A Global Representation of the Carbohydrate Structures: A Tool for the Analysis of Glycan

Kosuke Hashimoto[1]
khashimo@kuicr.kyoto-u.ac.jp

Shin Kawano[1]
kawano@kuicr.kyoto-u.ac.jp

Susumu Goto[1]
goto@kuicr.kyoto-u.ac.jp

Kiyoko F. Aoki-Kinoshita[1]
kiyoko@kuicr.kyoto-u.ac.jp

Masayuki Kawashima[2]
kawasima.m@jp.fujitsu.com

Minoru Kanehisa[1]
kanehisa@kuicr.kyoto-u.ac.jp

[1] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan

[2] Fujitsu Kyushu System Engineering Ltd, Momochihama 2-2-1 Sawara-ku, Fukuoka 814-8589, Japan

Abstract

Glycan resources have been developed of late, such as carbohydrate databases, analysis tools, and algorithms for analysis of carbohydrate features. With this background, bioinformatics approaches to carbohydrate research have recently begun using a large amount of protein and carbohydrate data. This paper introduces one of these projects that elucidates the range of carbohydrate structures.

In this study, the variety of carbohydrate structures have been enumerated in a global tree structure called variation trees, using the KEGG GLYCAN database, which is a public-domain glycan resource for bioinformatics analysis. Additionally, a glycosyltransferase mapping list of glycosyltransferases and their catalyzing glycosidic linkages was constructed. From this, we present the composite structure map (CSM), which is a structural variation map integrating its variation trees and glycosyltransferase map list. CSM is able to display, for example, expression data of glycosyltransferases in a compact manner, illustrating its versatility as a new bioinformatics resource and tool capable of analyzing carbohydrate structures on a global scale. These resources are available at `http://www.genome.jp/kegg/glycan/`.

Keywords: carbohydrate structure, glycome informatics, glycoinformatics, database, glycan, structure map

1 Introduction

Carbohydrates, following genes and proteins, play important roles in many fundamental biological processes such as protein folding, oligomerization, quality control, sorting and transport [7]. For example, it is well known through experimentation that misfolded proteins are generated and fail to reach a functional state when the N-linked glycosylation of a glycoprotein is inhibited [6]. Additionally, glycoproteins are key components of the immune system affecting stability, recognition, and regulation of the proteins [16]. Recent reports have also clarified that proteoglycans, which are cell-surface and extracellular matrix macromolecules, have essential functions in development [11, 17].

Carbohydrate-structure determination has had many challenges and difficulties, mainly due to the fact that carbohydrates are branched structures, and this branching occurs with many different types of linkages and anomeric configurations. This complex carbohydrate structure also requires the development of new computational methods for analysis [18]. Despite their complexity, because of the recent improvement of experimental techniques such as mass spectrometry, nuclear magnetic resonance (NMR), and knockout mice analysis, much knowledge about carbohydrate structures and functions has been accumulated [4, 10, 13]. Glycan resources have been developed such as carbohydrate databases, analysis tools, and algorithms for analysis of carbohydrate structures [1, 2, 12]. With this background, bioinformatics approaches to carbohydrate research have recently begun using a large amount of protein and carbohydrate data, such as studies on protein binding locations on carbohydrates [3, 8, 15]. This paper introduces one of these projects that elucidates the range of carbohydrate structures.

Carbohydrate structures have not been comprehensively investigated to date. The carbohydrate structures in each organism, tissue, and individual have differences that crucially affect phenotype. In fact, carbohydrate structures are assumed to potentially form extremely complicated structures due to the wide diversity of glycosidic linkages. That is, while DNA and proteins have only one kind of linkage for connecting two elements, carbohydrate structures have eight, comprised of four hydroxyl positions of linkages, and two types of anomeric configurations, (alpha and beta). When real carbohydrate structure data is investigated, the combinations of monosaccharides with glycosidic linkages were actually found to be limited to some extent. In addition, the number of monosaccharides that comprises a carbohydrate structure is much less than the number of amino acids that comprises a protein. Based on these facts, we claim that the variations in carbohydrate structures are limited. Thus, it is possible to represent all possible variations of carbohydrate structures within organisms in a single structure, which we call a variation tree. Due to this fact, it was determined that a glycome informatics resource enumerating these structural variations would be useful.

Carbohydrate biosynthesis is different from protein biosynthesis using mRNA as a template; carbohydrate structures are extended by the glycosyltransferase by adding monosaccharides individually. The variation of carbohydrate structures are thus produced by glycosyltransferases. There have been no reports of any glycosyltransferase replacing or inserting monosaccharides in the middle of a carbohydrate chain (or adding an oligosaccharide to an acceptor other than an oligosaccharide transferase (OST) transferring the core structure of N-Glycan to the asparagine residue of a protein). Therefore each glycosidic linkage in a carbohydrate structure is produced by a specific glycosyltransferase. Many glycosyltransferases have been identified by cloning technology, and the glycosidic linkages that some glycosyltransferases catalyze have been clarified [14]. This information, which we call a glycosyltransferase mapping list, is useful for investigating specific features of carbohydrate structures.

In this paper, we first clarified all the variations of carbohydrate structures based on the KEGG GLYCAN database, which is a public-domain glycan resource for bioinformatics analysis including more than 10,000 carbohydrate structures [5, 20]. In addition, glycosyltransferases were matched to glycosidic linkages catalyzed by their genes. We then present the composite structure map (CSM), which is a structural variation map integrating the variation trees and its glycosyltransferase map list [19]. CSM is constructed as a stepping stone for relating carbohydrate functions and structures; it is a bridge between carbohydrate structures and relevant genes. CSM is able to display, for example, expression data of glycosyltransferases in a compact manner, illustrating its versatility as a new bioinformatics instrument capable of analyzing carbohydrate structures on a global scale.

2 Method

Here, we describe the data processing and calculations used in constructing CSM. First, a resource of carbohydrate structures comprising variation trees is explained, and then the details of constructing variation trees are given. Next, we explain the method of matching glycosyltransferases and glycosidic linkages. Finally, we describe the procedure to integrate the variations trees and the glycosyltransferase list for building CSM.

2.1 Data Set

Starting with the entire KEGG GLYCAN database as of 03/30/2005 as our data set, first, the entries of those glycans that have bonds between two reducing terminals were removed. Molecules and non-carbohydrate nodes such as amino acids and lipids were also removed from the data set. Modification molecules, which are attached to monosaccharides after the completion of carbohydrate biosynthesis (e.g., methyl and O-acetyl groups), were further deleted. Then, the non-redundant data set of carbohydrate structures was derived by a tree structure alignment program [2]. In the end, our data set consisted of about 8000 non-redundant structures that consisted purely of monosaccharides.

2.2 Construction of the Variation Tree

The variation trees were constructed by taking all the carbohydrate structures containing a particular monosaccharide at its root and merging them together into one unified structure. For example, the variation tree having 'Glc' as the root was build as follows. (1) All the carbohydrate structures whose roots are 'Glc' were extracted from the data set. (2) All the branched structures were divided into linear structures traced from each terminal monosaccharide to the root (Figure 1a). (3) The non-redundant linear structures were obtained by removing redundant structures. (4) Finally, common substructures were merged from the root (Figure 1b). The variation trees rooted by other monosaccharides were similarly constructed.

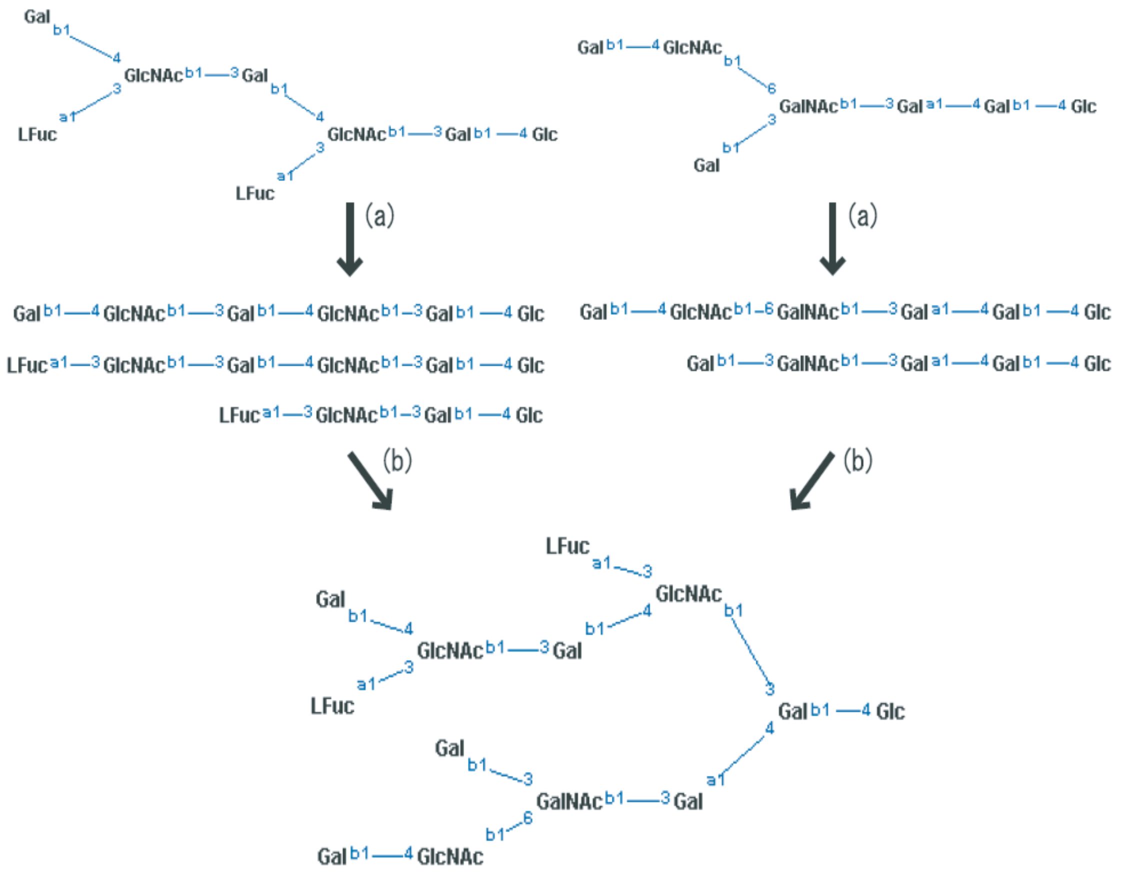

(a)
(a)
Galb1—4GlcNAcb1—3Galb1—4GlcNAcb1—3Galb1—4Glc
LFuca1—3GlcNAcb1—3Galb1—4GlcNAcb1—3Galb1—4Glc
LFuca1—3GlcNAcb1—3Galb1—4Glc
Galb1—4GlcNAcb1—6GalNAcb1—3Gala1—4Galb1—4Glc
Galb1—3GalNAcb1—3Gala1—4Galb1—4Glc
(b)
(b)
Gal
LFuc
GlcNAc
GalNAc
Gal
Glc
b1
a1
3
4
6

Figure 1: The flow merging carbohydrate structures.

2.3 Construction of the Mapping List between Glycosidic Linkages and Glycosyltransferases

We describe here the method of constructing the mapping list of glycosyltransferases and its catalyzing glycosidic linkages. First, all the entries for the reaction that synthesize glycosidic linkages were extracted from the KEGG REACTION database, which is a curated database of chemical (mostly enzymatic) reactions. Next, by comparing the structure before a reaction with itself after the reaction, glycosidic linkages composed of the following three elements were built from these reactions: the acceptor monosaccharide residue, the donor monosaccharide residue, and the linkage between them (e.g., Gal b1-4 Glc). Finally, the glycosidic linkages were mapped to KO entries if its catalyzing reaction was found in the KEGG KO database, which is a manually curated set of orthologous gene groups in the complete genomes [9]. The mapping between glycosidic linkages and ortholog groups was thus obtained (Table 2).

2.4 Construction of CSM as a Tool

The variation trees constructed in section 2.2 were developed into CSM as a web application. A monosaccharide and a glycosidic linkage in a variation tree were represented as a node and an edge in CSM, respectively. Information corresponding to the nodes and edges was incorporated. Any node on the CSM tree corresponds to a list of carbohydrate structures that are located on the single path from the root to that node. Thus, each node is hyperlinked to its corresponding list. Each structure in the resulting list is also hyperlinked to their GLYCAN entries. Since glycosyltransferases catalyze the biosynthesis of carbohydrate structures by the addition of individual linkages, each edge in the CSM is hyperlinked to its corresponding glycosyltransferase-related information, if known.

3 Results and Discussion

3.1 Variation of the Carbohydrate Structures

A variation tree represents the entire set of known comprehensive carbohydrate structures as a tree with a monosaccharide at its root. Different trees represent different root monosaccharides. Table 1 presents the number of branches appearing in each variation tree. The tree of ‘Glc’ has the highest diversity. This reflects the fact that the tree includes varied structures of glycolipid. By the same token, the ‘GlcNAc’ tree is also highly diverse due to the fact that the tree includes the core structure of N-glycan (Man b1-4 GlcNAc b1-4 GlcNAc).

There seems to be a wide variety of carbohydrate structures considering the thousands of unique structures in the database. The complexity is caused by the combination of different monosaccharides with the different glycosidic linkages. However, we find that the majority of the structures include core structures and common structures such as the root of N-glycan, O-glycan, and glycolipid. That is, the structures are not constructed randomly, and their variety is rather limited. Merging the common core structures, the maximum possible number of carbohydrate structures is represented compactly in the variation trees.

3.2 Correspondence of Glycosyltransferases to Glycosidic Linkages

Table 2 is the mapping between glycosidic linkages and KO entries. The representation of these glycosidic linkages comprises acceptor monosaccharide residue, the donor monosaccharide residue, and the linkage. Many KO entry names are EC (enzyme commission) numbers, while others are specific names based on the catalytic activity. Forty glycosidic linkages were thus mapped to sixty KO entries.

Table 1: Branch variation in each variation tree.

Root	Number of branches
Glc	1010
GlcNAc	752
Gal	380
GalNAc	310
Man	318

3.3 Features of CSM

CSM is a resource which illustrates the possible variations of carbohydrate structures. A monosaccharide and a glycosidic linkage are respectively represented as a node and an edge. An example is given in Figure 2, where the right-most node ‘GlcNAc’ in the figure is the root of the structures, and it is connected to the nodes to its left at the 2nd level from the right, which represents possible monosaccharide variations that may bind to it. Nodes connected at each consecutive level represent these variations similarly. Because every type of glycosidic linkage is distinguished, different glycosidic linkages are represented by different edges even if the attached monosaccharide is the same. For example, when a Gal can be attached to Glc by a b1-3 and a b1-4 glycosidic linkage, CSM will contain different edges ‘Gal b1-3 Glc’ and ‘Gal b1-4 Glc.’ In this map, selecting an organism from the pull-down menu at the top colors the rectangles of enzymes and changes the hyperlinks to the enzymes. After selecting ‘All organisms in KEGG’ and pushing the ‘Go’ button, enzymes that (1) correspond to glycosyltransferases that catalyze the synthesis of the reaction and (2) have references in KO will be hyperlinked to its corresponding KO entry. On the other hand, if a particular organism is selected under the same pull-down menu such as ‘Homo sapiens’ in the Figure 2, the rectangles of enzymes will be hyperlinked to the genes corresponding to the selected organism. In addition, selecting three monosaccharides in the three pull-down menus labeled ‘Root,’ ‘Next1,’ ‘Next2’ will display the CSMthat containing the designated chain of monosaccharides as its core. In the case of Figure 2, ‘GlcNAc’, ‘GlcNAc’, ‘Gal’ are respectively selected. Moreover, several options are available. For example, a ‘Threshold’ value may be specified to indicate the minimum number of times a particular edge should appear in the tree in order to be displayed. The ‘Position’ may be adjusted to display the CSM with the root at the top, the right (default), the left, or the bottom. The ‘Node size’ may be adjusted to increase the displayed size of the nodes in the CSM from the default minimum value of 5.

Because a carbohydrate structure is synthesized by glycosyltransferases that add monosaccharides one by one, if the entire set of glycosyltransferases can be elucidated, the CSM can display all possible carbohydrate structures. In addition to structural variations, CSM also illustrates the relationship between a glycosidic linkage and the glycosyltransferase that catalyzes it. The known 60 KO entries in KEGG, which include 176 organisms, are assigned to the edges in CSM. Twenty-four percent of the edges are related to ortholog groups. The edges that are not related to any ortholog groups correspond to either unknown genes or genes whose function has not been discovered in experiments.

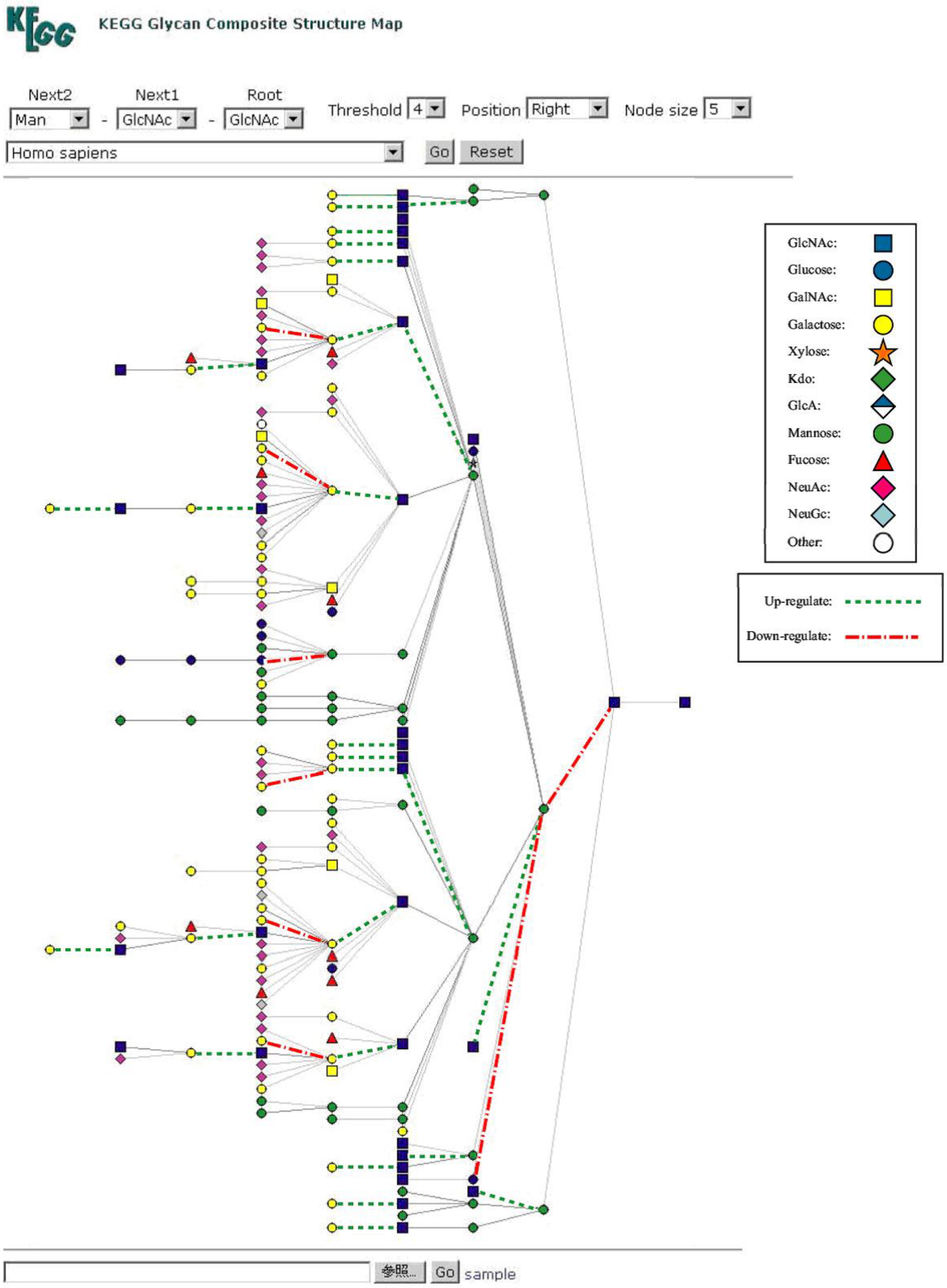
3.4 Using CSM for Glycan Expression Analysis

Furthermore, it is also possible to color the edges corresponding to a specification of colors and genes stored in a local file. This file can be uploaded to color the map, which may be useful for visualizing microarray results. Thus, the CSM provides a useful tool for glycome informatics analysis as it provides a comprehensive resource for all carbohydrate structures. Figure 2 represents the result of the microarray experiment in which a cancer cell was innervated by immunoglobulin type M (IgM). IgM infrequently recognizes ganglioside, a glycan, and then causes autoimmune diseases. This expression

Table 2: The relevant map between glycosidic linkages and KO entries.

Glycosidic linkage	KO entry name	Glycosidic linkage	KO entry name
Gal a1-3 Gal	ko:E2.4.1.37,	GlcNAc b1-3 GalNAc	ko:E2.4.1.147
	ko:E2.4.1.87	GlcNAc b1-4 Man	ko:E2.4.1.144,
Gal a1-4 Gal	ko:E2.4.1.228		ko:E2.4.1.145
Gal a1-6 Gal	ko:E2.4.1.67	GlcNAc b1-4 MurNAc	ko:MURG
Gal b1-3 Gal	ko:E2.4.1.134	GlcNAc b1-6 Gal	ko:E2.4.1.150
Gal b1-3 GalNAc	ko:B3GALT5,	GlcNAc b1-6 GalNAc	ko:E2.4.1.102,
	ko:E2.4.1.122, ko:E2.4.1.62		ko:E2.4.1.148
Gal b1-4 GlcNAc	ko:E2.4.1.38, ko:E2.4.1.69	GlcNAc b1-6 Man	ko:E2.4.1.155
Gal b1-4 Xyl	ko:E2.4.1.133	Kdo a2-4 Kdo	ko:KDTA
GalNAc a1-3 Gal	ko:E2.4.1.40	LFuc a1-2 Gal	ko:E2.4.1.69
GalNAc a1-3 GalNAc	ko:E2.4.1.88	LFuc a1-3 GlcNAc	ko:E2.4.1.152
GalNAc b1-3 Gal	ko:E2.4.1.79	LFuc a1-4 GlcNAc	ko:E2.4.1.65
GalNAc b1-4 Gal	ko:E2.4.1.92	LFuc a1-6 GlcNAc	ko:E2.4.1.68
GalNAc b1-4 GlcA	ko:E2.4.1.174,	Man a1-2 Man	ko:ALG11, ko:ALG9
	ko:E2.4.1.175	Man a1-3 Man	ko:ALG2, ko:ALG3
Glc a1-2 Glc	ko:ALG10	Man a1-6 Man	ko:ALG12
Glc a1-3 Glc	ko:ALG8	Man b1-4 GlcNAc	ko:ALG1
Glc a1-3 Man	ko:ALG6	Neu5Ac a2-3 Gal	ko:E2.4.99.10,
GlcA b1-3 Gal	ko:E2.4.1.135		ko:E2.4.99.6,
GlcA b1-3 GalNAc	ko:E2.4.1.226		ko:E2.4.99.9, ko:SIAT4A,
GlcA b1-4 GlcNAc	ko:E2.4.1.225		ko:SIAT4B
GlcNAc a1-4 GlcA	ko:EXTL1, ko:EXTL2,	Neu5Ac a2-6 Gal	ko:E2.4.99.1
	ko:EXTL3	Neu5Ac a2-6 GalNAc	ko:SIAT7A, ko:SIAT7C,
GlcNAc b1-2 Man	ko:E2.4.1.101,		ko:SIAT7E, ko:SIAT7F
	ko:E2.4.1.143	Neu5Ac a2-8 Neu5Ac	ko:SIAT8A, ko:SIAT8E
GlcNAc b1-3 Gal	ko:E2.4.1.149		

data is displayed on the CSM in this figure. Some up-regulated and down-regulated edges were observed using the tree including the core structures of ganglioside. On the other hand, in the case of N-glycan, many glycosyltransferases around the middle of the tree are highly expressed; however a core branch near the root was down-regulated in Figure 2, which represents the tree including the core structures of N-glycan. Although we have not gotten into the biological detail of these results, there is reason to believe that there may be some relationship between IgM and N-glycans to some extent.



KEGG Glycan Composite Structure Map

Next2 Man - Next1 GlcNAc - Root GlcNAc

Threshold 4

Position Right

Node size 5

Homo sapiens

Figure 2: Composite Structure Map.

4 Conclusion

We have demonstrated in this paper the fact that all the possible variations of carbohydrate structures are actually limited to only a certain set of structures that is possible to be displayed in a single global tree structure. This is in spite of the fact that, theoretically, the linkages and anomeric configurations of carbohydrate bonds make possible many more variations. With this knowledge, we have created the Composite Structure Map, which, along with the resources of KEGG, provide a useful tool for biologists to investigate glycan structures on a comprehensive scale. We have demonstrated this with an example of microarray expression data for IgM, finding that expression patterns are found in various areas of different variation trees, a point for possible further research. Needless to say, this is indeed the opening of many more possibilities in carbohydrate structure research.

Acknowledgments

We thank Dr. Yasushi Okuno for providing his suggestions and specialized knowledge of glycans. The computational resource was provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University. This work was supported by the grants from the Ministry of Education, Culture, Sports, Science and Technology of Japan, Japan Society for the Promotion of Science, and Japan Science and Technology Corporation.

Abbreviations

Gal, galactose; GalNAc, N-acetyl-galactosamine; Glc, glucose; GlcA, glucuronic acid; GlcNAc, N-acetyl-glucosamine; Kdo, keto-deoxyoctulosonic acid; LFuc, fucose; Man, mannose; Neu5Ac, N-acetyl-neuraminic acid (sialic acid); Xyl, xylose

References

[1] Aoki, K.F., Yamaguchi, A., Okuno, Y., Akutsu, T., Ueda, N., Kanehisa, M., and Mamitsuka, H., Efficient tree-matching methods for accurate carbohydrate database queries, *Genome Informatics,* 14:134–143, 2003.

[2] Aoki, K.F., Yamaguchi, A., Ueda, N., Akutsu, T., Mamitsuka, H., Goto, S., and Kanehisa, M., KCaM (KEGG Carbohydrate Matcher): A software tool for analyzing the structures of carbohydrate sugar chains, *Nucleic Acids Res.,* 32:W267–272, 2004.

[3] Ben-Dor, S., Esterman, N., Rubin, E., and Sharon, N., Biases and complex patterns in the residues flanking protein N-glycosylation sites, *Glycobiology,* 14:95–101, 2004.

[4] Dell, A. and Morris, H.R., Glycoprotein structure determination by mass spectrometry, *Science,* 291:2351–2356, 2001.

[5] Hashimoto, K., Goto, S., Kawano, S., Aoki-Kinoshita, K.F., Ueda, N., Miyazaki, S., Kawasaki, T., and Kanehisa, M., submitted.

[6] Helenius, A., How N-linked oligosaccharides affect glycoprotein folding in the endoplasmic reticulum, *Mol. Biol. Cell,* 5:253–265, 1994.

[7] Helenius, A. and Aebi, M., Intracellular functions of N-linked glycans, *Science,* 291:2364–2369, 2001.

[8] Julenius, K., Molgaard, A., Gupta, R., and Brunak, S., Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites, *Glycobiology,* 15:153–164, 2005.

[9] Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M., The KEGG resource for deciphering the genome, *Nucleic Acids Res.,* 32:D277–280, 2004.

[10] Kogelberg, H., Solis, D., and Jimenez-Barbero, J., New structural insights into carbohydrate-protein interactions from NMR spectroscopy, *Curr. Opin. Struct. Biol.,* 13:646–653, 2003.

[11] Lin, X., Functions of heparan sulfate proteoglycans in cell signaling during development, *Development,* 131:6009–6021, 2004.

[12] Marchal, I., Golfier, G., Dugas, O., and Majed, M., Bioinformatics in glycobiology, *Biochimie.,* 85:75–81, 2003.

[13] Miyakis, S., Robertson, S.A., and Krilis, S.A., Beta-2 glycoprotein I and its role in antiphospholipid syndrome-lessons from knockout mice, *Clin. Immunol.,* 112:136–143, 2004.

[14] Narimatsu, H., Construction of a human glycogene library and comprehensive functional analysis, *Glycoconj. J.,* 21:17–24, 2004.

[15] Petrescu, A.J., Milac, A.L., Petrescu, S.M., Dwek, R.A., and Wormald, M.R., Statistical analysis of the protein environment of N-glycosylation sites: implications for occupancy, structure, and folding, *Glycobiology,* 14:103–114, 2004.

[16] Rudd, P.M., Elliott, T., Cresswell, P., Wilson, I.A., and Dwek, R.A., Glycosylation and the immune system, *Science,* 291:2370–2376, 2001.

[17] Schwartz, N.B. and Domowicz, M., Proteoglycans in brain development, *Glycoconj. J.,* 21:329–341, 2004.

[18] Von Der Lieth, C.W., Bohne-Lang, A., Lohmann, K.K., and Frank, M., Bioinformatics for glycomics: status, methods, requirements and perspectives, *Brief. Bioinform.,* 5:164–178, 2004.

[19] CSM - `http://www.genome.jp/kegg-bin/draw\_csm`

[20] KEGG GLYCAN - `http://www.genome.jp/kegg/ligand.html`