**MINI REVIEW**

Conserved domains of glycosyltransferases

Dmitri Kapitonov and Robert K. Yu

Department of Biochemistry and Molecular Biophysics, Medical College of Virginia Campus, Virginia Commonwealth University, Richmond, VA

23298–0614, USA

Received on October 12, 1998; revised on March 2, 1999; accepted on March 8, 1999

1To whom correspondence should be addressed

Glycosyltransferases catalyze the synthesis of glycoconjugates by transferring a properly activated sugar residue to an appropriate acceptor molecule or aglycone for chain initiation and elongation. The acceptor can be a lipid, a protein, a heterocyclic compound, or another carbohydrate residue. A catalytic reaction is believed to involve the recognition of both the donor and acceptor by suitable domains, as well as the catalytic site of the enzyme. To elucidate the structural requirements for substrate recognition and catalytic reactions of glycosyltransferases, we have searched the databases for homologous sequences, identified conserved amino acid residues, and proposed potential domain motifs for these enzymes. Depending on the configuration of the anomeric functional group of the glycosyl donor molecule and of the resulting glycoconjugate, all known glycosyltransferases can be divided into two major types: retaining glycosyltransferases, which transfer sugar residue with the retention of anomeric configuration, and inverting glycosyltransferases, which transfer sugar residue with the inversion of anomeric configuration. One conserved domain of the inverting glycosyltransferases identified in the database is responsible for the recognition of a pyrimidine nucleotide, which is either the UDP or the TDP portion of a donor sugar-nucleotide molecule. This domain is termed “Nucleotide Recognition Domain 1,” or NRD1α, since the type of nucleotide is the only common structure among the sugar donors and acceptors. NRD1β is present in 140 glycosyltransferases. The central portion of the NRD1β domain is very similar to the domain that is present in one family of retaining glycosyltransferases. This family is termed NRD1α to designate the similarity and stereochemistry of sugar transfer, and it consists of 77 glycosyltransferases identified thus far. In the central portion there is a homologous region for these two families and this region probably has a catalytic function. A third conserved domain is found exclusively in membrane-bound glycosyltransferases and is termed NRD2; this domain is present in 98 glycosyltransferases. All three identified NRDs are present in archaeabacterial, euabacterial, viral, and eukaryotic glycosyltransferases. The present article presents the alignment of conserved NRD domains and also presents a brief overview of the analyzed glycosyltransferases which comprise about 65% of all known sugar-nucleotide dependent (Leloir-type) and putative glycosyltransferases in different databases. A potential mechanism for the catalytic reaction is also proposed. This proposed mechanism should facilitate the design of experiments to elucidate the regulatory mechanisms of glycosylation reactions. Amino acid sequence information within the conserved domain may be utilized to design degenerate primers for identifying DNA encoding new glycosyltransferases.

**Key words:** glycosyltransferase/domain structure/classification/nucleotide-binding domain

**Introduction**

Glycosyltransferases catalyze the synthesis of glycoconjugates, including glycolipids, glycoproteins, and polysaccharides, by transferring an activated mono- or oligosaccharide residue to an existing acceptor molecule for the initiation or elongation of the carbohydrate chain. Because the glycosylation reaction is highly specific with respect to both the anomeric configuration of the sugar residue and the site of the addition, it is expected that unique domain structures for substrate recognition and nucleotide-sugar binding are located within the enzyme molecule. Indeed, common amino acid sequences have been deduced for homologous binding sites. For example, sialyltransferases have sialylmotifs that may participate in the recognition of the donor substrate, CMP-sialic acid (Paulson and Colley, 1985; Datta and Paulson, 1995; Katsutoshi, 1996). The hexapeptide RDKKND in Galβ1–3 galactosyltransferase and RDKKNE in GlcNAcβ1–4 galactosyltransferase have been suggested as the binding site for UDP-Gal (Joziassé et al., 1985, 1989; Joziassé, 1992). These domains have been extensively reviewed in the literature and therefore are not a subject of this review. During the last few years, the amino acid sequences of a number of glycosyltransferases have been identified using sequence data provided by the complete genomic sequences obtained for such organisms as C.elegans, yeasts, several bacteria, and ongoing human and mouse genome projects. These advances have made it possible to deduce conserved domain structures in glycosyltransferases of diverse specificity.

Our interest in this area stems from our recent systematic analysis of the cDNAs encoding UDP-galactose: ceramide galactosyltransferases or CGT (EC 2.4.1.45), the enzyme responsible for the biosynthesis of galactosylceramide. This enzyme has been suggested to play a critical role in myelin formation (Costantino-Ceccarini and Suzuki, 1975; Costantino-Ceccarini and Poduslo, 1989), signal transduction (Dyer and Benjamins, 1989, 1991; Joshi and Mishra, 1992), viral (Baccetti et al., 1991, 1994; Bhat et al., 1991, 1994; Harouse et al., 1991; Yahi et al., 1992, 1995; Brogi et al., 1995; Toniolo et al., 1995) microbial adhesion (Garcia Monco et al., 1992; Khan et al., 1996; Kaneda et al., 1997), and oligodendrocyte development (Mirskey et al., 1980; Bansal and Pfeiffer, 1989). In our previous studies we have described the isolation of human, mouse, chicken, and bovine...
cDNAs for CGT and also the expression of human CGT cDNA (Kapitonov and Yu, 1997a,b). The CGT cDNA sequence appears to be evolutionarily conserved among higher vertebrates; the similarity of CGT from two different systematic classes of vertebrates exceeds 80% and from two different systematic orders of mammals is about 90%. CGT also has a high degree of sequence similarity with glucuronosyltransferases (UGT) (Schulte and Stoffel, 1993; Stahl et al., 1994). All species studied so far have only one copy of the CGT gene. The deduced amino acid sequence of human CGT reveals that it is a 61 kDa protein of 541 amino acid residues, 21 of which are charged. The protein has a calculated pI of 10.36. It maintains all the conserved features of other vertebrate CGTs: a C-terminal transmembrane hydrophobic domain, an endoplasmic reticulum retention signal, and three N-linked glycosylation sites at positions 83, 338, and 447 (none of these are conserved in glucuronosyltransferases). There are six potential sites for casein kinase 2 (CK2) phosphorylation (positions 85, 104, 234, 368, 408, and 410), two potential PKC phosphorylation sites (positions 100 and 426), and a potential N-terminal myristoylation site at position 16 that could be exposed upon cleavage of the signal peptide. Using the amino acid sequence information obtained from the four animal species, we deduced a consensus sequence of CGT, which we used to analyze the database for conserved domains of other glycosyltransferases. In this review, since we analyzed both known and putative glycosyltransferases in order to provide adequate description, all known glycosyltransferases are described by their proper protein names and organisms, while for putative glycosyltransferases from the database, an accession number immediately follows the particular organism. When several genes are located on the same cosmid, the accession number of the cosmid is followed by a dot and a gene number to indicate potential participation in the same biosynthetic pathway in lower eukaryotes. When one protein is chosen to represent a group of highly homologous enzymes, the name of the protein is followed by the number of the family members.

The NRD1β family of glycosyltransferases

When comparing CGT and UGT to some microbial glycosyltransferases, we identified a region with a high degree of similarity, which we have named “Nucleotide Recognition Domain type β” or NRD1β, to account for the inversion of anomic configuration and the resulting configuration of added sugar. This domain is located close to the C-terminal end before the C-terminal transmembrane domain, between residues 357–396 in CGT and between residues 369–407 in UGT. In both CGT and UGT this domain is encoded by one exon. A much smaller area of sequence similarity located closer to the N-terminus is identified as NRD1β S (S for small). In order to understand the functional significance of the NRD1β, we generated a homologous amino acid sequence pattern and searched several DNA databases translated in all 6 frames using the programs FindPattern and TFastA (Wisconsin Package, GCG). We found 140 proteins containing this domain. Some representative sequences are shown in Figure 1 together with the enzyme class (if known), gene or protein name, accession numbers and the organisms. If the genes are grouped together on the same chromosome, the accession number of the cosmid followed by dot and the gene number are given. Colocalization of the genes on the same cosmid may give a clue to their function since genes associated with the same biosynthetic pathway tend to group together in the C.elegans genome. Most of these proteins are putative glycosyltransferases encoded in the open reading frames (ORF) found in the genome projects and detected with programs such as Genefinder that utilize criteria such as codon preference and third position bias. The existence of common domains has been described for UGT (Mackenzie et al., 1997). The evolutionary relationship among all major groups of NRD1β glycosyltransferases is shown in Figure 2. Most of the eukaryotic members of NRD1β family have C-terminal transmembrane domain (with the exception of several putative glycosyltransferases from C.elegans, where the exon encoding such a domain could be missed by the program). The topology of NRD1β glycosyltransferases has been studied extensively for UGTs and CGT. All known UGTs have type I transmembrane topology with N-terminus and catalytic domain inside ER and microsomes (Meech and Mackenzie, 1997). Two different topologies were described for CGT biosynthesis: non-hydroxy fatty acid galactosylceramide is synthesized in the cytosolic leaflet of the Golgi, which would be possible only with type II orientation of CGT, while preferred substrate hydroxy-fatty acid galactosyl ceramide is synthesized in the luminal leaflet of the ER (Burger et al., 1996). Since there is only one CGT (Busio et al., 1996b; Coetze et al., 1996) that can synthesize both types of GalCer (Schaeren-Wiemers et al., 1995; van der Bijl et al., 1996) and that resides at the density of ER on a sucrose gradient, type I topology with ER localization is probably more accurate, similar to homologous UGT.

A potential Ser/Thr residue phosphorylation site located in the middle of the NRD1β domain that follows the conserved S/TXXE/D pattern for casein kinase 2 (CK2) phosphorylation (where X can be any amino acid residue other than P). Phosphorylation of this site could potentially regulate enzyme activity.

All NRD1β glycosyltransferases can be divided into five classes based on the degree of sequence similarity. Class I is a group of relatively large proteins (950–1050 amino acid residues) that are homologous to glycosyltransferases from Saccharomyces cerevisiae and Caenorhabditis elegans. Class II constitutes proteins of intermediate size (about 700 amino acid residues) from C.elegans. All bacterial glycosyltransferases fall into Class III. Plant glycosyltransferases constitute Class IV. Class V consists of the remaining glycosyltransferases from C.elegans, glucuronosyltransferases (UGT), and ceramide galactosyltransferases (CGT). Proteins with high degrees of sequence similarity are found in these groups, NRD1β glycosyltransferases are also found in archaeabacteria Methanococcus jannaschii, indicating that these enzymes evolved prior to the separation of archaeabacteria, eubacteria, and eukaryotes. Protozoa seldom have more than one NRD1β glycosyltransferase, while metazoa have multiple groups of these enzymes.

The most studied metazoan organism, whose genome sequence was about 70% completed at the time of this analysis is C.elegans (Sulston et al., 1992; Watson et al., 1993; Wilson et al., 1994). C.elegans is a free-living nematode comprised of a total of 959 cells with a well-defined pattern of cell division. Based on total protein similarity, we have divided putative glycosyltransferases identified using sequence data from the C.elegans genome into 25 groups, 8 of which contain three proteins or more. Some of the ORF products have been previously identified as similar to UGT. Our analysis indicates that these proteins are not strictly glucuronosyltransferases, but rather glycosyltransferases that are potentially able to catalyze the transfer of either galactose, glucose, or glucuronic acid. In most of the C.elegans putative glycosyltransferases, the NRD1β domain is encoded by one exon.
Fig. 1. Multiple alignment of the NRD1β family of proteins. All conservatively substituted residues that are present in more than 60% of the glycosyltransferases (GT) compared are shown as a NRD1β Motif. Signal peptide (SIG), transmembrane domain (TM), potential endoplasmic reticulum retention signal (RETENTION), all truncated, as well as small NRD1β S domain, and large NRD1β L domain, are indicated in Domains. P* indicates the S/T residue in the middle of NRD1β site that could be potentially phosphorylated by CK2. 1 and 2, residues that can potentially participate in the catalytic reaction (E at the position 2). The group number is followed by the type of glycosyltransferase (if known), then by the gene name (if given), then by organism and by accession number. For the multimember group the number of group members is indicated in brackets at the end. For C.elegans the accession number of a cosmid is followed by the gene number. S.c., Saccharomyces cerevisiae; C.e., Caenorhabditis elegans, nematoda; M.l., Mycobacterium leprae; M.t., Mycobacterium tuberculosis; Stc., Streptomyces capreolus; Ol, oleandomycin; S.a., Streptomyces antibioticus; B.s., Bacillus subtilis; Z, zeaxanthin; E.u., Erwinia uredovora; A.o., Amycolatopsis orientalis; RhuT, rhamnose/transferase; Pa., Pseudomonas aeruginosa; Ba, baumycin; Stc., Streptomyces C5; S.p., Streptomyces puercus; DaurT, daunosaminetransferase; GlcT, glucosyltransferase; E.u., Erwinia uredovora; C.e., Caenorhabditis elegans; C.c., Cuonarhabditis elegans; nema, homologues; M.e., Methanococcus jannaschii, archaebacteria; G, glucose; N.t., Nicotiana tabacum; L.esc., Lycopersicon esculentum; F, flavonol; Z.m, Zea mais; Pet., Petunia hybrida; I, indole-3-acetate; A.t., Arabidopsis thaliana; C.e.g., Caenorhabditis elegans; UGT, glucuronosyltransferase; Pplatessa, Pleurodectes platessa; CGT, ceramide galactosyltransferase; E, ecdysteroid; NPV, nuclear polyhedrosis virus. The transmembrane domain and signal peptide are truncated. Motif, amino acids residues occurring in more than 60% of UGT1, UGT2, CGT and NRD1β. The conserved α-helical portion is underlined. Putative proton donor and nucleophile involved in reaction mechanism are marked as 1 and 2 correspondingly. Numbers on the left column indicate the position of the first amino acid residue of the domain; numbers on the right column indicate the total length of the protein. P, Total length of the protein. Psgene, possible pseudogene; altern, alternative splice sites have been chosen.

Function of the NRD1β family members

As a result of our database analysis, we have discovered a new family of glycosyltransferases which we name as the NRD1β family. Class I NRD1β glycosyltransferases include two proteins of 1198 and 949 amino acid residues from S.cerevisiae (Q06321) and C.elegans (Z71177.2), respectively. Class II NRD1β proteins consist of six putative glycosyltransferases from C.elegans of 687–795 amino acid residues. Class III (microbial glycosyltransferases) NRD1β proteins include three proteins from C.elegans that catalyze the conversion of zeaxanthin (Figure 3A) into zeaxanthin-β-diglucoside using UDP-glucose as a substrate. This is a step in the biosynthesis of carotenoid, which is used for protection against ultraviolet light. All known zeaxanthin glucosyltransferases are compared in Figure 4A, and one member of this group is shown in Figure 2.

Oleandomycin glycosyltransferase from Streptomyces antibioticus catalyzes the transfer of a glucosyl moiety from UDP-glucose to the 2′-hydroxyl group of desosamine attached to the oleandomycin (Hernandez et al., 1993) (Figure 3B). Macrolide glycosyltransferase from Streptomyces lividans catalyzes the transfer of a glucose or galactose residue from UDP-glucose or UDP-galactose, respectively, to the 2′-hydroxyl group of desosamine or the mycaminose moiety at the C5 position of the lactone.
Fig. 2. The evolutionary relationship between the NRD1β family members. For the explanation of abbreviations used see the legend for Figure 1. All sequences are grouped in five classes. Members of Class I are underlined, members of Class II are shown in italics, Class III of microbial NRD1β glycosyltransferases is shown dark gray, Class IV of plant NRD1β glycosyltransferases is underlined, Class V of animal glycosyltransferases is shown in bold italic. C.elegans sequences are grouped according to similarity into 25 groups. Note that CGT, UGT1 and UGT2 appear as one group of Class V (shown in black). The corresponding division into classes is shown on the left margin by a dotted line.
Class III microbial glycosyltransferases from *Mycobacteria*: (Q49841) from *Mycobacterium leprae* and (AD000002) from *Mycobacterium tuberculosis* (Figure 2).

Class IV (plant glycosyltransferases) of the NRD1β proteins is composed of several groups. Flavonol 3-O-glucosyltransferases from plants catalyze the transfer of the glucosyl residue from UDP-glucose to flavonol (Figure 3F) to produce flavonol 3-O-β-glucoside in one of the last steps of anthocyanin biosynthesis. All known flavonol glucosyltransferases are compared in Figure 4D and one member is shown in Figure 2. Our sequence analysis indicates that the glucosyltransferase identified as CGT7 glucose:glucosyltransferase (X77464) and a glycosyltransferase from *Solanum molongena* are probably flavonol glucosyltransferases. A group of indole glucosyltransferases is responsible for the transfer of a glucose residue from UDP-glucose to the indole-3-acetate moiety of plant growth hormone (Figure 3G). All known indole-3-acetate glucosyltransferases are compared in Figure 4E, and one member is shown in Figure 2. Our analysis indicates that jasmonate-induced glucosyltransferase from *Nicotiana tabacum* (GenBank AB000623) is probably indole-3-acetate glucosyltransferase.

A group of glucose-glucosyltransferases that catalyzes the transfer of a glucosyl residue from UDP-glucose to glucose is compared in Figure 4F, and one member is shown in Figure 2. Our analysis indicates that a glycosyltransferase from *Nicotiana tabacum* (U32643) and a glycosyltransferase from *Lycopersicum*...
In addition to CGT and UGT, there are other glycosyltransferases with known functions that belong to Class V of the NRD1β family. Among them are the bacularoviral (nucleopolyhedrovirus) edysteroid glycosyltransferases (EGTs) that catalyze the transfer of a glucosyl residue from UDP-glucose to ec dysteroids; the latter are insect molting hormones, e.g., ec dysone (Figure 3H) (O’Reilly and Miller, 1989). The expression of EGTs interferes with normal insect development and blocks molting. All known edysteroid glycosyltransferases are combined into one group (Figure 4G) of which one member, ACMNPV, is shown in Figure 2. However, one member of the EGT group conjugates ec dysteroids with galactose rather than glucose (O’Reilly et al., 1992).

Group 9 of *C. elegans* glycosyltransferases contains two proteins, one with 440 amino acid residues and the other with 435 amino acid residues. Both are encoded on the same cosmids: nine *C. elegans* (Z78200.8) and nine *C. elegans* (Z78200.7). UGT is a family of microsomal enzymes that transfer glucuronic acid (GlcA) to bilirubin, phenol, and several xenobiotics and endogenous steroid compounds. Currently there are two groups of highly homologous enzymes: UGT 1 and UGT 2. The UGT1 group consists of eight members: human bilirubin-specific UGT1A (P22309) and UGT1D (P22310), UGT1B (P36509), UGT1C (P35503), UGT1E (P35504), phenol-specific UGT1F (P19224) (Ritter et al., 1992a,b), rat p-nitrophenol-specific UGT1F (P08430) (Iyamagi et al., 1986), and bilirubin-specific UGT1A (P20720) (Sato et al., 1991). Most of these members are derived as a result of alternative splicing with a common C-terminal domain of 245 amino acid residues. The UGT2 group consists of 16 homologous members that include glucuronidate steroid hormones such as androgen, estradiol, and testosterone. Among the UGT2 group are rat 17β-hydroxysteroid-specific UGT2B6 (P19488) (Mackenzie, 1990), and UGT2B3 (P08542) (Mackenzie, 1987), 3-hydroxyandrogen-specific UGT2B2 (P08541), UGT2B1 (P09875) (Mackenzie, 1986), monoterpenoid alcohol-specific UGT2B12 (P36511) (Green et al., 1995), odorants-specific olfactorial UGT2A1 (P36510) (Lazard et al., 1990), mouse UGT2B5 (P17717) (Kimura and Owens, 1987), rabbit 4-hydroxyphenyl-specific UGT2B13 (P36512) and UGT2B14 (P36513), UGT2X (P36514) (Tukey et al., 1993), human estriol-specific UGT2B28 (P2376) (Coffman et al., 1990) and UGT2B11 (P36538), UGT2B10 (P36537) (Jin et al., 1993), UGT2B15 (P5855), hydroxycholesterol-specific UGT2B4 (P06133) (Jackson et al., 1987), 3,4-catechol-estrogen-specific UGT2B7 (P16662) (Ritter et al., 1990), None of the Tetrapoda UGT has a potential phosphorylation site in the NRD1β domain, except the UGT from fish species. The CGT group transfer galactose residue to ceramide. Currently rat (Schulte and Stoffel, 1993), mouse (Bosio et al., 1996b; Coetzee et al., 1996), human (Bosio et al., 1996a; Kapitonov and Yu, 1997a), chicken, and bovine (Kapitonov and Yu, 1997c) CGTs are cloned. All members of UGT and CGT families are compared in Figure 5A. Group 10 of *C. elegans* has only one known protein, which is probably a *C. elegans* analogue of UGT. Group 11 of the NRD1β proteins contains three proteins of 526–530 amino acid residues. These proteins are compared in Figure 5B.

Group 12 of the NRD1β proteins consists of two proteins, Z75554.5 with 535 amino acid residues and Z75554.6 with 534 amino acid residues. Splice sites other than those predicted with the Genefinder Program are used in alignment on a coding negative strand for (Z75554.6):

```
12378-ATC GTGGT...agACT CCA-12327;12195-AAA AA[gt...ag]A TAT CAA-12130
```
Conserved domains of glycosyltransferases

Fig. 5. The evolutionary relationship between UGT superfamily (A), members of Groups 11, 13, 15, 16, 17 of C.elegans glycosyltransferases (B), and between members of the DAGaT family (C).

nucleotides. Comparison of Group 15 members is shown in Figure 5B.

Group 16 of C.elegans NRD1β proteins consists of four members with 520 to 542 amino acid residues. All members are compared in Figure 5B. Group 17 consists of seven members with 520–536 amino acid residues, two of which are pseudogenes. Comparison of Group 17 members is shown in Figure 5B. Group 17 C.elegans (U39851.6) is a pseudogene with a point deletion resulting in a frame shift at the amino acid position of 261. In 17 C.elegans (U88311.6) the frame shift is caused by insertion of one nucleotide at the position corresponding to the amino acid residue 268. For Group 17 C.elegans (U97003.6), splicing sites different from those predicted by Genefinder are used. Group 19 consists of two proteins: 19 C.elegans (Z34802.1) with 531 amino acid residues and (AF016424) with 508 amino acid residues.

The NRD1β family of glycosyltransferases encompasses enzymes from eukaryotes, prokaryotes and archaea. Thus, the NRD1β domain probably appeared prior to the separation of these three kingdoms. The acceptor substrates of these glycosyltransferases are either lipids, carbohydrates, proteins, or polycyclic compounds such as steroids, antibiotics, and pigments. These aglycones show little structural similarity (Figure 3). The donor substrates, however, are similar in their nucleotide portion. Therefore, we propose that this domain plays a role in recognition of the nucleotide portion of the sugar–nucleotide substrates, when the nucleotide portion is either UDP or TDP, and the sugar portion is either glucose, galactose, glucuronic acid, xylose, rhamnose, or daunosamine. Occasionally, close members of one group can utilize a different sugar residue, as exemplified by the utilization of galactose by one of the ecdysteroid glucosyltransferases. Potential results of the activity of the NRD1β glycosyltransferases include inactivation of biologically active compounds, as seen with ecdysone, indole acetate, and antibiotics, or the synthesis of new biologically active compounds. All of the enzymes in the NRD1β family probably share a common reaction mechanism, a possibility that could be verified by x-ray crystallographic and/or enzyme kinetic studies.

All C.elegans NRD1β members in the database were detected with the help of the Genefinder Program. It should be noted that this program does make occasional mistakes in selecting the correct splice sites due probably to the lack of information about codon preferences and the third nucleotide bias in C.elegans. The program cannot discriminate between genes and pseudogenes. Occasionally, alternative splicing sites and alternative transcription initiation sites were used based on homology with close NRD1β members. Apart from the complete genes, we found four examples of exon shuffling in the C.elegans genome, where the NRD1β-encoding exon was relocated to other parts of the genome and had undergone several mutations. Gene duplications, deletions, frame shifts and exon shuffling have evolutionary significance in terms of the generation of new genes by recombining domains, point mutations, insertions, and deletions.

After we finished our initial analysis (Kapitonov and Yu, 1997a,b), a similar domain was described (Mackenzie et al.,
1997), where the nomenclature for a UDP-glucuronosyltransferase gene superfamily was proposed. Our classification differs from theirs in the following aspects. (1) The NRD1β family of glycosyltransferases is able to utilize not only a UDP-sugar as a donor molecule, but also a TDP-sugar. Therefore, rather than naming these enzymes as UDP glycosyltransferases, we identified them as NRD1β-type glycosyltransferases in order to indicate the presence of a nucleotide recognition region that can recognize either UDP or TDP. (2) We divide the NRD1β proteins into the following five major classes based on protein sequence homology (Figure 2): Class III of microbial NRD1β glycosyltransferases, Class IV of plant NRD1β glycosyltransferases, animal (including both vertebrate and invertebrate animals) NRD1β glycosyltransferases of typical (440–570 amino acid residues, Class V), medium (about 700 amino acid residues, class II), and large (>800 amino acid residues, Class I) sizes. According to our analysis, some of the genes used in the proposed nomenclature (Mackenzie et al., 1997) are pseudogenes or just the NRD1β exons moved to other parts of the genome; therefore, the assignment of a particular number to those sequences may lead to multiple reassignments later on. Also, for many gene products we used splice sites and initiation codons different from those generated by Genefinder; this resulted in a different proposed evolutionary relationship of the members analyzed.

All C. elegans NRD1β family members have a C-terminal transmembrane domain and C-terminal motif similar to the endoplasmic reticulum retention signal. They can potentially transfer carbohydrate residues onto lipids, carbohydrates, oligopeptides, and polycyclic compounds such as steroid hormones. They also can take part in the biosynthesis of extracellular matrices, hormones, or certain extracellular signal factors. These factors may play an important role in the early stages of multicellular organism development.

The NRD1α family of glycosyltransferases

CGT can transfer a galactose residue from UDP-galactose to diacylglycerol (DAG) and produce monogalactosyldiacylglycerol (Schraen-Wiemers et al., 1995). The same function is performed in plants by DAG galactosyltransferase (DAGGalT) (EC 2.4.1.46, 522 amino acid residues), which produces the major structural lipid of chloroplast (Shimojima et al., 1997). DAGGalT shares homology with the murG gene from E. coli and Bacillus subtilis, whose product N-acetylglucosaminyltransferase (GlcNAcT) catalyzes the transfer of N-acetylgalactosamine to N-acetylmuramyl-(pentapeptide) phosphoryl-undecaprenol in the last step of peptidoglycan synthesis (Mengin-Lecreulx et al., 1991). This homology implies that the chloroplast membrane biosynthesis mechanism is evolutionarily derived from that of the bacterial cell wall, supporting the endosymbiotic theory (Shimojima et al., 1997). Our computer analysis indicates the existence of two more putative glycosyltransferases from Bacillus subtilis (P54166) and Methanococcus jannaschii (Q58652). These sequences are compared in Figure 5C. The largest area of homology shown in Figure 6 closely resembles that of NRD1β.

DAGGalT does not have a terminal transmembrane domain; the first 103 amino acid residues are cleaved upon translocation to the chloroplast. It appears that the murG gene product of E. coli and B. subtilis also lacks an identifiable terminal transmembrane domain although it is membrane associated.

Separation of the NRD1β family members and the NRD1 domain of the DAGGalT family probably occurred prior to the separation of archaea bacteria and bacteria; however, we still attribute NRD of the DAGGalT family to NRD1β because of the significant similarity in the central portion of NRD1 domain. The central portion of the NRD1β domain is very close to the conserved domain of a large group of glycosyltransferases. Since all known members of this group are retaining glycosyltransferases, we termed all NRD1α family to reflect the resulting anomeric configuration. This family is characterized by the presence of two conserved E residues, located at positions 1 and 2 (Figure 6), along with conserved surroundings, spanning about 37 amino acid residues. DAGGalT family is in a way an intermediate family between NRD1β and NRD1α (closer to NRD1β). Members of this family do not have characteristics for the NRD1α E residue at position 1 (Figures 1, 6). DAGGalT and B. subtilis (P54166) have K, MurG proteins have R, and M. jannaschii (Q58652) has H at position 1, typical for NRD1β.

However, the surrounding regions are not as conserved as for NRD1β, where the largest area of homology spans about 60 amino acid residues. Some of the NRD1α family members are shown in Figure 6. Among them are the family of 17 sucrose synthases, 5 sucrose-phosphate synthases from plants, microbial mannosyltransferases, galactosyltransferases, glucosyltransferases, and galacturonate transferases, as well as several putative glycosyltransferases from C. elegans. All bacterial and some eukaryotic NRD1α family members apparently lack a transmembrane domain; however, two out of three enzymes from C. elegans (P53993 and Q22698) and one from S. cerevisiae (P53954) have an N-terminal transmembrane domain. None of the above described glycosyltransferases have definable topology.

Function of NRD1α family members

All of the bacterial glycosyltransferases participate in the formation of the exopolysaccharide (EPS), lipopolysaccharide (LPS), and capsular polysaccharide structures (CPS). Mannosyltransferase C (MftC) from E. coli is the first ManT that transfers mannose residue to Gly-protophosphoryl-undecaprenol, forming an α(1→3) bond in the O9 antigen biosynthesis pathway. ManT B (MftB) is the second mannosyltransferase that transfers two mannose residues to form α-Man(1→3)-α-Man(1→3)-α-Man(1→3)-Glc-PP-Undecaprenol intermediate. ManT A (MftA) transfers three mannose residues to form α-Man(1→2)-α-Man(1→2)-α-Man(1→3)-α-Man(1→3)-α-Man(1→3)-Glc-PP-Undecaprenol intermediate. MftB and MftA then finish the formation of O9 mannan-repeating unit →(2)-α-Man(1→2)-α-Man(1→2)-α-Man(1→2)-α-Man(1→3)-Glc-PP-Undecaprenol in the last step of peptidoglycan synthesis: (Mengin-Lecreulx et al., 1995). TrsD, E, and H glycosyltransferases participate in the biosynthesis of the O-antigen LPS outer core, serotype O3 (Skurnik et al., 1995). E. coli K12 GaIT is responsible for the formation of an α-Gal(1→6) bond in the outer core biosynthesis: β-GlcNAc(1→6)-α-GlcIII(1→2)-α-GlcII(1→3)-|α-Gal(1→6)|-α-GlcI(1→ inner core (Pradel et al., 1992). CapH of Staphylococcus aureus (Lin et al., 1994; Lee, 1995; Sau and Lee, 1996). Cap1E and Cap1G of Streptococcus pneumoniae (Garcia and Lopez, 1997) are glycosyltransferases involved in the biosynthesis of the capsular polysaccharide. Cap1E is a galacturonosyltransferase that catalyzes the formation of an α(1→3) bond. EpsF and EpsG from Streptococcus thermophilus are involved in the exopolysaccharide (EPS) biosynthesis (Stinge et al., 1996). The family of RfbB galactosyltransferases has three members: GaIT from Klebsiella pneumoniae serotypes O1 (Q48487; domain starts at aa 277) and O8, and RfbB GaIT from Serratia marcescens (Q54481; domain starts at amino acid residue 277). RfbB is capable of forming both α and β bonds in the LPS core biosynthesis: β-Gal(1→3)-α-Gal(1→3)-β-GlcNAc(1→lipid (Clarke et al., 1995; Kelly and
Fig. 6. Members of the NRD1 family. The type of glycosyltransferase (if known) is followed by the gene name, organism and accession number. ManT, Mannosyltransferase; GaT, galactosyltransferase; GlcT, glucosyltransferase; GalAT, galacturonate transferase; GiT, glycosyltransferase; GlcNAcT, N-acetylgalcosaminyl transferase; CPS, capsular polysaccharide biosynthesis protein; EPS, exopolysaccharide biosynthesis protein; LPS, lipopolysaccharide biosynthesis protein; E.c., Escherichia coli; S.c., Salmonella enterica; S.e., Acetobacter xylinum; Pr.m., Proteus mirabilis; S.p., Proteus mirabilis; C.e., Caenorhabditis elegans; 5, group of five sucrose-phosphate synthases (the one from rice is shown), 17, group of 17 sucrose synthases from both dicotyledonae, shown are sucrose synthase from bean and barley; 5, group of five sucrose-phosphate synthases. Shown is sucrose phosphate synthase from rice, TRSD. E. H – LPS biosynthesis proteins; S.c., Saccharomyces cerevisiae; C.e., Caenorhabditis elegans; N-GlyT ALG2, asparagine-linked oligosaccharides biosynthesis protein ALG2; B.s., Bacillus subtilis. Motif, amino acid residues present in more than 60% of the family members, NRD1β members express Escherichia coli, diacylglycerol (DAG) galactosyltransferase from cucumber, oleandomycin glucosyltransferase, indole-3-acetate glucosyltransferase and CGT Motif are shown for comparison. 1 and 2, amino acid residues that we propose as a part of enzyme reaction mechanism. Residues predicted to be a part of α-helix in the middle of the proposed catalytic domain are underlined. The three-digit number on the left margin of the alignment indicates position of the first amino acid of the conserved domain; the number on the right margin indicates the total length of the protein.

Whitfield, 1996). CpsF from Proteus vulgaris participates in the formation of capsular polysaccharide (Gygi et al., 1995; Rahman et al., 1997). RfbF from Campylobacter hyoilei is a GaT involved in lipo-oligosaccharide (LOS) biosynthesis (Korolik et al., 1997). RfbP is a GaT that participates in the formation of O:8-antigen repeating unit (Jiang et al., 1991). Rfu is ManT responsible for the synthesis of O:8-antigen repeating unit of B serogroup: →[α-Glc-(1→4)]-β-Gal-(1→) RfbW and RfbZ are both ManT responsible for the addition of the second and first mannose to the repeat unit of O-antigen of C2 serogroup correspondingly: →[2-O-Ac-α-Abe-(1→3)]-α-L-Rha-(1→2)-α-Man-(1→3)-α-Gal-(1→) (Brown et al., 1992; Liu et al., 1993; Reeves, 1993; Xiang et al., 1993, 1994). RfbO ManT which catalyzes the transfer of mannose in E1 serogroup O-antigen biosynthesis forming a β-(1→4) bond has no homology within the NRD1 family. The remaining enzymes of O-antigen repeat unit biosynthesis RfbW—αequoseotransferase and RfbN—rhamnolysotransferase belong to the NRD2 family (see below). AceC from Acetobacter xylinum is a ManT which adds a third carbohydrate residue in the exopolysaccharide acetan biosynthesis (Rha-Glu-Glu-Glu-Man-OAc-Glu-Glu-PP-Prenol) (Griffin et al., 1994). Not shown are two very closely related ManT from Acetobacter xylinum (Q64571; starts at amino acid residue 272) and GumH from Vibrio cholerae (Q66476). The evolutionary relationship among members of the NRD1β family is shown for comparison. The evolutionary relationship among the NRD1 family is shown in Figure 7. Note that all eukaryote glycosyltransferases are placed in one group with two subgroups of plant and animal glycosyltransferases according to homology.
Therefore, NRD1\(\alpha\) glycosyltransferases can be further divided into three groups: prokaryotic, eukaryotic plant, and animal glycosyltransferases (very much like NRD1\(\beta\) glycosyltransferases). Apart from the domain shown in Figure 6, some, but not all, retaining glycosyltransferases share an additional domain, where the RXXXXK motif is following a hydrophobic area of five to seven amino acid residues.

The NRD2 family of glycosyltransferases

Human ceramide glycosyltransferase (CGlcT) catalyzes the transfer of the glucosyl residue from UDP-glucose onto ceramide (Ichikawa et al., 1996). In the NRD1\(\beta\) family of enzymes, some members are able to use UDP-glucose as well as UDP-galactose as a donor substrate (e.g., ecdysteroid glycosyltransferase from Spodoptera frugiperda (O'Reilly et al., 1992)). Therefore, we attempted to determine whether CGlcT belongs to the NRD1 family. Computer analysis indicated no significant similarity between CGT and CGlcT. Unlike CGT, CGlcT has an N-terminal transmembrane domain. Further analysis indicates that CGlcT belongs to another family of glycosyltransferases that share a common domain, which we have named NRD2. The NRD2 domain is located either immediately adjacent to the N-terminal transmembrane domain, or at the N-terminus. This recognition domain is bipartite: a large NRD2L, part of about 50 amino acid residues, located near the N-terminus (amino acid residues 51–93 in the human CGlcT), and a small NRD2S part of 11–17 amino acids (amino acids 136–153 in the human CGlcT). In different NRD2 family members, NRD2L and NRD2S are separated by 27–41 amino acid residues. Human CGlcT has three close homologs, one from Synechocystis and two from C.elegans. Based on the conserved sequence, we have generated a motif and searched the database in order to find additional members of the NRD2 family. We found very similar domains in the NodC protein, cellulose synthase, rhamnosyltransferase, hyaluronate synthase, and some other proteins (Figure 8). All NRD2 family members have a higher degree of similarity within NRD2L than within NRD2S. The evolutionary relationships among the NRD2 family members are shown in Figure 9. Apart from having a common NRD2 domain, all NRD2 family members share other common structures in that they all are membrane-bound proteins and all potentially make contact with the membrane several times as judged by hydropathy plots and turn probabilities. The topology was characterized for CGlcT (Jeckel et al., 1992), DPMarT, DPGlcT (Bossuyt and Blanckaert, 1993), and NodC (Barny et al., 1996) enzymes. All of these enzymes have catalytic domains (possibly an NRD2L domain) oriented toward the cytosolic site. The products of these enzymes, such as GalCer, are transported across the membrane either into the Golgi (Lannert et al., 1994) into the ER such as Dol-P-Glc and Dol-P-Man (Schutzbach and Zimmerman, 1992), or outside the cell such as the Nod factors. The N-terminal transmembrane domain may cross the membrane into the extracytoplasmic space. Most of the bacterial members of the NRD2 family lack an N-terminal transmembrane domain, but still have several potential transmembrane domains throughout the enzyme. The existence of homologous enzymes sharing the same topology may provide an evolutionary link between bacterial cell surface membranes and the eukaryotic ER and Golgi apparatus.

Function of NRD2 family members

Based on sequence similarities, the NRD2 family of glycosyltransferases can be divided into the following groups. Group 1 consists of GDP-mannose: dolichol-phosphate O-β-D-mannosyl-
transfcerase (EC 2.4.1.83). This is an essential membrane-bound enzyme responsible for the synthesis of dolichol-phosphate-mannose (DPM), a key glycosyl donor for the synthesis of the N-linked oligosaccharide chains of glycoproteins (Orlean et al., 1988; Orlean, 1990; Schutzbach et al., 1993). Other enzymes of the NRD2 family that belong to this group are DPM from yeast, DPM1 from Trypanosoma brucei (Q26732) and Ustilago maydis (P54856), hypothetical proteins from Synechocystis sp (P74505) and Mycobacterium tuberculosis (P71781), and two proteins from E.coli (P77293 and P77757). Dolichol-phosphate glucosyltransferase from Saccharomyces cerevisiae transfers glucose from UDP-glucose to dolichol phosphate (Heesen et al., 1994).

An absence of this enzyme leads to underglycosylation of the secreted proteins.

Group 2 of the NRD2 glycosyltransferase family consists of NODC proteins. These proteins are produced by rhizobial microorganisms that fix nitrogen inside specialized nodule structures in the root. NODC proteins possess N-acetylglucosaminyl transferase activities (chitin synthases) that are involved in the formation of lipo-chito-oligosaccharide Nod factors that initiate root nodule morphogenesis in legumes. NODC proteins are responsible for the synthesis of trimers to pentamers of \( \beta \)-1,4-linked GlcNAc residues called Nod factors (Figure 10A) (Geremia et al., 1994), which are also acylated (Debelle et al., 1994). Further modifications by acetylation or sulfation are achieved at R and R'' which determine host specificity (Schultze et al., 1992). Nod factors can be further fucosylated, and one microorganism can produce several Nod factors to broaden the host range (Lerouge et al., 1990; Price et al., 1992). At concentrations of \( 10^{-7} \) to \( 10^{-6} \) M, secreted Nod factors can induce root hair curling and the formation of nodules (Relic et al., 1994; Dazzo et al., 1996). So far, nine NODC proteins from different rhizobial strains of the genuses Rhizobium (P04341, P50357, P50356, P17862, P04340, P24151, and P04678) Azorhizobium (Q07755), and Bradirhizobium (P53417 and P04677) have been described (Ueda et al., 1995). All these proteins maintain a conserved NRD2 domain.

Another example of cell to cell interaction in prokaryotes can be found in the development of the myxobacterium Stigmatella aurantiaca, which undergoes a multicellular cycle of development resulting in the formation of fruiting bodies. The protein FBFA (Q53680), which participates in the fruiting body formation, is a glycosyltransferase that shares a higher degree of homology with the NODC protein than it does with any other protein of the NRD2 family. Inactivation of FBFA leads to the formation of nonstructured clumps rather than the structured fruiting body (Silakowski et al., 1996). Together with two other putative enzymes from Bacillus subtilis (P96587) and Synechocystis sp (P74165), FBFA forms a third group of NRD2 glycosyltransferases.

Several groups of the NRD2 enzymes participate in cell wall formation. Hyaluronan synthase, or HASA, from group A streptococci (e.g., Streptococcus pyogenes), produces the anti-phagocytic EPS hyaluronate capsule by alternating the addition of UDP-N-acetyl-d-glucosamine and UDP-d-glucuronic acid. The enzyme apparently has both glycosyltransferase activities (Dougherty and van de Rijn, 1992; DeAngelis et al., 1993).
Streptococcal hyaluronate is chemically indistinguishable from that found in animal connective tissues (Figure 10B).

EXO is a group of glycosyltransferases responsible for EPS type I succinoglycan biosynthesis in *Rhizobium meliloti*. EPS plays an important role during interaction with its host alfalfa. EXO glycosyltransferases participate in the biosynthesis of the repeating unit of succinoglycan (Figure 10C). EXOM adds the 4th, EXOO the 5th, EXOU the 6th, and EXOW the 7th sugar residue in the succinoglycan chain (Glucksmann et al., 1993; Reuber and Walker, 1993). EXOA, which adds the 2nd sugar, has a high similarity with the NRD2 motif in the NRD2L portion, but a lower similarity in the NRD2S portion. The 8th sugar is modified by addition of pyruvate, the 7th sugar is linked to succinate, and the 3rd is acetylated. The whole structure is anchored to the membrane by a lipid carrier (Reinhold et al., 1994).

The ICAA protein from *Staphylococcus epidermidis* participates in the synthesis of the polysaccharide intercellular adhesin (PIA), which is located on the cell surface and is responsible for the formation of the biofilm and the large cell clusters. PIA is a linear β-1,6-linked glycosaminoglycan composed of at least 130 2-deoxy-2-amino-D-glucopyranosyl residues, most of which are N-acetylated. ICAA has been shown to be a membrane-bound N-acetylglicosaminyltransferase that forms β-1,6-linked N-acetyl-d-glucosaminyl polymers (Heilmann et al., 1996).

Group 4 of the NRD2 glycosyltransferase family is the group of 13 enzymes responsible for cell surface polysaccharide (CSP) biosynthesis. CSP can be attached to the cell membrane as the O-antigen of lipopolysaccharides (LPS) to form a capsule around the cell as capsular polysaccharides (CPS). It can also be completely excreted as exopolysaccharides (EPS). The biological functions of the polysaccharides are diverse, including roles in pathogenesis and symbiosis, protection of the cell from desiccation or other environmental stresses, and/or facilitating adherence of bacteria to cell surfaces.

The EPSI glycosyltransferase from *Streptococcus thermophilus* Sf6 (Q56046) participates in the synthesis of excreted EPS having the repeating unit \(\rightarrow 3\)-β-D-Galp-(1→3)-[α-D-Galp-(1→6)]-β-D-Glcp-(1→3)-α-D-GalpNAc-(1→) which is responsible for the extremely slimy texture of some yogurt products (Stingele et al., 1996). The EPSG glycosyltransferase from *Lactococcus lactis* (O06035) participates in the synthesis of similar EPS (van Kranenburg et al., 1997).

CPSJ and I proteins of *Streptococcus pneumoniae* (O07340 and O07339) are glycosyltransferases that transfer the 4th sugar

---

**Fig. 10. Chemical structure of the acceptors of the NRD2 glycosyltransferases.**

---

**A**

Nod factor

**B**

Hyaluronate

**C**

**Repellent unit of Succinoglycan**

**D**

Repetitive unit of cellulose
(galactose) and 3rd sugar (N-acetylglucosamine), respectively, to the glycolipid intermediate (L):

\[-\beta-D-\text{GlcN}Ac(1\rightarrow3)\beta-D-\text{Galp}(1\rightarrow4)\beta-D-\text{Glp}-[\beta-D-\text{Galp}(1\rightarrow4)]+L\]

during the synthesis of capsular polysaccharides.

Minor teichoic acid synthesis proteins GGAA and GGAB from Bacillus subtilis (P46917 and P46918) are glycosyltransferases that are involved in the synthesis of poly(3-O-β-6-glucopyranosyl N-acetylglucosamine-1-phosphate), a secondary anionic polymer of gram-positive bacteria (Mauel et al., 1994; Lazarevic et al., 1995). In E.coli KFC protein (Q47330) is responsible for the synthesis of the capsular polysaccharide, K5 antigen, a polymer of 4)-β-GlcA-(1,4)-α-GlcNAc-1. KFC can use both UDP-glucuronic acid (UDP-GlcA) and UDP-N-acetylglucosamine (UDP-GlcNAc) as substrates (Petit et al., 1995).

Other proteins that belong to Group 4 of the NRD2 glycosyltransferases are Haemophilus influenzae (Q57022 and Q56869) (Fleischmann et al., 1995), Yersinia enterocolitica (Q56869, Q57022), and TRSB protein (Q56914). The latter is responsible for the biosynthesis of LPS O:3 antigen (homopolymer of 6-deoxy-l-altrose repeating units linked by 1,2 linkages) (Skurnik et al., 1995); Bacillus subtilis (P71059 and P71057) (Glaser et al., 1993). E.coli (P11290), Synecocystis sp. spore coat polysaccharide biosynthesis protein SPSA (P73983), and Pseudomonas aeruginosa mucosa-induced protein MIGA (P95448).

Group 5 of the NRD2 glycosyltransferases consists of enzymes involved in lipopolysaccharide (LPS) synthesis. The LGTA glycosyltransferase from Neisseria meningitidis (Q51151) is responsible for the synthesis of meningoococal LPS oligosaccharide of L3 immunotype. It transfers an N-acetylglucosamine residue to the terminal galactose (4th sugar) of the inner core lacto-N-neotetraose (the last four sugars): β-D-Gal(1→4)β-D-GlcNAc(1→3)β-D-Gal(1→4)β-D-Glc(1→4)α-Hept(1→5)α-KDO (Jennings et al., 1995). Its counterpart in Neisseria gonorrhoeae (Q50946) is encoded by the isi-2 gene (Danaher et al., 1995). Other members of the group 5 glycosyltransferases include putative proteins from Haemophilus influenzae (Q57287), Bacillus subtilis (P71054) and LGTD protein from Neisseria gonorrhoeae (Q50949) and meningitidis (P96649) that transfer GlcNAC to the terminal Gal of the lacto-N-neotetraose portion of the lipo-oligosaccharide (LOS) inner core (Gotschlich, 1994). Q50951 from Neisseria gonorrhoeae is a putative protein. In the fish pathogen Vibrio anguillarum VIRB protein takes part in the synthesis of the LPS, which is a major surface flagella sheath antigen (Norqvist and Wolf-Watz, 1993).

Group 6 of the NRD2 glycosyltransferases includes RFBV, an abequose transferase from Salmonella enterica, serogroup C2 (P26401), which takes part in the formation of the O-antigen repeat unit (Reeves, 1993):

\[\text{Abe}\]_{2,3}^\alpha \rightarrow [\text{L-Rha}(1\rightarrow2)\alpha-\text{Man}(1\rightarrow2)\alpha-\text{Man}(1\rightarrow3)\alpha-\text{Gal}]^\rightarrow 1\]

RFBV is a rhamnosyltransferase of Salmonella enterica serogroup B (Q54129) that participates in the formation of O-antigen: (1→2)-α-Man-[(1→3)-α-2OAceβH-(1→2)-α-L-Rha-(1→3)-α-Gal-(1→2)-α-Glc](1→α). The same function is performed in the Salmonella enterica serogroup E1 by the orf11.9 product (Liu et al., 1993). Other enzymes in this group include the product orf14.1 of Salmonella enterica (Q99192), which probably performs the same role as that encoded by orf11.9 of serogroup, two proteins from Mycoplasma genitalium (P47306 and P47271) (Fraser et al., 1995), one protein from Yersinia enterocolitica (Q56866), and two proteins from Synecocystis sp. (P72899 and P73996).

Rhamnosyltransferase RFBQ (Q03581) and RFBR (Q03582) from Shigella dysenteriae and RFBG from Shigella flexneri (P37782) belong to Group 7 of the NRD2 family. RFBQ transfers the 2nd sugar residue (Rha 1) and RFBR transfers the 3rd sugar residue (Rha II) to the O-antigen repeating unit of the cell surface LPS: (→2)-α-L-Rha II-(1→3)-α-L-RhaI-(1→2)-α-D-Gal-(1→3)-α-D-GlcNAc(1→)Acyl Carrier Lipid (ACL). RFBF transfers a Rha III residue in the Y serotype biosynthesis reaction to form O-antigen: (→2)-α-L-RhaIII-(1→2)-α-L-RhaII-(1→3)-α-L-RhaI-(1→3)-β-D-GlcNAc(1→)ACL (Morona et al., 1995). Other glycosyltransferases in this group are the following: RFBM from Yersinia enterocolitica (S53296) (Zheng et al., 1993) that participates in 6-deoxy-l-altrose (C3 epimer of rhamnose) polymerization, RFBF rhamnosyltransferase from Leptospira interrogans (P71441) (Mitchison et al., 1997), rhamnosyltransferase from Streptococcus pneumoniae (Q07866), and RFBG glycosyltransferase from Klebsiella pneumoniae (Q48482) (Kelly and Whitfield, 1996). EXP2 from Synorhizobium melliloti participates in the formation of EPS II, which consists of alternating glucose and galactose residues joined by α-1,3 and β-1,2 linkages.

Cellulose synthases from Acetobacter xylinum (P19449 and P21877) (Wong et al., 1990; Standal et al., 1994) and Paramecium bursaria Chlorella virus 1 (U42580) (Kutish et al., 1996) belong to the 8th group of NRD2 glycosyltransferases. These enzymes are responsible for the synthesis and crystallization of cellulose (Figure 10D).

The last and the most distant group of the NRD2 glycosyltransferases is formed by ceramide glucosyltransferases from humans, two putative proteins from C.elegans, and one putative protein from Synecocystis. The existence of two proteins in C.elegans may indicate that these enzymes utilize different substrates, such as hydroxy ceramide and nonhydroxy ceramide.

Similar to members of the NRD1 family, members of the NRD2 family of glycosyltransferases are found in eukaryotes, prokaryotes, bacteria, and archaeabacteria, indicating that this domain appeared prior to separation of these three kingdoms. All NRD2 members are membrane-bound enzymes that can potentially span the membrane several times. It is likely that the evolution of this family of enzymes was closely connected with the evolution of the membrane structure. In addition, we have identified a conserved region of these proteins that is located near the N-terminal or immediately following the N-terminal transmembrane domain (if such a domain is present).

Results of our computer analysis allow us to predict that it is possible to identify new members of the membrane-bound NRD2 family of glycosyltransferases through the following strategies: (1) using degenerate primers based on the conserved amino acid sequence of the NRD domains, (2) hybridization with the NRD portion of known genes, or (3) screening expression libraries (e.g., the AgtI1 library with polyclonal antibodies against the NRD2 domain). We can also speculate that glycosyltransferases needed for eukaryotic glycolipid and bacterial LPS biosynthesis evolved from a common predecessor. Therefore, based on the sequence homology with glycosyltransferases involved in the later steps of LPS biosynthesis, one can expect to discover new glycosyltransferase members of the NRD1 or NRD2 family that are membrane-bound and catalyze the addition of sugar residues.

Conserved domains of glycosyltransferases
to glycoconjugates. Since glycoconjugates play a crucial role in embryonic development, future research should be directed toward identifying those glycoconjugates by isolating NRD1 and NRD2 glycosyltransferases from higher eukaryotes and, in particular, mammals, and analyzing their function.

**Catalytic domain and proposed reaction mechanisms**

When we analyzed the NRD1α family of glycosyltransferases, we determined that all the members of that family transfer carbohydrate residues with the retention of configuration. Comparison of the NRD1β and NRD1α glycosyltransferase family members shows that one of the distinctive features of the two families is the presence of two conserved glutamic acid residues (marked with numbers 1 and 2 in Figure 6) with E at position 2 in the middle of a short and mostly uncharged α-helical domain (about 3 turns). We propose that these E residues may play a functional role as a base in a nucleophilic substitution reaction that proceeds through a double displacement mechanism with an oxocarbenium ion transition state (Davies et al., 1998a,b) (Figure 11A). These E residues would be negatively charged under physiological conditions. E at position 2 is located above the plane of the sugar and serves as a nucleophile, while E at the position 1 is located beneath the plane and serves as a proton donor at the first step of the reaction. The pK difference between D and E side chains (3.96 and 4.32, respectively) may account for the exclusive selection of E as a nucleophile since it would be expected to create a more stable transition bond than D. The hydrophobic environment of a catalytic domain may further increase the pK (up to 8.2 for E; Fersht, 1977). In the case of NRD1α glycosyltransferases, the donor substrate is always a sugar-nucleotide (either purine or pyrimidine), whereas the acceptor can be a growing carbohydrate chain, lipid, or protein.

For simplicity, only a carbohydrate is shown as an acceptor in Figure 11. The family of retaining β-glycoside hydrolases also utilizes two E residues in the catalytic domain in the reaction proceeding in the opposite direction through acid/base catalyzed formation and subsequent hydrolysis of a covalent glycosyl-enzyme intermediate (Davies et al., 1998a).

All known NRD1β glycosyltransferases transfer sugar residue with the inversion of configuration that would require one nucleophile. We noticed that E residue at position 2 and the α-helical domain are conserved in the NRD1β (Figure 1) and NRD1α (Figure 6) families of glycosyltransferases. We propose that this glutamic acid residue can play a role as a single nucleophile in the glycosyltransferase reaction with inversion of configuration. In a single nucleophile mechanism, S_{02}↑ nucleophilic substitution will result in an inversion of the anomeric configuration (Figure 11B). At position 1, eight amino acid residues away from the conserved E residue at position 2, is an H residue in the NRD1β family or R/K in DAG GaT. These residues are located at exactly the same distance from the conserved E, and therefore could be located underneath the sugar plane such as E at position 1 in the NRD1α family. H and R/K may play a role in the donation of a proton (Figure 11B). Another motif that is conserved among all NRD1 family members is the PQ and/or DQ motif (Figure 1). Since all NRD1 glycosyltransferases can accept only pyrimidine-sugar, Q residue may play a role in the recognition of pyrimidine residue (like in EQQN motif; Boegeman and Qasba, 1998).

Since NRD1α glycosyltransferases can accept both purine- and pyrimidine-containing sugar nucleotides, they are lacking conserved nucleotide recognition motifs. If our model is correct, substitution of E residue at position 2 for D in any NRD1β or NRD1α family member should cause a dramatic reduction in the reaction rate, while substitution for any other amino acid should completely eliminate the enzyme activity. Substitution of H residue for R/K at position 1 in NRD1 family members could possibly be tolerated with some changes in the reaction rate. Substitution of H/R/K residues at the position 1 for E may convert inverting glycosyltransferase into retaining and convert resulting anomeric configuration from β to α. Glycosyltransferases with S/T residues located two amino acid residues before residue E at position 2 could be potentially regulated by phosphorylation since placing a strong negative charge in the catalytic domain between two amino acid residues involved in the reaction should affect enzyme activity. Some microbial NRD1 glycosyltransferases apparently do not have a conserved E residue at position 2 (Figure 1). Since the homology in the surrounding area is high enough to attribute these enzymes to the NRD1β family, we believe that the catalytic domain of these enzymes evolved from the NRD1β catalytic domain to accommodate large molecules such as antibiotics that some of these enzymes inactivate or produce (such as vancomycin (Figure 3E) or baumycin). Therefore these glycosyltransferases utilize alternative nucleophile spaced further away.

The NRD2 family includes inverting glycosyltransferases that can accept both purine- and pyrimidine-containing sugar nucleotides. At this point we are unable to propose any particular amino acid residue as a candidate for the nucleophile or for nucleotide recognition, but the YN...E motif, E at the position 1 and two of the DD motifs (at positions 2 and 3) are good candidates for the...
beginning of site-directed mutagenesis experiments (Figure 8). Some of the conserved D residues at the position 3 fall into the conserved motif DXD proposed for metal binding (Boegeman and Qasba, 1998).

Classification of glycosyltransferases and concluding remarks

Based on our analysis, the most logical way to classify the glycosyltransferases that we have analyzed is according to the type of bond they form and the reaction mechanisms they carry out their catalytic function. Thus, all glycosyltransferases analyzed fall into two types: retaining glycosyltransferases that transfer sugar with the retention of configuration of anomer carbon, and inverting glycosyltransferases that transfer sugar with the inversion of configuration of anomer carbon. The retaining glycosyltransferases that we have analyzed are characterized by the presence of two E residues in the proposed catalytic domain and are further divided into the class of prokaryotic glycosyltransferases and eukaryotic glycosyltransferases within two subclasses of animal and plant glycosyltransferases. The inverting glycosyltransferases that we have analyzed are divided into 2 subclasses: H/R(K)/E subclass with H or R/K residue at position 1 and E residue at position 2 and NRD2 subclass. Each of these subtypes is further divided into the class of prokaryotic and eukaryotic glycosyltransferases. The latter is further divided into two subclasses of plant and animal enzymes. The division of NRD1β glycosyltransferases into five classes described in Figure 2 is based on the comparison of the full length proteins. Comparison of the catalytic or entire NRD1β domain will bring all animal NRD1β glycosyltransferases together by eliminating the variation in amino acid chain lengths. The second subtype of inverting glycosyltransferases (with the NRD2 domain) is also divided into prokaryotic and eukaryotic glycosyltransferases. According to recent classification (Campbell et al., 1997) most of NRD1β (H/R(K)-E) domain-containing glycosyltransferases fall into family 1, NRD1α (E-E) domain-containing glycosyltransferases fall into family 4, NRD2 glycosyltransferases fall into family 2 and 21 of nucleotide-diphospho-sugar glycosyltransferases (Campbell et al., 1997). There are numerous glycosyltransferases without E-E (NRD1α, H/R(K)-E (NRD1β), or NRD2 domains. Therefore, additional subtypes will be introduced upon accumulation of the data.

In conclusion, we have described glycosyltransferases that belong to three different families according to the similarity and proposed catalytic domain based on the most conserved amino acid residues. Using computer alignment of known glycosyltransferases, we generated patterns that we used to search the database for additional putative glycosyltransferases. Based on the amino acid residues present in the catalytic domain and the type of bond these glycosyltransferases catalyze, we have proposed reaction mechanisms and classified glycosyltransferases. Further studies by site-directed mutagenesis and/or x-ray structure analysis are required for the confirmation of our model. The crystallographic information can be used for the design of enzyme inhibitors. Because some of the NRD1 and NRD2 enzymes are indispensable proteins in bacteria, these inhibitors could lead to a new class of antibiotics. Since microbial glycosyltransferases are much closer to each other than to eukaryotic glycosyltransferases (Figures 2, 7, 9), it is possible to design inhibitors specific only for microorganisms. Additionally, suicide inhibitors can be designed that can be used for modification of NRD1 or NRD2 enzymes in order to purify and characterize new glycosyltransferases. Another way to search for new glycosyltransferases is to use PCR with degenerate primers synthesized based on the amino acid sequence of the conserved domains. The complete listing of all analyzed glycosyltransferases with conserved domains is available on the web site (www.views.vcu.edu/~glyco).

Acknowledgments

This work was supported by a USPHS Grant NS11853–25. We thank Drs. Cara-Lynne Schengrund and H. Tonie Wright for their helpful discussion of the manuscript.

Abbreviations

NRD, nucleotide recognition domain; CGT, ceramide galactosyltransferase; UGT, glucuronosyltransferase; DAG, diacylglycerol; DAGGAT, diacylglycerol galactosyltransferase; GlcNAcT, N-acetylgalactosaminyltransferase; GalNAcT, N-acetylgalactosaminyltransferase; EPS, exopolysaccharide; LPS, lipopolysaccharide; CPS, capsular polysaccharide structures; ManT, mannosyltransferase; LOS, lipooligosaccharide; GaT, galactosyltransferase; GlcAT, glucuronosyltransferase; CGl, ceramide glucosyltransferase; DPM, dolichol phosphate mannosate; DPMnT, dolichol-phosphate mannosyltransferase; ER, endoplasmic reticulum.

References


bium species strain NGR234 secretes a family of carbamoylated and fucosylated, nodulation signals that are O-acetylated or sulphated. Mol. Microbiol., 6, 3575–3584.


Relic,B., Perez,X., Estrada-Garcia,M.T., Kopcinska,J., Golinowski,W., Krish-

narachch,E.B., Pueppke,S.G. and Broughton,J. (1994) Nod factors of Rhizo-

bium are a key to the legume door. Mol. Microbiol., 13, 171–178.

Reuber,T.L. and Walker,G.C. (1993) Biosynthesis of succinoglycan, a symbioti-

cally important exopolysaccharide of Rhizobium meliloti. Cell, 74, 269–280.


Yahi,N., Sabater,J.M., Baghdiguian,S., Gonzalez-Scarano,F. and Fantini,J. (1995) Synthetic multimeric peptides derived from the principal neutraliz-

dation domain (V3 loop) of human immunodeficiency virus type 1 (HIV-1) gp120 bind to galactosylceramide and block HIV-1 infection in a human CD4-negative mucosal epithelial cell line. J. Virol., 69, 320–325.
