The beginnings of mucin biosynthesis: The crystal structure of UDP-GalNAc:polypeptide α-*N*-acetylgalactosaminyltransferase-T1

Timothy A. Fritz*, James H. Hurley†, Loc-Ba Trinh‡, Joseph Shiloach‡, and Lawrence A. Tabak*§

*Section on Biological Chemistry, †Laboratory of Molecular Biology, and ‡Biotechnology Unit, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Department of Health and Human Services, Bethesda, MD 20892

Edited by Robert L. Hill, Duke University Medical Center, Durham, NC, and approved September 10, 2004 (received for review August 3, 2004)

UDP-GalNAc:polypeptide α-*N*-acetylgalactosaminyltransferases (ppGaNTases) initiate the formation of mucin-type, *O*-linked glycans by catalyzing the transfer of α-*N*-acetylgalactosamine from UDP-GalNAc to Ser or Thr residues of core proteins to form the Tn antigen (GalNAc-α-1-*O*-Ser/Thr). ppGaNTases are unique among glycosyltransferases in containing a C-terminal lectin domain. We present the x-ray crystal structure of a ppGaNTase, murine ppGaNTase-T1, and show that it folds to form distinct catalytic and lectin domains. The association of the two domains forms a large cleft in the surface of the enzyme that contains a Mn^{2+} ion complexed by invariant D209 and H211 of the "DXH" motif and by invariant H344. Each of the three potential lectin domain carbohydrate-binding sites (α, β, and γ) is located on the active-site face of the enzyme, suggesting a mechanism by which the transferase may accommodate multiple conformations of glycosylated acceptor substrates. A model of a mucin 1 glycopeptide substrate bound to the enzyme shows that the spatial separation between the lectin α site and a modeled active site UDP-GalNAc is consistent with the *in vitro* pattern of glycosylation observed for this peptide catalyzed by ppGaNTase-T1. The structure also provides a template for the larger ppGaNTase family, and homology models of several ppGaNTase isoforms predict dramatically different surface chemistries consistent with isoform-selective acceptor substrate recognition.

glycosyltransferase | mucin

Mucin-type *O*-glycoprotein biosynthesis is initiated by the transfer of α-*N*-acetylgalactosamine from UDP-GalNAc to Ser or Thr residues of core proteins. The enzymes catalyzing this reaction are UDP-GalNAc:polypeptide α-*N*-acetylgalactosaminyltransferases (ppGaNTases, EC 2.4.1.41) and define the family 27 retaining glycosyltransferases (1). Genomic database analyses indicate that ppGaNTase expression is conserved from nematodes to humans with mammals expressing up to 24 distinct members, whereas *Drosophila* express 14, and *Caenorhabditis elegans* express 9 (2). At least one isoform is essential for *Drosophila* development (3, 4), and a recent report indicates that loss of ppGaNTase-T3 function may cause familial tumoral calcinosis (5). The biological importance of mucin glycans is further highlighted by the finding that disruption of the core 1 β1,3-galactosyltransferase (the enzyme that elongates sugar chains initiated by ppGaNTases to form the T antigen Gal-β1,3-GalNAc-α-1-*O*-Ser/Thr) is embryonic lethal in mice (6). However, mice in which the individual expression of ppGaNTase-T1, -T4, -T5, and -T13 has been ablated remain viable, suggesting possible functional redundancy among isoforms (2, 7, 8).

ppGaNTases are type II Golgi membrane proteins (9, 10) whose primary structure consists of a short cytoplasmic tail, a single transmembrane domain, a variable length stem region, a catalytic domain of ≈350 amino acids, and an ≈130-aa C-terminal lectin domain. The presence of this lectin domain is unique among glycosyltransferases (11). *In vitro* analyses have defined two ppGaNTase activities based on the nature of acceptor peptide substrates. Most isoforms, termed peptide transferases, transfer GalNAc to both unmodified peptides and glycopeptides, whereas a few such as mammalian ppGaNTase-T7 and -T10 and *Toxoplasma gondii* ppGaNTase-T3 appear to require the prior addition of GalNAc before they transfer additional GalNAc residues to the peptides (12, 13). These are termed glycopeptide transferases. Biochemical analyses suggest that lectin domain function is required for the transfer of GalNAc to glycopeptide but not peptide substrates (compare ref. 14 with ref. 15). To gain an understanding of the structure/function relationship of ppGaNTases, we have determined the crystal structure of murine ppGaNTase-T1.

Methods

Murine ppGaNTase-T1 (mppGaNTase-T1) was expressed as a fusion protein with maltose-binding protein at the N terminus and residues 42–559 of mppGaNTase-T1 at the C terminus separated by a tobacco etch virus protease recognition sequence. The fusion protein was expressed in *Pichia pastoris* and purified by using a combination of ion-exchange and hydrophobic interaction chromatography. The details of the cloning, expression, and purification are included as *Supporting Text*, which is published as supporting information on the PNAS web site. All chemicals were from Sigma–Aldrich unless otherwise noted.

Crystals of mppGaNTase-T1 (without maltose-binding protein) were grown by hanging drop vapor diffusion at room temperature. Crystal growth was initiated by mixing 1 μl of protein solution containing 5.8 mg/ml mppGaNTase-T1, 1 mM Mes (pH 6.5), 0.5 mM EDTA, 10 mM 2-mercaptoethanol (β-ME), 1 mM 4-(2-aminoethyl)benzenesulfonyl fluoride (AEBSF), 1 mM UDP-GalNAc, and 10 mM $MnCl_2$ with 1 μl of precipitant solution containing 14–16% polyethylene glycol 8000, 100 mM Mes (pH 6.0–6.5), and 0.2 M calcium acetate. Crystals were grown over 0.3 ml of precipitant solution in 48-well plates and appeared in 3–4 days. Samarium derivatives were prepared by transferring crystals for 1 h to precipitant solution containing 10 mM β-ME, 1 mM AEBSF, 1 mM UDP-GalNAc, and 10 mM $MnCl_2$, in which ≈0.1 M samarium acetate was substituted for 0.2 M calcium acetate. Crystals were transferred briefly (30–60 sec) to the same solution containing 15% polyethylene glycol 400 before cooling in a 95–100 K N_2 stream.

Native and derivative diffraction intensities from single crystals were collected by using 1° oscillations on a Raxis-IV detector and a rotating anode generator (Rigaku/MSC, The Woodlands, TX). Intensities from 60 (native) or 240 (derivative) frames were integrated and scaled by using the program DENZO/SCALEPACK (16). Samarium sites were located, and single-wavelength anom-

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: ppGaNTase, UDP-GalNAc:polypeptide α-*N*-acetylgalactosaminyltransferase; mppGaNTase-T1, murine ppGaNTase-T1; MUC*n*, mucin *n*.

Data deposition: The atomic coordinates have been deposited in the Protein Data Bank, www.pdb.org (PDB ID code 1XHB).

§To whom correspondence should be addressed. E-mail: tabakl@mail.nih.gov.

alous dispersion (SAD) phases to 2.71 Å were determined by using the programs SOLVE (17) and RESOLVE (18). A partial structure was automatically built and refined by using the programs RESOLVE-AUTOBUILD (19) and REFMAC (20). Manual model building was done by using XTALVIEW (21). A partial model was refined by using several rounds of torsional simulated annealing in CNS (22) by using experimental phase information from the samarium SAD data set before simulated annealing and final refinement against the native data set by using model phases. Domain interface residues were identified by using the program CONTACT from the CCP4 program suite (23), and solvent accessibilities were calculated by using the program AREAIMOL from the same suite. Topology diagrams were created by using TOPS and EDITTOPS (24). Electrostatic surface potentials were calculated by using SWISSPDBVIEWER (25) with partial atomic charges and a solvent ionic strength of 0.15 M. Secondary structures were also determined with SWISSPDBVIEWER. Homology models were created by using LOOK/GENEMINE (26) and were energy minimized by using the "full refine" option. Protein sequence alignments were created by using CLUSTALX (27) and manually edited with SEAVIEW (28).

Results

Expression and Purification. mppGaNTase-T1 was expressed as a fusion protein with maltose-binding protein at the N terminus and residues 42–559 of mppGaNTase-T1 separated by a tobacco etch virus protease recognition sequence (ENLYFQS). During purification, maltose-binding protein was released from mppGaNTase-T1 by copurifying protease activity that quantitatively cleaved the fusion construct between mppGaNTase-T1 residues L50 and V51. Additional proteolysis occurred during crystallization, yielding two smaller fragments differing by ≈1.9 kDa that cocrystallized. Sequence analysis of the larger of the two fragments indicated that it began at or near A88, and assay of washed crystals showed that they were enzymatically active (data not shown).

Electron density of the refined structure of mppGaNTase-T1 encompassed residues 95–553 with the exception of residues 347–358. Electron density corresponding to mppGaNTase-T1 residues 347–358 was also absent or poor in the α-*N*-acetylhexosaminlytransferase EXTL2 (residues 275–286, PDB ID code 1OMZ), bacterial SpsA (residues 218–231, PDB ID code 1QGQ), and bovine β1,4-galactosyltransferase (residues 347–354, PDB ID code 1FGX) structures, indicating that these residues form a flexible loop in several glycosyltransferases. The side chain of W316 also lacked electron density. ppGaNTase-T1 was crystallized in the presence of Mn^{2+} and UDP-GalNAc, but density was observed only for the manganese ion. Density for 2 GlcNAc and 1 Man residue was observed on N552, which is part of a NXS/T sequon for N-linked glycosylation. ϕ and ψ angles of all residues were in allowed regions of the Ramachandran plot, except for T308, R479, and K501. Because the electron density was well defined for each of these residues, their ϕ and ψ torsion angles were not changed. Crystallographic data are shown in Table 1.

Catalytic Domain Structure. The mppGaNTase-T1 catalytic domain (residues 95–426) folds as a central eight-stranded β-sheet flanked by α-helices (Fig. 1), as predicted for family 27 glycosyltransferases (29). All of the secondary structural elements referred to as the "SCG" domain (30) or UDP-binding domain (31) are preserved in mppGaNTase-T1 (Fig. 2). Two disulfide bonds, C106–C339 and C330–C408, are formed from four invariant cysteine residues. Mutating any of these Cys to Ala significantly reduces expression and abolishes enzyme activity (32). The three remaining catalytic domain Cys are moderately conserved (conservation was determined by comparison of 38 known enzymatically active ppGaNTases described to date) and do not form disulfide bonds. Two of these (C212 and C214) have been proposed to interact with UDP-GalNAc based upon chemical modification and mutagenesis studies. Mutating either Cys to Ala reduces activity to 6% and 17%, respectively, of wild-type levels (32). The third Cys (C235) is in a hydrophobic pocket and forms a close contact with M291. Mutation of C235 to Ala does not significantly affect enzyme function or expression (32).

The mppGaNTase-T1 catalytic domain tertiary structure is further strengthened by several invariant or highly conserved

Table 1. Crystallographic data and phasing and refinement statistics

	Native	Samarium acetate
Data collection and phasing		
Space group	$P4_3$	$P4_3$
Unit cell, Å*	$a = b = 65.605$	$a = b = 65.390$
	$c = 125.947$	$c = 125.597$
Resolution, Å	2.5 (2.66–2.5)	2.7 (2.8–2.7)
Unique reflections	18334 (1788)	14545 (1447)
Completeness, %	99 (97.3)	99.9 (100)
R_{merge}, %†	13.9 (63.7)	12.8 (42.9)
Figure of merit-SOLVE‡		0.32 (0.19)
Refinement statistics		
Resolution range	46.4–2.5	
No. of reflections	18279	
R, %§	22.8	
R_{free}, %¶	25.5	
rms deviations		
Bond length, Å	0.007	
Bond angle, °	1.5	
Average B factor, Å^2	40.4	
No. of protein atoms	3617	
No. of solvent atoms	37	

*Statistics shown in parentheses are for the highest-resolution shell.
†$R_{merge} = \Sigma|I(k) - \langle I(k)\rangle|\Sigma/I(k)$.
‡Figure of merit is before density modification.
§$R = \Sigma|F_{obs} - kF_{calc}|/\Sigma|F_{obs}|$.
¶R_{free} is the *R* value calculated for a randomly selected 7.9% of the data not used for refinement.

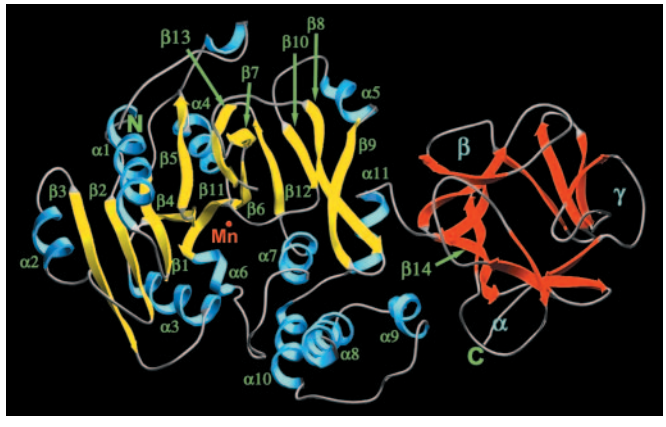

Fig. 1. Ribbon drawing of the mppGaNTase-T1 crystal structure. α-Helices and β-strands of the catalytic domain are colored blue and yellow, respectively, and random coil structures are colored gray. β-strands and random coils of the lectin domain are colored red and gray, respectively. The active-site Mn^{2+} ion is shown by the red sphere. Strands and helices are numbered from the N to C termini. The putative carbohydrate-binding sites of the α, β, and γ repeats of the lectin domain are indicated by the corresponding Greek letters. Except for β-strand 14, the β-strands of the lectin domain were left unnumbered for clarity.

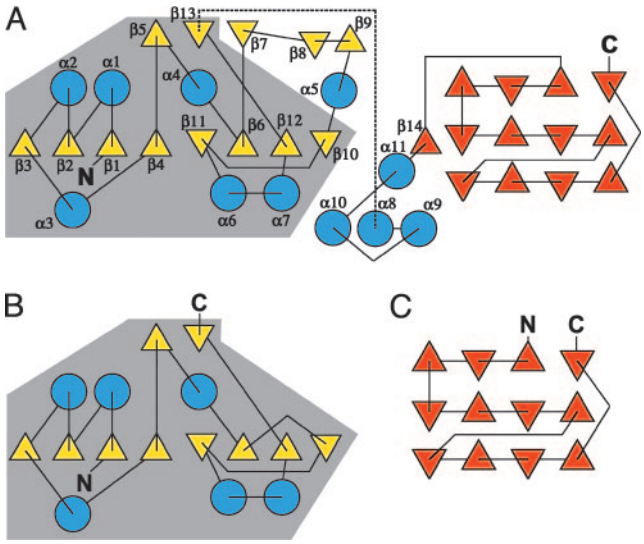

Fig. 2. The mppGaNTase-T1 structure preserves the uridine-binding domain (UBD) and ricin B chain topologies. (*A*) Topology diagram for mppGaNTase-T1. Coloring and numbering for the helices and strands are as in Fig. 1. The dashed line represents residues 347–358, for which electron density was not observed. (*B*) Topology diagram of the UBD of β1,4-*N*-acetylglucosaminyltransferase (PDB ID code 1FO9). The gray shaded areas in *A* and *B* denote the topology comprising the UBD. (*C*) Topology diagram for residues 1–135 of the ricin B chain.

(≥95% identity) salt bridges/hydrogen bonds between secondary structural elements. These bonds include S119-E150 between β-strands 1 and 2, R182 in β-strand 3, and the peptide oxygen of K164 near α-helix 2, D310-R368 between a random coil and α-helix 8, and N365-amide N between the same helix and random coil. An E150Q mutant retains near wild-type activity, whereas a D310N mutant is only 2% as active as wild-type (14). No mutants of R182 or N365 have been described. An additional hydrogen bond is formed between highly conserved D375 in α-helix 9 and R403 in α-helix 11. R403 is less well conserved (84% identity), but variants (either K or Q) maintain hydrogen bonding capability. Mutation of D375 to N or A has no effect on activity (14). Three additional invariant residues (D155, S158, and E322) also participate in hydrogen bonds, but these bonds are between residues near in sequence space. Aspartate 155 pairs with S158 and E322 bonds with R326, which is less well conserved but whose only variant is K. No effect on activity was observed for a D155N mutant, but an E322Q mutant had only 1% of wild-type activity (14). Finally, residues R225-E335 between α-helix 4 and β-sheet 11 and D239-H341 between β-sheets 7 and 13 also form salt bridges. Instead of strict conservation, these residues show isoform-dependent chemical covariance to preserve their interaction. Mutants of R225 or E335 have not been described, but mutation of H341 to A, L, V, K, or R reduces activity only 2-fold (14).

A Mn^{2+} ion is coordinated by direct hydrogen bonds to invariant D209 and H211 of the ppGaNTase catalytic domain "DXH" motif and by invariant H344 (Fig. 3). The ppGaNTase DXH motif corresponds to the active-site DXD motif of other glycosyltransferases, but mutation of H211 to D abolishes enzyme activity (14) as does mutation of H344 to A (33). The Mn^{2+} ion of the retaining galactosyltransferase LgtC (34) is also directly coordinated by three residues (D103, D105, and H244), but that of the retaining bovine α-1,3-galactosyltransferase (PDB ID code 1K4V) is coordinated only by two residues (D225 and D227).

Lectin Domain Structure. The C-terminal domain of mppGaNTase-T1 (residues 427–553) is homologous to the B chain of ricin and has thus been termed the lectin or $(QxW)_3$ domain (11, 14). This domain contains three repeat sequences (α, β, and γ) believed to have arisen from triplication and fusion events of an ≈40 residue "galactose-binding peptide" (35). Figs. 1 and 2 show that the mppGaNTase-T1 lectin domain adopts a β-trefoil fold topologically identical to that of the ricin B chain. Six invariant cysteine residues form three disulfide bonds, and each of the putative sugar-binding sites of the α, β, and γ repeats is located on the same face of the enzyme as the active site Mn^{2+} ion and is accessible for binding. The lectin domain is located on the side of the catalytic domain opposite that of the N terminus (Fig. 1), suggesting that it extends further into the Golgi lumen than the catalytic domain.

The catalytic and lectin domains are connected by a short random coil and β-strand 14, which adds an additional strand to one of the β-sheets of the β-trefoil fold (Fig. 2). The association of the two domains forms a deep cleft ≈40 Å long in the surface of the enzyme perpendicular to the long axis of the enzyme. Previous results have suggested that peptide acceptors bind to ppGaNTases in a primarily extended conformation (36, 37) and that the active site spans roughly nine acceptor residues (38). The length of the cleft is more than sufficient to accommodate nine amino acids in an extended conformation.

Interaction between the catalytic and lectin domains is me-

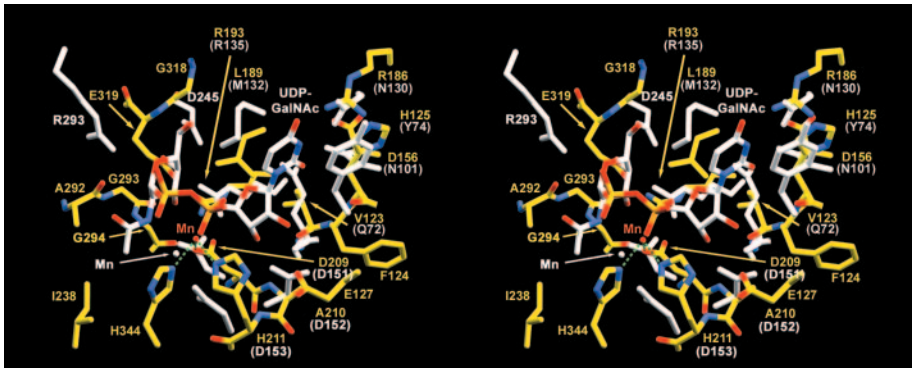

Fig. 3. Identification of potential UDP-GalNAc-binding residues. The figure shows selected residues of superimposed structures of mppGaNTase-T1 and EXTL2 (PDB ID code 1OMZ). Residues 207–213 of mppGaNTase-T1 containing the DXH motif and residues 149–155 of EXTL2 containing the DXD motif were aligned, followed by the Improve Fit option of SWISSPDBVIEWER. Residues of mppGaNTase-T1 within 4 Å of the modeled UDP-GalNAc (atomic colors with white carbons) were identified by using the program CONTACT and are shown in atomic coloring with yellow carbons. Residues of EXTL2 are shown in white with white numbering. The Mn^{2+} ion bound to mppGaNTase-T1 is shown in red, and hydrogen bonds between it and residues D209 (2.42 Å), H211 (2.04 Å), and H344 (2.34 Å) are shown by the green dashed lines.

Table 2. Catalytic/lectin domain interface residues

Residues	Catalytic domain	Lectin domain
Hydrophobic core	W261, F380, I383, P425	Y428, L431, V467, F468, S469
Peripheral	N260, K262, Y268, N379, I384, P386, P420, D421, S422, Q423	H427, G432, M462, G463, G464, T471, R477, T478, D479, D480, C497, H499, V553

diated by 32 residues from noncontiguous segments of both domains and buries ≈645 Å^2 of each domain. The architecture of the interacting surfaces consists of a hydrophobic core surrounded by more hydrophilic residues (Table 2). The hydrophobic core is comprised of residues W261, F380, I383, and P425 of the catalytic domain and Y428, L431, V467, and S469 of the lectin domain. The hydrophilic interactions include hydrogen bonds between W261 N and D479 OD2 (2.9 Å), D421 O and H499 NE2 (2.9 Å), and Q423 OE1 and R477 NH1 (2.6 Å). Only residues W261 and C497 are invariant, suggesting that catalytic–lectin domain interactions are isoform-specific.

UDP-GalNAc Binding. A DALI (39) search of the Protein Data Bank identified the glycosyltransferase SpsA (PDB ID code 1QG8) and the α-*N*-acetylhexosaminyltransferase EXTL2 (PDB ID code 1OMX) as structurally most similar to the mppGaNTase-T1 catalytic domain (DALI *z* scores of 16.1 and 13.4, respectively). The DALI *z* score is a statistical measure of structural similarity and equals the number of standard deviations away from an expected value. A structural alignment of mppGaNTase-T1 with EXTL2 containing a bound UDP-GalNAc was used to model the binding of the sugar nucleotide to ppGaNTase-T1, and residues within 4 Å of the UDP-GalNAc atoms were identified (Fig. 3). The Mn^{2+} ions of the aligned structures are 2.0 Å apart, and only ppGaNTase E319 clashes with the modeled UDP-GalNAc. Several residues of ppGaNTase-T1, including V123, H125, D156, R186, L189, and R193, align well and are chemically similar to those of EXTL2 that bind the uridine moiety of UDP-GalNAc. Of these residues, D156 and R193 are invariant, whereas R186 differs in a single isoform (*Drosophila* pGant 8 to Q), and L189 differs conservatively in two isoforms (human ppGaNTase-T15 to A and *T. gondii* ppGaNTase-T3 to I). Mutation of H125 to Q or F has a minor effect on enzyme activity, but the D156Q enzyme retains only 0.1% of wild-type activity (14). A R193W mutation in a *Drosophila* ppGaNTase homolog (pgant35A) is lethal, although the recombinant enzyme expresses well and retains minimal activity (3, 4). No mutational analyses have been reported for V123, R186, or L189. Invariant E127 (which, when mutated to Q, reduces activity 500-fold) aligns well with EXTL2 R76, which binds to two phosphate oxygens. Although E127 would not be expected to coordinate phosphate oxygens directly, it may do so through intervening water molecules. The two cysteine residues (C212 and C214) proposed to bind to UDP-GalNAc make their closest approach to an oxygen atom on the α-phosphate of UDP-GalNAc. C214 is located 14 Å from the phosphate oxygen and is completely buried inside a hydrophobic pocket, suggesting that it does not interact with UDP-GalNAc. C212 is positioned 7 Å from the phosphate oxygen, suggesting that it also does not make direct contact with UDP-GalNAc, although it may do so via an intervening water molecule.

Forty-two of the 459 residues observed in the crystal structure of mppGaNTase-T1 are invariant across all 38 active ppGaNTase isoforms described to date, and the likely functions of several of these residues are described above. The mppGaNTase-T1 crystal structure provides insight into possible roles of the remaining conserved residues (Table 3). Putative functions were assigned on the basis of solvent accessibility and proximity to the modeled UDP-GalNAc. Residues with no solvent accessibility were considered to be primarily structural, whereas mppGaNTase-T1 residues that aligned in real space with UDP-GalNAc-binding residues of EXTL2 were assigned putative sugar nucleotide-binding functions. Assigning roles for several additional invariant or highly conserved residues was less certain. Three invariant (W328, G332, and W413) and a highly conserved residue (G331) form a patch on the surface of the enzyme opposite the active site, possibly for mediating protein–protein interactions. Three additional residues, S130 (89% identity), R134 (invariant), and S138 (97% identity), line the same face of α-helix 1 near the N terminus of the structure and may bind to residues near and including N95 that also show strong conservation.

Acceptor Peptide Recognition. Mutational studies of the α lectin site of human ppGaNTase-T4 implicate the involvement of the lectin domain in glycopeptide recognition. Specifically, mutation of D459 (equivalent to mppGaNTase-T1 D444 and D22 of the ricin B chain, a residue critical for galactose recognition) to H or the inclusion of high concentrations of GalNAc (but not other monosaccharides) abolishes the ability of ppGaNTase-T4 to act on glycopeptides (15). In contrast, transfer of GalNAc to a peptide substrate by a D444H mutant of ppGaNTase-T1 was only moderately reduced (2-fold) compared to wild type (14). An NMR study of the glycosylation of a mucin 1 (MUC1) tandem repeat peptide (PAPGSTAPPAHGVTSAPDTR) by ppGaNTase-T1 revealed that the enzyme sequentially adds GalNAc to T14 and then eight residues away to T6 of the newly generated glycopeptide (40). Similar findings were obtained by using a different peptide (GTTPSPVPTTSTTSAP) from MUC5AC. Threonines 13 and 3 separated by 10 residues are the first to be modified by ppGaNTase-T1, although the order of addition has not been established (12, 41).

To determine whether the structure presented in this paper is consistent with lectin binding of glycopeptide after initial GalNAc transfer, the MUC1 peptide mentioned above containing a covalently bound GalNAc on T14 was modeled into the ppGaNTase-T1 structure (Fig. 4). The lectin domain of a xylanase (*Streptomyces olivaceoviridis*, PDB ID 1XYF), highly similar

Table 3. Function/putative function of conserved residues

Residues*	Defined or proposed function†
D209, H211, H344	Mn^{2+} coordination
V123, H125, **E127, D156,** R186, L189, **R193, E319**	UDP-GalNAc Binding
G196, W218, I241, P289, S340, W373 (F259–F325–**F411**)	Structural
W328, G331, **G332, W413**	Protein–protein interactions
S130, **R134,** S138	N-terminal interactions
G188, P221, P236, G293, G307, G317	Unknown

*Residues in bold are invariant in the 38 active ppGaNTase isoforms described to date. Residues in plain text are less conserved (>89% identity), with the exception of V123 (45% identity) and H125 (66% identity), and are included because of their association with invariant residues or their proximity to the modeled UDP-GalNAc. Because of mutual interaction, residues F259, F325, and F411 were grouped together.

†Residues assigned a primarily structural role were defined as being completely solvent inaccessible, with the exception of W218, which had 1 Å^2 of accessible surface area.

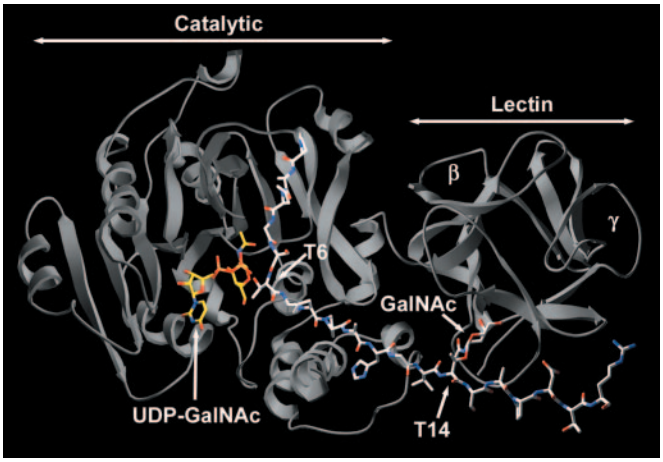

Fig. 4. Modeled binding of a MUC1 glycopeptide to ppGaNTase-T1 is consistent with its *in vitro* pattern of glycosylation. Structural alignment of the ppGaNTase-T1 and *S. olivaceoviridis* β-xylanase (PDB ID code 1XYF) lectin domains was used to model GalNAc covalently attached to Thr-14 of a MUC1 peptide (PAPGSTAPPAHGVTSAPDTR, white carbons) into the α site of the ppGaNTase-T1 lectin domain. This docking allowed Thr-6 of the peptide to be positioned within 2.5 Å of the anomeric carbon of the modeled UDP-GalNAc (yellow carbons). The remainder of the peptide was positioned to avoid overlap with the enzyme. All peptide ϕ and ψ angles are in allowed regions of the Ramachandran plot.

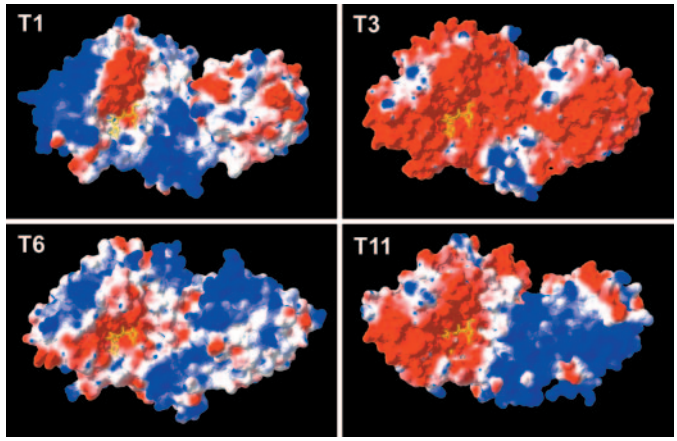

Fig. 5. Electrostatic surface potentials of homology-modeled murine ppGaNTases suggest isoform-specific substrates. The orientation of the transferases is the same as in Fig. 1. Residues of individual isoforms corresponding to residues 347–358 of mppGaNTase-T1 for which no electron density was observed were removed from the PDB files before surface potential calculations to facilitate comparison between isoforms. Potentials are colored by using a gradient varying from red (negative) through white (neutral) to blue (positive). The modeled UDP-GalNAc is shown in yellow. Activity for mppGaNTase-T6 has not been demonstrated but its sequence is 88% identical to active human ppGaNTase-T6.

structurally to that of ppGaNTase-T1 (DALI z score = 18.0) and with galactose molecules bound to the α and γ lectin sites, was used to model the GalNAc into the α site of the ppGaNTase-T1 lectin domain. The α site was chosen because mutations to this site diminish enzyme activity more than mutations to the β and γ sites (42). The hydroxyl group of T6 could be positioned within 2.5 Å of the anomeric carbon of the modeled UDP-GalNAc while maintaining proper geometric constrains on the peptide. Allowing for uncertainties in the modeled positions of donor and acceptor due to deviations in the superimposed structures [3.2 Å rms deviation (rmsd) for the *N*-acetylglucosaminyltransferase and 1.6 Å rmsd for the xylanase lectin domain), it is reasonable that the current crystal structure is consistent with the ppGaNTase-T1 experimental glycosylation pattern of the MUC1 peptide. Because the first two ppGaNTase-T1 acceptor sites of the MUC5AC peptide are spaced two residues farther apart than in the MUC1 peptide, the structure may also explain the glycosylation pattern of the MUC5AC peptide, assuming it adopts a somewhat less extended conformation than the MUC1 peptide.

ppGaNTase Isoform Surface Potential Variability. Literature reports demonstrate that certain parent-protein-derived acceptor peptides are glycosylated by specific ppGaNTases. For example, ppGaNTase-T3 was the only isoform tested that efficiently glycosylated a peptide from the HIV gp-120 protein both *in vitro* (43) and *in vivo* (44), whereas a P-selectin glycoprotein ligand-1 peptide was ppGaNTase-T4-specific (45). Of six transferases assayed, ppGaNTase-T2 has been reported to be responsible for glycosylating the IgA hinge region (46). The molecular basis for this selectivity remains unknown, and important caveats of these studies are that only short peptides from much larger proteins were assayed and a limited set of isoforms was tested.

The strong conservation of multiple residues mediating interactions between ppGaNTase-T1 secondary structural elements, including the two disulfide bonds in the catalytic domain and three in the lectin domain, indicate that the structure of other ppGaNTase isoforms will be very similar to that of ppGaNTase-T1. We therefore created homology models of three additional murine ppGaNTases and examined their electrostatic surface potential to gain insight into the molecular basis of isoform substrate specificities (Fig. 5). Only murine isoforms were used to minimize differences due to species variation. The surface potential of each isoform is similar near the modeled UDP-GalNAc, reflecting the conserved molecular basis of sugar nucleotide recognition. However, the potential becomes more divergent and unique to each isoform at sites away from the sugar nucleotide pocket. Thus, depending on the area of interaction between enzyme and acceptor, this finding predicts differential affinities for a given acceptor and provides support for isoform-dependent acceptor substrate specificity. Two of the most closely related isoforms, T3 and T6, show striking differences in their surface potentials.

Discussion

We have determined the x-ray crystal structure of a ppGaNTase, murine ppGaNTase-T1. The structure shows that the catalytic domain belongs to the GT-A superfamily fold, and that the lectin domain folds as the B chain of ricin, as was previously predicted (11, 29). The two domains form a deep cleft in which a Mn^{2+} ion is coordinated by D209 and H211 of the DXH motif and by H344. The structure provides a molecular basis for understanding the results of numerous mutational analyses, and the orientation and accessibility of the putative binding sites of the lectin domain suggest a mechanism for accommodating multiple orientations of glycopeptide substrates. The structure also revealed a patch on the surface of the enzyme comprised of the invariant or highly conserved W328, G331, G332, and W413. Although these residues may be primarily structural and/or may participate in the catalytic mechanism, another possibility is that they may mediate protein–protein interactions. Several glycosyltransferases and glycan-modifying enzymes within the same biosynthetic pathway associate with each other (see ref. 47 for a recent review).

An important question about this family of retaining glycosyltransferases unresolved by the structure concerns the mechanism of catalysis. It was proposed that the reaction proceeds by means of a double displacement mechanism with a glycosyl-enzyme intermediate by analogy with retaining glycosidases (48). Failed attempts at finding this intermediate in other retaining glycosyltransferases led to the hypothesis that catalysis might occur by means of a S_Ni-like mechanism (34). However, the recent isolation of a glycosyl-

enzyme intermediate for the retaining galactosyltransferase LgtC has revived the double displacement theory (49). In this study, the proposed catalytic nucleophile (Q189) was mutated to a more nucleophilic Glu in an attempt to accumulate the elusive intermediate. Surprisingly, the Asp adjacent to Q189 was covalently modified with Gal, implicating D190 as the catalytic nucleophile. Structure-based alignment of ppGaNTase-T1 with LgtC shows that ppGaNTase-T1 residues E319 and N320 are positioned near the space occupied by LgtC Q189 and D190. Mutation of E319 to Q abolishes enzyme activity with little effect on expression, whereas an N320A mutant expresses well and retains ≈50% activity (14). A Glu residue (E322) is also nearby, and mutation to Q reduces activity to ≈1% of wild-type (14). The lack of electron density for UDP-GalNAc in the ppGaNTase-T1 structure precludes the assignment of the catalytic nucleophile, but the mutational analyses would seem to implicate invariant E319. A more thorough analysis of E319 mutants seems warranted in light of the LgtC results.

The model presented in Fig. 4 is consistent with the pattern of glycosylation of MUC1 and possibly MUC5AC peptides catalyzed by ppGaNTase-T1, but other important questions about ppGaNTase-T1 acceptor substrate recognition remain unanswered. For example, how does the enzyme bind unglycosylated peptides? Additionally, ppGaNTase-T1 (and other isoforms) must use alternate means of glycopeptide binding in addition to that postulated in Fig. 4 as evidenced by ppGaNTase-T1 transfer of GalNAc to Thr residues of a MUC2 peptide separated by only a single amino acid (50). When presented with the EA2 peptide PTTDSTTPAPTTK glycosylated at Thr-7, the glycopeptide transferases T7 and T10 add GalNac to Thr-6, immediately N-terminal to the extant GalNAc (12). Clearly, docking the MUC2 or EA2 GalNAc-glycosylated peptides into any of the lectin-binding sites would position adjacent Thr or Thr one residue removed from the glycosylated Thr, too far from the active site for the next GalNAc transfer unless the enzyme undergoes a major conformational change. Alternatively, the transferases may form dimers with a lectin domain from a separate molecule positioning the glycopeptide for transfer. However, no studies of ppGaNTase oligomerization have been reported. A third possibility is that there is a cryptic lectin site in the catalytic domain near the active site. A recent study highlights the complexity of ppGaNTase substrate recognition (51). These questions of substrate recognition must be addressed through biochemical and structural experiments.

We thank the National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) crystallographic community for helpful discussions and Dr. Susan Buchanan (NIDDK) for critical reading of the manuscript. We also thank Emmanuel Gonzalez, Polina Shcherbatov, Shino Shimoji, and Hua Mao for technical assistance. Research was supported by the National Institutes of Health, NIDDK intramural program.

1. Campbell, J. A., Davies, G. J., Bulone, V. V. & Henrissat, B. (1997) *Biochem. J.* **326,** 929–939, and correction (1998) **329,** 719.
2. Ten Hagen, K. G., Fritz, T. A. & Tabak, L. A. (2003) *Glycobiology* **13,** 1–16.
3. Ten Hagen, K. G. & Tran, D. T. (2002) *J. Biol. Chem.* **277,** 22616–22622.
4. Schwientek, T., Bennett, E. P., Flores, C., Thacker, J., Hollmann, M., Reis, C. A., Behrens, J., Mandel, U., Keck, B., Schafer, M. A., *et al.* (2002) *J. Biol. Chem.* **277,** 22623–22638.
5. Topaz, O., Shurman, D. L., Bergman, R., Indelman, M., Ratajczak, P., Mizrachi, M., Khamaysi, Z., Behar, D., Petronius, D., Friedman, V., *et al.* (2004) *Nat. Genet.* **36,** 579–581.
6. Xia, L., Ju, T., Westmuckett, A., An, G., Ivanciu, L., McDaniel, J. M., Lupu, F., Cummings, R. D. & McEver, R. P. (2004) *J. Cell Biol.* **164,** 451–459.
7. Hennet, T., Hagen, F. K., Tabak, L. A. & Marth, J. D. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 12070–12074.
8. Zhang, Y., Iwasaki, H., Wang, H., Kudo, T., Kalka, T. B., Hennet, T., Kubota, T., Cheng, L., Inaba, N., Gotoh, M., *et al.*. (2003) *J. Biol. Chem.* **278,** 573–584.
9. Rottger, S., White, J., Wandall, H. H., Olivo, J. C., Stark, A., Bennett, E. P., Whitehouse, C., Berger, E. G., Clausen, H. & Nilsson, T. (1998) *J. Cell Sci.* **111,** 45–60.
10. Roth, J., Wang, Y., Eckhardt, A. E. & Hill, R. L. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 8935–8939.
11. Hazes, B. (1996) *Protein Sci.* **5,** 1490–1501.
12. Ten Hagen, K. G., Bedi, G. S., Tetaert, D., Kingsley, P. D., Hagen, F. K., Balys, M. M., Beres, T. M., Degand, P. & Tabak, L. A. (2001) *J. Biol. Chem.* **276,** 17395–17404.
13. Ten Hagen, K. G., Tetaert, D., Hagen, F. K., Richet, C., Beres, T. M., Gagnon, J., Balys, M. M., VanWuyckhuyse, B., Bedi, G. S., Degand, P., *et al.* (1999) *J. Biol. Chem.* **274,** 27867–27874.
14. Hagen, F. K., Hazes, B., Raffo, R., deSa, D. & Tabak, L. A. (1999) *J. Biol. Chem.* **274,** 6797–6803.
15. Hassan, H., Reis, C. A., Bennett, E. P., Mirgorodskaya, E., Roepstorff, P., Hollingsworth, M. A., Burchell, J., Taylor-Papadimitriou, J. & Clausen, H. (2000) *J. Biol. Chem.* **275,** 38197–38205.
16. Otwinowski, Z. & Minor, W. (1997) *Methods Enzymol.* **276,** 307–326.
17. Terwilliger, T. C. & Berendzen, J. (1999) *Acta Crystallogr. D* **55,** 849–861.
18. Terwilliger, T. C. (2000) *Acta Crystallogr. D* **56,** 965–972.
19. Terwilliger, T. C. (2003) *Acta Crystallogr. D* **59,** 38–44.
20. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997) *Acta Crystallogr. D* **D53,** 240–255.
21. McRee, D. E. (1999) *J. Struct. Biol.* **125,** 156–165.
22. Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., *et al.* (1998) *Acta Crystallogr. D* **54,** 905–921.
23. Collaborative Computation Project, N (1994) *Acta Crystallogr. D* **50,** 760–763.
24. Michalopoulos, I., Torrance, G. M., Gilbert, D. R. & Westhead, D. R. (2004) *Nucleic Acids Res.* **32,** D251–D254.
25. Guex, N. & Peitsch, M. C. (1997) *Electrophoresis* **18,** 2714–2723.
26. Lee, C. & Irizarry, K. (2001) *IBM Systems J.* **40,** 592–603.
27. Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G. & Thompson, J. D. (2003) *Nucleic Acids Res.* **31,** 3497–3500.
28. Galtier, N., Gouy, M. & Gautier, C. (1996) *Comput. Appl. Biosci.* **12,** 543–548.
29. Breton, C., Heissigerova, H., Jeanneau, C., Moravcova, J. & Imberty, A. (2002) *Biochem. Soc. Symp.* 23–32.
30. Unligil, U. M., Zhou, S., Yuwaraj, S., Sarkar, M., Schachter, H. & Rini, J. M. (2000) *EMBO J.* **19,** 5269–5280.
31. Gastinel, L. N., Bignon, C., Misra, A. K., Hindsgaul, O., Shaper, J. H. & Joziasse, D. H. (2001) *EMBO J.* **20,** 638–649.
32. Tenno, M., Toba, S., Kezdy, F. J., Elhammer, A. P. & Kurosaka, A. (2002) *Eur. J. Biochem.* **269,** 4308–4316.
33. Wragg, S., Hagen, F. K. & Tabak, L. A. (1997) *Biochem. J.* **328,** 193–197.
34. Persson, K., Ly, H. D., Dieckelmann, M., Wakarchuk, W. W., Withers, S. G. & Strynadka, N. C. (2001) *Nat. Struct. Biol.* **8,** 166–175.
35. Rutenber, E., Ready, M. & Robertus, J. D. (1987) *Nature* **326,** 624–626.
36. Kirnarsky, L., Nomoto, M., Ikematsu, Y., Hassan, H., Bennett, E. P., Cerny, R. L., Clausen, H., Hollingsworth, M. A. & Sherman, S. (1998) *Biochemistry* **37,** 12811–12817.
37. Gerken, T. A., Owens, C. L. & Pasumarthy, M. (1997) *J. Biol. Chem.* **272,** 9709–9719.
38. Sugahara, T., Pixley, M. R., Fares, F. & Boime, I. (1996) *J. Biol. Chem.* **271,** 20797–20804.
39. Holm, L. & Sander, C. (1993) *J. Mol. Biol.* **233,** 123–138.
40. Brokx, R. D., Revers, L., Zhang, Q., Yang, S., Mal, T. K., Ikura, M. & Gariepy, J. (2003) *Biochemistry* **42,** 13817–13825.
41. Tetaert, D., Ten Hagen, K. G., Richet, C., Boersma, A., Gagnon, J. & Degand, P. (2001) *Biochem. J.* **357,** 313–320.
42. Tenno, M., Kezdy, F. J., Elhammer, A. P. & Kurosaka, A. (2002) *Biochem. Biophys. Res. Commun.* **298,** 755–759.
43. Wandall, H. H., Hassan, H., Mirgorodskaya, E., Kristensen, A. K., Roepstorff, P., Bennett, E. P., Nielsen, P. A., Hollingsworth, M. A., Burchell, J., Taylor-Papadimitriou, J., *et al.* (1997) *J. Biol. Chem.* **272,** 23503–23514.
44. Nehrke, K., Hagen, F. K. & Tabak, L. A. (1998) *Glycobiology* **8,** 367–371.
45. Bennett, E. P., Hassan, H., Mandel, U., Mirgorodskaya, E., Roepstorff, P., Burchell, J., Taylor-Papadimitriou, J., Hollingsworth, M. A., Merkx, G., van Kessel, A. G., *et al.* (1998) *J. Biol. Chem.* **273,** 30472–30481.
46. Iwasaki, H., Zhang, Y., Tachibana, K., Gotoh, M., Kikuchi, N., Kwon, Y. D., Togayachi, A., Kudo, T., Kubota, T. & Narimatsu, H. (2003) *J. Biol. Chem.* **278,** 5613–5621.
47. Young, W. W., Jr. (2004) *J. Membr. Biol.* **198,** 1–13.
48. Sinnott, M. L. (1990) *Chem. Rev.* **90,** 1171–1202.
49. Lairson, L. L., Chiu, C. P., Ly, H. D., He, S., Wakarchuk, W. W., Strynadka, N. C. & Withers, S. G. (2004) *J. Biol. Chem.* **279,** 28339–28344.
50. Kato, K., Takeuchi, H., Miyahara, N., Kanoh, A., Hassan, H., Clausen, H. & Irimura, T. (2001) *Biochem. Biophys. Res. Commun.* **287,** 110–115.
51. Pratt, M. R., Hang, H. C., Ten Hagen, K. G., Rarick, J., Gerken, T. A., Tabak, L. A. & Bertozzi, C. R. (2004) *Chem. Biol.* **11,** 1009–1016.