doi:10.1006/jmbi.2001.5151 available online at http://www.idealibrary.com on IDEAL® *J. Mol. Biol.* (2001) **314**, 365–374

JMB

COMMUNICATION

# Homology Between O-linked GlcNAc Transferases and Proteins of the Glycogen Phosphorylase Superfamily

**James O. Wrabl[1] and Nick V. Grishin[1,2]***

[1]*Howard Hughes Medical Institute, and*

[2]*Department of Biochemistry University of Texas Southwestern Medical Center 5323 Harry Hines Blvd, Dallas TX 75390-9050, USA*

**Corresponding author*

The O-linked GlcNAc transferases (OGTs) are a recently characterized group of largely eukaryotic enzymes that add a single β-*N*-acetylglucosamine moiety to specific serine or threonine hydroxyls. In humans, this process may be part of a sugar regulation mechanism or cellular signaling pathway that is involved in many important diseases, such as diabetes, cancer, and neurodegeneration. However, no structural information about the human OGT exists, except for the identification of tetratricopeptide repeats (TPR) at the N terminus. The locations of substrate binding sites are unknown and the structural basis for this enzyme's function is not clear. Here, remote homology is reported between the OGTs and a large group of diverse sugar processing enzymes, including proteins with known structure such as glycogen phosphorylase, UDP-GlcNAc 2-epimerase, and the glycosyl transferase MurG. This relationship, in conjunction with amino acid similarity spanning the entire length of the sequence, implies that the fold of the human OGT consists of two Rossmann-like domains C-terminal to the TPR region. A conserved motif in the second Rossmann domain points to the UDP-GlcNAc donor binding site. This conclusion is supported by a combination of statistically significant PSI-BLAST hits, consensus secondary structure predictions, and a fold recognition hit to MurG. Additionally, iterative PSI-BLAST database searches reveal that proteins homologous to the OGTs form a large and diverse superfamily that is termed GPGTF (glycogen phosphorylase/glycosyl transferase). Up to one-third of the 51 functional families in the CAZY database, a glycosyl transferase classification scheme based on catalytic residue and sequence homology considerations, can be unified through this common predicted fold. GPGTF homologs constitute a substantial fraction of known proteins: 0.4% of all non-redundant sequences and about 1% of proteins in the *Escherichia coli* genome are found to belong to the GPGTF superfamily.



*Keywords:* glycogen phosphorylase; glycosyl transferase; O-linked GlcNAc transferase; homology detection; fold prediction

The O-linked GlcNAc transferases (OGTs) catalyze the addition of O-linked *N*-acetylglucosamine monomers (O-GlcNAc) to Ser and Thr side-chains of cytosolic and nuclear proteins.[1] The precise function of the O-GlcNAc modification is not known, but it has been observed that the OGT gene is essential for completion of embryogenesis in mice.[2] O-GlcNAc modification is apparently widespread among many different protein types, and targets of this modification are sometimes reciprocally phosphorylated,[3,4] i.e. a single residue on a particular protein has been observed to contain either a phosphoryl or a GlcNAc group.[5–7] These data, in combination with other studies, imply that OGTs are mediators of signal transduction or cellular regulatory pathways.[1,3,8]

The importance of the human OGT is underscored by the suggestion that deficiencies in O-GlcNAc regulation play a role in diabetes, cancer, and neurodegeneration.[1] Despite the growing

Abbreviations used: OGT, O-linked GlcNAc transferase; TPR, tetratricopeptide; GPGTF, glycogen phosphorylase/glycosyl transferase; CAZY, carbohydrate active enzymes; GP, glycogen phosphorylase.

E-mail address of the corresponding author: grishin@chop.swmed.edu

0022-2836/01/030365–10 $35.00/0


Iβ1
Iβ2
Iβ3
Iβ4
Iβ5
Iβ6
Iβ7
IIβ1
IIβ2
IIβ3
IIα3
IIβ4
IIα4
IIβ5
IIα5
IIβ6
Predicted
Reliability
PDB Consensus
Group 1
Group 2
Group 3
Group 4
GPGTF Motif

need for structural and functional information concerning the OGTs, only a few members of the family have been cloned, expressed, and experimentally characterized. No structure has yet been determined for a member of the OGT family, and the locations and identities of the active site residues are unknown. Part of the difficulty in identification of functional sites is due to the inability of automatic database searches to uncover significant homology to sequences outside the immediate family.[1]

It has become possible to extend automatic database search methods using manual analysis of sequence data.[9–11] In favorable cases, correct structural and functional information about an enzyme can be predicted by detection of remote homology that is not discovered using completely automated approaches. Here, we describe the use of such techniques to infer remote homology between the OGT family and a large group of glycosyl transferases and other sugar-processing enzymes. Among the homologous sequences are several proteins that are known to share structural similarity. These data suggest that the C-terminal part of the human OGT molecule adopts a fold that consists of two Rossmann-like domains. The presence of a conserved acidic residue in helix 4 of the second domain identifies this region as a UDP-GlcNAc binding site.

A current classification scheme groups the glycosyl transferases into 51 families according to experimental determination of mechanism and conservation of active site residues (http://afmb.cnrs-mrs.fr/ ~ pedro/CAZY/gtf_table.html).[12] The OGTs, for example, are placed in family 41 of that database. However, a limited but growing amount of crystallographic data suggest that the separation into functional families masks more fundamental structural, and possibly evolutionary, relationships.[13,14] The present work supports this idea by unifying approximately one-third (15) of these functional families to a single superfamily on the basis of sequence homology and a common predicted fold. This superfamily of glycosyl transferases is typified by the glycogen phosphorylase structure, the first structure available for a superfamily representative.[15] Thus, the superfamily is termed GPGTF (glycogen phosphorylase/glycosyl transferase).

## Database searches and multiple sequence alignment

A PSI-BLAST[16] search initiated with the C-terminal portion of the human OGT, (gi|4505499, residues 370-920), was iterated to convergence against

**Figure 1.** Multiple sequence alignment of O-linked GlcNAc transferases with significant iterative PSI-BLAST search hits. Sequences from the NCBI GenBank database (http://www.ncbi.nlm.nih.gov) are labeled with gi identifier on the extreme left side of the Figure. Group 1 sequences are the O-linked GlcNAc transferases; group 2 sequences are homologous *E. coli* and *S. cerevisiae* sequences found during iterative PSI-BLAST searches (including the phage T4 β glucosyl transferase); group 3 are glycogen phosphorylase and maltodextrin phosphorylase, and group 4 are ribosomal S2 protein homologs. For clarity, the most conserved regions of secondary structure and putative functional significance are shown, and the number of residues separating each region is shown in parentheses. The total number of residues in each sequence is given in the second to last column, and the CAZY glycosyl transferase family number is given in the last column on the right-hand side of the Figure. A boldface family number indicates that the exact sequence existed in the CAZY database of 3 May 2001 (http://afmb.cnrs-mrs.fr/ ~ pedro/CAZY/gtf_table.html). A family number in regular font indicates that this sequence was provisionally assigned to a family through a significant PSI-BLAST hit to a true family member, as described in footnote[a] of Table 1. NC indicates that the sequence was not a member of any glycosyl transferase family. Conservation is indicated by color pattern: hydrophobic/non-charged side-chains are highlighted in yellow, small side-chains are highlighted in light gray, and charged/polar residues are highlighted in light blue. Highly conserved residues are shown as white lettering on black background. Possible functional residues, discussed in the text, at positions 1 and 2 are marked. Two conserved Asp residues at the crossover position in each domain are shown as black letters on dark gray background. The predicted consensus secondary structure of the 12 O-linked GlcNAc transferases is given above the sequences (β-strands as arrows and α-helices as cylinders), with the Jpred[19] (http://jura.ebi.ac.uk:8888) reliability estimate given just below. Gray cylinders outlined with broken lines indicate approximate positions of predicted α-helices that were observed but not explicitly shown in the Figure due to space considerations and complexity of alignment (Supplementary Material and ftp://www.iole.swmed.edu/usr1/ftproot/pub/GPGTF/). Sequences corresponding to PDB structures are indicated by an underlined gi identifier. The PDB consensus secondary structure given on the last line of the Figure resulted from four out of the six PDB structures having identical secondary structure at the indicated sequence position, and the secondary structure labels are given for each element at the top of the Figure. Colors for secondary structure elements correspond to colors in Figure 2. *E. coli* sequences are indicated by blue gi identifiers, and *S. cerevisiae* sequences are indicated by red gi identifiers; these sequences are ordered so as to approximately group sequences of greater similarity together. Italicized residues indicate alternative translation start points, which may represent translation start errors in the NCBI GenBank database. The source organism for each sequence is given by a two letter abbreviation: Hv (*Hordeum vulgare*); Ph (*Petunia x hybrida*); At (*Arabidopsis thaliana*); Ce (*Caenorhabditis elegans*); Dm (*Drospohila melanogaster*); Ss (*Synechocystis* sp.); Hs (*Homo sapiens*); Rn (*Rattus norvegicus*); Nc (*Neurospora crassa*); Rc (*Rhodobacter capsulatus*); Hi (*Haemophilus influenzae*); Sc (*Saccharomyces cerevisiae*); Ec (*Escherichia coli*); Tt (*Thermus thermophilus*); T4 (T4 bacteriophage).

the NCBI non-redundant database (3/8/01). The C-terminal portion of the sequence was chosen for analysis because the N-terminal portion was known to be a tetratricopeptide repeat (TPR) domain.[8,17] The *E*-value cutoff for this search was 0.1. Despite the permissive *E*-value cutoff, the automatic search converged at three iterations within the immediate OGT family. Eighteen sequences were found by this procedure. After removal of redundant sequences and fragments, the remaining 12 sequences formed the OGT set (Figure 1, group 1). An automatic sequence alignment (T-Coffee[18]) was manually improved using predicted secondary structural information (Jpred,[19] http://jura.ebi.ac.uk:8888).

One sequence found just below the *E*-value cutoff, the *Streptococcus pneumoniae* glycosyl transferase cpoA[20] (gi|2108333), was used to search for remote homologs. Three rounds of iterative PSI-BLAST searches were performed starting from cpoA (residues 1-330) with an *E*-value cutoff of 0.001 against the NCBI non-redundant database (3/29/01). Hits found from one round were grouped using single-linkage clustering at a score of 1.0 bit per site (SEALS[21]) and used as queries for the next round as described.[11] Each query was run until convergence or for ten iterations, whichever occurred first. Approximately 2700 sequences were found by this procedure (Supplementary Material). The first member of the OGT family, gi|7469165 (residues 414-564), was found during the second round with an *E*-value of $2 \times 10^{-5}$ on the second iteration from the query gi|6690126 (residues 53-382). The majority of the PSI-BLAST hits were sugar processing proteins, as inferred from their NCBI database definition lines. To reduce this multiple sequence alignment to a manageable size while still retaining sequence diversity, only sequences from a representative eukaryote (*Saccharomyces cerevisiae*) and prokaryote (*Escherichia coli*) were aligned (Figure 1, groups 2 and 3). The S2 ribosomal proteins (group 4) were found through an alternative method and will be discussed below. Information from secondary structure predictions and known tertiary structures found during the iterative searches was used at this stage to construct the final alignment between the OGTs and PSI-BLAST hits (Figure 1).

### Fold prediction of O-linked GlcNAc transferases

A total of four distinct PDB structures, all representing two-domain Rossmann-type folds with C-terminal helices, were among the significant PSI-BLAST hits found during the searches. These structures were yeast glycogen phosphorylase, (PDB code 1ygpA, gi|130174; Figure 2(a)), bacterial maltodextrin phosphorylase (1e4oA, gi|10120893), bacterial glycosyl transferase MurG (1f0kA, gi|9955024, Figure 2(a)), and bacterial UDP-GlcNac 2-epimerase (1f6dA, gi|148189). The sequence homology and structural similarity between several of these structures have been described.[22–24] The quality of the PSI-BLAST hits to these structures and the presence of a conserved motif (described below) made it likely that the non-TPR portion of the human OGT, as well as most of the sequences retrieved by the database searches, adopted a similar fold. A fifth protein, phage T4 β-glucosyl transferase (1c3jA, gi|121691), was added to the alignment because of its documented homology and structural similarity to glycogen phosphorylase.[25,26]

To confirm the validity of this prediction, it was necessary to evaluate the multiple sequence alignment in greater detail (Figure 1). It was observed that significant sequence similarity persisted between the OGTs and the PSI-BLAST hits along the entire length of the alignment. The similarity was completely consistent with the Rossmann fold domains in at least two respects. First, secondary structure predictions of seven β strands in the first domain and six β strands in the second domain matched well with the consensus structure from the experimentally characterized Rossmann folds found during the database searches (Figure 1). Locations and lengths of predicted helices in both domains exhibited the $(\beta\alpha)_6$ alternation typical of the Rossmann fold (Figure 1 and Supplementary Material). C-Terminal recognition helices similar to those observed in the glycogen phosphorylase, maltodextrin phosphorylase, and MurG structures were predicted for most of the sequences (Figure 2(a) and (b) and Supplementary Material). Second, patterns of side-chain hydrophobicity, size, and charge were well-conserved in almost all sequences. Noteworthy here was the predominantly hydrophobic character of the buried β strands, and the presence of a conserved small side-chain (Asp) at the start of strand 4 in each domain. This latter observation was characteristic of the Rossmann fold because of the tight turn necessary for the chain to cross over to the other side of the domain.[27]

In addition, a conserved GPGTF motif was observed in the second domain (Figure 1).[8,24,28,29] This motif could be loosely defined by the presence of a hydrophobic strand 4, a functional acidic residue in the middle of helix 4, followed by a glycine-mediated turn into hydrophobic strand 5. The presence of this conserved motif in the correct context of the rest of the alignment provided further evidence that these sequences were homologous, as the six known structures in the multiple sequence alignment of Figure 1 superimposed well in this region (Figure 2(d)).

Finally, the sequence of the human OGT (residues 330-920) was submitted to a fold-recognition server (http://www.cs.bgu.ac.il/ ~ bioinbgu/).[30] This server was fully automatic and used a battery of sequence homology, threading and secondary structure prediction methods to assess the compatibility between a sequence and a library of structures. Thus, this method was independent of, yet complementary to, the sequence database searches

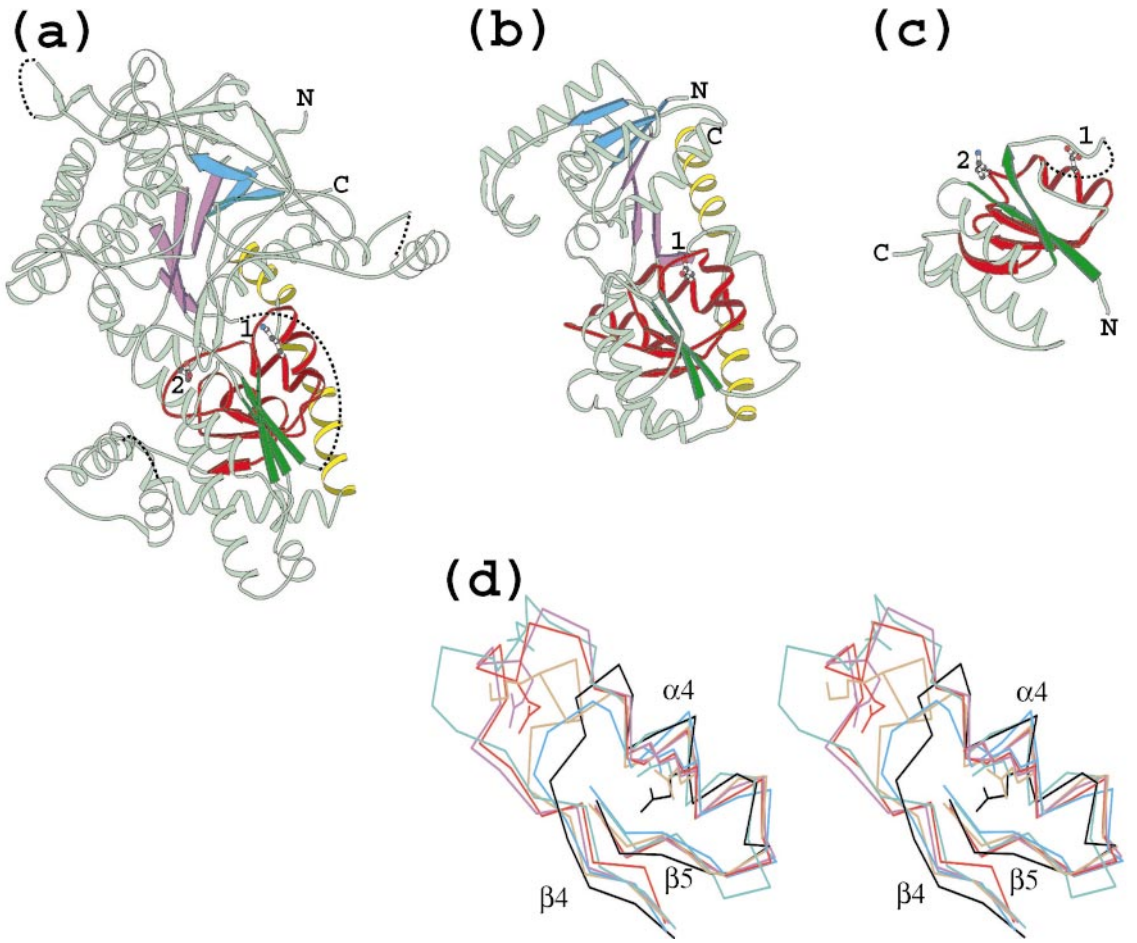


**Figure 2.** Structures of three proposed O-linked GlcNAc transferase homologs. (a) Yeast glycogen phosphorylase (PDB 1ygpA[42]); (b) bacterial glycosyl transferase MurG (PDB 1f0kA[24]); (c) *Thermus thermophilus* ribosomal protein S2 (PDB 1fjfB[39]). β-Strands featured in the multiple sequence alignment are color coded according to Figure 1. Additional helices near the conserved GPGTF motif in domain II are colored red. Two conserved helices at the extreme C terminus of 1ygpA and 1f0kA (shown schematically in Figure 1) are colored yellow. Side-chains of conserved residues 1 and 2, discussed in the text, are indicated by numbers. Coordinate gaps in the PDB file and some deleted non-homologous regions are connected by dotted lines to facilitate chain tracing. Regions of the structures not explicitly aligned in Figure 1 were slightly narrowed to facilitate viewing of the overall fold of each molecule. Figures were composed with BobScript.[43,44] (d) Stereo view of superimposition of the GPGTF (glycogen phosphorylase/glycosyl transferase) motif region (IIβ4 to IIβ5) for six PDB structures discussed in the text. Portions of PDB structures of T4 phage β-glucosyltransferase (green, 1c3jA[34]), bacterial maltodextrin phosphorylase (red, 1e4oA[22]), bacterial MurG (black, 1f0kA[24]), bacterial UDP-GlcNAc 2-epimerase (blue, 1f6dA[23]), *T. thermophilus* ribosomal protein S2 (brown, 1fjfB[39]), yeast glycogen phosphorylase (purple, 1ygpA[42]). The orientation of the structural fragment is identical with the orientation of the complete structures in (a)-(c). Conserved side-chains 1 and 2, discussed in the text, are shown explicitly. The Figure was composed using InsightII (Molecular Simulations, Inc.).

used in this work. One of the top hits returned was the bacterial glycosyl transferase MurG structure (1f0kA), which had a significant consensus score of 10.3. In addition, the alignment for this hit produced by the PRFSEQ component of the server was largely identical in the second domain with that displayed in Figure 1 (data not shown). The agreement of this independent, automatic fold-recognition server with the results of the manually adjusted sequence alignment further supported the homology found in PSI-BLAST searches. Taking all the evidence together, it is confidently predicted that the human OGT belongs to the GPGTF superfamily and folds as two Rossmann fold domains, whose mutual orientation is partly determined by C-terminal α-helices.

## Conserved residues from multiple sequence alignment of O-linked GlcNAc transferases

Despite a shared predicted fold, the proteins aligned in Figure 1 exhibited diverse biological functions and catalytic mechanisms. Examples of both inverting (the anomeric configuration of the product sugar is inverted relative to the configuration of the starting sugar, e.g. human OGT, bac-

terial MurG) and retaining (the anomeric conformation of the product sugar is retained relative to the starting configuration, e.g. bacterial MalP, yeast GP) glycosyl transferases, as well as sugar-processing enzymes with other functions (*e.g.* bacterial GlcNAc-2-epimerase), were represented. The catalytic mechanisms for each of these enzymes are known to varying degrees of understanding. Well understood, and supported by structural evidence, is the mechanism for inverting glycosyl transferases, which is thought to involve attack of an activated sugar donor by an acceptor species in a single displacement reaction.[14,31] Less well understood is the glycosyl transferase mechanism leading to retention of configuration, but one candidate (by analogy with retaining glycoside hydrolases) is a double-displacement mechanism which requires a side-chain catalytic nucleophile in the vicinity of the sugar donor.[31,32] One complication in the quest for prediction of the structural basis of the retaining mechanism is the observation that the first structure of a retaining NDP-sugar glycosyl transferase, IgtC, exhibits a different fold than glycogen phosphorylase, and yet the correspondence within the active sites of these enzymes suggests that they have similar reaction mechanisms.[31,33]

Consequently, the locations and identities of functional residues are under active investigation, and several groups have described the presence of conserved residues in the functional domains of glycosyl transferases.[8,13,24,28,29] These conserved residues might have structural, binding, or catalytic importance. For example, the conserved Asp residues at the start of strand 4 in each domain play crucial structural roles in the Rossmann fold, as discussed above. Two other conserved residues were apparent in domain II of the multiple sequence alignment of Figure 1: a conserved D/E/K in the middle of helix 4 and a conserved D/E between strand 4 and helix 4. For brevity, the following discussion refers to the conserved helix 4 position as ''position 1'' and the conserved position between strand 4 and helix 4 as ''position 2''.

In the bacterial GlcNAc-2-epimerase, the bacterial glycosyl transferase MurG, and the T4 β-glucosyl transferase, Asp and Glu residues at position 1 have been implicated in substrate binding, with the carboxyl group of the acidic side-chain hydrogen bonding to the ribose ring of a UDP-sugar moiety.[23,24,34] The presence of the GPGTF motif and of the Asp at position 1 predict this location to be part of the UDP-GlcNAc binding site in the human OGT. Most sequences displayed in Figure 1 contained an Asp or Glu at position 1, but several sequences contained a basic or even hydrophobic side-chain. It is important to note that the absence of an acidic side-chain at position 1 did not invalidate the fold prediction nor did it discount a transferase function; it merely may have indicated a different mechanism for a similar reaction.[35] For example, yeast glycogen phosphorylase and bacterial maltodextrin phosphorylase contained a Lys residue at this position (covalently linked to PLP cofactors that are involved in catalysis) but function as, and are classified as, sugar transferases.[12,36,37] It has also been proposed that the side-chains of His or Tyr, the latter observed at position 1 in *E. coli* glycogen synthase (gi|1169908 of Figure 1), could hydrogen bond to the UDP ribose.[28,38] In contrast, the ribosomal protein S2 contained Glu at position 1 but the function of this residue is unknown.[39]

Many, but not all, sequences in Figure 1 contained position 2 in an inserted loop between strand 4 and helix 4. Some of these sequences lacked the acidic side-chain. For example, the OGTs contained the loop but did not have an Asp or Glu at position 2. Acidic residues were present at position 2 in the ribosomal protein S2, phage T4 β-glucosyl transferase, bacterial maltodextrin phosphorylase, and yeast glycogen phosphorylase. The function, if any, of position 2 in ribosomal protein S2 is unknown, and in structures of the phage T4 enzyme position 2 has not been associated with substrate binding or catalytic activity.[34,40] In maltodextrin phosphorylase (and presumably glycogen phosphorylase), the side-chain carboxyl of E672 at position 2 hydrogen bonds to an oligosaccharide substrate.[22]

However, in some glycosyl transferases classified with retaining mechanism, experimental evidence suggests that position 2 is associated with catalytic activity. For example, the *Acetobacter xylinium* α-mannosyltransferase AceA (closely related to the *E. coli* mannosyltransferase A, gi|2125945, in Figure 1) completely loses activity, but retains structural integrity, when E287 at position 2 is replaced by Ala.[28] In addition, similar results have been obtained for the E510A mutation at position 2 in human muscle glycogen synthase (closely related to the yeast glycogen synthase, isoform 2, gi|6323287 in Figure 1).[38] For those glycosyl transferase sequences in Figure 1 that had been assigned to a mechanistic family in the CAZY database, a strong correlation was observed between the presence of an Asp or Glu at position 2 and the classification of the mechanism of the enzyme as retaining: every retaining glycosyl transferase had an acidic residue at position 2. It is important to note that the presence of the inserted loop containing position 2 was not itself diagnostic of retaining mechanism, as the inverting family 41 of the OGTs exhibited the loop but no acidic residue at position 2. Additional structural and kinetic information is necessary to unambiguously identify the catalytic residues in these enzymes.

Domain I is necessary to complete the active site cleft for the enzymes whose structures are known, but the details of the functional residues in this domain are unclear. Proposed catalytic residues E22 and D100 of T4 β-glucosyl transferase are located in this domain.[34] Domain I is thought to contain the acceptor binding site of MurG and probably contributes to DNA binding in T4 β-glucosyl transferase.[24,34] Less residue conservation indicative of function was observed in domain I of

the Figure 1 alignment, and a region comparable to the GPGTF motif of domain II was not found. However, we did find evidence (Supplementary Material) of G-rich loops in the helical regions of some sequences, similar to those seen in MurG as variants of the phosphate binding G-loops found in classical Rossmann fold proteins.[24]

In summary, the active center of most of the enzymes shown is Figure 1 is formed by the cleft between domains I and II, with at least one residue from the GPGTF motif contributing to substrate binding and possibly one residue from the GPGTF motif contributing to catalysis in retaining glycosyl transferases. It is likely that residues from domain I also contribute to additional (phosphate) binding sites and perhaps to catalysis, but the latter cannot be generalized for these diverse enzymes from the data in Figure 1.

## Remote homologs of the O-linked GlcNAc transferases

The GPGTF superfamily's large size suggested that its evolutionary roots are ancient. Consequently, additional remote homologs outside of the superfamily may await discovery. PSI-BLAST profile searches of the *S. cerevisiae* and *E. coli* genomes found the functional motif of the GPGTF superfamily in the ribosomal proteins S0A (gi|417730) and S2 (gi|133908). Remarkably, the GPGTF motif was also conserved in this family of ribosomal proteins (Figure 1, group 4). In addition, the acidic residues at the GPGTF motif positions 1 and 2 were conserved, despite the fact that these ribosomal proteins were not known to function as sugar transferases. The *Thermus thermophilus* homolog (gi|10835564, PDB code 1fjfB) showed that many of the secondary structural elements in the second domain of glycogen phosphorylase and MurG were similar in length and orientation to those of the ribosomal proteins (Figure 2(c)). The GPGTF motif residues 1 and 2 in the ribosomal protein S2 structure were also positioned similarly to those of the other structures. These potential remote homologs raise the possibility that the GPGTF superfamily might have evolved from ribosomal proteins, which are themselves thought to be ancient.[41]

To gauge the relationship between each of the sequences aligned in Figure 1, the evolutionary distances between them were estimated. This was accomplished by construction of a Euclidian space in which sequences were represented by points with Euclidian distances between them approximating evolutionary distances (V.N. Grishin and N.V.G., unpublished results). Thus, proteins that are more divergent in evolutionary sequence space are separated by greater distances in the multidimensional space, as shown in Figure 3. Three results of this work were rationalized by this plot. First, the OGTs were well-separated from the rest of the sequences, especially the glycogen phosphorylases, explaining the difficulty of connecting these groups directly through automatic database searches. Second, the sequence gi|2108333 (cpoA) was located near the bulk of *E. coli* and *S. cerevisiae* sequences, suggesting that it was a good linking sequence which connected the more divergent sequences together. Third, the positioning of the ribosomal S2 sequences was consistent with their proposed relationship to the GPGTF superfamily.

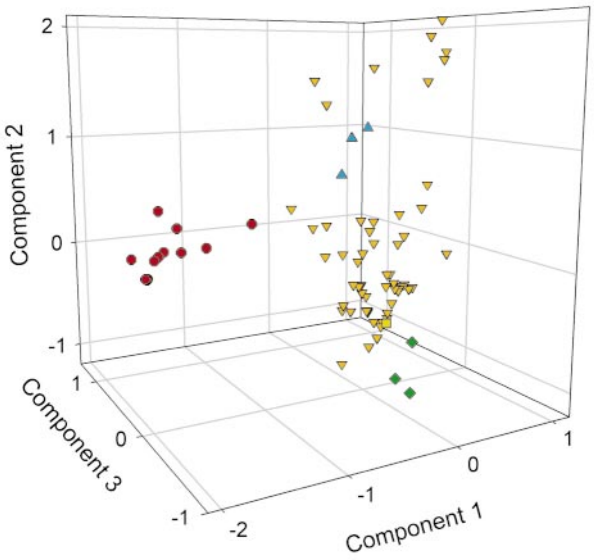


**Figure 3.** Estimates of evolutionary distances between the O-linked GlcNAc transferases and homologous sequences. The three principal dimensions (the most information-rich components) of a multidimensional Euclidian space approximating evolutionary distances are shown. Points on this plot represent sequences in Figure 1; greater distances between points correspond to greater predicted evolutionary distance between sequences. O-linked GlcNAc transferases are displayed as red circles, glycogen phosphorylases and maltodextrin phosphorylase are shown as green diamonds, ribosomal S2 proteins as blue triangles, and the remaining sequences of Figure 1 as inverted orange triangles. The linker sequence gi|2108333 (cpoA) is shown as a yellow square. This plot was generated from the following algorithm (V.N. Grishin & N.V.G., unpublished results). The conserved segments of the multiple sequence alignment of Figure 1 were used to calculate pairwise identity fractions $q_{ij}$ between each sequence pair $i$ and $j$. The identity fractions were converted to evolutionary distances using the formula $d_{ij} = -\ln[(q_{ij} - q_{ij}^{\text{random}})/(1 - q_{ij}^{\text{random}})]$, where $q_{ij}^{\text{random}}$ is an expected identity percentage of two random sequences with the same amino acid composition as sequences $i$ and $j$. Each sequence was represented as a point in a multidimensional Euclidian space in such a way that Euclidian distances $\bar{d}_{ij}$ between the points optimally approximated the estimated distances $d_{ij}$ between the sequences using:

$$\min\left[\sum_{ij}(\bar{d}_{ij}^2 - d_{ij}^2)^2/d_{ij}^4\right]$$

The first three dimensions of this space formed the axes of the Figure.

## Structural unification of glycosyl transferase functional families

The CAZY classification scheme groups all known glycosyl transferases into 51 families based on catalytic mechanism and conservation of active site residues.[12] It was found that of the approximately 2700 hits retrieved from the iterative PSI-BLAST searches, 2300 of these covered 15 of the 51 families (Table 1). These results suggested that up to one-third of glycosyl transferase families shared significant sequence homology and were evolutionarily related to the GPGTF superfamily. Representatives from 12 of these 15 families exist in the *E. coli* and *S. cerevisiae* genomes (Figure 1). Hits were observed to families exhibiting both inverting and retaining mechanisms. Only three of the 15 families were currently represented by a known structure, but these structures turned out to be the two-domain Rossmann folds found during the iterative PSI-BLAST searches (Table 1). It was therefore likely that the fold prediction for the OGTs extends to the remaining 12 families as well.

This hypothesis was tested by probing the extent of inter-family sequence relationships using PSI-BLAST. For 13 out of the 15 CAZY families listed in Table 1, it was demonstrated that randomly chosen members of one family used as PSI-BLAST queries against the NCBI non-redundant database (6/27/01) easily recovered members of the other 14 families. Manual inspection confirmed that the GPGTF motif in the second Rossmann domain was typically aligned in the hit (data not shown). For two families, 10 and 47, it was difficult to directly recover hits to the GPGTF motif from other families, and more extensive searches were needed. Links were eventually found between these two families and the remaining 13 using intermediate sequences. For example, the intermediate sequence query gi|9631683, residues 199-451, found PSI-BLAST hits in the GPGTF motif region to both family 10 (gi|11290272, residues 234-319, *E*-value $4e^{-6}$) and family 47 (gi|4263719, residues 368-440, *E*-value $3e^{-4}$) during the first iteration (*E*-value cutoff of 0.001, expect 100 in all cases). Family 10 was then linked to additional families through family 4. Family 4 query gi|7427928, residues 62-377, recovered family 10 sequence gi|4587296, residues 200-359, on the third iteration with an *E*-value of 0.54 for a hit to the GPGTF motif region. These results demonstrated that the 15 CAZY families given in Table 1 are homologous to each other and strongly suggest that proteins of these families share the two-domain Rossmann fold similar to that first described for glycogen phosphorylase.

In addition, stronger cross-family relationships were found merely through simple BLAST searches against known glycosyl transferases from the CAZY database (Table 1). Queries from ten families did not register significant BLAST hits (*E*-value < 0.0001) with any other family. However, sequence queries from families 1, 4, 5, 28, and 33 exhibited statistically significant BLAST hits to sequences belonging to a different family (Table 1). Inspection of the sequence regions involved in the

**Table 1.** Glycosyl transferase families found during iterative PSI-BLAST searches and significant BLAST hits between glycosyl transferase families

| Glycosyl transferase family[a] | Representative sequence | PDB | Number of sequences[b] | Mechanism[a] | BLAST hits to different glycosyl transferase family[c] |
|---|---|---|---|---|---|
| 1 | 7477437 | 1iir[13] | 630 | Inverting | 28 (17)[d] |
| 3 | 6321127 | | 21 | Retaining | None |
| 4 | 2125946 | | 656 | Retaining | 5 (891); 33 (1) |
| 5 | 146139 | | 341 | Retaining | 4 (3278) |
| 9 | 585816 | | 100 | Inverting | None |
| 10 | 4587296 | | 65 | Inverting | None |
| 19 | 126464 | | 25 | Inverting | None |
| 20 | 730244 | | 64 | Retaining | None |
| 28 | 6685640 | 1f0kA[24] | 62 | Inverting | 1 (13) |
| 30 | 3821833 | | 43 | Unknown | None |
| 32 | 1362173 | | 1 | Retaining | None |
| 33 | 9989060 | | 8 | Inverting | 4 (3) |
| 35 | 130174 | 1e4oA[19], *et al.* | 126 | Retaining | None |
| 41 | 4505499 | | 18 | Inverting | None |
| 47 | 3435314 | | 55 | Bifunctional | None |

[a] Family classifications and mechanisms for hits from the iterative PSI-BLAST searches were assigned according to the CAZY glycosyl transferase database of 3 May 2001 (http://afmb.cnrs-mrs.fr/ ~ pedro/CAZY/gtf.html).[12] If an iterative PSI-BLAST hit did not exactly match a CAZY sequence, it was nonetheless assigned to a family if its first hit from a BLAST search of the CAZY database had an *E*-value of less than 0.0001 and there was evidence that the hit contained the conserved GPGTF region mentioned in the text.

[b] The number of hits from the iterative PSI-BLAST searches that could be assigned to the family using the criteria described above.

[c] A significant BLAST hit was defined as having an *E*-value of less than 0.0001 to a member of a different CAZY family than the query and there was evidence that the hit contained the conserved GPGTF region.

[d] The first number is the CAZY family number of the alternative hit. The number in parentheses is the number of hits from the alternative family found from all queries from the family in the first column.

BLAST hits revealed that nearly all cross-family hits listed in Table 1 involved the GPGTF motif. Two other pieces of evidence contributed to the significance of these relationships. First, some of the interfamily relationships in Table 1 had been independently hypothesized, as it had been previously suggested that families 3, 4, and 5 were related.[12] Second, the links between families were reciprocal, as family pairs 4/5, 1/28, and 4/33 detected each other.

### Conclusions

It is proposed that the human OGT adopts a two-domain Rossmann fold with an N-terminal TPR region. The presence of a conserved motif in the OGT family identifies a substrate binding site in the second domain of these enzymes and links this group to the GPGTF superfamily. The observation of sequence and structural similarity between the ribosomal S2 protein family and the GPGTF superfamily may point to the latter's evolutionary origins. These results enlarge the GPGTF superfamily and suggest that up to one-third of the CAZY glycosyl transferase functional families are homologous within the GPGTF superfamily. It is hoped that this information will help to design experiments and select glycosyl transferases for future structural characterization.

## Acknowledgements

This work was supported, in part, by the Welch foundation grant I-1505 to N.V.G.

## References

1. Comer, F. I. & Hart, G. W. (2000). O-Glycosylation of nuclear and cytosolic proteins. Dynamic interplay between O-GlcNAc and O-phosphate. *J. Biol. Chem.* **275**, 29179-29182.
2. Shafi, R., Iyer, S. P., Ellies, L. G., O'Donnell, N., Marek, K. W., Chui, D. *et al*. (2000). The O-GlcNAc transferase gene resides on the X chromosome and is essential for embryonic stem cell viability and mouse ontogeny. *Proc. Natl Acad. Sci. USA,* **97**, 5735-5739.
3. Wells, L., Vosseller, K. & Hart, G. W. (2001). Glycosylation of nucleocytosplasmic proteins: signal transduction and O-GlcNAc. *Science,* **291**, 2376-2378.
4. Hart, G. W. (1997). Dynamic O-linked glycosylation of nuclear and cytoskeletal proteins. *Annu. Rev. Biochem.* **66**, 315-335.
5. Haltiwanger, R. S., Busby, S., Grove, K., Li, S., Mason, D., Medina, L.*et al*. (1997). O-glycosylation of nuclear and cytoplasmic proteins: regulation analogous to phosphorylation? *Biochem. Biophys. Res. Commun.* **231**, 237-242.
6. Comer, F. I. & Hart, G. W. (2001). Reciprocity between O-GlcNAc and O-phosphate on the carboxy terminal domain of RNA polymerase II. *Biochemistry,* **40**, 7845-7852.
7. Kelly, W. G., Dahmus, M. E. & Hart, G. W. (1993). RNA polymerase II is a glycoprotein. *J. Biol. Chem.* **268**, 10416-10424.
8. Roos, M. D. & Hanover, J. A. (2000). Structure of O-linked GlcNAc transferase: mediator of glycan-dependent signaling. *Biochem. Biophys. Res. Commun.* **271**, 275-280.
9. Aravind, L. & Koonin, E. V. (1999). Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches. *J. Mol. Biol.* **287**, 1023-1040.
10. Copley, R. R. & Bork, P. (2000). Homology among (betaalpha)(8) barrels: implications for the evolution of metabolic pathways. *J. Mol. Biol.* **303**, 627-641.
11. Pei, J. & Grishin, N. V. (2001). GGDEF domain is homologous to adenylyl cyclase. *Proteins: Struct. Funct. Genet.* **42**, 210-216.
12. Campbell, J. A., Davies, G. J., Bulone, V. & Henrissat, B. (1997). A classification of nucleotide-diphospho-sugar glycosyltransferases based on amino acid sequence similarities. *Biochem. J.* **326**, 929-939.
13. Mulichak, A. M., Losey, H., Walsh, C. T. & Garavito, R. M. (2001). Structure of the UDP-glucosyltransferase GtfB that modifies the heptapeptide aglycone in the biosynthesis of vancomycin group antibiotics. *Structure,* **9**, 547-557.
14. Unligil, U. M. & Rini, J. M. (2000). Glycosyltransferase structure and mechanism. *Curr. Opin. Struct. Biol.* **10**, 510-517.
15. Fletterick, R. J. & Madsen, N. B. (1980). The structures and related functions of phosphorylase a. *Annu. Rev. Biochem.* **49**, 31-61.
16. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389-3402.
17. Lubas, W. A., Frank, D. W., Krause, M. & Hanover, J. A. (1997). O-Linked GlcNAc transferase is a conserved nucleocytoplasmic protein containing tetratricopeptide repeats. *J. Biol. Chem.* **272**, 9316-9324.
18. Notredame, C., Higgins, D. G. & Heringa, J. (2000). T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205-217.
19. Cuff, J. A., Clamp, M. E., Siddiqui, A. S., Finlay, M. & Barton, G. J. (1998). JPred: a consensus secondary structure prediction server. *Bioinformatics,* **14**, 892-893.
20. Grebe, T., Paik, J. & Hakenbeck, R. (1997). A novel resistance mechanism against beta-lactams in *Streptococcus pneumoniae* involves CpoA, a putative glycosyltransferase. *J. Bacteriol.* **179**, 3342-3349.
21. Walker, D. R. & Koonin, E. V. (1997). SEALS: a system for easy analysis of lots of sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **5**, 333-339.
22. Watson, K. A., McCleverty, C., Geremia, S., Cottaz, S., Driguez, H. & Johnson, L. N. (1999). Phosphorylase recognition and phosphorolysis of its oligosaccharide substrate: answers to a long outstanding question. *EMBO J.* **18**, 4619-4632.
23. Campbell, R. E., Mosimann, S. C., Tanner, M. E. & Strynadka, N. C. (2000). The structure of UDP-*N*-acetylglucosamine 2-epimerase reveals homology to phosphoglycosyl transferases. *Biochemistry,* **39**, 14993-15001.
24. Ha, S., Walker, D., Shi, Y. & Walker, S. (2000). The 1.9 Å crystal structure of *Escherichia coli* MurG, a

membrane-associated glycosyltransferase involved in peptidoglycan biosynthesis. *Protein Sci.* **9**, 1045-1052.
25. Artymiuk, P. J., Rice, D. W., Poirrette, A. R. & Willett, P. (1995). beta-Glucosyltransferase and phosphorylase reveal their common theme. *Nature Struct. Biol.* **2**, 117-120.
26. Holm, L. & Sander, C. (1995). Evolutionary link between glycogen phosphorylase and a DNA modifying enzyme. *EMBO J.* **14**, 1287-1293.
27. Horvath, M. M. & Grishin, N. V. (2001). The C-terminal domain of HPII catalase is a member of the type I glutamine amidotransferase superfamily. *Proteins: Struct. Funct. Genet.* **42**, 230-236.
28. Abdian, P. L., Lellouch, A. C., Gautier, C., Ielpi, L. & Geremia, R. A. (2000). Identification of essential amino acids in the bacterial alpha - mannosyltransferase aceA. *J. Biol. Chem.* **275**, 40568-40575.
29. Kapitonov, D. & Yu, R. K. (1999). Conserved domains of glycosyltransferases. *Glycobiology,* **9**, 961-978.
30. Fischer, D. (2000). Hybrid fold recognition: combining sequence derived properties with evolutionary information. *Pac. Symp. Biocomput.* 119-130.
31. Davies, G. J. (2001). Sweet secrets of synthesis. *Nature Struct. Biol.* **8**, 98-100.
32. Saxena, I. M., Brown, R. M., Jr, Fevre, M., Geremia, R. A. & Henrissat, B. (1995). Multidomain architecture of beta-glycosyl transferases: implications for mechanism of action. *J. Bacteriol.* **177**, 1419-1424.
33. Persson, K., Ly, H. D., Dieckelmann, M., Wakarchuk, W. W., Withers, S. G. & Strynadka, N. C. (2001). Crystal structure of the retaining galactosyltransferase LgtC from *Neisseria meningitidis* in complex with donor and acceptor sugar analogs. *Nature Struct. Biol.* **8**, 166-175.
34. Morera, S., Imberty, A., Aschke-Sonnenborn, U., Ruger, W. & Freemont, P. S. (1999). T4 phage beta-glucosyltransferase: substrate binding and proposed catalytic mechanism. *J. Mol. Biol.* **292**, 717-730.
35. Henrissat, B. & Davies, G. J. (2000). Glycoside hydrolases and glycosyltransferases. Families, modules, and implications for genomics. *Plant Physiol.* **124**, 1515-1519.
36. Johnson, L. N., Acharya, K. R., Jordan, M. D. & McLaughlin, P. J. (1990). The refined crystal structure of the phosphorylase-heptulose 2-phosphate-oligosaccharide-AMP complex. *J. Mol. Biol.* **211**, 645-661.
37. Palm, D., Klein, H. W., Schinzel, R., Buehner, M. & Helmreich, E. J. M. (1990). The role of pyridoxal-5′-phosphate in glycogen phosphorylase catalysis. *Biochemistry,* **29**, 1099-1107.
38. Cid, E., Gomis, R. R., Geremia, R. A., Guinovart, J. J. & Ferrer, J. C. (2000). Identification of two essential glutamic acid residues in glycogen synthase. *J. Biol. Chem.* **43**, 33614-33621.
39. Wimberly, B. T., Brodersen, D. E., Clemons, W. M., Jr, Morgan-Warren, R. J., Carter, A. P. & Vonrhein, C., *et al.* (2000). Structure of the 30S ribosomal subunit. *Nature,* **407**, 327-339.
40. Morera, S., Lariviere, L., Kurzeck, J., Aschke-Sonnenborn, U., Freemont, P. S., Janin, J. & Ruger, W. (2001). High resolution crystal structures of T4 phage beta-glucosyltransferase: induced fit and effect of substrate and metal binding. *J. Mol. Biol.* **311**, 569-577.
41. Ramakrishnan, V. & White, S. W. (1998). Ribosomal protein structures: insights into the architecture, machinery and evolution of the ribosome. *Trends Biochem. Sci.* **23**, 208-212.
42. Lin, K., Rath, V. L., Dai, S. C., Fletterick, R. J. & Hwang, P. K. (1996). A protein phosphorylation switch at the conserved allosteric site in GP. *Science,* **273**, 1539-1542.
43. Kraulis, P. J. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein stucture. *J. Appl. Crystallog.* **24**, 946-950.
44. Esnouf, R. M. (1997). An extensively modified version of MolScript that includes greatly enhanced coloring capabilities. *J. Mol. Graph. Model,* **15**, 132-134, 112-133.

***Edited by J. Thornton***

*(Received 25 July 2001; received in revised form 2 October 2001; accepted 4 October 2001)*

**http://www.academicpress.com/jmb**

Supplementary Material is available on IDEAL or by anonymous ftp at ftp://iole.swmed.edu/usr1/ftproot/pub/GPGTF/