Carbohydrate Research 344 (2009) 881–887

Contents lists available at ScienceDirect

Carbohydrate Research

journal homepage: www.elsevier.com/locate/carres

Comprehensive analysis of glycosyltransferases in eukaryotic genomes for structural and functional characterization of glycans

Kosuke Hashimoto[a], Toshiaki Tokimatsu[a], Shin Kawano[a,†], Akiyasu C. Yoshizawa[a,‡], Shujiro Okuda[a,§], Susumu Goto[a], Minoru Kanehisa[a,b,*]

[a] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan
[b] Human Genome Center, Institute of Medical Science, University of Tokyo, Minato-ku, Tokyo 108-8639, Japan

ARTICLE INFO

Article history:
Received 2 October 2008
Received in revised form 24 February 2009
Accepted 3 March 2009
Available online 9 March 2009

Keywords:
Bioinformatics
Evolution
Genome analysis
Glycan structures
Glycosyltransferase

ABSTRACT

Glycosyltransferases comprise highly divergent groups of enzymes, which play a central role in the synthesis of complex glycans. Because the repertoire of glycosyltransferases in the genome determines the range of synthesizable glycans, and because the increasing amount of genome sequence data is now available, it is essential to examine these enzymes across organisms to explore possible structures and functions of the glycoconjugates. In this study, we systematically investigated 36 eukaryotic genomes and obtained 3426 glycosyltransferase homologs for biosynthesis of major glycans, classified into 53 families based on sequence similarity. The families were further grouped into six functional categories based on the biosynthetic pathways, which revealed characteristic patterns among organism groups in the degree of conservation and in the number of paralogs. The results also revealed a strong correlation between the number of glycosyltransferases and the number of coding genes in each genome. We then predicted the ability to synthesize major glycan structures including N-glycan precursors and GPI-anchors in each organism from the combination of the glycosyltransferase families. This indicates that not only parasitic protists but also some algae are likely to synthesize smaller structures than the structures known to be conserved among a wide range of eukaryotes. Finally we discuss the functions of two large families, sialyltransferases and β4-glycosyltransferases, by performing finer classifications into subfamilies. Our findings suggest that universality and diversity of glycans originate from two types of evolution of glycosyltransferase families, namely conserved families with few paralogs and diverged families with many paralogs.

1. Introduction

The glycosyltransferase is a key enzyme for the synthesis of complex glycans, which play various roles, including development,[1] immune response,[2] cell–cell interaction,[3] and cell invasion.[4] Thus, much effort has been expended on the cloning and molecular identification of genes encoding glycosyltransferases.[5] In particular, human genes related to glycan biosynthesis, including glycosyltransferases, sugar–nucleotide synthetases, and sugar–nucleotide transporters, have been studied as targets for comprehensive identification. In this way, more than 180 human genes have been characterized.[6] In addition, several studies have recently demonstrated the reconstitution of the pathways for mammalian glycosylation in transgenic fungi and plants to overcome differences of the glycosylation system between human and the organisms.[7,8]

Although the three-dimensional architecture of the glycosyltransferase is well conserved and is classified into only a few fold groups, amino acid primary sequences are quite diverse.[9,10] Thus, glycosyltransferases necessarily consist of numerous families. A total of 91 families containing eukaryotic and prokaryotic sequences are defined in the CAZy database.[11,12] Until now, several multifamily analyses have been performed[13–15] as well as phylogenetic analyses focused on a single glycosyltransferase family.[16–18] In the next stage, it will be useful to analyze glycosyltransferases across a wide range of families and also across organisms for further understanding of the glycosylation system. In other words, once a comprehensive picture is obtained for the genomic repertoires of glycosyltransferases, this will help understand the architecture of biosynthetic pathways and the repertoire of synthesizable glycans in each organism.

The strategy and the resources to achieve this objective are provided by KEGG, a knowledge base for understanding higher-level

[*] Corresponding author. Tel.: +81 774 38 3270; fax: +81 774 38 3269.
E-mail address: kanehisa@kuicr.kyoto-u.ac.jp (M. Kanehisa).

[†] Present address: Database Center for Life Science, Research Organization of Information and Systems, 2-11-16 Yayoi, Bunkyo-ku, Tokyo 113-0032, Japan.

[‡] Present address: Division of Disease Proteomics, Institute for Enzyme Research, The University of Tokushima, 3-18-15, Kuramoto-cho, Tokushima 770-8503, Japan.

[§] Present address: Department of Bioinformatics, College of Life Sciences, Ritsumeikan University, 1-1-1 Nojihigashi, Kusatsu, Shiga 525-8577, Japan.

0008-6215/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.carres.2009.03.001

functions of cellular processes and organism behaviors from genomic information.[19] Together with KEGG we have been developing bioinformatics approaches to integrated analysis of genomic and chemical information. The best example of this type of approach is the prediction of glycan structures from gene expression data using the inventory of glycosyltransferases with their substrate specificities.[20,21] Another example of the integrative analysis of genomic and chemical information is an attempt to elucidate fatty acid variations in various eukaryotes from the inventory of their modification enzymes, namely desaturases and elongases.[22] In addition to KEGG GLYCAN,[23] several glycomics resources are currently available[24] including those at the Consortium for Functional Glycomics (CFG),[25] and Glycosciences.de.[26] We expect that the application of a similar strategy to glycosyltransferases with the glycomics resources can allow us to gain insight into the glycosylation system.

The purpose of this study is to obtain the structural perspective through a global view of glycosyltransferases, based on the fact that glycan structures are synthesized by a particular combination of glycosyltransferases. We first investigated 36 eukaryotic genomes to obtain all the glycosyltransferases for the synthesis of major glycans, excluding those for the synthesis of storage polysaccharides and structural polysaccharides, such as glycogen synthases, hyaluronan synthases, cellulose synthases, and chitin synthases, and defined 53 families based on the sequence similarity. We then examined differences of major glycan structures among the organisms by using the repertoire of the families. Our results suggest that universality and diversity of glycans originate from two types of evolution of glycosyltransferase families, namely conserved families with few paralogs and diverged families with many paralogs. Finally, we discuss some controversial families including plant sialyltransferases in the context of the biosynthesis pathway for the glycosyl donor.

2. Results

2.1. The global view of glycosyltransferases

A total of 3426 glycosyltransferase homologs were obtained from 36 eukaryotic genomes through sequence and phylogenetic analyses. This number represents the enzymes for the synthesis of the major classes of glycans and excludes those for the synthesis of storage or structural polysaccharides and the synthesis of *O*-galactose. In this definition, the number of glycosyltransferases in each organism was strongly correlated with the total number of genes in its genome (Fig. S1). The correlation coefficient is 0.88 for all the organisms and 0.935 for the organisms exclusive of protists. The glycosyltransferases accounted for about 0.5–1.1% of the total number of genes in animal, plant, and fungi genomes. In contrast, fewer glycosyltransferases were detected in parasitic protists, where the average number was only 24 genes (0.26%) per organism.

Glycosyltransferases form multiple families, and no single motif is conserved across all the glycosyltransferases, at least in their primary sequences. Our analyses based on sequence similarities indicate that glycosyltransferases can be classified into 53 families, each of which contains at least one conserved region. Detailed numbers of glycosyltransferases in each family and each organism are shown in Table S1. These families roughly correspond to the CAZy classification,[12] except for approximately 30 CAZy families that consist exclusively of prokaryotic sequences. In some cases, multiple families defined in this study correspond to a single family in CAZy, and vice versa. For example, three families in our analysis, namely ALG13, ALG14, and UGT, correspond to GT1, which is one of the largest families in CAZy. Conversely, a single family, POFUT, corresponds to two GT families, GT65 and GT68.

Figure 1 is an abridged version of Table S1, illustrating the families grouped into six functional categories and with the distribution of genes in the four kingdoms, animals, plants, fungi, and protists. A clear characteristic is shown in glycosyltransferases involved in the asparagine-linked glycan (N-glycan) precursor or the GPI-anchor biosynthesis, which are conserved in a wide range of organisms, especially in the initial steps of the pathways. In other functional categories, only the β3-glycosyltransferase (β3) family, the uridine diphosphate glycosyltransferase (UGT) family, and the UDP-glucose:glycoprotein glucosyltransferase (HUGT) family are widely distributed across different kingdoms and organisms, whereas many of the other families are found in a limited number of organisms. Similarly, several families exist in a kingdom-specific manner, such as the ABO family in animals, the *Arabidosis thaliana*

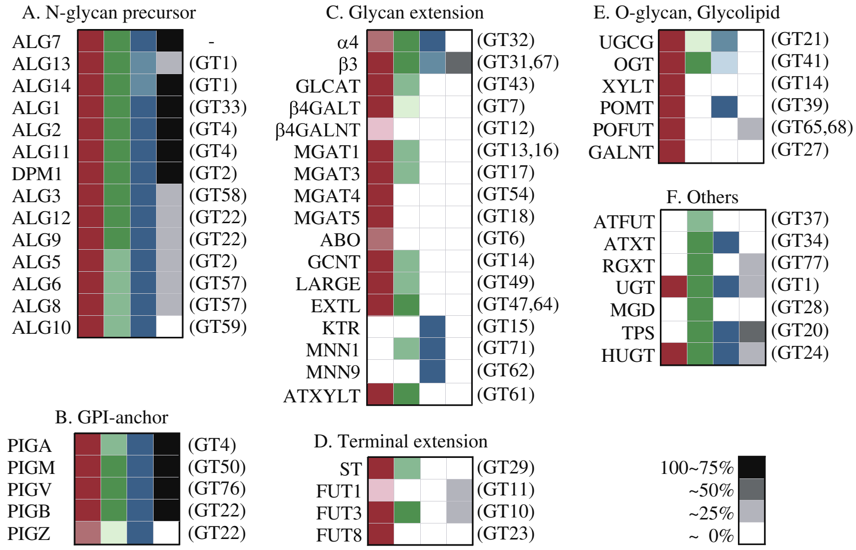

Figure 1. Fifty-three families of glycosyltransferases in six functional categories and the distribution in four kingdoms: animals, plants, fungi, and protists. Fifty-three families are separated into 6 functional categories, namely glycosyltransferases for (A) N-glycan precursor biosynthesis, (B) GPI-anchor biosynthesis, (C) extension of glycans, (D) terminal extension, (E) O-glycan and glycolipid biosynthesis, and (F) others. From left to right, the four horizontal blocks represent animals, plants, fungi, and protists. The density of colors indicates what percent of organisms possess glycosyltransferases in the family. For example, ALG7, the top of the category of N-glycan precursor, is distributed in more than 75% of the organisms in all the four kingdoms.

fucosyltransferases (ATFUT) family in plants, and the α3-mannosyltransferase (MNN9) family in fungi.

In addition to this biased distribution, the number of paralogs is also characteristic. In each family related to the N-glycan precursor biosynthesis or the GPI-anchor biosynthesis, the paralog is few; only one or occasionally two sequences were found in each organism (Fig. 2). In contrast, many of the other families contain a large number of paralogs; in particular, more than 20 paralogs were found in most of animals, land plants, and *Trypanosoma cruzi* in the β3 family and more than 100 paralogs were found in land plants in the UGT family. The UDP-GalNAc:polypeptide *N*-acetylgalactosaminyltransferase (GALNT) family and the sialyltransferase (ST) family are also large families, in which animal genomes contain about 20 paralogs.

2.2. The prediction of glycan structures from the repertoire of glycosyltransferases

We predicted synthesizable glycan structures in each organism from a set of glycosyltransferases based on the fact that the glycan structures are synthesized by glycosyltransferases in the stepwise manner. Our target structures are N-glycan precursor, GPI-anchor, root structures of O-glycan, a root structure of glycolipids, and extension structures of N-glycan. The predictions were conducted by mapping the combination of enzymes onto their biosynthesis pathways.

For the N-glycan precursor, structures were inferred from the existence of the 14 enzymes listed in Figure 1A. As a result, the N-glycan precursor is Glc3Man9GlcNAc2 in most animals, plants, and fungi because there is a complete set of biosynthesis enzymes, whereas smaller structures such as Glc2Man9GlcNAc2, Man9GlcNAc2, and Man5GlcNAc2 are likely to be synthesized in algae and parasites because they lack glucosyltransferases and mannosyltransferases (Fig. 3A). These structures correspond to previous findings; in particular, Man5GlcNAc2 in *Entamoeba histolytica* was experimentally proved by a pioneering work.[27]

GPI-anchor structures were inferred from the existence of PIG-A, M, V, B, and Z (Fig. 1B) without regard to associated proteins, such as PIG-H and PIG-P.[28] Similar to the case of the N-glycan precursor, GPI-anchors are likely to be standard structures of Man4GlcN or Man3GlcN in most of animals, plants, and fungi and, conversely, to be smaller structures in algae and parasites (Fig. 3B). In particular, *Cyanidioschyzon merolae*, a small unicellular red alga, seems to have none of the 5 PIG enzymes and appears to be incapable of synthesizing GPI-anchors. In most cases our method was able to correctly predict glycan structures, but there was an exception. The GPI-anchor structures containing the forth mannose have been found in *Plasmodium falciparum*[29] despite the apparent absence of PIG-Z in its genome. Additional database searches showed that PIG-Z homologs were also absent in other *Plasmodium* genomes, such as *Plasmodium yoelii* and *Plasmodium vivax*. This inconsistency indicates that the PIG-Z sequence of *Plasmodium* has been too changed to be detected by the sequence similarity search or the forth mannose may be transferred not by PIG-Z but by an unknown glycosyltransferase in *Plasmodium*.

The structures of N-glycan precursors and GPI-anchors were generally conserved among organisms except for algae and parasites; however, O-glycans and glycolipids had less common structures. The analysis of six enzymes (Fig. 1E) responsible for the first step of their biosynthesis revealed clear differences between lineages (Fig. S2). Approximately half of the enzymes were lacking in plants and fungi, in contrast to animals, which contained almost all of the enzymes. For example, GALNTs, conserved from human to sea anemone, were not detected in plants and fungi, the *O*-GlcNAc transferases (OGTs) were not detected in any of the three saccharomycetes, and the protein *O*-mannosyltransferases (POMTs) were not detected in the three land plants, indicating that structures of O-glycan are the most diverged in animals.

Further differences among organisms appear on N-glycan extension structures. We predicted extension patterns of N-glycan by using a set of seven families of MGAT, KTR, and MNN, which transfer GlcNAc or Man to N-glycan structures (Fig. 1C). The results

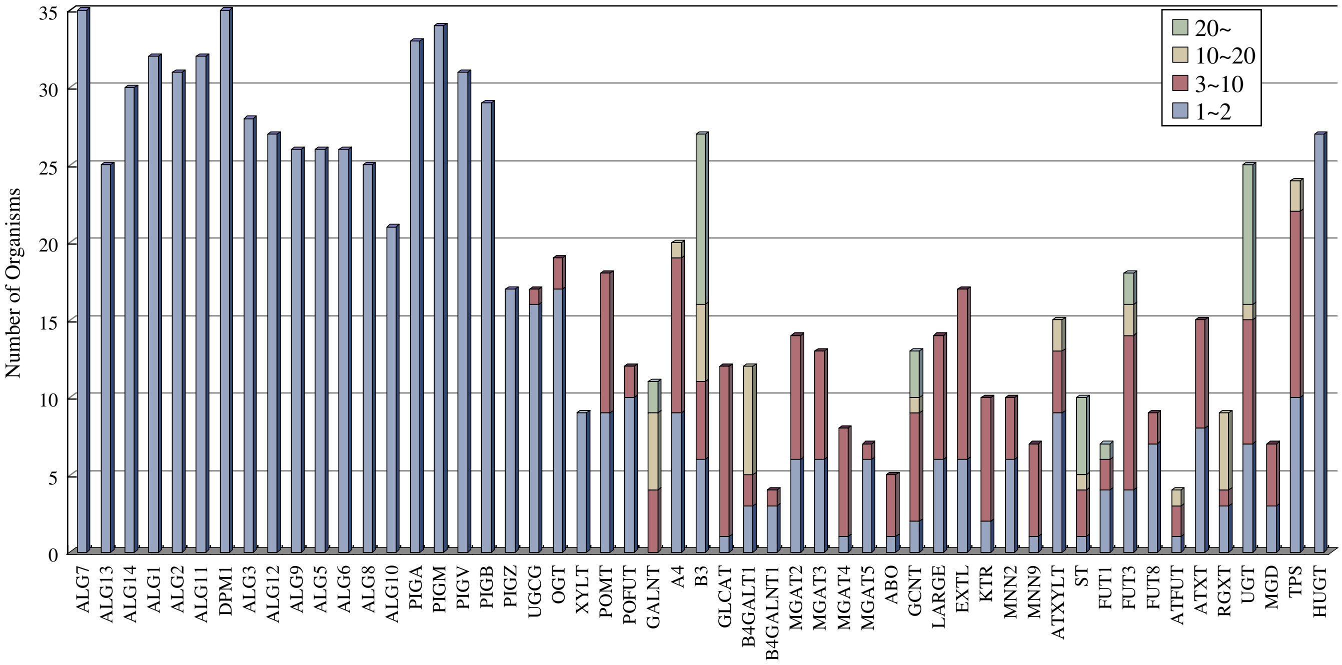

Figure 2. The number of paralogs in the 53 glycosyltransferase families. The *x*-axis indicates the families, *y*-axis indicates the number of organisms, and the four colors indicate the number of paralogs. For example, the first bar represents that 35 organisms possess one or two sequences in the ALG7 family and the third bar from the right represents that three organisms possess 1–2 sequences and four organisms possess 3–10 paralogs in the MGD family.

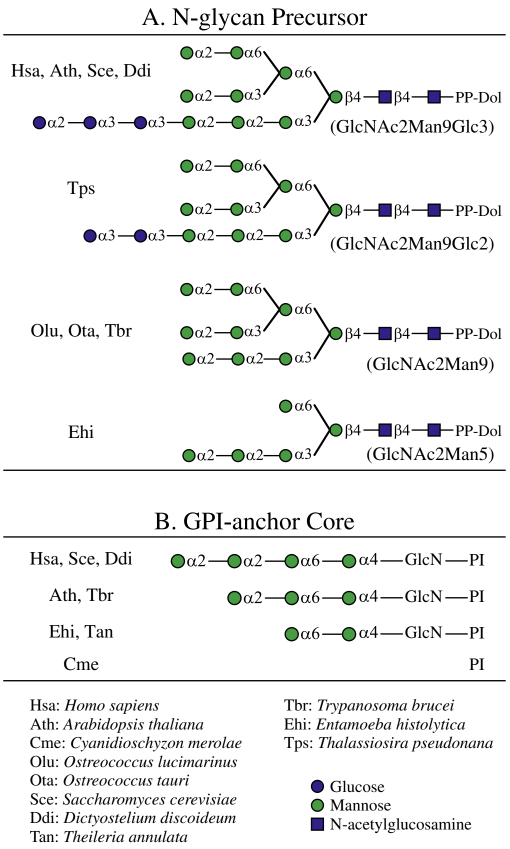

Figure 3. Predicted structures of N-glycan precursor and GPI-anchor. (A) Four representative structures of predicted N-glycan precursors are displayed. The prediction was conducted by using combination of 14 ALG or DPM proteins present in each organism. (B) Four representative structures of predicted GPI-anchors are displayed. The prediction was conducted by using combination of 5 phosphatidylinositol glycan (PIG) proteins present in each organism. Monosaccharides are represented by the symbols defined by the Nomenclature Committee Consortium for Functional Glycomics. Names of organisms are represented by three-letter codes defined by KEGG.

indicate that the extension patterns are different among animals as well as among kingdoms. For example, vertebrates and *Nematostella vectensis* are likely to add GlcNAc with α1-2, α1-4, and α1-6 linkages to Man; *Caenorhabditis elegans* and *Drosophila melanogaster* can catalyze fewer types of glycosylations; land plants add GlcNAc with α1-2 and α1-4; and fungi have several enzymes to transfer Man to Man but no enzymes to transfer GlcNAc to Man (Fig. 4). Although most of the protists do not have any enzymes in the families of MGAT, KTR, and MNN, only *Dictyostelium discoideum* and the marine diatom *Thalassiosira pseudonana* possess MGAT homologs, indicating that a few protists can extend N-glycan core structures with GlcNAc.

2.3. Subfamily analyses of sialyltransferases and β4-glycosyltransferases

The repertoire of glycosyltransferase families revealed differences in basic glycan structures among organisms. To gain more detailed structural insight, it seems that families including different substrate specificities should be investigated at the subfamily level. We thus attempt to define subfamilies of two large families, namely the sialyltransferase family and the β4-glycosyltransferase family.

We constructed a phylogenetic tree of the sialyltransferase family using the length of about 200 amino acids containing sialylmotifs. The phylogenetic tree was divided into eight clusters that were well supported by bootstrap values >98% (Fig. S3). Two of the clusters consist of only plant sequences (clusters G and H), all of which remain uncharacterized (see Section 3). The other six clusters (A–F), which consist of only animal sequences, were assigned substrate specificities based on the functions of the known vertebrate sialyltransferases. From the viewpoint of enzymatic function, clusters B and F (ST6GalNAc) may form a single subfamily as well as clusters D and E (ST3Gal) because of their common glycosidic linkages. Nevertheless, we defined them as different subfamilies due to their low bootstrap values and because specific motifs in two ST6GalNAc subfamilies had been identified.[17] Vertebrates have all six types of sialyltransferases, whereas invertebrates have some or none of them. This classification corresponds to experimental results of vertebrate sialyltransferases. However, note that there is an exception in which a Drosophila sialyltransferase that exhibited highest activity toward GalNAcβ1-4GlcNAc[30] fell into the ST6Gal subfamily. Thus, further experiments with invertebrates need to be performed in order to apply the classification beyond vertebrates.

A phylogenetic tree of the β4 family was constructed using the length of about 140 amino acids containing the β4 motif in the C-terminal region. The phylogenetic tree comprises four clusters (Fig. S4), each with a distinct function, namely the β4Gal-T subfamily (A), the β4GalNAc-T subfamily (B), and two chondroitin synthase (CSS) subfamilies (C, D). The β4Gal-T subfamily (A) is the largest subfamily, including seven human β4-galactosyltransferases. Both vertebrates and invertebrates possess many paralogous sequences in this subfamily. In addition, the subfamily contains three sequences of algae, in contrast to the other subfamilies that consist of only animal sequences. The second subfamily (B) contains two human β4-*N*-acetylgalactosaminyltransferases, which synthesize LacdiNAc (GalNAcβ1-4GlcNAc). The other two clusters (C, D) are the CSS subfamilies including six human enzymes, responsible for elongation of chondroitin sulfate. It is known that four glycosyltransferases that synthesize a core tetrasaccharide of chondroitin sulfate are present in *C. elegans* and *D. melanogaster*[31] as well as vertebrates. Our results show that invertebrates also possess enzymes that are likely to be involved in elongation of chondroitin sulfate. In particular, their homologs are present even in the genome of *N. vectensis*, belonging to the oldest eumetazoan phylum, Cnidaria. These indicate that the biosynthesis of chondroitin sulfate is of ancient origin and is conserved among a wide range of animals.

3. Discussion

3.1. Universal glycosyltransferases and diverged glycosyltransferases

Unexpectedly, our results indicate that glycosyltransferases are present at an approximately constant rate in most eukaryotic genomes, especially in the genomes of free-living organisms, rather than human and mammals possessing a significantly larger number of glycosyltransferases. At the same time, glycosyltransferases have extremely diverse sequences, comprising many families. However, from a wider evolutionary perspective, these families are divided into two groups. One is a group of universal glycosyltransferases, which are widely conserved among organisms, and the other is a group of diverged glycosyltransferases, which are distributed exclusively in specific kingdoms or lineages. These two groups seem to be directly related to two types of glycans: universal glycans and diverged glycans.

Universal glycosyltransferases include, for example, the asparagine-linked glycosylation (ALG) families and the PIG families,

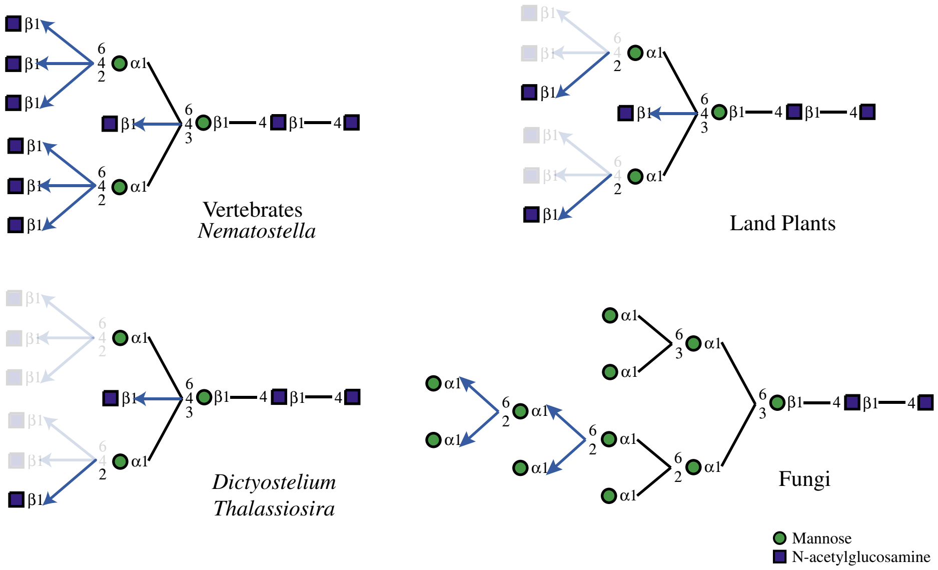

Figure 4. N-Glycan extension patterns in lineages. The prediction was conducted by using combination of seven families of MGAT, KTR, and MNN, which transfer GlcNAc or Man. The blue arrow indicates possible extensions of the N-glycan core in each lineage. Monosaccharides are represented by the symbols defined by the Nomenclature Committee of the Consortium for Functional Glycomics.

which contain few paralogs even in organisms whose genomes have been duplicated during the eukaryotic evolutionary process, such as vertebrates[32] and yeasts[33] (Fig. 2). These characteristics are most likely to reflect the necessity of maintaining the structures they synthesize. In other words, an increase in the number of paralogs does not readily occur in these families because paralogous enzymes with different substrate specificities may alter the structures of the N-glycan precursor or the GPI-anchor, thereby affecting an extremely wide range of proteins. To our knowledge, there are few reports of a monosaccharide or a glycosidic linkage being replaced by another monosaccharide or another linkage in N-glycan precursors or GPI-anchor structures. Furthermore, some of the smaller structures in Figure 3 have been experimentally demonstrated to lack the quality control functions of glycoprotein folding and the endoplasmic reticulum (ER)-associated degradation of proteins, indicating that N-glycan precursor length has profound effects on these functions.[34]

In contrast, diverged glycosyltransferases, which are often found in kingdom- or lineage-specific manners, have many paralogs that have different types of substrate specificities. Large families detected predominantly in animals are the GALNT family, the β3 family, the β4 family, and the ST family, mostly containing more than ten paralogs in each organism. These families contribute to the diversity of O-glycosylation sites,[35] middle structures,[6] and terminal structures,[36] respectively. In land plants, a notable family is the UGT family, containing 112 sequences of *A. thaliana*, 140 sequences of *Oryza sativa*, and 155 sequences of *Populus trichocarpa*. The family, which accounts for about 25% of glycosyltransferases, appears to play a major role in land plant glycosylation. Although many of them are functionally uncharacterized, several studies revealed that some of them recognize and catalyze various secondary metabolites and hormones such as flavonoids, auxins, and cytokinins.[37,38] In fungi, the distribution of glycosyltransferases showed less commonality with that of animals and plants. Our results indicate that more than half of glycosyltransferases in *Saccharomyces cerevisiae* are mannosyltransferases. Taken together, lineage specific evolution has led to paralogs that have various substrate specificities, yielding the apparent differences of glycans and compounds among kingdoms.

3.2. Inferences of newly detected glycosyltransferase functions

We identified approximately 3000 new putative glycosyltransferases from 36 eukaryotic genomes, which were classified into 53 families. We deal here with some of the families in detail by incorporating extra information such as biosynthetic pathways and related proteins. First, we discuss glycosyltransferases of parasitic protists, which have rarely been studied and characterized, compared with those of mammals. Such organisms have been revealed to contain unusual glycan structures[39–41] suggesting that unknown glycosyltransferases are encoded in their genomes. Our analysis detected a total of 244 sequences from 10 parasitic protists (Table S1). To our knowledge, many of the detected proteins remain uncharacterized except for β3-galactosyltransferases in *Leishmania major*[42] and the glycosyltransferases synthesizing GPI-anchor,[43,44] which are believed to play crucial roles in parasite infection and pathogenesis.[45] Nevertheless, indirect evidence allows us to infer the presence of glycosyltransferases belonging to a certain family; for example, a fucosylated oligosaccharide was detected in *T. cruzi*[46] as well as GDP-fucose, which is the donor substrate for all known fucosyltransferases.[47] Further, a very recent report identified functional GDP-fucose synthetases in *Trypanosoma brucei* and demonstrated that the fucose metabolism was essential for cell growth.[48] Thus, fucosyltransferases should be present in trypanosomatids. We detected seven fucosyltransferase-like sequences of trypanosomatids in two families. They are promising candidates for their fucosyltransferases, because the four sequences belonging to the FUT1 family contain two of the three motifs that are conserved across subfamilies.[14] In addition, three sequences belonging to the FUT3 family also clearly contain two motifs, which were identified in the previous work.[49] Another example is the GALNT family, responsible for the O-glycan biosynthesis. The polypeptide GalNAc transferase is well characterized in several animals such as human, mouse, fly, and worm,[50] whereas, in the other kingdoms, homologs were found only in the apicomplexans *Cryptosporidium parvum* and *Cryptosporidium hominis*. These sequences contained a motif including Mn^{2+} coordination residues,[51] which are conserved among animal sequences. In addition, functional polypeptide GalNAc transferases were identified in the apicomplexan *Toxoplasma gondii*,[52] not included in our analysis. The enzyme is thus likely to be distributed into two deeply divergent lineages, namely the animal and the apicomplexa.

Next, we discuss sialic acids in plants, the existence of which is controversial. Sialic acids and sialyltransferases were confirmed to be present in animals,[53] whereas sialic acids are firmly believed to be absent in plants, although a report showed the presence of sialylated glycoproteins in *A. thaliana* cells.[54] Our analysis found several homologs in land plant genomes (see plant clusters in Fig. S3). An additional search against plant EST data (see Section 4) detected a total of 67 similar sequences in 29 genomes of angiosperm, gymnosperm, and moss. Comparison of sequences between animals and plants shows that sialylmotif L, involved in binding to the donor substrate,[55] is well conserved (Fig. S5), strongly suggesting that they have diverged from a common ancestor. In addition, a recent report demonstrated that some of the homologs in *O. sativa* have sialyltransferase-like activity.[56] Nevertheless, it remains unclear whether or not the primary activity of these enzymes is the sialyltransferase activity, because the plant EST sequences lack the two other motifs (sialylmotif S and VS), which are conserved in animal sequences, and other recent studies have re-evaluated the data and concluded that plants do not synthesize sialic acids.[57,58]

Another possibility is that the plant homologs may be involved in the transfer of 3-deoxy-D-manno-octulosonic acid (Kdo), which is a component of rhamnogalacturonan II in the plant primary cell walls. Evolutionary relationships between Kdo and sialic acids are indicated because both of them form unusual activated sugar nucleotide, namely CMP–sialic acid and CMP–Kdo, and also their synthetases have weak sequence similarity.[59] Additionally, using the ATTED-II database,[60] we found that one gene (AT1G08660) encoding a sialyltransferase-like protein in *A. thaliana* is highly co-expressed with the genes encoding Kdo-8-phosphate synthetases (AtkdsA1, AtkdsA2), involved in CMP–Kdo biosynthesis. In any case, sialyltransferase homologs found in plants are probably glycosyltransferases and their enzymatic activities should be further examined.

3.3. The application of our strategy and analysis

In this study, we focused on the eukaryotic genomes and attempted to describe the whole view of glycosyltransferases. Examination of the repertoire of glycosyltransferases elucidated differences in major glycan structures among eukaryotic organisms. Furthermore, our results are also available to predict the repertoire of glycosyltransferases from newly sequenced genomes. As an example, we searched the genome of *Monosiga brevicollis*, which is one of the closest known relatives of metazoans,[61] using HMM profiles constructed from each glycosyltransferase family. The combination of detected glycosyltransferases allows us to readily infer that N-glycan precursors synthesized by *M. brevicollis* are likely to be Man9GlcNAc2, the same as the composition of *T. brucei*, and the organism can synthesize all the five types of O-glycan cores like animals. This strategy would also be applicable to the analysis of bacterial glycans by focusing on bacterial specific glycosyltransferases. The classification system for glycosyltransferases, which is continuously updated and expanded, is available in the KEGG BRITE database (http://www.genome.jp/kegg/brite.html).

4. Materials and methods

We first comprehensively extracted all sequences similar to known glycosyltransferases including slightly or partially similar sequences from a genomic dataset using PSI-BLAST (version 2.2.17).[62] We then manually discarded any false-positive hits by investigating the results of multiple alignments and phylogenetic analyses. Finally, we defined glycosyltransferase families according to phylogenetic tree clusters. Various computational operations were performed with the BioRuby library version 1.1 for the Ruby language (http://bioruby.org/). Methods described here are based on modifications of previously published methods for comprehensive analyses of large protein families.[22,63]

4.1. Searching for similar sequences with PSI-BLAST

As PSI-BLAST targets, we used amino acid sequences from 36 complete or draft-quality whole eukaryotic genomes including nine animals, nine fungi, six plants, and 12 protists (Table S1). These data were derived from KEGG GENES and DGENES Release 44.0. As the query sequences for PSI-BLAST search, we compiled a list of 182 experimentally known glycosyltransferase sequences from representative organisms, namely *Homo sapiens*, *S. cerevisiae*, and *A. thaliana* (Table S2) using the literature and databases, such as KEGG GLYCAN,[23] GGDB,[64] and CAZy.[65]

PSI-BLAST was performed sequentially for each query sequence using an expectation value of 0.1 and a maximum number of passes of 10. We selected the one that returns the maximum number of glycosyltransferase-like proteins in the PSI-BLAST iterations. Note that when query sequences are similar to each other, results of PSI-BLAST searches are also similar. For example, whichever human sialyltransferase is selected as a query, results of the searches are roughly the same. We thus merged such similar results into a single file and removed redundant sequences.

4.2. Discarding false-positive sequences

To remove false-positive sequences, we generated a multiple alignment using MAFFT version 6.24[66] and calculated phylogenetic trees with the neighbor-joining method[67] using ClustalW version 1.83.[68] The reliability of the tree topologies was assessed by 1000 bootstrap iterations. Trees were visualized with MEGA[69] and Interactive Tree Of Life (iTOL).[70] We then manually checked all the alignments and the trees, and subsequently determined false-positive sequences using two criteria, literature information and motif information. If one or more proteins in a cluster of a phylogenetic tree were annotated as non-glycosyltransferase proteins according to the literature or database annotations, the cluster was discarded. We also discarded the sequences lacking strongly conserved regions among the other members in each alignment. The procedure for discarding sequences was repeated gradually to avoid discarding real glycosyltransferases.

4.3. Defining glycosyltransferase families

We obtained a number of sequence groups, each of which was composed of non-redundant glycosyltransferase-like sequences that were similar to each other. In some cases, a group was defined as a single family since the members share a common function. On the other hand, the group in which sequences formed distinct clusters of different functions was divided into particular families such as ALG6, 8, and PIG-M. The most standard names in the literature and databases are adopted as the family name.

4.4. Searching for sialyltransferases in the plant EST data

Because only a few plant genomes were available, we used plant EST consensus contigs derived from KEGG EGENES[71] including angiosperm, gymnosperm, and moss sequences to confirm the existence of sialyltransferase-like proteins in plants (see Section 3). The nucleotide sequences were translated in all six reading frames, using their amino acid translations as the target dataset. We searched in the dataset using hidden Markov model profiles

(HMM profiles) built from sialylmotifs in 146 obtained animal sialyltransferases with HMMER version 2.3.2 (http://hmmer.janelia.org/). Graphical representations of the conservation patterns of consensus sequences were generated by WebLogo.[72]

Acknowledgments

We thank Dr. Nelson Hayes for critical reading of our manuscript. This work was supported by grants from the Institute for Bioinformatics Research and Development of the Japan Science and Technology Agency, the Ministry of Education, Culture, Sports, Science and Technology of Japan, and the Consortium for Functional Glycomics. K.H. was supported by Research Fellowship for Young Scientists from the Japan Society for the Promotion of Science. The computational resource was provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.carres.2009.03.001.

References

1. Inatani, M.; Irie, F.; Plump, A. S.; Tessier-Lavigne, M.; Yamaguchi, Y. *Science* **2003**, *302*, 1044–1046.
2. Crocker, P. R.; Paulson, J. C.; Varki, A. *Nat. Rev. Immunol.* **2007**, *7*, 255–266.
3. Collins, B. E.; Paulson, J. C. *Curr. Opin. Chem. Biol.* **2004**, *8*, 617–625.
4. Mayer, D. C.; Jiang, L.; Achur, R. N.; Kakizaki, I.; Gowda, D. C.; Miller, L. H. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 2358–2362.
5. Kikuchi, N.; Narimatsu, H. *Biochim. Biophys. Acta* **2006**, *1760*, 578–583.
6. Narimatsu, H. *Curr. Opin. Struct. Biol.* **2006**, *16*, 567–575.
7. Amano, K.; Chiba, Y.; Kasahara, Y.; Kato, Y.; Kaneko, M. K.; Kuno, A.; Ito, H.; Kobayashi, K.; Hirabayashi, J.; Jigami, Y.; Narimatsu, H. *Proc. Natl. Acad .Sci. U.S.A.* **2008**, *105*, 3232–3237.
8. Castilho, A.; Pabst, M.; Leonard, R.; Veit, C.; Altmann, F.; Mach, L.; Glossl, J.; Strasser, R.; Steinkellner, H. *Plant Physiol.* **2008**, *147*, 331–339.
9. Breton, C.; Snajdrova, L.; Jeanneau, C.; Koca, J.; Imberty, A. *Glycobiology* **2006**, *16*, 29R–37R.
10. Hu, Y.; Walker, S. *Chem. Biol.* **2002**, *9*, 1287–1296.
11. Coutinho, P. M.; Deleury, E.; Davies, G. J.; Henrissat, B. *J. Mol. Biol.* **2003**, *328*, 307–317.
12. Cantarel, B. L.; Coutinho, P. M.; Rancurel, C.; Bernard, T.; Lombard, V.; Henrissat, B. *Nucleic Acids Res.* **2009**, *37*, D233–238.
13. Oriol, R.; Martinez-Duncker, I.; Chantret, I.; Mollicone, R.; Codogno, P. *Mol. Biol. Evol.* **2002**, *19*, 1451–1463.
14. Martinez-Duncker, I.; Mollicone, R.; Candelier, J. J.; Breton, C.; Oriol, R. *Glycobiology* **2003**, *13*, 1C–5C.
15. Kaneko, M.; Nishihara, S.; Narimatsu, H.; Saitou, N. *Trends Glycosci. Glycotech.* **2001**, *13*, 147–155.
16. Ross, J.; Li, Y.; Lim, E.; Bowles, D. J. *Genome Biol.* **2001**, *2*. REVIEWS3004.
17. Patel, R. Y.; Balaji, P. V. *Glycobiology* **2006**, *16*, 108–116.
18. Turcot-Dubois, A. L.; Le Moullac-Vaidye, B.; Despiau, S.; Roubinet, F.; Bovin, N.; Le Pendu, J.; Blancher, A. *Glycobiology* **2007**, *17*, 516–528.
19. Kanehisa, M.; Araki, M.; Goto, S.; Hattori, M.; Hirakawa, M.; Itoh, M.; Katayama, T.; Kawashima, S.; Okuda, S.; Tokimatsu, T.; Yamanishi, Y. *Nucleic Acids Res.* **2008**, *36*, D480–484.
20. Kawano, S.; Hashimoto, K.; Miyama, T.; Goto, S.; Kanehisa, M. *Bioinformatics* **2005**, *21*, 3976–3982.
21. Suga, A.; Yamanishi, Y.; Hashimoto, K.; Goto, S.; Kanehisa, M. *Genome Inform.* **2007**, *18*, 237–246.
22. Hashimoto, K.; Yoshizawa, A. C.; Okuda, S.; Kuma, K.; Goto, S.; Kanehisa, M. *J. Lipid Res.* **2008**, *49*, 183–191.
23. Hashimoto, K.; Goto, S.; Kawano, S.; Aoki-Kinoshita, K. F.; Ueda, N.; Hamajima, M.; Kawasaki, T.; Kanehisa, M. *Glycobiology* **2006**, *16*, 63R–70R.
24. Lutteke, T. *Chembiochem* **2008**.
25. Raman, R.; Venkataraman, M.; Ramakrishnan, S.; Lang, W.; Raguram, S.; Sasisekharan, R. *Glycobiology* **2006**, *16*, 82R–90R.
26. Lutteke, T.; Bohne-Lang, A.; Loss, A.; Goetz, T.; Frank, M.; von der Lieth, C. W. *Glycobiology* **2006**, *16*, 71R–81R.
27. Samuelson, J.; Banerjee, S.; Magnelli, P.; Cui, J.; Kelleher, D. J.; Gilmore, R.; Robbins, P. W. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 1548–1553.
28. Murakami, Y.; Siripanyaphinyo, U.; Hong, Y.; Tashima, Y.; Maeda, Y.; Kinoshita, T. *Mol.. Biol. Cell* **2005**, *16*, 5236–5246.
29. Delorenzi, M.; Sexton, A.; Shams-Eldin, H.; Schwarz, R. T.; Speed, T.; Schofield, L. *Infect. Immun.* **2002**, *70*, 4510–4522.
30. Koles, K.; Irvine, K. D.; Panin, V. M. *J. Biol. Chem.* **2004**, *279*, 4346–4357.
31. Ueyama, M.; Takemae, H.; Ohmae, Y.; Yoshida, H.; Toyoda, H.; Ueda, R.; Nishihara, S. *J. Biol. Chem.* **2008**, *283*, 6076–6084.
32. Dehal, P.; Boore, J. L. *PLoS Biol.* **2005**, *3*, e314.
33. Kellis, M.; Birren, B. W.; Lander, E. S. *Nature* **2004**, *428*, 617–624.
34. Banerjee, S.; Vishwanath, P.; Cui, J.; Kelleher, D. J.; Gilmore, R.; Robbins, P. W.; Samuelson, J. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 11676–11681.
35. Ten Hagen, K. G.; Fritz, T. A.; Tabak, L. A. *Glycobiology* **2003**, *13*, 1R–16R.
36. Harduin-Lepers, A.; Vallejo-Ruiz, V.; Krzewinski-Recchi, M. A.; Samyn-Petit, B.; Julien, S.; Delannoy, P. *Biochimie* **2001**, *83*, 727–737.
37. Bowles, D.; Isayenkova, J.; Lim, E. K.; Poppenberger, B. *Curr. Opin. Plant Biol.* **2005**, *8*, 254–263.
38. Hou, B.; Lim, E. K.; Higgins, G. S.; Bowles, D. J. *J. Biol. Chem.* **2004**, *279*, 47822–47832.
39. Mendonca-Previato, L.; Todeschini, A. R.; Heise, N.; Previato, J. O. *Curr. Opin. Struct. Biol.* **2005**, *15*, 499–505.
40. Atrih, A.; Richardson, J. M.; Prescott, A. R.; Ferguson, M. A. *J. Biol. Chem.* **2005**, *280*, 865–871.
41. Sernee, M. F.; Ralton, J. E.; Dinev, Z.; Khairallah, G. N.; O'Hair, R. A.; Williams, S. J.; McConville, M. J. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 9458–9463.
42. Dobson, D. E.; Scholtes, L. D.; Myler, P. J.; Turco, S. J.; Beverley, S. M. *Mol. Biochem. Parasitol.* **2006**, *146*, 231–241.
43. Weber, C.; Blazquez, S.; Marion, S.; Ausseur, C.; Vats, D.; Krzeminski, M.; Rigothier, M. C.; Maroun, R. C.; Bhattacharya, A.; Guillen, N. *PLoS Negl. Trop. Dis.* **2008**, *2*, e165.
44. Basagoudanavar, S. H.; Feng, X.; Krishnegowda, G.; Muthusamy, A.; Gowda, D. C. *Biochem. Biophys. Res. Commun.* **2007**, *364*, 748–754.
45. Ferguson, M. A. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 10673–10675.
46. Barboza, M.; Duschak, V. G.; Fukuyama, Y.; Nonami, H.; Erra-Balsells, R.; Cazzulo, J. J.; Couto, A. S. *Febs J.* **2005**, *272*, 3803–3815.
47. Turnock, D. C.; Ferguson, M. A. *Eukaryot. Cell* **2007**, *6*, 1450–1463.
48. Turnock, D. C.; Izquierdo, L.; Ferguson, M. A. *J. Biol. Chem.* **2007**, *282*, 28853–28863.
49. Oriol, R.; Mollicone, R.; Cailleau, A.; Balanzino, L.; Breton, C. *Glycobiology* **1999**, *9*, 323–334.
50. Ten Hagen, K. G.; Tran, D. T.; Gerken, T. A.; Stein, D. S.; Zhang, Z. *J. Biol. Chem.* **2003**, *278*, 35039–35048.
51. Fritz, T. A.; Hurley, J. H.; Trinh, L. B.; Shiloach, J.; Tabak, L. A. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 15307–15312.
52. Wojczyk, B. S.; Stwora-Wojczyk, M. M.; Hagen, F. K.; Striepen, B.; Hang, H. C.; Bertozzi, C. R.; Roos, D. S.; Spitalnik, S. L. *Mol. Biochem. Parasitol.* **2003**, *131*, 93–107.
53. Harduin-Lepers, A.; Mollicone, R.; Delannoy, P.; Oriol, R. *Glycobiology* **2005**, *15*, 805–817.
54. Shah, M. M.; Fujiyama, K.; Flynn, C. R.; Joshi, L. *Nat. Biotechnol.* **2003**, *21*, 1470–1471.
55. Datta, A. K.; Paulson, J. C. *J. Biol. Chem.* **1995**, *270*, 1497–1500.
56. Takashima, S.; Abe, T.; Yoshida, S.; Kawahigashi, H.; Saito, T.; Tsuji, S.; Tsujimoto, M. *J. Biochem.* **2006**, *139*, 279–287.
57. Seveno, M.; Bardor, M.; Paccalet, T.; Gomord, V.; Lerouge, P.; Faye, L. *Nat. Biotechnol.* **2004**, *22*, 1351–1352. author reply 1352–1353.
58. Zeleny, R.; Kolarich, D.; Strasser, R.; Altmann, F. *Planta* **2006**, *224*, 222–227.
59. Angata, T.; Varki, A. *Chem. Rev.* **2002**, *102*, 439–469.
60. Obayashi, T.; Kinoshita, K.; Nakai, K.; Shibaoka, M.; Hayashi, S.; Saeki, M.; Shibata, D.; Saito, K.; Ohta, H. *Nucleic Acids Res.* **2007**, *35*, D863–D869.
61. King, N.; Westbrook, M. J.; Young, S. L.; Kuo, A.; Abedin, M.; Chapman, J.; Fairclough, S.; Hellsten, U.; Isogai, Y.; Letunic, I.; Marr, M.; Pincus, D.; Putnam, N.; Rokas, A.; Wright, K. J.; Zuzow, R.; Dirks, W.; Good, M.; Goodstein, D.; Lemons, D.; Li, W.; Lyons, J. B.; Morris, A.; Nichols, S.; Richter, D. J.; Salamov, A.; Sequencing, J. G.; Bork, P.; Lim, W. A.; Manning, G.; Miller, W. T.; McGinnis, W.; Shapiro, H.; Tjian, R.; Grigoriev, I. V.; Rokhsar, D. *Nature* **2008**, *451*, 783–788.
62. Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. *Nucleic Acids Res.* **1997**, *25*, 3389–3402.
63. Yoshizawa, A. C.; Kawashima, S.; Okuda, S.; Fujita, M.; Itoh, M.; Moriya, Y.; Hattori, M.; Kanehisa, M. *Traffic* **2006**, *7*, 1104–1118.
64. Narimatsu, H. *Glycoconj J* **2004**, *21*, 17–24.
65. Coutinho, P. M.; Henrissat, B. Carbohydrate-active Enzymes: An Integrated Approach. In *Recent Advances in Carbohydrate Bioengineering*; Gilbert, H. J., Davies, G., Henrissat, B., Svensson, B., Eds.; The Royal Society of Chemistry: Cambridge, UK, 1999; pp 3–14.
66. Katoh, K.; Kuma, K.; Toh, H.; Miyata, T. *Nucleic Acids Res.* **2005**, *33*, 511–518.
67. Saitou, N.; Nei, M. *Mol. Biol. Evol.* **1987**, *4*, 406–425.
68. Aiyar, A. *Methods Mol. Biol.* **2000**, *132*, 221–241.
69. Tamura, K.; Dudley, J.; Nei, M.; Kumar, S. *Mol. Biol. Evol.* **2007**, *24*, 1596–1599.
70. Letunic, I.; Bork, P. *Bioinformatics* **2007**, *23*, 127–128.
71. Masoudi-Nejad, A.; Goto, S.; Jauregui, R.; Ito, M.; Kawashima, S.; Moriya, Y.; Endo, T. R.; Kanehisa, M. *Plant Physiol.* **2007**, *144*, 857–866.
72. Crooks, G. E.; Hon, G.; Chandonia, J. M.; Brenner, S. E. *Genome Res.* **2004**, *14*, 1188–1190.