-2)[^]Man(a1-6)	GlcNAc(b1-6)	On the core Man(a1-6). No bisecting GlcNAc

Table B.1: Glycosyltransferase catalog (reaction patterns)

No.	Substrate	Donor residue	Condition rule
35	GlcNAc(b1-2)[GlcNAc(b1-4)][^]Man(a1-6)	GlcNAc(b1-6)	On the core Man(a1-6). No bisecting GlcNAc
36	GlcNAc(b1-2)[^]Man(a1-3)	GlcNAc(b1-6)	On the core Man(a1-3). No bisecting GlcNAc in the parent glycan
37	GlcNAc(b1-2)[GlcNAc(b1-4)][^]Man(a1-3)	GlcNAc(b1-6)	On the core Man(a1-3). No bisecting GlcNAc in the parent glycan
38	^Gal(b1-3)	Neu5Ac(a2-3)	-
39	^Gal(b1-4)GlcNAc	Neu5Ac(a2-3)	-
40	^Gal(b1-4)GlcNAc	Neu5Ac(a2-6)	-
41	^Neu5Ac(a2-3)Gal(b1-	Neu5Ac(a2-8)	-
42	^Neu5Ac(a2-8)Neu5Ac(a2-3)	Neu5Ac(a2-8)	-
43	^Neu5Ac(a2-6)Gal(b1-	Neu5Ac(a2-8)	-
44	^Neu5Ac(a2-8)Neu5Ac(a2-6)	Neu5Ac(a2-8)	-

Table B.2: Glycosyltransferase catalog (protein and gene name)

No.	UniProt entry	UniProt protein name	UniProt gene name
1	Q92685	Dol-P-Man:Man(5)GlcNAc(2)-PP-Dol alpha-1,3-mannosyltransferase	ALG3
2	Q9Y672	Dolichyl pyrophosphate Man9GlcNAc2 alpha-1,3-glucosyltransferase	ALG6
3	Q9BVK2	Probable dolichyl pyrophosphate Glc1Man9GlcNAc2 alpha-1,3-glucosyltransferase	ALG8
4	Q9H6U8	Alpha-1,2-mannosyltransferase ALG9	ALG9
5	Q9H6U8	Alpha-1,2-mannosyltransferase ALG9	ALG9
6	Q5BKT4	Dol-P-Glc:Glc(2)Man(9)GlcNAc(2)-PP-Dol alpha-1,2-glucosyltransferase	ALG10
7	Q2TAA5	GDP-Man:Man(3)GlcNAc(2)-PP-Dol alpha-1,2-mannosyltransferase	ALG11
8	Q2TAA5	GDP-Man:Man(3)GlcNAc(2)-PP-Dol alpha-1,2-mannosyltransferase	ALG11
9	Q9BV10	Dol-P-Man:Man(7)GlcNAc(2)-PP-Dol alpha-1,6-mannosyltransferase	ALG12

Table B.2: Glycosyltransferase catalog (protein and gene name)

No.	UniProt entry	UniProt protein name	UniProt gene name
10	O43505	Beta-1,4-glucuronyltransferase 1	B4GAT1
11	-	-	-
12	Q9Y5Z6. O43825	Beta-1,3-galactosyltransferase 1. Beta-1,3-galactosyltransferase 2	B3GALT1. B3GALT2
13	P15291. O60909	Beta-1,4-galactosyltransferase 1. Beta-1,4-galactosyltransferase 2	B4GALT1. B4GALT2
14	P15291. O60909. O60512. O60513	Beta-1,4-galactosyltransferase 1. Beta-1,4-galactosyltransferase 2. Beta-1,4-galactosyltransferase 3. Beta-1,4-galactosyltransferase 4.	B4GALT1. B4GALT2. B4GALT3. B4GALT4
15	Q00973	Beta-1,4 N-acetylgalactosaminyltransferase 1	B4GALNT1
16	Q6L9W6. Q76KP1	Beta-1,4-N-acetylgalactosaminyltransferase 3. N-acetyl-beta-glucosaminyl-glycoprotein 4-beta-N-acetylgalactosaminyltransferase 1	B4GALNT3. B4GALNT4
17	P16442	Histo-blood group ABO system transferase	ABO
18	P16442	Histo-blood group ABO system transferase	ABO
19	P19526. Q10981	Galactoside 2-alpha-L-fucosyltransferase 1. Galactoside 2-alpha-L-fucosyltransferase 2	FUT1. FUT2
20	P21217. Q11128. P51993	Galactoside 3(4)-L-fucosyltransferase. Alpha-(1,3)-fucosyltransferase 5. Alpha-(1,3)-fucosyltransferase 6	FUT3. FUT5. FUT6
21	P21217. Q11128. P51993. Q11130	Galactoside 3(4)-L-fucosyltransferase. Alpha-(1,3)-fucosyltransferase 5. Alpha-(1,3)-fucosyltransferase 6. Alpha-(1,3)-fucosyltransferase 7	FUT3. FUT5. FUT6. FUT7
22	Q9BYC5	Alpha-(1,6)-fucosyltransferase	FUT8
23	Q9Y231	Alpha-(1,3)-fucosyltransferase 9	FUT9
24	Q495W5	Alpha-(1,3)-fucosyltransferase 11	FUT11
25	O95395	Beta-1,3-galactosyl-O-glycosyl-glycoprotein beta-1,6-N-acetylglucosaminyltransferase 3	GCNT3
26	O95395	Beta-1,3-galactosyl-O-glycosyl-glycoprotein beta-1,6-N-acetylglucosaminyltransferase 3	GCNT3
27	P26572	Alpha-1,3-mannosyl-glycoprotein 2-beta-N-acetylglucosaminyltransferase	MGAT1
28	Q10469	Alpha-1,6-mannosyl-glycoprotein 2-beta-N-acetylglucosaminyltransferase	MGAT2
29	Q09327	Beta-1,4-mannosyl-glycoprotein 4-beta-N-acetylglucosaminyltransferase	MGAT3

Table B.2: Glycosyltransferase catalog (protein and gene name)

No.	UniProt entry	UniProt protein name	UniProt gene name
30	Q9UM21. Q9UQ53. Q9UBM8	Alpha-1,3-mannosyl-glycoprotein 4-beta-N-acetylglucosaminyltransferase A. Alpha-1,3-mannosyl-glycoprotein 4-beta-N-acetylglucosaminyltransferase B. Alpha-1,3-mannosyl-glycoprotein 4-beta-N-acetylglucosaminyltransferase C	MGAT4A. MGAT4B. MGAT4C
31	Q9UM21. Q9UQ53. Q9UBM8	Alpha-1,3-mannosyl-glycoprotein 4-beta-N-acetylglucosaminyltransferase A. Alpha-1,3-mannosyl-glycoprotein 4-beta-N-acetylglucosaminyltransferase B. Alpha-1,3-mannosyl-glycoprotein 4-beta-N-acetylglucosaminyltransferase C.	MGAT4A. MGAT4B. MGAT4C
32	Q9UM21. Q9UQ53. Q9UBM8	Alpha-1,3-mannosyl-glycoprotein 4-beta-N-acetylglucosaminyltransferase C	MGAT4C
33	Q9UM21. Q9UQ53. Q9UBM8	Alpha-1,3-mannosyl-glycoprotein 4-beta-N-acetylglucosaminyltransferase C	MGAT4C
34	Q09328	Alpha-1,6-mannosylglycoprotein 6-beta-N-acetylglucosaminyltransferase A	MGAT5
35	Q09328	Alpha-1,6-mannosylglycoprotein 6-beta-N-acetylglucosaminyltransferase A	MGAT5
36	-	-	-
37	-	-	-
38	Q11201. Q16842. Q11206	CMP-N-acetylneuraminate-beta-galactosamide-alpha-2,3-sialyltransferase 1. CMP-N-acetylneuraminate-beta-galactosamide-alpha-2,3-sialyltransferase 2. CMP-N-acetylneuraminate-beta-galactosamide-alpha-2,3-sialyltransferase 4.	ST3GAL1. ST3GAL2. ST3GAL4
39	Q11203. Q11206. Q9Y274	CMP-N-acetylneuraminate-beta-1,4-galactoside alpha-2,3-sialyltransferase. CMP-N-acetylneuraminate-beta-galactosamide-alpha-2,3-sialyltransferase 4. Type 2 lactosamine alpha-2,3-sialyltransferase	ST3GAL3. ST3GAL4. ST3GAL6
40	P15907. Q96JF0	Beta-galactoside alpha-2,6-sialyltransferase 1. Beta-galactoside alpha-2,6-sialyltransferase 2	ST6GAL1. ST6GAL2

Table B.2: Glycosyltransferase catalog (protein and gene name)

No.	UniProt entry	UniProt protein name	UniProt gene name
41	Q92185. O43173. Q92187. O15466. P61647	Alpha-N-acetylneuraminide alpha-2,8-sialyltransferase. Sia-alpha-2,3-Gal-beta-1,4-GlcNAc-R:alpha 2,8-sialyltransferase. CMP-N-acetylneuraminate-poly-alpha-2,8-sialyltransferase. Alpha-2,8-sialyltransferase 8E. Alpha-2,8-sialyltransferase 8F	ST8SIA1. ST8SIA3. ST8SIA4. ST8SIA5. ST8SIA6
42	Q92186. O43173. Q92187. O15466	Alpha-2,8-sialyltransferase 8B. Sia-alpha-2,3-Gal-beta-1,4-GlcNAc-R:alpha 2,8-sialyltransferase. CMP-N-acetylneuraminate-poly-alpha-2,8-sialyltransferase. Alpha-2,8-sialyltransferase 8E	ST8SIA2. ST8SIA3. ST8SIA4. ST8SIA5
43	Q92186. O43173. Q92187. O15466. P61647	Alpha-2,8-sialyltransferase 8B. Sia-alpha-2,3-Gal-beta-1,4-GlcNAc-R:alpha 2,8-sialyltransferase. CMP-N-acetylneuraminate-poly-alpha-2,8-sialyltransferase. Alpha-2,8-sialyltransferase 8E. Alpha-2,8-sialyltransferase 8F	ST8SIA2. ST8SIA3. ST8SIA4. ST8SIA5. ST8SIA6
44	Q92186. O43173. Q92187. O15466	Alpha-2,8-sialyltransferase 8B. Sia-alpha-2,3-Gal-beta-1,4-GlcNAc-R:alpha 2,8-sialyltransferase. CMP-N-acetylneuraminate-poly-alpha-2,8-sialyltransferase. Alpha-2,8-sialyltransferase 8E	ST8SIA2. ST8SIA3. ST8SIA4. ST8SIA5

Table B.3: Glycosyltransferase catalog (resource databases (1))

No.	EC number	KO id	GGdb name	BRENDA name
1	2.4.1.258	K03845	ALG3	dolichyl-P-Man:Man5GlcNAc2-PP-dolichol alpha-1,3-mannosyltransferase
2	2.4.1.267	K03848	ALG6	dolichyl-P-Glc:Man9GlcNAc2-PP-dolichol alpha-1,3-glucosyltransferase
3	2.4.1.265	K03849	ALG8	dolichyl-P-Glc:Glc1Man9GlcNAc2-PP-dolichol alpha-1,3-glucosyltransferase
4	2.4.1.259	K03846	ALG9	dolichyl-P-Man:Man6GlcNAc2-PP-dolichol alpha-1,2-mannosyltransferase
5	2.4.1.261	K03846	ALG9	dolichyl-P-Man:Man8GlcNAc2-PP-dolichol alpha-1,2-mannosyltransferase

Table B.3: Glycosyltransferase catalog (resource databases (1))

No.	EC number	KO id	GGdb name	BRENDA name
6	2.4.1.256	K03850	ALG10	dolichyl-P-Glc:Glc2Man9GlcNAc2-PP-dolichol alpha-1,2-glucosyltransferase
7	2.4.1.131	K03844	ALG11	GDP-Man:Man3GlcNAc2-PP-dolichol alpha-1,2-mannosyltransferase
8	2.4.1.131	K03844	ALG11	GDP-Man:Man3GlcNAc2-PP-dolichol alpha-1,2-mannosyltransferase
9	2.4.1.260	K03847	ALG12	dolichyl-P-Man:Man7GlcNAc2-PP-dolichol alpha-1,6-mannosyltransferase
10	2.4.1.-	K00741	B3GNT1	N-acetyllactosaminide beta-1,3-N-acetylglucosaminyltransferase
11	-	K09664	B3GNT7	-
12	2.4.1.-	K07819. K07820	B3GALT1. B3GALT2	-
13	2.4.1.-	K07966. K07967	B4GALT1. B4GALT2	lactose synthase
14	2.4.1.-	K07966. K07967. K07968. K07969	B4GALT1. B4GALT2. B4GALT3. B4GALT4	N-acetyllactosamine synthase. beta-N-acetylglucosaminylglycopeptide beta-1,4-galactosyltransferase
15	2.4.1.92	-	B4GALNT1	(N-acetylneuraminyl)-galactosylglucosylceramide N-acetylgalactosaminyltransferase
16	2.4.1.244	-	B4GALNT3. B4GALNT4	N-acetyl-beta-glucosaminyl-glycoprotein 4-beta-N-acetylgalactosaminyltransferase
17	2.4.1.40	K00709	A(ABO)	glycoprotein-fucosylgalactoside alpha-N-acetylgalactosaminyltransferase. fucosylgalactoside 3-alpha-galactosyltransferase
18	2.4.1.37	K00709	A(ABO)	glycoprotein-fucosylgalactoside alpha-N-acetylgalactosaminyltransferase. fucosylgalactoside 3-alpha-galactosyltransferase
19	2.4.1.69	K00718	FUT1. FUT2	galactoside 2-alpha-L-fucosyltransferase
20	2.4.1.65	K00716. K07633. K07634	FUT3. FUT5. FUT6	3-galactosyl-N-acetylglucosaminide 4-alpha-L-fucosyltransferase

Table B.3: Glycosyltransferase catalog (resource databases (1))

No.	EC number	KO id	GGdb name	BRENDA name
21	2.4.1.65	K00716. K07633. K07634. K07635	FUT3. FUT5. FUT6. FUT7	3-galactosyl-N-acetylglucosaminide 4-alpha-L-fucosyltransferase
22	2.4.1.68	K00717	FUT8	glycoprotein 6-alpha-L-fucosyltransferase
23	2.4.1.-	K03663	FUT9	3-galactosyl-N-acetylglucosaminide 4-alpha-L-fucosyltransferase
24	2.4.1.-	K00753	FUT11	4-galactosyl-N-acetylglucosaminide 3-alpha-L-fucosyltransferase
25	2.4.1.150	K00742. K09662	GCNT3	N-acetyllactosaminide beta-1,6-N-acetylglucosaminyl-transferase
26	2.4.1.102	K00742. K09662	GCNT3	beta-1,3-galactosyl-O-glycosyl-glycoprotein beta-1,6-N-acetylglucosaminyltransferase
27	2.4.1.101	K00726	MGAT1	alpha-1,3-mannosyl-glycoprotein 2-beta-N-acetylglucosaminyltransferase
28	2.4.1.143	K00736	MGAT2	alpha-1,6-mannosyl-glycoprotein 2-beta-N-acetylglucosaminyltransferase
29	2.4.1.144	K00737	MGAT3	beta-1,4-mannosyl-glycoprotein 4-beta-N-acetylglucosaminyltransferase
30	2.4.1.145	K00738	MGAT4A	alpha-1,3-mannosyl-glycoprotein 4-beta-N-acetylglucosaminyltransferase
31	2.4.1.145	K00738	MGAT4B	alpha-1,3-mannosyl-glycoprotein 4-beta-N-acetylglucosaminyltransferase
32	2.4.1.145	-	-	alpha-1,6-mannosyl-glycoprotein 4-beta-N-acetylglucosaminyltransferase
33	2.4.1.145	-	-	alpha-1,6-mannosyl-glycoprotein 4-beta-N-acetylglucosaminyltransferase
34	2.4.1.155	K00744. K09661	MGAT5. MGAT5B	alpha-1,6-mannosyl-glycoprotein 6-beta-N-acetylglucosaminyltransferase
35	2.4.1.155	K00744. K09661	MGAT5. MGAT5B	alpha-1,6-mannosyl-glycoprotein 6-beta-N-acetylglucosaminyltransferase
36	2.4.1.155	K00744. K09661	MGAT5. MGAT5B	alpha-1,6-mannosyl-glycoprotein 6-beta-N-acetylglucosaminyltransferase

Table B.3: Glycosyltransferase catalog (resource databases (1))

No.	EC number	KO id	GGdb name	BRENDA name
37	2.4.1.155	K00744. K09661	MGAT5. MGAT5B	alpha-1,6-mannosyl-glycoprotein 6-beta-N-acetylglucosaminyltransferase
38	2.4.99.4	K03368	ST3GAL2	beta-galactoside alpha-2,3-sialyltransferase
39	2.4.99.-	K00780. K03368. K00781. K03494. K03370. K03792	ST3GAL1. ST3GAL2. ST3GAL3. ST3GAL4. ST3GAL5. ST3GAL6	neolactotetraosylceramide alpha-2,3-sialyltransferase. N-acetyllactosaminide alpha-2,3-sialyltransferase
40	2.4.99.1	K00778	ST6GAL1. ST6GAL2	beta-galactoside alpha-2,6-sialyltransferase
41	2.4.99.8. 2.4.99.-	K03371. K03369	ST8SIA1. ST8SIA2. ST8SIA4. ST8SIA5. ST8SIA6	alpha-N-acetylneuraminate alpha-2,8-sialyltransferase
42	2.4.99.-	K03371. K03369	ST8SIA1. ST8SIA2. ST8SIA4. ST8SIA5. ST8SIA6	alpha-N-acetylneuraminate alpha-2,8-sialyltransferase
43	2.4.99.-	K03371. K03369	ST8SIA1. ST8SIA2. ST8SIA4. ST8SIA5. ST8SIA6	alpha-N-acetylneuraminate alpha-2,8-sialyltransferase
44	2.4.99.-	K03371. K03369	ST8SIA1. ST8SIA2. ST8SIA4. ST8SIA5. ST8SIA6	alpha-N-acetylneuraminate alpha-2,8-sialyltransferase

Table B.4: Glycosyltransferase (GT) catalog (resource databases (2))

No.	CAZy class	CFG id	Krambeck model (2009)
1	GT58	-	-
2	GT57	-	-
3	GT57	-	-
4	GT22	-	-
5	GT22	-	-
6	GT59	-	-
7	GT4	-	-
8	GT4	-	-
9	GT22	-	-
10	GT31	gt hum 536	iGnT
11	GT31	gt hum 562	-
12	GT31	gt hum 429	b3GalT
13	GT7	-	-
14	GT7	gt hum 460. gt hum 436	b4GalT
15	GT12	gt hum 482	-
16	-	gt hum 475	-
17	GT6	gt hum 450	GalNAcT-A
18	GT6	gt hum 450	GalT-B
19	GT37. GT10. GT11	gt hum 598	FucTH
20	GT37. GT10	gt hum 600	FucTLe
21	GT37. GT10	gt hum 600	FucTLe
22	GT23	gt hum 605	a6FucT
23	GT10	-	-
24	GT10	gt hum 601	a3FucT
25	GT14	gt hum 548. gt hum 544	IGnT
26	GT14	gt hum 548. gt hum 544	IGnT
27	GT13	gt hum 535	GnTI
28	GT16	gt hum 534	GnTII
29	GT17	gt hum 540	GnTIII
30	GT54	gt hum 545	GnTIV
31	GT54	gt hum 545	GnTIV
32	-	-	-
33	-	-	-
34	GT18	gt hum 553	GnTV

Table B.4: Glycosyltransferase (GT) catalog (resource databases (2))

No.	CAZy class	CFG id	Krambeck model (2009)
35	GT18	gt hum 553	GnTV
36	GT18	gt hum 553	GnTV
37	GT18	gt hum 553	GnTV
38	GT29	gt hum 625	-
39	GT29	gt hum 627. gt hum 624	a3SiaT
40	GT29	gt hum 629	a6SiaT
41	GT29	gt hum 639	-
42	GT29	gt hum 639	-
43	GT29	gt hum 639	-
44	GT29	gt hum 639	-