A Molecular Model of the Human UDP-Glucuronosyltransferase 1A1, Its Membrane Orientation, and the Interactions between Different Parts of the Enzyme

Lisa Laakkonen and Moshe Finel
Centre for Drug Research, Faculty of Pharmacy, University of Helsinki, Helsinki, Finland
Received December 22, 2009; accepted March 9, 2010

ABSTRACT

The vertebrate UDP-glucuronosyltransferases (UGTs) are membrane-bound enzymes of the endoplasmic reticulum that process both endogenous and exogenous substrates. The human UGTs are well known biologically, but biophysical understanding is scarce, largely because of problems in purification. The one resolved crystal structure covers the C-terminal domain of the human UGT2B7. Here, we present a homology model of the complete monomeric human UGT1A1, the enzyme that catalyzes bilirubin glucuronidation. The enzyme can be seen as composed of four different domains: two large ones, the N- and C-terminal domains, and two small ones, the “envelope” helices and the transmembrane segment that includes the cytoplasmic tail. The hydrophobic core of the N-terminal domain and the two envelope helices that connect the large domains are shown to be structurally well conserved even among distant homologs and can thus be modeled with good certainty according to plant and bacterial structures. We consider alternative solutions for the highly variable N-terminal regions that probably contribute to substrate binding. The bilirubin binding site, known pathological mutations in UGT1A1, and other specific residues have been examined in the context of the model with regard to available experimental data. A putative orientation of the protein relative to the membrane has been derived from the location of predicted N-glycosylation sites. The model presents extensive interactions between the N- and C-terminal domains, the two envelope helices, and the membrane. Together, these interactions could allow for a concerted large-scale conformational change during catalysis.

UDP-glycosyltransferases catalyze the transfer of sugar groups, activated by uridine diphosphate (UDP-sugar), to small-molecule acceptor substrates. The most commonly used sugar is glucose, and the range of acceptor substrates is seemingly infinite. Glycosyltransferases are found in all kingdoms of life and are involved in biosynthesis, biontransformations, signal transduction, and metabolism. Despite this tremendous functional variability, all structurally known glycosyltransferases that use sugar nucleotides as donors adopt one of only two different folds, GT-A or GT-B, and are readily identified by a 44-residue-long “signature sequence” at the UDP-sugar binding site (Prosite pattern PS00375; http://au.expasy.org/prosite). Both GT-A and GT-B folds comprise two ß/ß sandwich domains that, in the GT-A fold, are packed closely together to form a continuous central ß-sheet. In the GT-B fold enzymes, including the human UDP-glucuronosyltransferases (UGTs), the two domains are connected by a flexible linker (Unligil and Rini, 2000; Lairson et al., 2008). The contrast between multiple biological roles and very similar structures highlights the functional importance of local structural differences.

The vertebrate UGTs transfer the glucuronic acid moiety from UDP-GlcUA to lipophilic acceptor substrates in an S_n2 reaction. The reaction proceeds stepwise: UDP-GlcUA binds first to the enzyme and primes it for the binding of the acceptor substrate. After the sugar transfer, the ternary complex resolves and the product dissociates from the enzyme. Substrate inhibition is often observed, and it has been interpreted as formation of a nonproductive enzyme–UDP–substrate complex (Luukkanen et al., 2005).

A single C-terminal transmembrane helix anchors the UGTs to the endoplasmic reticulum membrane, whereas most of their mass is on the luminal side. Secondary membrane attachment sites have been proposed (Ciotti et al., 1998; Ouzzine et al., 1999). The cellular localization and membrane attachment separate the vertebrate UGTs from

ABBR EVIATIONS: UGT, UDP-glucuronosyltransferase; GT, glycosyltransferase; PDB, Protein Data Bank; CN-I, Crigler-Najjar type I; CN-II, Crigler-Najjar type II.
the bacterial and plant glycosyltransferases, which are mostly water-soluble proteins in the cytoplasm. The transmembrane helix of the vertebrate UGTs is important for activity (Meech et al., 1996; Kurkela et al., 2004a,b), and UGTs probably exist as oligomers (Finel and Kurkela, 2008), but the physical details of these features are currently unknown.

The crystal structure of the C-terminal domain of the human UGT2B7 has been resolved (Miley et al., 2007) and agrees well with other known GT-B structures. In complete GT-B structures (see Fig. 1), the two major domains are connected by a central extended linker of approximately 10 residues and two C-terminal “envelope” helices, each approximately 20 residues long, that fold back over both domains (e.g., Brazier-Hicks et al., 2007).

The human UGTs have different, but partly overlapping, substrate specificities. By sequence similarity and gene structure, the human UGTs can be clustered into two groups, UGT1A and UGT2 (Mackenzie et al., 2005). The glucuronidation of bilirubin, the neurotoxic breakdown product of heme, is catalyzed by UGT1A1 and is essential for its biliary excretion (Bosma et al., 1994). Mutations that lower UGT1A1 expression level or reduce its activity result in elevated levels of free bilirubin in the serum, leading to Crigler-Najjar type I (CN-I) or II (CN-II) syndrome or Gilbert syndrome.

Six homolog models have been published of two-domain UGTs: two of the human UGT1A1 (Li and Wu, 2007; Locuson and Tracy, 2007), one of the human UGT1A9 (Fujiwara et al., 2009a), and three of the plant enzymes UGT73A5 (Hans et al., 2004), UGT85B1 (Thorsøe et al., 2005), and UGT94B1 (Osmani et al., 2008). The two UGT1A1 models and the UGT94B1 model are built on the same template UGT71G1 of Medicago truncatula (PDB ID code 2acv), the UGT1A9 model is based on the structure of GtA from Amycolatopsis orientalis (PDB ID code 1pn3), and the slightly older UGT73A5 and UGT85B1 models use GtB of A. orientalis as the template (PDB ID code 1iir). All of these models lack some parts of the enzyme: the envelope helices or some central loops in the N-terminal domain.

We have constructed an all-atom model of the monomeric human UGT1A1. The model highlights extensive interactions between different parts of the enzyme, proposes an orientation with respect to the membrane, and suggests molecular explanations for pathological mutations.

Materials and Methods

The UGT1A1 sequence P22309 from the SwissProt database (http://www.uniprot.org) was used in this study. Protein structures related to the human UGT1A1 sequence were searched by both sequence and structure comparisons, using the program PSI-BLAST (Altschul et al., 1997) against the Protein Data Bank (Berman et al., 2000) and the program Dali (Holm et al., 2008). Structurally annotated sequence alignments were rendered with Alscript (Barton et al., 1993). Once identified, 13 structural relatives of UGT1A1 were superimposed with a tcl script (http://www.tcl.tk) in the program VMD (Humphrey et al., 1996). For the N-terminal domain, the atoms selected for superposition were 66 Ca from the seven β-strands and helices Na1 and Na4 (see Supplemental Fig. 1); and for the C-terminal domain, 73 Ca atoms were selected from the β-strands and helices Co3, Co4, and Co5. Secondary structure was predicted for
the target sequence with the program PredictProtein (Rost et al., 2004). The same prediction program was run for four selected templates as a test, and the results agreed well with the known structures (data not shown).

The secondary structure elements were named by the domain, N or C; β-strands in each domain core were numbered (β1–β7), and helices were given the number of the strand they follow, supplemented by a sequential subnumber when needed. For example, Nu1 stands for the α-helix in the N-terminal domain that follows strand β1 of this domain, and Nu2–1 stands for the first segment in the α-helix that follows strand β3 of the N-terminal domain. The pre-strand helix in the C domain was called Ca0. The nomenclature resembles those used by others (Unigil and Rini, 2000; Li and Wu, 2007; Miley et al., 2007). Amino acid numbers refer to the full-length human UGT1A1, unless mentioned otherwise.

The human UGT sequences were aligned to each other by using ClustalX 2.0.5 (Larkin et al., 2007). The UGT alignment was matched to the structural alignment of the selected GT-B proteins, including the predicted secondary structure elements as weight factors. Two alignments were constructed for Nj4, aligning either Asp146 or Asp151 as the aspartate situated right behind the catalytic histidine of the known structures. For residues 83 to 143, three sequence-to-structure matches were considered. The rotational orientation of helix Nα3-2 relative to the rest of the protein is the same in all of them, but the position of the sequences relative to the structures differs by one helical turn. The envelope helices were modeled at residues 434 to 448 and 454 to 467 according to the templates, and the transmembrane helix was constructed as a standard α-helix at residues 491 to 516.

Homology models were constructed by using the program Modeler 9v6 with a standard modeling scheme, very thorough annealing, and molecular dynamics optimization (Sali and Blundell, 1993). Three templates were used for the N-terminal domain (PDB ID codes 2vc8, 2vc9, and 2vc7) and one for the C-terminal domain (PDB ID code 2o7h). The nonhydrolyzable UDP-glucose analog (residue name UF2) was copied from the 2vc8 structure and included in the model as a block residue. Likewise, an N-linked branched sugar [N-acetyl-D-glucosamine]-(N-acetyl-Dglucosamine)β-mannose-(α-mannose)-(β-mannose), was copied from 3d12 and attached to the residues predicted to be glycosylated by NetNGlyc (http://www.cbs.dtu.dk/services/NetNGlyc; R. Gupta, E. Jung, and S. Brunak, manuscript in preparation). The residues corresponding to the predicted UGT-glycosylation sites were changed to asparagines in structure 1irr (Ala44 and Asp235) to allow for the binding of the N-glycosylamine. Helical conformation was constrained between residues Val491 and Lys516 for the predicted transmembrane helix (Rost et al., 2004). Fifty structures were modeled and optimized. The majority of these jobs finished successfully, and the 10 with the best values of the target function were studied in detail. The stereochemical quality of the models is good, as analyzed by MolProbity (data not shown) (Davis et al., 2007), whereas optimization violations occur in the long loops, as expected. The model coordinates are available on request (liisa.laakkonen@helsinki.fi).

A molecular structure for (Z,Z)-bilirubin was retrieved from the PubChem database (http://pubchem.ncbi.nlm.nih.gov). It was modified to correspond to the experimentally resolved ridge-tile configurations E,E and E,Z (Nogales and Lightner, 1995) and optimized in MOPAC2009 (http://openmol.net) with PM3 methodology (Stewart, 2007). The substrate structures were docked manually to the final enzyme model between the active-site histidine, His39, and the nonhydrolyzable UDP-glucose analog. A list of nonsynonymous CN-I mutations was gathered from the UGT allele site (http://www.pharmacogenomics.pha.ulaval.ca), and the mutations were localized into the protein model.

The new UGT1A1 model was compared with previous ones done by Locucion and Tracy (2007) and Li and Wu (2007). The basis for the comparison was a structural alignment between UGT71G1 from M. truncatula (PDB ID code 2acev) and GtfD from A. orientalis (PDB ID code 1rrv), onto which the UGT1A1 sequences were added manually as described in the previous studies. A direct structural comparison could not be performed because the coordinates of the earlier models were not available.

**Results**

**Structurally Related Proteins.** We divided the human UGT into four different substructures: the N-terminal domain, the C-terminal domain, the envelope helices, and the transmembrane segment (Fig. 1). For the C-terminal domain and the envelope helices there are good structural templates, and the single transmembrane helix can be safely modeled de novo, but the N-terminal domain requires special attention. It has no high-homology templates; therefore, we first analyzed the available low-homology templates to identify structurally conserved features that could guide the construction. To avoid biasing the modeling toward either plant or bacterial templates, several representative templates were used for the modeling.

Structural relatives of UGT1A1 were searched by both sequence and structure comparisons. The sequence similarity search with PSI-BLAST against the Protein Data Bank converged in five rounds onto 15 hits with expectation values below 10−30. The expectation values of the best proteins were >1. The hit structures also were found to be true homologs by visual evaluation. It is noteworthy that all local similarities between the query and hit sequences correspond to the respective C-terminal domains.

The structural similarity search on Dali could not be done with UGT1A1, but the partial structure of UGT2B7 was used as a query. Nevertheless, the outcome of the Dali search overlapped with the PSI-BLAST results, as expected. The separation of glycosyltransferases from the next group of enzymes, glycosylpyrophosphates, was not as clear-cut in Dali as in the sequence search. Dali returns both chains of crystallographic dimers as independent hits. In such cases only the first of the two monomers was recorded. Another issue we had to address was sequence redundancy among the search results. By sequences, structures 2acev and 2acew are 100% identical as are the 2vg8 (or 2vev/2vev) and 3d0z/3d0z groups. Only one structure of each set was considered in the structural alignment.

Finally, 13 unique structures were considered representative of the GT-B fold and potentially good templates for modeling the human UGT1A1 (Table 1). The 13 proteins were superimposed for structural analysis (see Supplementary Fig. 1 for structural alignment and original references). Overlaying complete proteins by core β-strands did not give good results, because the interdomain angles vary between structures, but aligning either the N- or C-terminal domains separately yielded good matches. The pairwise root-mean-square deviation values between the templates for 66 Ca atoms in the N-terminal domains ranged from 0.58 to 2.24 Å, with a mean of 1.42 Å. The corresponding values for 73 Ca atoms in the C-terminal domains were 0.51, 1.81, and 1.25 Å. The same sets of atoms were used for all pairwise comparisons.

The GT-B structures cluster by their biological roles, sequence identities, and root-mean-square deviations into four mutually similar groups: 1) human UGT2B7 (PDB ID code
2o6l), 2) plant UGT enzymes (PDB ID codes 2vce, 2acv, 2c1x, and 2pq6), 3) bacterial glycosyltransferases (PDB ID codes 1iir, 1rrv, 1pn3, 1f0k, and 2pq6), and 4) macrolide glycosyltransferases from Streptomyces (PDB ID codes 2iya, 2iyf, and 3d0q). Three structures were selected to serve as a template for modeling the N-terminal domain of UGT1A1, one each from groups 2 to 4 (PDB ID codes 2vce, 1iir, and 2iya), and the 2O6L structure was the template for the C-terminal domain in addition, helix Na3-1 was constructed according to structure 2p6p of group 2.

### Analysis of the Template Structures

The template proteins fold into canonical GT-B structures of two α/β/α domains with parallel β-strands. Strands and helices alternate throughout the structure, with a few exceptions (see Fig. 1). Both the N- and C-terminal β-sheets follow similar topology, 3-2-1-4-5-6-7, but they differ in that the former are strongly twisted, whereas the latter are rather flat. The angles between the strands at opposite edges are approximately 110 and 40° in the N- and C-terminal domains, respectively. Despite the similar folds, the two-domain structures vary in length by more than 70 amino acids among the 13 templates. One reason is that in the shorter bacterial enzymes the last helix in the N-terminal domain, Na6, is replaced by a coil. Major length variations are also found in the segment between Nβ5 and Nβ6.

The core β-strands are structurally well conserved in both domains, but the best-conserved helices occur asymmetrically. In the N-terminal domains, the conserved helices are Na1 and N4 that embrace strands Nβ1, Nβ2, Nβ4, and Nβ5 at the center of the twisted sheet. In the C-terminal domains the best-conserved helices are Ca3, Ca4, and Ca5, lying above the flat β-sheet and facing the active site at the center of the protein (see Fig. 1b). These helices were used in the structural superpositions in addition to the core β-strands.

Eight of the studied GT-B protein structures are complete, whereas the other five end at various positions within the envelope helices. The two envelope helices are connected by a short linker that adopts different conformations in the structures. In addition to variations in the angle between the N- and C-terminal domains, the structures exhibit pronounced variation at two other sites: between strands Nβ3 and Nβ4 and between strands Nβ5 and Nβ6. Instead of folding into single helices such as most interstrand segments, these two long stretches contain several ordered helices. Here, we call the complete segments loops 3 and 5 and number the included helices sequentially (Fig. 1).

Loop 3 is 42 to 57 residues long. After Nβ3 the polypeptide chain first continues along the direction of that strand, either as a random coil or a helix, but 10 to 15 residues later the chain turns and folds back as a helical structure. The early part of loop 3 is disordered in many crystal structures, implying mobility, but the shared helices are well resolved and pack to their surroundings. The first helix interacts with a helix in loop 5, whereas the second packs to Nα4. The overall path of loop 3 helices is similar in all template proteins, even though the helical feature is straight and continuous in the bacterial proteins, but is kinked in the plant proteins. The end helix interacts with the segment between Nβ3 and Nβ6.

### Table 1

<table>
<thead>
<tr>
<th>PDB ID Code</th>
<th>Source Species</th>
<th>Source</th>
<th>PSI-Blast z-score</th>
<th>Dali z-score</th>
<th>Used as a Template (Source)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2acv</td>
<td>Medicago truncatula</td>
<td>Triterpene UGT71G</td>
<td>10^-51</td>
<td>21.2</td>
<td>Li and Wu, 2007; Locunson and Tracy, 2007; Osmani et al., 2008</td>
</tr>
<tr>
<td>2c1x</td>
<td>Vitis vinifera</td>
<td>UDP-glucose flavonoid GT</td>
<td>2 x 10^-50</td>
<td>20.7</td>
<td>Hans et al., 2004; Thorssøe et al., 2005; current study</td>
</tr>
<tr>
<td>1iir</td>
<td>Amycolatopsis orientalis</td>
<td>GtfB</td>
<td>5 x 10^-50</td>
<td>19.3</td>
<td></td>
</tr>
<tr>
<td>2pq6</td>
<td>Medicago truncatula</td>
<td>Multifunctional GT</td>
<td>7 x 10^-46</td>
<td>21.3</td>
<td></td>
</tr>
<tr>
<td>2iya</td>
<td>Streptomyces antibioticus</td>
<td>Macrolide GT</td>
<td>2 x 10^-45</td>
<td>21.3</td>
<td>Current study</td>
</tr>
<tr>
<td>2iyf</td>
<td>Streptomyces antibioticus</td>
<td>Macrolide GT</td>
<td>6 x 10^-45</td>
<td>20.3</td>
<td></td>
</tr>
<tr>
<td>1rrv</td>
<td>Amycolatopsis orientalis</td>
<td>Gfd</td>
<td>10^-44</td>
<td>19.6</td>
<td>Locuson and Tracy, 2007</td>
</tr>
<tr>
<td>2o6l</td>
<td>Homo sapiens,</td>
<td>UGT2B7</td>
<td>10^-44</td>
<td>36.1</td>
<td>Thorssøe et al., 2005; Fujiwara et al., 2005</td>
</tr>
<tr>
<td>1pn3</td>
<td>Amycolatopsis orientalis</td>
<td>GtfA</td>
<td>2 x 10^-39</td>
<td>18.5</td>
<td></td>
</tr>
<tr>
<td>2vce</td>
<td>Arzobacter thaliana</td>
<td>UGT72B1</td>
<td>10^-42</td>
<td>20.3</td>
<td>Osmani et al., 2008; current study</td>
</tr>
<tr>
<td>2pq6</td>
<td>Streptomyces fradiae</td>
<td>C-C bond forming GT</td>
<td>2 x 10^-39</td>
<td>16.8</td>
<td>Current study for Na3–1</td>
</tr>
<tr>
<td>3d0q</td>
<td>Micromonospora echinospora</td>
<td>CALG3</td>
<td>9 x 10^-39</td>
<td>17.8</td>
<td></td>
</tr>
<tr>
<td>1f0k</td>
<td>Escherichia coli</td>
<td>MurG</td>
<td>4 x 10^-36</td>
<td>14.1</td>
<td></td>
</tr>
</tbody>
</table>
UGTs: from Sequences to Structures. The sequence identities of the mature proteins are approximately 70% within the UGT1As, 75% among the UGT2s, and 45% between the two groups. The corresponding values for the N-terminal domains are 40, 50, and 30%. Aligning the mature UGT sequences is simple because of their high mutual similarity. Their lengths vary by merely two residues, 504 to 506, and gaps are required in the multiple alignment at only seven sites: after residues 77, 97, 102, 175, 226, 508, and 516 (Supplemental Fig. 1). The four latter sites are single gaps. Of the 504 to 506 residues, 145 are identical in all human UGTs. Against this overall similarity, the sequence variation around residues 77, 97 to 102, and 175 is striking, and the local sequence alignment becomes ambiguous.

Matching the 19 human UGT sequences to the 13 GT-B structures is trivial for the C-terminal domain because UGT2B7 is included in both the structural and the sequence alignment. The residues in the “signature sequence” form a central part of the C-terminal domain, covering Ca3, Cβ4, Ca4, and Cβ5. As has been discussed earlier (Miley et al., 2007), residues involved in binding the shared sugar donor are found conserved in this segment (Trp354, Gln357, His372, Glu380, Phe394, Asp396, and Gln396). The additional conserved residues in the segment (Leu360, Leu361, Phe369, Ile370, Thr371, Ser381, Ile382, and Met388) form the tightly packed hydrophobic core of the C domain.

All β-strands, and helices Na1 and Ne4 and the two envelope helices, align well to the template structures (see Fig. 1 and Supplemental Fig. 1). On the contrary, the variable loops 3 and 5 could be aligned in several different ways. Sequence homology is insufficient for smooth modeling without additional input. We have included secondary structure predictions, physical considerations derived from the templates, and biological data on the UGTs in the modeling. The area of Nβ4 was studied in detail because of recent conflicting data (Li et al., 2007; Patana et al., 2008). In this area, the most striking feature of the GT-B structural alignment is a fully conserved stretch of hydrophobic amino acids at the center of Nβ4, with aspartates at both ends. Neither of the acids is conserved among the 13 template proteins, and neither one lacks matching acidic residues. In our preferred sequence-to-structure alignment (Supplemental Fig. 1), the two UGT aspartates are located at the beginning and end of Nβ4, 16 Å apart from each other. This places Asp151 next to the active site, in hydrogen-bonding distance from His39, whereas Asp146 is located at the beginning of strand Nβ4. In the GT-B proteins studied, aspartates at the proposed position for Asp146 of UGT1A1 form hydrogen bonds to the beginning of Nβ1 and the end of the second envelope helix. In line with this, Asp146 in the model comes into contact with Lys29 and Lys469. In the alternative alignment the sequence is moved forward by five residues, bringing Asp146 next to the catalytic histidine. The second aspartate, Asp151, is now situated in the loop between Nβ4 and Ne4, where it can interact with the substrates. This change is feasible because loop 3 is of sufficient length and low homology, but it leads to an alignment in which the center of Nβ4 is no longer universally hydrophobic, and the fork position before Nβ4 is occupied by Ala141 (data not shown). In addition, the size of the substrate binding cavity is reduced.

Loop 3 of UGT1A1, with its 65 residues, is longer than the same loop in any of the templates (42–57 amino acids). It is predicted to contain three helices, Na3-1 (Arg85/Val99), Na3-2 (Phe105/Leu130), and Na3-3 (Lys134/Ala141). Most probably, two of these correspond to the helical stretch seen in all the templates. If Na3-1 and Na3-2 correspond to the shared helices, the extra helix will fold at the far end of the protein, close to the N terminus. And, if Na3-2 and Na3-3 make for the shared feature, Na3-1 will lie at the interface between the large domains. We find both alternative alignments more likely. First, the templates show structural variability before the long helical stretch, but none after it. Second, many residues in loop 3 affect activity in various UGTs and supposedly contribute to the binding site at the domain interface (Lewis et al., 2007; Nishiyama et al., 2008; Fujiiwara et al., 2009a,b). Accordingly, the predicted helices Na3-2 and Na3-3 of UGT1A1 are matched to the two shared loop 3 helices of the templates (Supplemental Fig. 1). The additional helix, Na3-1, is built at the domain interface according to an analogous extra helix in structure 2p6p.

The resulting helical segment Na3-2/3 is longer than what is observed in the templates (37 versus 23–30 amino acids) and reveals no obvious sequence-to-structure match. The single conserved property of Na3s in the studied GT-B proteins is amphiphility and, indeed, a periodic separation of hydrophobic and hydrophilic residues is also visible in the UGT alignment. This dictates the rotational orientation of the sequences to the structure. The gap in the Na3-2/3 helix predictions (residues 131 and 132) would naturally match the interhelical kink of many templates, in which case N131K132E133 of UGT1A1 would align to N89P90E91 of structure 2vce. However, lacking any clear sequence similarity between human UGTs and the template proteins in this region, three different alignments were constructed, varying the position by one helical turn at a time. The first match is as suggested above, the next matches NPE of structure 2vce to the L130L131H132 of UGT1A1, and the third to C127S128H129 of the human enzyme. The resulting model options are called c, b, and a, respectively. All three models yield stable structures of comparable energies. There are large structural variations in loop 3 between individual model structures, however, reflecting the poor homology of UGTs to any template and conflicting template structures. Again, several polar interactions are observed between residues in loop 3 and its surroundings.

The other highly variable segment, loop 5, is 77 residues long in UGT1A1, well within the observed length variation in the template structures. Sequence analysis strongly predicts one long helix from Phe206 to Val226, followed by two shorter ones at 230 to 237 and 244 to 248 (Fig. 1b). This pattern of three helices close to Nβ6 agrees with the structures of the bacterial proteins, but not the plant proteins. Hence, structure 1iiir was chosen as the sole template for modeling loop 5. Some sequence similarity can be observed between the UGTs and the bacterial templates at the end of this segment. An additional, weakly predicted helix was built de novo for the early part of the loop, at Leu179 to Glu182.

Figure 2 shows a superposition of the 10 lowest-energy model structures. They are practically indistinguishable from each other at the C-terminal domain, where a good template was available. It is noteworthy that the structures match reasonably well also for the core of the N-terminal domain and the envelope helices. Reflecting both the struc-
tural and sequence analyses, the path of the peptide chains vary remarkably in loops 3 and 5, while always folding back to a compact structure. No nonphysical strongly extended loops out of the protein core were observed.

**Nonhomology Modeling.** To get a complete view of the functional protein, even if some parts are more hypothetical, we also modeled the transmembrane segment and the N-linked complex sugars. They were constructed by combining sequence data to molecular knowledge on other proteins. In general, transmembrane helices show well characterized properties, especially in single-pass proteins (White, 2009). The predicted transmembrane segment of UGT1A1 was built as a standard α-helix from Val491 to Lys516, as predicted. This is longer than the 17 hydrophobic residues, Val491 to Phe507, that are often taken as the membrane-intercalated segment of the UGTs. Without additional data on intramolecular interactions, the orientation of the helix with respect to the globular part was left undetermined.

The orientation of UGT1A1 relative to the membrane was addressed from another, unrelated point of view: the N-linked glycosylation sites. The very hydrophilic and mobile branched sugar groups cannot exist but in an aqueous environment. The high sequence similarity among the human UGTs strongly suggests that they all fold very similarly and, therefore, are oriented to the membrane in the same way. Accordingly, we have analyzed the glycosylation sites in all human UGTs and located these sites on the structural model of UGT1A1. The likelihood of the NXS/T sequence motifs to be glycosylated was predicted by NetNGlyc (http://www.cbs.dtu.dk/services/NetNGlyc), and the results are shown in Fig. 3. Only one glycosylation site was predicted for UGT1A1, Asn295 in Ca0, that is shared by all UGT1A enzymes. In the model, Asn295 points out of the protein and could easily accommodate a large and mobile sugar unit. The UGT2 enzymes lack the glycosylation site at Ca0, but there is an alternative site eight residues later, corresponding to Glu303 in UGT1A1. It lies before the start of Ca1, close to the domain interface.

In all other human UGTs, with the exception of UGT1A1 and UGT2B4, at least one glycosylation site is predicted within the N-terminal domain. In UGT1A3–UGT1A5 the residues to be glycosylated are Asn119 and Asn142 (UGT1A3 numbering), both of which are in loop 3. These residues correspond to Lys118 and Ala141 in UGT1A1, which are located in helix Na3 and right after it. Both are on the surface of the protein and could be decorated with sugars. It is noteworthy that Asn189 of UGT1A3 in loop 5, corresponding to Asn188 of UGT1A1, is predicted not to be glycosylated because it is followed by a proline. Compared with the positively predicted sites, Asn189 is found on the opposite side of the globular domain, suggesting a possible lumen-membrane orientation for the protein. UGT1A7–UGT1A10 exhibit one strongly predicted glycosylation site at Asn71 (corresponding to Phe73 of UGT1A1), in the loop right before Nβ3. UGT2A1 contains a strongly predicted glycosylation site at Asn49 before the start of Nβ2 (corresponding to Glu56 of UGT1A1). It is noteworthy that the three predicted sites in three different UGTs (UGT1A3: Asn142; UGT1A7: Asn71; UGT2A1: Asn49) all fall spatially quite close to each other in the model, even though they originate from different parts of the sequences. Taken together, the data suggest a spatially conserved glycosylation site before strands Nβ2, Nβ3, or Nβ4.

We modeled an N-linked complex sugar to the loop between Na2 and Nβ3 of UGT1A1.

**Bilirubin Binding.** With several protein–substrate complexes of GT-Bs available, the localization of the active site at the N/C-domain interface is unproblematic. The UDP-sugar analogs found in six GT-B proteins studied overlap very well; the maximal difference in the positions of the phosphates between any two structures is 1.2 Å (data not shown). The same position is expected to be valid for the UGTs, too.

Fig. 2. Internal variation among the 10 lowest-energy UGT1A1 models of alignment c. In the C-terminal domain (thin lines at bottom) the models superimpose perfectly, reflecting high similarity between the template and target. The structurally conserved parts of the N-terminal domain (gray thick lines on the top) and the envelope helices, on the left, agree reasonably well among the models, but the variation is much more pronounced for loops 3 and 5 (thin black lines on the right).

Fig. 3. Location of N-glycosylation sites and CN-I mutations in the UGT1A1 model. The N-terminal domain is shown in white, the C-terminal domain is shown in green, and the envelope helices and transmembrane helix are shown in blue. The gray space-filling molecule in the middle is the UDP-sugar. The pink stick structures represent the N-linked sugars. The one on the right is attached to Asn295 of UGT1A1, and the one on the left, shown by thin lines, shows the position where many other UGTs are predicted to be glycosylated. Red spheres highlight CN-I mutations that are located within 5 Å from the UDP-sugar, and the orange spheres sites are further from it.
Hence, a UDP-glucose analog was copied from one template, structure 2vce, to the UGT1A1 model. Bilirubin is highly flexible, but is known to exist in two internally hydrogen-bonded configurations in aqueous solutions: (Z,Z)-bilirubin and (Z,E)-bilirubin (Nogales and Lightner, 1995). Docking either of the free carbonyls in bilirubin between the active-site histidine and the anomeric carbon of the UDP-glucose places the rest of the molecule between loops 3 and 5. Not only His39 but the whole helix Na1 lies quite close to the UDP-sugar and forms an active site wall. Asp36, Ser38, and the middle of the protein. Two identical copies of bilirubin in shown in blue on the left for orientation. Loop 3 is colored orange, and from the top of the N-terminal domain, and the second envelope helix is in line with the N-terminal domain are marked by red spheres.

Comparison to Earlier Work. One way to evaluate the current molecular model is to compare it with the earlier ones by Locuron and Tracy (2007) and Li and Wu (2007).

They were published so close in time that no previous comparison exists. We have matched the template-target alignments from those studies to our own alignment and examined the results. The combination alignment is shown in Supplemental Fig. 2, and the differences in the positions of the secondary structure elements are listed in Table 2. Even without the UGT2B7 structure as a template, the models agree perfectly with each other between residues Glu288 and Arg467. The two earlier models do not continue past the C-terminal domain into the envelope helices and the transmembrane helix. In the N-terminal domain, Na1–Nβ1 and Nβ4–Na4–Nβ5 at the center of the domain are aligned identically to the template in the three models. For Nβ2 two models agree fully, whereas the third differs by a single residue. Our model resembles more closely that by Locuron and Tracy (2007). The average unsigned difference between the two models is 1.5 amino acids for the core domain. We were surprised to find that Locuron and Tracy had not aligned the structurally conserved Na2 with anything, even though they had assigned the helix in the template structure. In their model, this segment must be a random coil, or there is a mistake in their figure 1. The next largest difference between the two models is in helix Na6, by one helical turn. The similarity between the models is acceptable, with the exception of Na2, and differences of the same magnitude are probably caused by variations in construction and optimization schemes of the molecular models. The model by Li and Wu (2007) differs more from our model and that by Locuron and Tracy (2007). The average unsigned differences in the positions of the secondary structure elements in the core N-terminal domain are very similar for both comparisons, 2.4 and 2.8, respectively. The differences arise from Na2–Nβ3 and Nβ6–Na6, and for the short β-strands especially the variations are crucial. It remains to be seen which assignment is correct.

Discussion

Our goal in modeling the human UGT1A1 was to provide a structural framework in which to analyze biological data and formulate new functional hypotheses that could be tested experimentally. The common understanding is that the N-terminal half of the UGTs is variable, whereas the C-terminal half is conserved. According to our analysis this is an oversimplification. The C-terminal half is better seen as composed of three structural domains (Fig. 1), the largest of which, the C-terminal α/β/α sandwich, contains the UDP-GlcUA binding site. As for the degree of conservation, the hydrophobic core of the N-terminal α/β/α domain is equally well conserved as the core of the C-terminal domain. The variability of the N-terminal domain only holds for certain loop segments.

The differences between closely homologous human UGTs are found mainly in loop 3, in a segment called hypervariable region II (Ala61 to Tyr74) by Li and Wu (2007) in their

---

**Figure 4.** Bilirubin in the active site of the UGT1A1. The image is viewed from the top of the N-terminal domain, and the second envelope helix is shown in blue on the left for orientation. Loop 3 is colored orange, and loop 5 is yellow. Histidine 39 is depicted as a red molecular structure in the middle of the protein. Two identical copies of bilirubin in Z,Z conformation are shown, oriented either more toward loop 3 or loop 5. The CN-I mutation sites in the N-terminal domain are marked by red spheres.

**Table 2**

<table>
<thead>
<tr>
<th>Source</th>
<th>Nβ1</th>
<th>Nα1</th>
<th>Nβ2</th>
<th>Nα2</th>
<th>Nβ3</th>
<th>Nα3</th>
<th>Nβ4</th>
<th>Nα4</th>
<th>Nβ5</th>
<th>Nα5</th>
<th>Nβ6</th>
<th>Nα6</th>
<th>Nβ7</th>
<th>Average*“Average”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Li and Wu, 2007; current study</td>
<td>0</td>
<td>0</td>
<td>-2</td>
<td>-4</td>
<td>-5</td>
<td>+19</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>+8</td>
<td>+7</td>
<td>0</td>
<td>3.75/2.36</td>
<td></td>
</tr>
<tr>
<td>Locuron and Tracy, 2007; current study</td>
<td>0</td>
<td>0</td>
<td>-2</td>
<td>-9</td>
<td>0</td>
<td>+4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>-4</td>
<td>-1</td>
<td>1.75/1.54</td>
<td></td>
</tr>
<tr>
<td>Li and Wu, 2007/Locuson and Tracy, 2007</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>+5</td>
<td>+5</td>
<td>+15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>+9</td>
<td>+11</td>
<td>+1</td>
<td>3.83/2.82</td>
<td></td>
</tr>
</tbody>
</table>

* The averages are calculated from unsigned differences; “average” values are those calculated without helix Nα3.
extensive sequence analysis. Modeling this region is uncertain, but needed for understanding substrate binding. Helix Na3-3 packs to two helices of quite different character: the structurally well conserved Na4 and the poorly defined Na5-2 that corresponds to the hypervariable region IV (Iie215 to Pro229) of Li and Wu (2007). In the template structures, several interhelical contacts are seen within this helix triplet that seem to stabilize a second helical layer above the core αβ structure. This may allow for more coherent movements of the variable part of loop 3 that seems to form one wall of the substrate binding pocket. Small substrates would bind between His39 and UDP-sugar without reaching out to contact loops 3 and 5, whereas larger ones such as bilirubin would touch the loops. Helix Na3-1 would isolate the reaction site from bulk water and move concertedly to allow exit of the products.

The model drew our attention to two potentially interesting residues in loop 5, Cys177 and His173. Cys177 is conserved among UGT1As, and its replacement by arginine or tyrosine causes CN-I syndrome (Seppen et al., 1994; Ghosh et al., 2005). There is another conserved cysteine (Cys186) close by, but a previous study presented evidence against disulfide bonds in UGT1A1 (Ghosh et al., 2005). Neither are disulfides predicted for this segment by sequence analysis (Ceroni et al., 2006). Could Cys177 be palmitoylated and serve as an additional membrane attachment point within the N-terminal domain? Ciotti et al. (1998) predicted a membrane-embedded helix in UGT1A1 between cysteines 156 and 177. In our model, these residues fold into Na4 and Nb5 and continue for five residues into loop 5. A membrane attachment site after Nb5 would agree with all three molecular models of UGT1A1. Another clue to membrane orientation is provided by the predicted N-linked glycosylation sites. There are two commonly glycosylated segments in the UGTs, one in Co0 and another in the short loops between Nb2 and Nb4. When the molecule is oriented so that the putative glycosylation sites are fully solvated, some part of loop 5 probably contacts the membrane (Fig. 3). This proposal agrees with palmitoylated Cys177 forming the actual membrane contact. Nevertheless, because the corresponding residue in the UGTs of subfamilies 2A and 2B is either Ala or Gly, the possible acylation of Cys177 in UGT1As may not be essential for the membrane attachment of the enzymes.

His173 is located right at the beginning of loop 5, where the peptide chain runs past the bound UDP-sugar. His173 sits next to the glucuronic acid and could form a hydrogen bond with its carboxylic group. In most human UGTs the locus is occupied by an arginine. Recently, a serine at the corresponding site was shown to correlate with glucuronic acid specificity in the plant UGT88D, whereas the main determinant for sugar selectivity in that case was Arg350 (Noguchi et al., 2009). Other residues suggested to bind to the carbohydrate group are Arg53 and Arg254 (Radomska-Pandya et al., 2009).

The catalytic residues of the UGTs, a histidine at the beginning of Na1 and an aspartate at the end of Nb4, have been addressed in several studies (Li et al., 2007; Patana et al., 2008; Kerdpin et al., 2009). In UGT1A1, either Asp146 or Asp151 could be the critical acid. Data for corresponding aspartates in UGT1A6 (Li et al., 2007) and UGT1A9 (Patana et al., 2008) are conflicting. In UGT1A6, the Asp-to-Ala mutation abolished activity only at the latter position, whereas in UGT1A9 both mutants retain weak activity. The Km increase for several substrates was larger for the latter position. We have constructed two alternative structures for the Nb4 area. In the first structure, the aspartates are situated at the beginning and end of the sheet, and in the second, they are situated at the end of the sheet and in the following loop. The importance of the pre-Nb4 position (Asp146), faraway from the active site, may be explained by interactions with Nb1 and env2, whereas the post-Nb4 acid (Asp151) polarizes the catalytic histidine.

The UGT database lists 15 different point mutations that lead to a full-length UGT1A1 that is functionally inactive, causing CN-I syndrome. These 15 amino acids were localized in the new UGT1A1 model (Fig. 3), and the surrounding structure was analyzed for possible reasons for the functional failure (Table 3 and Fig. 3). Only three CN-I mutations fall within the N-terminal domain: H39D, C177R, and G276R. Histidine 39 is situated in the structurally well conserved first helix of the protein and plays a major role in the catalytic reaction (Miley et al., 2007; Patana et al., 2008). The C177R mutation falls in the most variable segment of the protein, loop 5. G276R occurs at the beginning of the interdomain linker. Strikingly, Gly276 is fully conserved in most known GT-B structures.

The majority (12 of 15) of the pathological CN-I sites are found in the C-terminal domain. Five of them, G308E, R336W/Q/L, Q357R, S375R, and G395V lie within 5Å of the UDP-sugar and can be considered as active-site mutations, together with H39D. Three additional ones (A368I, S381R, and P387R/S) are situated at the beginning and end of the sheet, and in the second, they are situated at the end of the sheet and in the following loop.

### Table 3

<table>
<thead>
<tr>
<th>Mutation</th>
<th>Localization in the Model</th>
<th>Suggested Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>H39D</td>
<td>First turn of Na1</td>
<td>Catalytic residue</td>
</tr>
<tr>
<td>C177R</td>
<td>Early in loop 5</td>
<td>Possible palmitoylation site</td>
</tr>
<tr>
<td>G276R</td>
<td>In Co0, facing the solvent</td>
<td>Conserved Gly site</td>
</tr>
<tr>
<td>A292V</td>
<td>In Co0, facing the solvent</td>
<td>Inhibits glycosylation at Asn295?</td>
</tr>
<tr>
<td>G308E</td>
<td>In Co0, facing the solvent</td>
<td>Disrupts the binding pocket</td>
</tr>
<tr>
<td>R336W/Q/L</td>
<td>Late in Cβ2, next to UDP</td>
<td>Interferes with the active site</td>
</tr>
<tr>
<td>Q357R</td>
<td>Starts Cα3, binds to uridine</td>
<td>Interferes with the active site</td>
</tr>
<tr>
<td>A368I</td>
<td>Middle of Cβ4, packs to Co6 and Cβ1</td>
<td>Interferes with packing</td>
</tr>
<tr>
<td>S375R</td>
<td>Starts Cα4, at domain interface</td>
<td>Interferes with packing</td>
</tr>
<tr>
<td>S381R</td>
<td>In Co4, packed to Cβ4 and Cβ5</td>
<td>Conserved small residue site</td>
</tr>
<tr>
<td>P387R/S</td>
<td>Starts Cβ5, next to Co6 and env1</td>
<td>Conserved Pro site</td>
</tr>
<tr>
<td>G395V</td>
<td>Before Co6, next to UDP</td>
<td>Destabilizes domain interface</td>
</tr>
<tr>
<td>A401P</td>
<td>In Co5, packs tight to Cβ5 and Cβ6</td>
<td>Interferes with packing</td>
</tr>
<tr>
<td>K425E</td>
<td>In Co6, contacts E424 and N425</td>
<td>Interferes with packing</td>
</tr>
<tr>
<td>W461R</td>
<td>In env2, next to Cβ7</td>
<td>Interferes with packing</td>
</tr>
</tbody>
</table>

Bold residues belong to the signature sequence.
occurs in the UGT signature sequence, Trp354 to Gln397. These residues are well conserved throughout all glycosyltransferases and participate in forming the hydrophobic core of the C domain. The remaining four CN-I mutations are A292V, A401P, K428E, and W461R. Ala292 is located in helix C0, three residues from the predicted N-glycosylation site Asn295. One may thus wonder whether a valine at this position could interfere with residues from the predicted N-glycosylation site Asn295. One may wonder whether a valine at this position could interfere with the hydrophobic packing of the protein. The last CN-I site, Lys428, is situated on the outer surface of the last helix, C6. In the model Lys428 forms salt bridges to Glu424 and Asn425 in the same helix and stabilizes the structure at the domain interphase. The 428E disease mutation could be linked to the interdomain motion.

Evaluating the functional importance of the CN-II or Gilbert syndrome sites is harder, specifically because their genetic background and possible related promoter mutations have not always been studied. Nevertheless, it may be interesting to note that the UGT database lists one Gilbert syndrome mutation in loop 3 (F83L) and one in loop 5 (P229Q). Three CN-II sites in loop 5 (L175Q, Q185P, and R209W) are also in summary. We have modeled the full-length human UGT1A1 and gained new insights to central functional questions. The added features in the present model, the two structurally conserved envelope helices and the N-linked glycosylation groups, focus the analysis to the extensive interdomain contact surfaces and the probing of the luminal-membrane orientation.

Acknowledgments

We thank Dr. J. Ravantti for advice and generous computer resources.

References


