Published in final edited form as:

Biochemistry. 2015 December 22; 54(50): 7326–7334. doi:10.1021/acs.biochem.5b01086.

Conservation and Covariance in Small Bacterial Phosphoglycosyltransferases Identifies the Functional Catalytic Core

Vinita Lukose¹, Lingqi Luo^{2,3}, Dima Kozakov³, Sandor Vajda³, Karen N. Allen^{*,2}, and Barbara Imperiali^{*,1}

¹Departments of Chemistry and Biology, Massachusetts Institute of Technology, Cambridge, MA 02139

²Department of Chemistry, Boston University, Boston, MA 02215

³Department of Department of Biomedical Engineering, Boston University, Boston, MA 02215

Abstract

Phosphoglycosyltransferases (PGTs) catalyze the transfer of a C1'-phosphosugar from a soluble sugar nucleotide diphosphate to a polyprenol-phosphate. These enzymes act at the membrane interface, forming the first membrane-associated intermediates in the biosynthesis of cell-surface glycans and glycoconjugates including glycoproteins, glycolipids and the peptidoglycan in bacteria. PGTs vary greatly in both in their membrane topologies and their substrate preferences. PGTs, such as MraY and WecA, are polytopic, while other families of uniquely prokaryotic enzymes have only a single predicted transmembrane helix. PglC, a PGT involved in the biosynthesis of *N*-linked glycoproteins in the enteropathogen *C. jejuni*, is representative of one of the structurally most simple members of the diverse family of small bacterial PGT enzymes. Herein, we apply bioinformatics and covariance-weighted distance constraints in geometry- and homology-based model building, together with mutational analysis to investigate monotopic PGTs. The pool of 15,000 sequences that are analyzed include the PglC-like enzymes, as well as sequences from two other related PGTs that contain a "PglC-like" domain embedded in their larger structures (namely, the bifunctional PglB family, typified by PglB from *N. gonorrhoeae* and WbaP-like enzymes, typified by WbaP from *S. enterica*). Including these two sub-families of PGTs in the analysis highlights key residues conserved across all three families of small bacterial PGTs. Mutagenesis analysis of these conserved residues provides further information on the essentiality of many of these residues in catalysis. Construction of a structural model of the cytosolic globular domain utilizing three-dimensional distance constraints, provided by

*To whom correspondence should be addressed: Barbara Imperiali, Department of Biology, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139. Phone: 617-253-1838; imper@mit.edu. Karen N. Allen, Department of Chemistry, 590 Commonwealth Avenue, Boston, MA 02215. Phone: 617-358-5544; drkallen@bu.edu.

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: The supporting document comprises four supporting tables (Tables S1–S4) and six supplementary figures (Figures S1–S6).

Notes

The authors declare no competing financial interest.

conservation covariance analysis, provides additional insight into the catalytic core of these families of small bacterial PGT enzymes.

Graphical abstract

Keywords

phosphoglycosyltransferase; glycoconjugate biosynthesis; membrane protein; conservation covariance; bioinformatics; modeling

Phosphoglycosyltransferases (PGTs) are a family of enzymes that catalyze the transfer of a sugar 1-phosphate from a nucleotide-activated donor to a polyprenol-phosphate acceptor substrate. This family encompasses the previously described polyisoprenyl-phosphate hexose-1-phosphate transferase (PHPT) and polyisoprenyl-phosphate *N*-acetylaminosugar-1-phosphate transferase (PNPT) family enzymes.^{1, 2} The products of PGT reactions are elaborated into complex polyprenol-diphosphate-linked glycans, which serve as glycosyl donors in pathways such as those in the biosynthesis of glycoproteins and glycolipids. Although most identified PGTs are bacterial, there are important eukaryotic members such as Alg7, which initiates the *N*-linked protein glycosylation pathway in eukaryotes from yeast to man, by catalyzing the biosynthesis of dolichol-PP-GlcNAc.³

PGTs catalyze interfacial reactions between a soluble sugar substrate and a membrane-bound polyprenol phosphate to form a membrane-bound product (Scheme 1). PGTs are integral membrane proteins and can be organized into subfamilies based on their membrane topologies. The best known PGTs include members of the WecA and MraY subfamilies.² These are polytopic membrane proteins with ten or more transmembrane helices (TMHs) (Figure 1A and B). To date, the only PGT with a known three-dimensional structure is MraY from *Aquifex aeolicus*.⁴ Based on conservation analysis together with the crystal structure, the active site of MraY was predicted to include three conserved aspartate residues and one conserved histidine residue. Mutagenesis of each of the three aspartate residues to asparagine eliminates translocase activity in MraY.⁵

In conjunction with studies on prokaryotic protein glycosylation, a distinct structural family of prokaryotic PGTs has been identified, which is typified by the *Campylobacter jejuni* PGT known as PglC. PglC catalyzes the reaction between UDP-diNAcBac and undecaprenol-phosphate, forming undecaprenol-P-P-diNAcBac with the release of UMP as a by-product.⁶ Homologs of PglC are predicted to be Type I membrane proteins with monotopic structures characterized by a short N-terminal periplasmic domain, a single predicted TMH, and a soluble globular C-terminal domain located in the cytosol (Figure 1C).⁶ In contrast to the polytopic PGTs, including MraY and WecA, the PglC subfamily of PGTs are small (approximately 200 amino acids in length) and, although they perform similar chemical transformations, they do not share significant sequence identity or possess any of the identified consensus motifs found in the larger PGTs.

Functional and bioinformatics analysis in our laboratories⁷ as well as the Valvano, group^{8, 9} reveals two other related families, which include the primary sequence of the globular cytosolic domain of PglC-type PGTs embedded into a more complex framework (Figure 1 C, D and E). One of these families is typified by the bifunctional PglB(*Ng*) from *N. gonorrhoeae*, which features an N-terminal PGT domain,⁷ with high sequence identity (53% identity) to the *C. jejuni* PglC, and a C-terminal amino-sugar acetyltransferase domain that catalyzes acetylation of the UDP-4-amino sugar precursor to UDP-diNAcBac.⁷ The second related family is exemplified by WbaP from *Salmonella enterica*. WbaP catalyzes the transfer of galactose 1-phosphate from UDP-Gal to undecaprenol-phosphate.⁸⁻¹⁰ Enzymes in this family are polytopic, with five to seven predicted TMHs;¹¹ however, several of the N-terminal helices are not essential for PGT activity.¹⁰⁻¹² The functional C-terminal domain comprises a predicted membrane-bound hydrophobic sequence and a cytosolic globular domain, which shows 34.7% sequence identity to the *C. jejuni* PglC. A recent study with WcaJ, an *E. coli* enzyme that is a member of the WbaP family, used PhoA/LacZ fusions and cysteine labeling to investigate the cellular localization of loops and the globular domain.¹¹ The results of this study suggest that the predicted TMH adjacent to the C-terminal globular domain is actually a re-entrant helix, forming a hairpin bend in the membrane, rather than spanning the membrane (Figure 1E). At the current time, similar experimental studies have not been performed with PglC-like or bifunctional PglB-like PGTs and therefore the topological models illustrated in Figure 1C and D are based solely on TMH predictions.

Herein, sequence analyses were used to gather the related cytosolic globular domains in all three small bacterial PGT families: PglC-like, bifunctional PglB, and WbaP-like. The resulting sequence alignments allowed the identification of highly conserved residues and construction of a structural model utilizing three-dimensional distance constraints was provided by covariance analysis. The conserved amino acids and the model together identified residues, which were validated by site-directed mutagenesis as critical for catalytic activity. In addition to the TMH, the predicted secondary structural features are consistent with the presence of a second highly conserved helix. The application of a helical wheel plot to this helix suggests a pattern of hydrophobicity consistent with placement at the protein-membrane interface or within a protein-docking interface. Together, these studies yield insight into the relationship between distant families that catalyze the phosphoglycosyltransferase reaction, identify and confirm catalytically essential amino acids, and begin to elucidate the structure/function connections in small bacterial PGTs.

Experimental Procedures

Cloning and site-directed mutagenesis of SUMO-PglC

Wild-type PglC was cloned into the pET-SUMO vector using forward primer 5'-CGCCGGTCTCCAGGTATGTATGAAAA-3' and reverse primer 5'-ATCGCTCGAGTTATGCCGTCCCGGTCTT-3' (BsaI and XhoI restriction site underlined in bold, respectively). Mutations were introduced into this sequence with the primers listed in Table S1, using the Quikchange protocol.

Overexpression of Wild-Type and SUMO-PglC Variants

The pET-SUMO-PglC plasmid was transformed into BL21-RIL cells (Agilent) for overexpression, using kanamycin and chloramphenicol for selection. Overexpression was performed using the Studier method.¹³ In this method, 1 mL of an overnight cell culture was added to expression media containing 30 µg/mL kanamycin and 30 µg/mL chloramphenicol in 1 L of autoinduction media (0.1% (w/v) tryptone, 0.05% (w/v) yeast extract, 2 mM MgSO₄, 0.05% (v/v) glycerol, 0.005% (w/v) glucose, 0.02% (w/v) α-lactose, 2.5 mM Na₂HPO₄, 2.5 mM KH₂PO₄, 5 mM NH₄Cl, 0.5 mM Na₂SO₄). Cells were allowed to grow with shaking for 3 h at 37 °C. After 3 h, the temperature was decreased to 16 °C, and the cells were incubated for 16 hours. Cells were harvested by centrifuging at 9000 × g, and cells were stored at -80 °C. Cell pellets were thawed in 10% of the original culture volume in 50 mM Tris pH 8.0, 150 mM NaCl, 40 µL protease inhibitor cocktail (Calbiochem). The cells were lysed by two rounds of sonication for 90 seconds each, at an amplitude of 50% with one-second on/off pulses. The cells were incubated on ice for ten minutes between rounds of sonication. Cellular debris was removed by centrifugation at 9000 × g for 45 minutes. The resulting supernatant was transferred to a clean centrifuge tube and subjected to centrifugation at 142,000 × g for 65 minutes to pellet the cell envelope fraction (CEF).

When the CEF was used for activity assays, it was homogenized into 1% of the original culture volume in 50 mM HEPES pH 7.5, 100 mM NaCl and stored at -80 °C. When protein was to be purified from the CEF, it was isolated and homogenized into 5% of the original culture volume in 50 mM HEPES pH 7.5, 100 mM NaCl, 1% n-dodecyl β-D-maltoside (DDM), using a glass homogenizer. An aliquot of 20 µL of protease inhibitor cocktail (Calbiochem) was added to prevent proteolysis. This sample was incubated at 4 °C with gentle rocking for 16 hours, after which it was centrifuged (145,000 × g) to remove insoluble material. The resulting supernatant was incubated with 1 mL Ni-NTA resin for 1–2 hours. The resin was washed with 30 ml buffer A (50 mM HEPES pH 7.5, 100 mM NaCl, 0.03% DDM, 20 mM imidazole), followed by a wash with 30 ml buffer B (50 mM HEPES pH 7.5, 100 mM NaCl, 0.03% DDM, 45 mM imidazole). PglC was eluted in 4 × 1 mL fractions of elution buffer (50 mM HEPES pH 7.5, 100 mM NaCl, 0.03% DDM, 300 mM imidazole). Gel filtration analysis was performed using a Superdex S200 10/300 column (GE Healthcare) equilibrated with 50 mM HEPES pH 7.5, 100 mM NaCl, 0.03% DDM (Figure S1).

Activity Assays for Wild-Type PglC and PglC Variants

Wild-type PglC and variants were assayed using a radioactive extraction-based assay. Assays contained 16 μ M Und-P, 2.75% DMSO, 0.2% Triton X-100, 30 mM Tris pH 8, 7.5 μ M [3 H]-UDP-diNAcBac (5.4 mCi/mmol), and 3 nM wild-type PglC and 3, 30, or 300 nM of the PglC variants in a final volume of 120 μ L. After initiation of the reaction with [3 H]-UDP-diNAcBac, aliquots (20 μ L) were taken at three, six and ten minute time points and quenched in 1 mL CHCl₃:MeOH (2:1 v/v). The organic layer was washed three times with 400 μ L PSUP (Pure Solvent Upper Phase, composed of 15 mL CHCl₃, 240 mL MeOH, 1.83 g KCl in 235 mL H₂O). The resulting aqueous layers were combined with 5 mL EcoLite (MP Biomedicals) liquid scintillation cocktail. Organic layers were combined with 5 mL OptiFluor (PerkinElmer). Both layers were analyzed on a Beckman Coulter LS6500 scintillation counting system. When CEFs were used instead of pure protein, the CEFs were added at a final concentration of 8% (v/v).

Construction of Sequence Similarity Network

Three programs (TMHMM,¹⁴ MEMSAT¹⁵ and TOPPRED¹⁶) were used to identify potential transmembrane helices (TMHs) in full length *C. jejuni* PglC (see Supporting Methods). A sequence similarity network was generated for each of the three families of PGTs, as well as a final alignment including all three families (Supporting Methods and Figure S2A for details). For each family, the soluble C-terminal phosphosugar transferase domain sequence (without the predicted TMH) was extracted and used as the query sequence in BLAST against the UniprotKB database. A set of 15,000 potential homologous sequences were retrieved and used as the input to construct a sequence similarity network. All sequences sharing greater than 65% identity (resulting in 4,554 sequences for all 3 families) were clustered using the program CD-HIT¹⁷ and represented by a single representative sequence. A sequence similarity network was built from the set of pairwise relationships denoted by e-value for each of the three PGT families and visualized using Cytoscape (Figure S3). Under the strict e-value cutoff of 1×10^{-50} , the number of nodes in both gross and direct neighboring child network are summarized in Table S2.

Based on the similarity network, a stringent set of homologous sequences was parsed out by selecting the original target and the neighboring nodes with direct links to the target. The selected sequences were used to generate a high-quality multiple sequence alignment (MSA) and to develop a final MSA including all three families. The MSAs were further edited to remove gaps, and subjected to sequence logo construction through WebLogo 3.¹⁸

Evolutionary Coupling Analysis and Structure Modeling

The web server EVFold, developed for evolutionary coupling analysis-based *de-novo* structural modeling¹⁹ was used to predict the structures of the *C. jejuni* PglC C-terminal soluble domain (see Figure S2B for the outline of the method). The server extracts a large number of homologous sequences in the corresponding protein family and uses a global maximum entropy model based algorithm, EVCoupling¹⁹ to predict a set of ‘direct’ residue couplings that are likely to represent spatial proximity. A subset of high-scoring residue pairs is converted to “evolutionary inferred contacts” (EICs) and utilized to predict three-dimensional structures. Since the number of distance constraints is a critical parameter for

accurate model generation, for each system, only those models with an adequate number of EICs (at least half of the total number of residues) are retained. The reliability of the approach was evaluated by application to two proteins with known structures and similar in size to *PglC*, 6-hydroxymethyl-7,8-dihydropterin pyrophosphokinase and human macrophage elastase (see Supporting Methods). In order to assess the reproducibility of the structure prediction, the homology modeling server I-TASSER²⁰ was also used to model the soluble domain of *C. jejuni* PglC with the same set of top 80 ranked EICs from EVFOLD as restraints.

Results and Discussion

Identification of conserved and functional residues

The functional and mechanistic similarity of the reactions catalyzed by the PglC-type PGTs accompanied by bioinformatics analysis^{7, 8} suggested that the PglC-like, bifunctional PglB, and WbaP-like subfamilies could be grouped into one family of PGTs and that conservation analysis could be utilized to determine critical residues. To this end, an alignment of 984 sequences (those sequences in the direct neighboring child network (see Table S2)), using one query sequence from each of the sub-families of the small bacterial PGTs (Uniprot IDs E8SR45, U6QR91, O86156), was utilized to generate a multiple sequence alignment (MSA), revealing that ~25% of the sequence is conserved. After editing to remove gaps, this high-quality MSA of the globular C-terminal domain from the PglC-like and the WbaP-like PGT families, and the homologous domain in the bifunctional PglB, was utilized to generate a sequence logo to display the patterns in the aligned sequences. The sequence logo of the small bacterial PGT families reveals highly conserved residues that would not have been apparent from alignments of any of the families individually (Figure 2). The conserved residues were compared to those previously identified for the WecA and MraY families of PGTs via bioinformatics and mutagenesis as being implicated in substrate binding and catalysis.^{2, 4, 5, 21–24} Notably, the WecA and MraY family motifs were not observed in the MSA of the small bacterial PGTs, confirming that these PGTs form two distinct structural classes.

Identification and delineation of the margins of any transmembrane segments and soluble domain(s) the *C. jejuni* PglC was performed via prediction and topology algorithms (see Methods). For PglC, a single N-terminal TMH was predicted; thus residues 34–200 comprise a soluble domain. Guided by the margins of the predicted TMH and the MSA, conserved residues were selected for analysis by site-directed mutagenesis in the full length PglC from *C. jejuni*. The selections for site-directed variants included highly conserved aspartate and glutamate residues (D92, E93, D156, D168) as these residues typically coordinate metal ion cofactors and have additionally been identified as potential active site residues in the WecA and MraY families of PGTs.^{5, 25} Also targeted were conserved arginine residues (R87, R111), which may be involved in interactions with the phosphate groups of substrates and products and a conserved methionine residue (M62). In addition, we identified a conserved proline in the TMH (*C. jejuni* PglC P24); the presence of a Pro in the center region of the membrane-bound sequence that is N-terminal to the globular domain has also been noted previously in the WbaP-like family PGTs.¹¹ In our study the extensive

alignment pool provided insight into the identity of other residues that might feature in this site of the protein and that might be investigated to assess the significance of the secondary amino acid, proline, in enzyme function. Finally, two non-conserved glutamate residues (E65, E116) were mutated to glutamine, as controls for mutagenesis analysis.

Wild-type and variants of the *C. jejuni* PglC were expressed as SUMO fusions to aid in maintaining protein solubility during expression and purification. The SUMO tag greatly improved the solubility of PglC relative to the wild-type construct, and the SUMO-PglC fusion protein exhibited catalytic activity that did not significantly differ from that of wild-type enzyme. All proteins were purified to homogeneity and analyzed by SDS-PAGE for purity, and were analyzed by size exclusion chromatography (SEC) to assess their levels of aggregation (Figure S1). It should be noted that although some of the SEC traces showed aggregation, in all cases a peak corresponding to the monomeric protein was observed. Enzyme activity was quantified by applying a radioactivity-based extraction assay using undecaprenol phosphate and [³H]-labeled UDP-diNAcBac. Purified PglC variants were tested at up to a hundred times the concentration of the wild-type enzyme to increase sensitivity towards low activity samples.

Deletion studies have been used previously to provide insight into the importance of the transmembrane domains in many families of PGTs. For example, WbaP can be expressed without the first four predicted transmembrane helices and still found to be catalytically active.¹⁰ Importantly, replacing the fifth predicted TMH of WbaP with the first TMH in the sequence resulted in non-functional protein, suggesting that the specific identity of the TMH is critical for catalysis.⁸ The PGT alignment revealed that proline is highly conserved at position 12/13 of the TMH in 95% of the PGTs in the analysis. This conserved proline has also been previously identified in a homologous position in the WbaP-family of PGTs.¹¹ The conserved proline (P24) in the *C. jejuni* PglC is part of a sequence (FILALVLLVLFSPVILITALLL) that is reminiscent of the polyisoprene recognition sequence (PIRS), LL(F/I)IXFXXIPFXFY identified in other enzymes that process polyprenol phosphate-linked substrates as the PIRS sequences also includes a centrally-located proline residue.^{4, 26} Additionally, a tryptophan is shown for the first time to replace this proline residue in approximately 5% of the PGT sequences in the analysis. Therefore, in the mutational analysis, Pro24 was replaced by both alanine and tryptophan to assess the importance of the proline and to observe whether tryptophan represents a functionally conservative substitution. Indeed, the P24A variant showed no activity; however, the P24W variant retained measureable activity, although at lower levels than the wild-type enzyme (Table 1). A helix-breaking proline has also been identified at a similar position (residue 12–13) in a transmembrane helix (residues of 413–432) of the *C. lari* oligosaccharyl transferase PglB, which has recently been structurally characterized. PglB uses undecaprenol diphosphate-linked oligosaccharide as its substrate, suggesting that this structural feature may be a binding determinant for the undecaprenol moiety.²⁷ While additional studies are clearly needed to further this hypothesis, analysis of the interactions of Pro, Trp and Ala-containing TMHs with undecaprenol phosphate will be of considerable interest.

Mutagenesis analysis of the conserved arginine residues, Arg87 and Arg111 resulted in catalytically inactive PglC. These positively charged residues may interact with the

negatively charged phosphate groups of the substrates and products of this reaction, or may participate in salt bridges required to maintain the structure of PglC. We also investigated the role of the conserved methionine residue M62, hypothesizing that it may play a role in maintaining protein structure. When replaced by glutamine, activity decreased significantly to approximately 1% of native levels. However, the more conservative mutation to isoleucine retained activity at approximately 10% of the wild-type enzyme (Table 1). In both cases, we believe that the structures of the variants were preserved since the proteins could be purified without major aggregation issues and size exclusion chromatography revealed discrete peaks corresponding to the monodisperse proteins (See Figure S1).

Aspartic acid residues are implicated in metal-ion cofactor coordination in the WecA and MraY families of proteins.²⁵ Sequences in both families contain adjacent aspartic acid residues (D90/91 in *E. coli* WecA and D115/D116 in *E. coli* MraY). The adjacent aspartate residues resemble the conserved DDXD motif found in other enzymes with diphosphate substrates, such as prenyl transferases, where the DD pair coordinates the Mg²⁺ cofactor.²⁸ The WbaP family does not contain this DD motif, but does possess adjacent aspartate and glutamate residues (D382/E383), which when mutated, abolishes enzyme activity.⁹ The PGT sequence alignment in this study (Figure 2) highlighted a similar sequence, corresponding to residues D92/E93 in PglC. Mutational analysis showed a complete loss of activity when E93 was replaced by glutamine, which replaces steric but not electronic properties. However, in the case of D92, the D92N variant failed to express at a useful level (data not shown) and thus the D92A variant was constructed. Although this protein expressed at wild-type level it was prone towards aggregation during solubilization and therefore was analyzed as the CEF. This approach was adopted because in general we have observed that extracting protein from the CEF into detergent micelles during purification can contribute significantly to protein instability *in vitro*. The D92A variant showed no activity even at protein concentration levels >100-fold compared to the CEF with the wild type protein (based on SDS-PAGE analysis, Table 1).

In addition to D92 two other aspartic acid residues, D156 and D168, were found to be strictly conserved and were also selected for mutagenesis since these might play key roles in catalysis. For example, aspartate residues may play a nucleophilic role in PGTs and it has been proposed that a conserved aspartic acid residue in the fourth cytoplasmic loop of MraY is a nucleophile that initiates the PGT reaction by forming a covalent phosphosugar-enzyme intermediate.⁵ Attempts to express and purify D156A and D168A variants yielded similar results as for the D92, and therefore the D156A and D168A variants were prepared and analyzed as CEFs. Replacement of either of these aspartic acid residues by alanine resulted in the complete loss of activity. Because of the poor stability properties of these enzymes, it is not yet clear whether the mutations of D156 and D168, altered the catalytic properties of the enzyme in a specific way by eliminating key acidic residues that are essential for catalysis or whether the loss of activity was due to a major structural perturbation.

The only branch of the minimal monotopic PGT family on which a bioinformatic analysis has been performed previously is that typified by the C-terminal domain of WbaP from *Salmonella enterica* (Figure 1). The study compared several hundred WbaP homologs to identify conserved residues.⁹ The most conserved charged or polar amino acids were

mutated to alanine residues, and variant proteins were evaluated using *in vivo* complementation assays to observe the formation of LPS O-antigen. There is overlap in the conserved residues discovered in those experiments and herein. Both studies identified the adjacent Asp-Glu pair (D92/E93 in *C. jejuni* PglC) as critical for activity, as well as an additional conserved aspartate residue (D168 in PglC, corresponding to D458 in WbaP). Additionally, the studies also revealed a homologous arginine residue (R111 in PglC, R401 in WbaP) that is essential for catalysis.

Evolutionary Couplings and Structure Prediction

Based on the conservation covariance, a predicted three-dimensional structure was constructed for the soluble domain of PglC (see Methods). The main tool is the EVfold method, which uses multiple sequence alignments and a maximum entropy model of the protein sequence to obtain direct residue pair couplings that might represent spatial proximity. The couplings are used to calculate residue-residue proximity in folded protein structures in conjunction with distance constraints, secondary structure predictions and molecular dynamics simulations, to yield predicted structures. Seven thousand PGT globular domain sequences were used to generate predicted covarying residues and structures of PglC in EVfold. The TMH was not included in the modeling, and the prediction was limited to the soluble domain in all cases. The multiple sequence alignment for the last 20 residues was of poor quality due to the large number of gaps (lack of information); therefore this segment appeared as an unstructured loop in the predicted model from EVfold. Since the number of distance constraints is a critical parameter in the generation of an accurate model, especially when there is not a native structure available for comparison, the only models kept were those with more than 80 evolutionary inferred contacts (EICs) (half of the protein length). If the top ranking model (Figure 3A) is used as the putative native structure, the RMSD value of that model compared to the other top 10 models, range from 4.4 Å to 7.5 Å (Table S3), with only two exceptions (11.4 Å and 12.7 Å). These top models are comparable to the control models generated for proteins of known structure in terms of model consistency, which lends confidence to the predicted PglC structure.

To control for the uncertainty inherent in computational modeling approaches, the program suite, I-TASSER was also used to model the three-dimensional structure of soluble PglC (see Methods). I-TASSER is a threading (fold recognition) based structure prediction method, which differs from the distance geometry based method in EVFold. Using the same set of EICs, the best model (largest structure cluster) from I-TASSER is very similar to the EVFold model in both the tertiary structure (RMSD = 4.7 Å) and in the local spatial arrangement of all those conserved, functionally critical residues (Figure 3B). This lends support for these models. Many of the residues identified as important for catalysis were found in the model proximal to one other, suggesting they may be involved in substrate binding (Figure 3A). To examine whether the distances amongst these residues are realistic with respect to the size of the soluble substrate UDP-Bac, we compared selected inter-residue distances and the overall dimensions of UDP-Bac (PDB 3BSS). The distance between the D92 (C α) to D156 (C α) is 12.4 Å and that between D92 and D168 is 16.4 Å compared to the extended conformer of UDP-Bac (14.0 Å). Additionally, the non-conserved residues used as controls in our mutagenesis studies were distal to the predicted active site of

the protein. Notably, a search of the structural database using the similarity server DALI²⁹ with the predicted PglC three-dimensional structure did not find any structural homologs.

The segment between residues 160–180 in the PglC model (the C-terminal helix) was predicted to be a hydrophobic helix in the three dimensional structure models (Figure 4A). However, based on multiple programs (see Methods), the helix is not predicted to be trans-membrane. To further investigate the role of the C-termini helix, the sequence was input into a helical-wheel projection tool,³⁰ showing it to have two hydrophobic surfaces (Figure 4B). The three-dimensional model predicts that this segment forms an α -helix, in which hydrophobic residues line one face of the helix facing away from the protein core and a second group of hydrophobic residues pack towards the protein core (Figure 4A). It is possible that such a helix provides a face to interact with the membrane surface or to act as the interface in a protein-protein interaction.

Conclusions

Proteins that span the membrane once, while abundant in nature (50% of integral membrane proteins), have proven to be relatively intractable to structural methods. Key issues have recently been highlighted in a paper by Stroud and coworkers.³¹ The difficulties lie in the need for a membrane surrogate in solution-based methods (NMR) and in the problems with ordered crystallization (in X-ray crystallography). These methods may suffer from the lack of order normally imposed by interaction with membrane and indeed the lack of the intermembrane portion of the protein, which is often truncated to provide the more soluble, tractable construct.

In the absence of a structural analysis, bioinformatics in concert with biochemical structure/function analysis provides a powerful approach to guide definition of the catalytic core of the small bacterial PGTs discussed herein. The sequence analysis was greatly improved by including the corresponding globular domains of PGTs of the PglC-like, bifunctional PglB, and WbaP-like families, as the alignment of all three groups emphasized residues that may not have been revealed by analyzing the sequences of the PglC-like family alone. Residues identified from this analysis correlate to residues proximal to one another in the predicted model, suggesting that they may participate in the activity of the enzyme. The diversity of PGT structures highlights intriguing questions concerning the evolution of PGTs in prokaryotes. All of the enzymes catalyze essentially the same reaction; formation of a polyprenol diphosphate-linked glycan by the reaction of a polyprenol phosphate with a UDP-sugar as a sugar-phosphate donor, yet the topologies supporting the catalytic cores are disparate. This analysis now sets the stage for future structural and functional investigations of the smallest bacterial PGTs, which will ultimately provide insight into the entire phosphoglycosyltransferase family of enzymes that perform this first committed step in the membrane-associated glycan assembly pathways.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The computational work was performed on the Shared Computing Cluster administered by Boston University's Research Computing Services. The authors would like to thank the Research Computing Services group for providing consulting support.

Funding: This research was supported by NIH grants R21 AI101807 (to BI and KNA) and GM039334 (to BI) and NIH R01 GM064700 and NIH R01 GM 061867 (to SV).

Abbreviations

diNAcBac	di- <i>N</i> -acetyl bacillosamine (2,4-diacetamido-2,4,6-trideoxyglucose)
CEF	cell envelope fraction
DDM	n-dodecyl β -D-maltoside
EICs	evolutionary inferred contacts
MSA	multiple sequence alignment
PGTs	phosphoglycosyltransferases
PHPT	polyisoprenyl-phosphate hexose-1-phosphate transferase
PNPT	polyisoprenyl-phosphate <i>N</i> -acetylaminosugar-1-phosphate transferase
PSUP	Pure Solvent Upper Phase
TMH	transmembrane helix
UDP-Bac	UDP-di- <i>N</i> -acetyl bacillosamine (2,4-diacetamido-2,4,6-trideoxyglucose)
UDP-Gal	UDP-galactose
Und-PP-Bac	undecaprenol-P-P-diNAcBac

References

1. Valvano MA. Export of O-specific lipopolysaccharide. *Front Biosci.* 2003; 8:s452–471. [PubMed: 12700099]
2. Price NP, Momany FA. Modeling bacterial UDP-HexNAc: polyprenol-P HexNAc-1-P transferases. *Glycobiology.* 2005; 15:29R–42R.
3. Burda P, Aepli M. The dolichol pathway of N-linked glycosylation. *Biochimica et Biophysica Acta (BBA) - General Subjects.* 1999; 1426:239–257. [PubMed: 9878760]
4. Chung BC, Zhao J, Gillespie RA, Kwon DY, Guan Z, Hong J, Zhou P, Lee SY. Crystal Structure of MraY, an Essential Membrane Enzyme for Bacterial Cell Wall Synthesis. *Science.* 2013; 341:1012–1016. [PubMed: 23990562]
5. Lloyd AJ, Brandish PE, Gilbey AM, Bugg TDH. Phospho-N-Acetyl-Muramyl-Pentapeptide Translocase from *Escherichia coli*: Catalytic Role of Conserved Aspartic Acid Residues. *J Bacteriol.* 2004; 186:1747–1757. [PubMed: 14996806]

6. Glover KJ, Weerapana E, Chen MM, Imperiali B. Direct biochemical evidence for the utilization of UDP-bacillosamine by PglC, an essential glycosyl-1-phosphate transferase in the *Campylobacter jejuni* N-linked glycosylation pathway. *Biochemistry*. 2006; 45:5343–5350. [PubMed: 16618123]
7. Hartley MD, Morrison MJ, Aas FE, Børud B, Koomey M, Imperiali B. Biochemical characterization of the O-linked glycosylation pathway in *Neisseria gonorrhoeae* responsible for biosynthesis of protein glycans containing N, N'-diacetyl bacillosamine. *Biochemistry*. 2011; 50:4936–4948. [PubMed: 21542610]
8. Saldías MS, Patel K, Marolda CL, Bittner M, Contreras I, Valvano MA. Distinct functional domains of the *Salmonella enterica* WbaP transferase that is involved in the initiation reaction for synthesis of the O antigen subunit. *Microbiology*. 2008; 154:440–453. [PubMed: 18227248]
9. Patel KB, Furlong SE, Valvano MA. Functional analysis of the C-terminal domain of the WbaP protein that mediates initiation of O antigen synthesis in *Salmonella enterica*. *Glycobiology*. 2010; 20:1389–1401. [PubMed: 20591829]
10. Patel KB, Ciepchal E, Swiezewska E, Valvano MA. The C-terminal domain of the *Salmonella enterica* WbaP (UDP-galactose:Und-P galactose-1-phosphate transferase) is sufficient for catalytic activity and specificity for undecaprenyl monophosphate. *Glycobiology*. 2012; 22:116–122. [PubMed: 21856724]
11. Furlong SE, Ford A, Albarnez-Rodriguez L, Valvano MA. Topological analysis of the *Escherichia coli* WcaJ protein reveals a new conserved configuration for the polyisoprenyl-phosphate hexose-1-phosphate transferase family. *Sci Rep*. 2015; 5:9178. [PubMed: 25776537]
12. Wang L, Liu D, Reeves PR. C-terminal half of *Salmonella enterica* WbaP (RfbP) is the galactosyl-1-phosphate transferase domain catalyzing the first step of O-antigen synthesis. *J Bacteriol*. 1996; 178:2598–2604. [PubMed: 8626328]
13. Studier FW. Protein production by auto-induction in high density shaking cultures. *Protein Expr Purif*. 2005; 41:207–234. [PubMed: 15915565]
14. Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes I. *J Mol Biol*. 2001; 305:567–580. [PubMed: 11152613]
15. Jones DT, Taylor WR, Thornton JM. A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry*. 1994; 33:3038–3049. [PubMed: 8130217]
16. Claros MG, von Heijne G. TopPred II: an improved software for membrane protein structure predictions. *Comput Appl Biosci*. 1994; 10:685–686. [PubMed: 7704669]
17. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*. 2012; 28:3150–3152. [PubMed: 23060610]
18. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. *Genome Res*. 2004; 14:1188–1190. [PubMed: 15173120]
19. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C. Protein 3D Structure Computed from Evolutionary Sequence Variation. *PLoS ONE*. 2011; 6
20. Zhang Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*. 2008; 9:40. [PubMed: 18215316]
21. Lehrer J, Vigeant KA, Tatar LD, Valvano MA. Functional Characterization and Membrane Topology of *Escherichia coli* WecA, a Sugar-Phosphate Transferase Initiating the Biosynthesis of Enterobacterial Common Antigen and O-Antigen Lipopolysaccharide. *J Bacteriol*. 2007; 189:2618–2628. [PubMed: 17237164]
22. Anderson MS, Eveland SS, Price NP. Conserved cytoplasmic motifs that distinguish sub-groups of the polyisoprenol phosphate:N-acetylhexosamine-1-phosphate transferase family. *FEMS Microbiol Lett*. 2000; 191:169–175. [PubMed: 11024259]
23. Amer AO, Valvano MA. Conserved amino acid residues found in a predicted cytosolic domain of the lipopolysaccharide biosynthetic protein WecA are implicated in the recognition of UDP-N-acetylglucosamine. *Microbiology*. 2001; 147:3015–3025. [PubMed: 11700352]
24. Furlong SE, Valvano MA. Characterization of the highly conserved VFMGD motif in a bacterial polyisoprenyl-phosphate N-acetylaminosugar-1-phosphate transferase. *Protein Sci*. 2012; 21:1366–1375. [PubMed: 22811320]

25. Amer AO, Valvano MA. Conserved aspartic acids are essential for the enzymic activity of the WecA protein initiating the biosynthesis of O-specific lipopolysaccharide and enterobacterial common antigen in *Escherichia coli*. *Microbiology*. 2002; 148:571–582. [PubMed: 11832520]
26. Albright CF, Orlean P, Robbins PW. A 13-amino acid peptide in three yeast glycosyltransferases may be involved in dolichol recognition. *PNAS*. 1989; 86:7366–7369. [PubMed: 2678101]
27. Lizak C, Gerber S, Numao S, Aebi M, Locher KP. X-ray structure of a bacterial oligosaccharyltransferase. *Nature*. 2011; 474:350–355. [PubMed: 21677752]
28. Marrero PF, Poulter CD, Edwards PA. Effects of site-directed mutagenesis of the highly conserved aspartate residues in domain II of farnesyl diphosphate synthase activity. *J Biol Chem*. 1992; 267:21873–21878. [PubMed: 1400496]
29. Holm L, Rosenström P. Dali server: conservation mapping in 3D. *Nucleic Acids Res*. 2010; 38:W545–549. [PubMed: 20457744]
30. Zidovetzki R, Rost B, Armstrong DL, Pecht I. Transmembrane domains in the functions of Fc receptors. *Biophys Chem*. 2003; 100:555–575. [PubMed: 12646391]
31. Monk BC, Tomasiak TM, Keniya MV, Huschmann FU, Tyndall JDA, O'Connell JD, Cannon RD, McDonald JG, Rodriguez A, Finer-Moore JS, Stroud RM. Architecture of a single membrane spanning cytochrome P450 suggests constraints that orient the catalytic domain relative to a bilayer. *Proc Natl Acad Sci USA*. 2014; 111:3865–3870. [PubMed: 24613931]

Figure 1.

Predicted and experimentally-determined membrane topologies of phosphoglycosyltransferase (PGT) families. The soluble C-terminal PglC-like domain is shown in red.

Figure 2.

Sequence alignment of the PglC-like domain from the PglC, PglB(*Ng*) and WbaP families of PGTs. The alignment was visualized using WebLogo 3.¹² Residues indicated with a red line were analyzed by site-directed mutagenesis and are marked above the red line using the *C. jejuni* (NCTC 11168) PglC numbering system.

Figure 3.

Predicted structure of the soluble domain of PgIC. (A) Stereo-view ribbon diagram of the highest ranked predicted structure from EVfold showing residues mutated in this study. Conserved residues are depicted in orange, and non-conserved residues are shown in green. (B) Superposition of the best models generated by EVFold (blue ribbon) and I-TASSER (yellow ribbon).

Figure 4.

Structure of the predicted hydrophobic helix in PglC (A) Focused view of the predicted hydrophobic helix shown in the highest ranked model structure from EVfold, with surface-exposed hydrophobic residues represented as sticks. (B) Helical wheel predicted for the hydrophobic helix. The most hydrophobic residue is shown in green, and the amount of green decreases proportionally to the hydrophobicity, with zero hydrophobicity coded as yellow. Hydrophilic residues are coded red with pure red being the most hydrophilic (uncharged) residue.

Scheme 1.

Phosphoglycosyltransferase reaction shown with a generic UDP-carbohydrate substrate and membrane-bound undecaprenol phosphate.

Table 1

Summary of activities of PglC variants.

Entry ^a	Mutation	Predicted Role of Mutated Residue	Activity ^b
1	WT		+++
2	P24A	Interactions with Und-P	–
3	P24W	Interactions with Und-P	++
4	M62Q	Structural	+
5	M62I	Structural	++
6	E93Q	Coordination of metal cofactor	–
7	R87Q	Interactions with phosphate groups	–
8	R111Q	Interactions with phosphate groups	–
9	E116Q	Control	+
10	E65Q	Control	++
11	D92A (CEF) ^c	Coordination of metal cofactor	–
12	D156A(CEF)	Coordination of metal cofactor/Structural	–
13	D168A (CEF)	Coordination of metal cofactor/Structural	–

^aAll assays were performed under identical conditions with the substrate concentrations set at the K_M values: 16 μ M Und-P and 7.5 μ M UDP-diNAcBac. Entries 1–10 were carried out with solubilized and purified enzyme and entries 11–13 were carried out with CEF.

^bThe wild-type level of activity in the presence of 3 nM enzyme, to which all mutant enzymes are compared, is defined as (+++). A designation of (++) represents the activity of a mutant enzyme that attains ~85% of wild-type activity with 30 nM enzyme. A designation of (+) represents the activity of a mutant enzyme that attains ~85% of wild-type activity with 300 nM of enzyme. A designation of (–) is used to describe the activity of a mutant enzyme that attains < 30% of wild-type activity with 300 nM of enzyme.

^cFor those variants labeled (CEF), wild-type levels of activity are designated as (+++) and mutant enzymes exhibiting less than 30% of the wild-type levels of activity with 100- fold more CEF than wild-type enzyme as estimated by gel densitometry analysis. This level of activity is designated as (–).