Molecular Pharmacology Fast Forward. Published on March 9, 2010 as doi:10.1124/mol.109.063289
This article has not been copyedited and formatted. The final version may differ from this version.

MOL #63289

# A molecular model of the human UGT1A1, its membrane orientation and the interactions between different parts of the enzyme

Liisa Laakkonen and Moshe Finel

Centre for Drug Research (CDR), Faculty of Pharmacy, University of Helsinki, Finland

Molecular Pharmacology Fast Forward. Published on March 9, 2010 as DOI: 10.1124/mol.109.063289
This article has not been copyedited and formatted. The final version may differ from this version.



RUNNING TITLE: Molecular model of the human UGT1A1

CORRESPONDING AUTHOR: Moshe Finel

PO Box 56 (Viikinkaari 5), 00014 University of Helsinki, Finland

Telephone: +358-9-191-59193; fax+358-9-191- 59556; email: moshe.finel@helsinki.fi



Number of text pages (excluding title pages, including references, etc.): 33

Number of tables: 3

Number of figures: 4

Number of references: 41

Numbers of words in Abstract: 240

Numbers of words in Introduction: 663

Numbers of words in Discussion: 1246

Non-standard abbreviations

ER, endoplasmic reticulum; GT, glycosyl transeferase; pdb, Protein Data Bank; RMSD, root mean square deviation; UGT, UDP-glucuronosyl transeferase

Molecular Pharmacology Fast Forward. Published on March 9, 2010 as DOI: 10.1124/mol.109.063289
This article has not been copyedited and formatted. The final version may differ from this version.



**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 due to 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 4 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-terminal 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.

Molecular Pharmacology Fast Forward. Published on March 9, 2010 as DOI: 10.1124/mol.109.063289
This article has not been copyedited and formatted. The final version may differ from this version.



(Introduction)

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 they are involved in biosynthesis, biotransformations, 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 residues 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 $\alpha/\beta/\alpha$ sandwich domains that, in the GT-A fold, are packed closely together to form a continuous central beta-sheet. In the GT-B fold enzymes, including the human UDP-glucuronosyltransferases (UGTs), the two domains are connected by a flexible linker (Unglil 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-glucuronic acid (UDP-GA) to lipophilic acceptor substrates in an $S_N2$ reaction. The reaction proceeds stepwise: UDP-GA 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 was interpreted as formation of a non-productive enzyme-UDP-substrate complex (Luukkanen et al., 2005).

A single C-terminal transmembrane helix anchors the UGTs to the endoplasmic reticulum (ER) membrane, while most of their mass is on the lumenal side. Secondary membrane

Molecular Pharmacology Fast Forward. Published on March 9, 2010 as DOI: 10.1124/mol.109.063289
This article has not been copyedited and formatted. The final version may differ from this version.



attachment sites have been proposed (Ciotti et al., 1998; Ouzzine et al.,1999). The cellular localization and membrane attachment separate the vertebrate UGTs from the bacterial and plant glycosyltransferases, which are mostly water-soluble proteins in the cytoplasm. The transmembrane helix of the vertebrate UGTs are important for activity (Meech et al., 1996; Kurkela et al., 2004) 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 it agrees well with other known GT-B structures. In complete GT-B structures (see Figure 1), the two major domains are connected by a central extended linker of about 10 residues, and by two C-terminal "envelope" helices, each about 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 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 or II syndrome, or Gilbert syndrome.

Six homology 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

Molecular Pharmacology Fast Forward. Published on March 9, 2010 as DOI: 10.1124/mol.109.063289
This article has not been copyedited and formatted. The final version may differ from this version.



code 2acv), the UGT1A9 model is based on the structure of GtfA from *Amycolatopsis orientalis* (1pn3), while the slightly older UGT73A5 and UGT85B1 models use GtfB of *Amycalotopsis orientalis* as the template (1iir). All 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.

Molecular Pharmacology Fast Forward. Published on March 9, 2010 as DOI: 10.1124/mol.109.063289
This article has not been copyedited and formatted. The final version may differ from this version.



## MATERIALS AND METHODS

The UGT1A1 sequence P22309 from the SwissProt database (http://www.uniprot.org/) was employed 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). 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 Cα's from the seven beta strands and from helices Nα1 and Nα4 (see Fig. S1 in the supplementary materials); and for the C-terminal domain, 73 Cα atoms from the beta strands and helices Cα3, Cα4 and Cα5. 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; beta 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, Nα1 stands for the alpha helix in the N-terminal domain that follows strand β1 of this domain, and Nα3-1 stands for the first segment in the alpha helix that follows strand β3 of the N-terminal domain. The pre-strand helix in the C-domain was called Cα0. The nomenclature resembles those used by others (Unglil and Rini, 2000; Miley et al., 2007; Li and Wu, 2007). Amino acid numbers refer to the full-length human UGT1A1, unless mentioned otherwise.

Molecular Pharmacology Fast Forward. Published on March 9, 2010 as DOI: 10.1124/mol.109.063289
This article has not been copyedited and formatted. The final version may differ from this version.



The 19 human UGT sequences were aligned to each other 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 Nβ4, aligning either D146 or D151 as the aspartate situated right behind the catalytic histidine of the known structures. For residues 83-143, 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 each of them, but the position of the sequences relative to the structures differs by one helical turn. The envelope helices were modeled at 434-448 and 454-467 according to the templates, and the transmembrane helix was constructed as a standard alpha helix at residues 491-516.

Homology models were constructed using the program Modeller 9v6 with a standard modeling scheme, very through annealing and molecular dynamics optimization (Sali and Blundell, 1993). Three templates were used for the N-terminal domain (2vce, 1iir, 2iya) and one for the C-terminal domain (2o6l). The nonhydrolysable UDP-glucose analog (residue name UF2) was copied from the 2vce structure and included in the model as a block residue. Similarly, an N-linked branched sugar (N-acetyl-D-glucosamine)-(N-acetyl-D-glucosamine)-(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/, Gupta et al., in preparation). The residues corresponding to the predicted UGT-glycosylation sites were changed to asparagines in 1iir (Ala44 and Asp235) to allow for the binding of the N-glycosylamine. Helical conformation was constrained between residues Val 491 and Lys 516 for the predicted transmembrane helix (Rost et al., 2004). Fifty structures were modeled and optimized. The majority of these jobs finished successfully and the ten with 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), while 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 (J.J.P. Stewart, Stewart Computational Chemistry, Colorado Springs, CO, USA, HTTP://OpenMOPAC.net_2008) with the PM3 methodology (Stewart, 2007). The substrate structures were docked manually to the final enzyme model between the active site histidine, H39, and the non-hydrolysable UDP-glucose analog.

A list of non-synonymous 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 to the previous ones by Locuson and Tracy (2007) and Li and Wu (2007). The basis for the comparison was a structural alignment between UGT71G1 from *Medicago truncatula* (2avc), and GtfD from *Amycolatopsis orientalis* (1rrv), onto which the UGT1A1 sequences were added manually as given in the publications. A direct structural comparison could not be performed because the coordinates of the earlier models were not available.

Molecular Pharmacology Fast Forward. Published on March 9, 2010 as DOI: 10.1124/mol.109.063289
This article has not been copyedited and formatted. The final version may differ from this version.



## RESULTS

*Structurally related proteins*. We divide the human UGT into 4 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 and, therefore, we first analyzed the available low-homology templates in order to indentify 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 1e-30. The expectation values of the next best proteins were >1. The hit structures were found true homologs also by visual evaluation. Notably, all local similarities between the query and the 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, glycosylepimerases 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

Molecular Pharmacology Fast Forward. Published on March 9, 2010 as DOI: 10.1124/mol.109.063289
This article has not been copyedited and formatted. The final version may differ from this version.



was sequence redundancy among the search results. By sequences, 2acv and 2acw are 100 % identical, and so are 2vg8/2vcu/2vce/2vch, 2c1x/2c9z/2c1z and 3d0q/3d0r groups, respectively. 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 Fig. S1 for structural alignment and original references). Overlaying complete proteins by core beta strands did not give good results, as the interdomain angles vary between structures, but aligning either the N or the C-terminal domains separately yielded good matches. The pairwise RMSD values between the templates for 66 Cα-atoms in the N-terminal domains ranged from 0.58 Å to 2.24 Å, with a mean of 1.42 Å. The corresponding values for73 Cα-atoms in the C-terminal domains were 0.51 Å, 1.81 Å and 1.25 Å, respectively. The same sets of atoms was used for all pairwise comparisons.

The GT-B structures cluster by their biological roles, sequence identities and RMS deviations into four mutually similar groups: i. human UGT2B7 (2o6l), ii. plant UGT enzymes (2vce, 2acv, 2c1x, 2pq6), iii. bacterial glycosyltransferases (1iir, 1rrv, 1pn3, 1f0k, 2p6p), and iv. macrolide glycosyltransferases from *Streptomyces* (2iya, 2iyf, 3d0q). Three structures were selected to serve as a template for modeling the N-terminal domain of UGT1A1, one from each groups ii-iv (2vce, 1iir, 2iya), while 2o6l structure was the template for the C-terminal domain. In addition, helix Nα3-1 was constructed according to structure 2p6p of group ii.

*Analysis of the template structures.* The template proteins fold into canonical GT-B structures of two α/β/α domains with parallel beta strands. Strands and helices alternate throughout the structure, but for few exceptions (see Figure 1). Both the N-terminal and C-terminal β-sheets

Molecular Pharmacology Fast Forward. Published on March 9, 2010 as DOI: 10.1124/mol.109.063289
This article has not been copyedited and formatted. The final version may differ from this version.



follow similar topology, 3-2-1-4-5-6(-7), but they differ in that the former are strongly twisted while the latter are rather flat. The angles between the strands at opposite edges are approximately 110 degrees and 40 degrees in the N-terminal and C-terminal domains, respectively. Despite the similar folds, the two-domain structures vary in length by over 70 amino acids among the thirteen templates. One reason is that in the shorter bacterial enzymes the last helix in the N-terminal domain, Nα6, is replaced by a coil. Major length variations are also found in the segment between Nβ5 and Nβ6.

The core beta strands are structurally well conserved in both domains, but the best conserved helices occur asymmetrically. In the N-terminal domains, the conserved helices are Nα1 and Nα4 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 Cα3, Cα4, and Cα5, lying above the flat beta-sheet and facing the active site at the center of the protein (see Figure 1 b). These helices were used in the structural superpositions in addition to the core beta strands.

Eight of the studied GT-B protein structures are complete, while 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. Besides variations of the angle between the N-terminal 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 like most inter-strand segments, these two long stretches contain several ordered helices. Here we call the complete segments “loop 3” and “loop 5”, and number the included helices sequentially (Fig. 1).

Loop 3 is 42-57 residues long. After Nβ3 the polypeptide chain first continues along the direction of that strand, either as a random coil or as a helix, but 10-15 residues later the chain

Molecular Pharmacology Fast Forward. Published on March 9, 2010 as DOI: 10.1124/mol.109.063289
This article has not been copyedited and formatted. The final version may differ from this version.



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 kinked in the plant proteins. The end halves overlap well structurally, but due to different helix geometries, the starting points differ by approximately 12 Å between the plant and bacterial proteins. The helices are clearly amphipathic: their hydrophobic faces pack to the protein core, and the hydrophilic faces toward the neighboring helices and the solvent. Interestingly, there are many specific interactions between residues in Nα3 and its surroundings in the known proteins. For example, the following pairs are observed in 2vce: S87-N200, T91-D122, R92-D75, R98-E129; in 1iir, T76-R165, F84-H173; and in 2p6p, R86-E71, R90-E7, R98-D55.

Loop 5 between strands Cβ5 and Cβ6 is highly variable; its scatter in length, from 37 to 85 residues, and structural features is larger than of any other part of the studied GT-B proteins. It folds into a compact shape and, like loop 3, it is discontinuous in several crystal structures. The 10 residues closest to the flanking beta strands follow similar paths in the plant and bacterial proteins, whereas the rest of the peptide chain adopts totally different structures in these two groups. In the plant enzymes, the loop starts and ends with ten-residue helices, between which there are tiny helices and strands. The mass of the loop is oriented toward the C-terminal domain of the enzyme. In the bacterial proteins, the body of the loop reaches to the N-terminal domain, and its main structural feature is a long bent helix in the latter part of the loop that packs to the latter half of helices Nα3 and to Nα4. The length of the loop before this helix varies from 5 to 25 amino acids, and reveals no shared characteristics.

Molecular Pharmacology Fast Forward. Published on March 9, 2010 as DOI: 10.1124/mol.109.063289
This article has not been copyedited and formatted. The final version may differ from this version.



*UGTs: from sequences to structures.* The sequence identities of the mature proteins are about 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 due to their high mutual similarity. Their lengths vary by merely two residues, 504-506, and only at seven sites are gaps required in the multiple alignment: after residues 77, 97, 102, 175, 226, 508 and 516 (Fig. S1). The four latter sites are single gaps. Of the 504-506 residues, 145 are identical in all human UGTs. Against this overall similarity, the sequence variation around residues 77, 97-102 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 both in the structural and in the sequence alignment. The residues in the "signature sequence" form a central part of the C-terminal domain, covering Cα3, Cβ4, Cα4 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 (W354, Q357, H372, E380, F394, D396, Q396). The additional conserved residues in the segment (L360, L361, F369, I370, T371, S381, I382, M388) form the tightly packed hydrophobic core of the C-domain.

All beta strands, as well as helices Nα1 and Nα4 and the two envelope helices, align well to the template structures (see Figs. 1 and S1). 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

Molecular Pharmacology Fast Forward. Published on March 9, 2010 as DOI: 10.1124/mol.109.063289
This article has not been copyedited and formatted. The final version may differ from this version.



fully conserved stretch of hydrophobic amino acids at the center of Nβ4, with aspartates at both ends. Neither of the acids is fully conserved among the 13 template proteins, and neither one lacks matching acidic residues. In our preferred sequence-to-structure alignment (Fig S1), the two UGT aspartates are located at the beginning and the end of Nβ4, 16 Å apart from each other. This places D151 next to the active site, in hydrogen bonding distance from H39, while D146 is located at the beginning of strand Nβ4. In the GT-B proteins studied, aspartates at the proposed position for D146 of UGT1A1 form hydrogen bonds to the beginning of Nβ1 and to the end of the second envelope helix. In line with this, D146 in the model comes into contact with K29 and K469. In the alternative alignment the sequence is moved forward by five residues, bringing D146 next to the catalytic histidine. The second aspartate, D151, is now situated in the loop between Nβ4 and Nα4, where it could interact with the substrates. This change is feasible because loop 3 is of sufficient length and low homology, but it leads to alignment in which the center of Nβ4 is no longer universally hydrophobic, and the fork position before Nβ4 is occupied by A141 (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, Nα3-1(R85-V99), Nα3-2 (F105-L130), and Nα3-3 (K134-A141). Most probably, two of these correspond to the helical stretch seen in all the templates. If Nα3-1 and Nα3-2 correspond to the shared helices, the extra helix will fold at the far end of the protein, close to the N-terminus. And, if Nα3-2 and Nα3-3 make for the shared feature, Nα3-1 will lie at the interface between the large domains. We find the latter alternative more likely. Firstly, the templates show structural variability before the long helical stretch, but none after it. Secondly, many residues in loop 3 affect activity in various UGTs and supposedly contribute to the binding site at the domain interface

Molecular Pharmacology Fast Forward. Published on March 9, 2010 as DOI: 10.1124/mol.109.063289
This article has not been copyedited and formatted. The final version may differ from this version.



(Nishiyama et al., 2008; Fujiwara et. al., 2009a and 2009b; Lewis et al., 2007). Accordingly, the predicted helices Nα3-2 and Nα3-3 of UGT1A1 are matched to the two shared loop 3 helices of the templates (Fig. S1). The additional helix, Nα3-1, is built at the domain interface according to an analogous extra helix in 2p6p.

The resulting helical segment Nα3-2/3 is longer than what is observed in the templates (37 versus 23-30 aa) and reveals no obvious sequence-to-structure match. The single conserved property of Nα3s in the studied GT-B proteins is amphiphilicity and, indeed, a periodic separation of hydrophobic and hydrophobic residues is also visible in the UGTs alignment. This dictates the rotational orientation of the sequences to the structure. The gap in the Nα3-2/3 helix predictions (residues 131 and 132) would naturally match the interhelical kink of many templates, in which case $N_{133}K_{134}E_{135}$ of UGT1A1 would align to $N_{89}P_{90}E_{91}$ of 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 2vce to the $L_{130}L_{131}H_{132}$ of UGT1A1, and the third to $C_{127}S_{128}H_{129}$ 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 F206 to V226, followed by two shorter ones at 230-237 and 244-248 (Fig. 1b). This pattern of three helices close to Nβ6 agrees with the structures of the bacterial

Molecular Pharmacology Fast Forward. Published on March 9, 2010 as DOI: 10.1124/mol.109.063289
This article has not been copyedited and formatted. The final version may differ from this version.



proteins, but not with the plant proteins. Hence, 1iir 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 L179-E182.

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

*Non-homology modeling.* To get a complete view of the functional protein, even if some parts are more hypothetical, we have 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 alpha helix from V491 to K516, as predicted. This is longer than the 17 hydrophobic residues, V491-F507, that is 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

Molecular Pharmacology Fast Forward. Published on March 9, 2010 as DOI: 10.1124/mol.109.063289
This article has not been copyedited and formatted. The final version may differ from this version.



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, N295 in Cα0, that is shared by all UGT1A enzymes. In the model, N295 points out of the protein and could easily accommodate a large and mobile sugar unit. The UGT2 enzymes lack the glycosylation site at Cα0, but there is an alternative site eight residues later, corresponding to E303 in UGT1A1. It lies before the start of Cα1, 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 UGT1A3-5 the residues to be glycosylated are N119 and N142 (UGT1A3 numbering), both of which are in loop 3. These residues correspond to K118 and A141 in UGT1A1, that are located in helix Nα3 and right after it. Both are on the surface of the protein, and could be decorated with sugars. Interestingly, N189 of UGT1A3 in loop 5, corresponding to N188 of UGT1A1, is predicted *not* to be glycosylated since it is followed by a proline. Compared to the positively predicted sites, N189 is found on the opposite side of the globular domain, suggesting a possible lumen-membrane orientation for the protein. UGT1A7-10 exhibit one strongly predicted glycosylation site at N71 (corresponding to F73 of UGT1A1), in the loop right before Nβ3. UGT2A1 contains a strongly predicted glycosylation site at N49 before the start of Nβ2 (corresponding to E56 of UGT1A1). Interestingly, the three predicted sites in three different UGTs - 1A3:N142, 1A7:N71, 2A1:N49 - all fall spatially quite close to each other in the model, even though they originate from different parts of the sequences. Taken together, the

Molecular Pharmacology Fast Forward. Published on March 9, 2010 as DOI: 10.1124/mol.109.063289
This article has not been copyedited and formatted. The final version may differ from this version.



data suggests a spatially conserved glycosylation site before either strand Nβ2, Nβ3 or Nβ4. We modeled an N-linked complex sugar to the loop between Nα2 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. Hence, a UDP-glucose analog was copied from one template, 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 loop 3 and loop 5. Not only His39 but the whole helix Nα1 lies quite close to the UDP-sugar and forms an active site wall. Asp36, Ser38 and L41 of Nα1 come within van der Waals distance from the UDP-sugar, and also His376 of the Cα4. There seems to be two alternative orientations for bilirubin binding, either toward loop 5 and the membrane, or toward loop 3 and the ER lumen (Fig. 4). In the former orientation, bilirubin is in contact with residues Q84, D87, K115 and I116, whereas in the latter with K115, I116, A174, V193 and D224. The contact sites are highly tentative, because loops 3 and 5 are largely non-homologous to the templates. It should be mentioned, though, that the substrates in the known glycosyltranseferase structures are found closer to loop 5 (data not shown).

*Comparison to earlier work.* One way to evaluate the current molecular model is to compare it to the earlier ones by Locuson and Tracy (2007) and Li and Wu (2007). These were published so close in time that no previous comparison exists. We have matched the template-target alignments from Li and Locuson to our own alignment, and examined the results. The

Molecular Pharmacology Fast Forward. Published on March 9, 2010 as DOI: 10.1124/mol.109.063289
This article has not been copyedited and formatted. The final version may differ from this version.



combination alignment is shown in the supplementary material (Fig S2) and the differences in the positions of the of secondary structure elements are listed in Table 3. Even without the UGT2B7 structure as a template, the models agree perfectly with each other between residues Q288 and R467. 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, Nα1-Nβ1 and Nβ4-Nα4-Nβ5 at the center of the domain are aligned identically to the template in the three models. For Nβ7 two models agree fully, while the third differs by a single residue. Our model resembles more closely that by Locuson and Tracy (2007). The average unsigned difference between the two models is 1.5 amino acids for the core domain. Surprisingly, Locuson and Tracy have not aligned the structurally conserved Nα2 with anything, even though they have 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 Nα6, by one helical turn. The similarity between the models is acceptable, with the exception of Nα2, and differences of the same magnitude are likely to be 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 from that by Locuson and Tracy. 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 Nα2-Nβ3 and Nβ6-Nα6, and especially for the short beta strands the variations are crucial. It remains to be seen which assignment is correct.

Molecular Pharmacology Fast Forward. Published on March 9, 2010 as DOI: 10.1124/mol.109.063289
This article has not been copyedited and formatted. The final version may differ from this version.



## 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 3 structural domains (Fig. 1) the largest of which, the C-terminal α/β/α sandwich, contains the UDP-GA 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 mainly found in loop 3, in a segment called hypervariable region II (A61-Y74) by Li and Wu in their extensive sequence analysis (2007). Modeling this region is uncertain, but needed for understanding substrate binding. Helix Nα3-3 packs to two helices of quite different character: the structurally well-conserved Nα4, and the poorly defined Nα5-2 that corresponds to the hypervariable region IV (I215-P229) of Li and Wu (2007). In the template structures, several inter-helical contacts are seen within this helix triplet which seem to stabilize a second helical layer above the core α/β structure. This may allow for more coherent movements of the variable part of loop 3 that appears 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, while larger ones like bilirubin would touch the loops. Helix Nα3-1 would isolate the reaction site from bulk water, and move concertedly to allow exit of the products.

Molecular Pharmacology Fast Forward. Published on March 9, 2010 as DOI: 10.1124/mol.109.063289
This article has not been copyedited and formatted. The final version may differ from this version.



The model drew our attention to 2 potentially interesting residues in loop 5, C177 and H173. Cysteine 177 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 close by, C186, but a previous study presented evidence against disulfide bonds in UGT1A1 (Ghosh et al., 2005). Could C177 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 Nα4 and Nβ5, and continue for 5 residues into loop 5. A membrane attachment site after Nβ5 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 Cα0, and another in the short loops between Nβ2-Nβ4. When the molecule is oriented so that the putative glycosylation sites are fully solvated, some part of loop 5 is likely to contact the membrane (Fig. 3). This proposal agrees with palmitoylated C177 forming the actual membrane contact. Nevertheless, since the corresponding residue in the UGTs of subfamilies 2A and 2B is either Ala or Gly, the possible acylation of C177 in UGT1As may not be essential for the membrane attachment of the enzymes.

H173 is located right at the beginning of loop 5, where the peptide chain runs past the bound UDP-sugar. H173 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, while the main determinant for sugar selectivity in that case was R350 (Noguchi et al., 2009). Other residues suggested to bind to the carboxylic group are R53 and R254 (Radominska-Pandya et al., 2009).

Molecular Pharmacology Fast Forward. Published on March 9, 2010 as DOI: 10.1124/mol.109.063289
This article has not been copyedited and formatted. The final version may differ from this version.



The catalytic residues of the UGTs, a histidine at the beginning of Nα1 and an aspartate at the end of Nβ4, have been addressed in several studies (Kerdpin et al., 2009; Li et al., 2007; Patana et al., 2008). In UGT1A1, either D146 or D151 could be the critical acid. Data for corresponding aspartates in UGT1A6 (Li et al., 2007) and UGT1A9 (Patana et al., 2008) are conflicting. In 1A6, the D toA mutation abolished activity only at the latter position, while in 1A9 both mutants retain weak activity while the $K_m$ increase for several substrates was larger for the latter position. We have constructed two alternative structures for the Nβ4 area. In the first structure, the aspartates are situated at the beginning and the end of the sheet, and in the second, at the end of the sheet and in the following loop. The importance of the pre-Nβ4 position (D146), far away from the active site may be explained by interactions with Nβ1 and env2, while the post-Nβ4 acid (D151) polarizes the catalytic histidine.

The UGT database lists 15 different point mutations that lead to a full-length UGT1A1 that is functionally inactive, causing Crigler-Najjar syndrome type I (CN-I). 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 2 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 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, glycine 276 is fully conserved in most known GT-B structures,.

The majority of the pathological CN-I sites, 12/15, are found in the C-terminal domain. Five of them, G308E, R336W/Q/L, Q357R, S375R, G395V lie within 5 Å of the UDP-sugar, and can be considered as active site mutations, together with H39D. Three additional ones,

Molecular Pharmacology Fast Forward. Published on March 9, 2010 as DOI: 10.1124/mol.109.063289
This article has not been copyedited and formatted. The final version may differ from this version.



A368I, S381R and P387R/S occur in the UGT "signature sequence", W354-Q397. These residues are well conserved throughout all glycosyltransferases, and they participate forming the hydrophobic core of the C-domain. The remaining four CN-I mutations are A292V, A401P, K428E and W461R. A292 is located in helix Cα0, three residues from the predicted N-glycosylation site N295. One may thus wonder if a valine at this position could interfere somehow with the N-glycosylation of N295. We propose that mutations of A401, situated in the middle of Cα5, and of W461, early in the second envelope helix, disrupt the hydrophobic packing of the protein. The last CN-I site, K428, is situated on the outer surface of the last helix, Cα6. In the model K428 forms salt bridges to E424 and N425 in the same helix, and stabilizes the structure at the domain interphase. The K428E disease mutation could be linked to the interdomain motion.

Evaluating the functional importance of the CN-II or Gilbert syndrome sites is harder, specifically as 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 are also present, L175Q, Q185P and R209W.

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 glycosyl groups - focus the analysis to the extensive interdomain contact surfaces and the probing of the lumenal-membrane orientation.

Molecular Pharmacology Fast Forward. Published on March 9, 2010 as DOI: 10.1124/mol.109.063289
This article has not been copyedited and formatted. The final version may differ from this version.



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

REFERENCES

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, and Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389-3402.

Barton GJ (1993) ALSCRIPT a tool to format multiple sequence alignments. *Protein Eng* **6**: 37-40.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov I N and, Bourne PE (2000) The Protein Data Bank, *Nucleic Acids Res* **28**: 235-242.

Bosma PJ, Seppen J, Goldhoorn B, Bakker C, Oude Elferink RP, Chowdhury JR, Chowdhury NR, and Jansen PL (1994) Bilirubin UDP-glucuronosyltransferase 1 is the only relevant bilirubin glucuronidating isoform in man. *J Biol Chem* **269**: 17960-17964.

Brazier-Hicks M, Offen WA, Gershater MC, Revett TJ, Lim E-K, Bowles DJ, Davies GJ, and Edwards R (2007) Characterization and engineering of the bifunctional N- and O-glucosyltransferase involved in xenobiotic metabolism in plants. *Proc Natl Acad Sci U S A* **104**: 20238-43.

Ceroni A, Passerini A, Vullo A, and Frasconi P (2006) DISULFIND: a Disulfide Bonding State and Cysteine Connectivity Prediction Server. *Nucleic Acids Res* **34**: W177-W181.

Ciotti M, Cho JW, George J, and Owens IS (1998) Required Buried a-Helical Structure in the Bilirubin UDP-Glucuronosyltransferase UGT1A1, Contains a Nonreplaceable Phenylalanine. *Biochemistry* **37**: 11018-11025.

Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, Murray LW, Arendall WB III, Snoeyink J, Richardson JS, and Richardson DC (2007) MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res* **35**: W375-W383.

Finel M and Kurkela M (2008) The UDP-Glucuronosyltransferases as Oligomeric Enzymes. *Curr Drug Metab* **9**: 70-76.

Fujiwara R, Nakajima M, Yamamoto T, Nagao H, Yokoi T (2009a) In silico and in vitro approaches to elucidate the thermal stability of human UDP-glucuronosyltransferase (UGT) 1A9. Drug Metab Pharmacokinet **24**:235-244.

Fujiwara R, Nakajima M, Yamanaka H, and Yokoi T (2009b) Key Amino Acid Residues Responsible for the Differences in Substrate Specificity of Human UDP-Glucuronosyltransferase (UGT)1A9 and UGT1A8. *Drug Metab Dispos* **37**: 41-46.

Ghosh SS, Lu Y, Lee SW, Wang X, Guha C, Roy-Chowdhury J, and Roy-Chowdhury N (2005) Role of cysteine residues in the function of human UDP glucuronosyltransferase isoform 1A1 (UGT1A1). *Biochem J* **392**: 685-692.

Hans J, Brandt W, and Vogt T (2004) Site-directed mutagenesis and protein 3D-homology modelling suggest a catalytic mechanism for UDP-glucose-dependent betanidin 5-O-glucosyltransferase from *Dorotheanthus bellidiformis. Plant J* **39***:* 319-33.

Holm L, Kääriäinen S, Rosenström P, and Schenkel A (2008) Searching protein structure databases with DaliLite v.3. *Bioinformatics* **24**: 2780-2781.

Humphrey W, Dalke A, and Schulten K (1996) VMD - Visual Molecular Dynamics. *J Mol Graphics* **14**: 33-38.

Kerdpin O, Mackenzie PI, Bowalgaha K, Finel M, and Miners JO (2009) Influence of N-terminal domain histidine and proline residues on the substrate selectivties of human UDP-glucuronosyltransferase (UGT) 1A1, 1A6, 1A9, 2B7 and 2B10. *Drug Metab Dispos* **37**: 1948-55.

Kurkela M, Hirvonen J, Kostiainen R, and Finel M (2004a) The interactions between the N-terminal and C-terminal domains of the human UDP-glucuronosyltransferases are partly isoform-specific, and may involve both monomers. *Biochem Pharmacol* **68***:* 2443–245.

Kurkela M, Mörsky S, Hirvonen J, Kostiainen R, and Finel M (2004b) An active and water-soluble truncation mutant of the human UDP-glucuronosyltransferase 1A9. *Mol Pharmacol* **65**: 826-31.

Lairson LL, Henrissat B, Davies GJ, and Withers SG (2008) Glycosyltransferases: Structures, Functions, and Mechanisms. *Annu Rev Biochem* **77**: 521–55.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, and Higgins DG (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* **23**: 2947-2948.

Lewis BC, Mackenzie PI, Elliot DJ, Burchell B, Bhasker CR, and Miners JO (2007) Amino terminal domains of human UDP-glucuronosyltransferases (UGT) 2B7 and 2B15 associated with substrate selectivity and autoactivation. *Biochem Pharmacol* **73**:1463-1473.

Li C and Wu Q, (2007) Adaptive evolution of multiple-variable exons and structural diversity of drug-metabolizing enzymes. *BMC Evol Biol* **7**: 69.

Li D, Fournel-Gigleux S, Barre L, Mulliert G, Netter P, Magdalou J, and Ouzzine M (2007) Identification of Aspartic Acid and Histidine Residues Mediating the Reaction Mechanism and the SubstrateSpecificity of the Human UDP-glucuronosyltransferases 1A. *J Biol Chem* **282**: 36514-24.

Locuson CW and Tracy TS (2007) Comparative modelling of the human UDP-glucuronosyltransferases: Insights into structure and mechanism. *Xenobiotica* **37***:* 155–168.

Luukkanen L, Taskinen J, Kurkela M, Kostiainen R, Hirvonen J, and Finel M (2005) Kinetic characterization of the 1A subfamily of recombinant human UDP-glucuronosyltransferases. *Drug Metab Dispos* **33**: 1017-26.

Mackenzie PI, Bock KW, Burchell B, Guillemette C, Ikushiro S, Iyanagi T, Miners JO, Owens IS, and Neberth DW (2005) Nomenclature update for the mammalian UDP glycosyltransferase (UGT) gene superfamily. *Pharmacogenet Genomics* **15**: 677–685.

Meech R, Yogalingam G, and Mackenzie PI (1996) Mutational analysis of the carboxy-terminal region of UDP-glucuronosyltransferase 2B1. *DNA Cell Biol* **15**:489-494.

Miley MJ, Zielinska AK, Keenan JE, Bratton SM, Radominska-Pandya A, and Redinbo MR (2007) Crystal Structure of the Cofactor-Binding Domain of the Human Phase II Drug-Metabolism Enzyme UDP-Glucuronosyltransferase 2B7. *J Mol Biol* **369**: 498–511.

Nishiyama T, Fujishima M, Masuda Y, Izawa T, Ohnuma T, Ogura K, and Hiratsuka A (2008) Amino acid positions 69–132 of UGT1A9 are involved in the C-glucuronidation of phenylbutazone. *Arch Biochem Biophys* **478**: 75–80.

Nogales D and Lightner DA (1995) On the structure of bilirubin in solution. 13C[1H] heteronuclear Overhauser effect NMR analyses in aqueous buffer and organic solvents. *J Biol Chem* **270**: 73-7.

Noguchi A, Horikawa M, Fukui Y, Fukuchi-Mizutani M, Iuchi-Okada A, Ishiguro M, Kiso Y, Nakayama T, and Ono E (2009) Local Differentiation of Sugar Donor Specificity of Flavonoid Glycosyltransferase in Lamiales. *Plant Cell* **37**: 1948-55.

Osmani SA, Bak S, Imberty A, Olsen CE, and Lindberg Møller B (2008) Catalytic Key Amino Acids and UDP-Sugar Donor Specificity of a Plant Glucuronosyltransferase, UGT94B Molecular Modeling Substantiated by Site-Specific Mutagenesis and Biochemical Analyses. *Plant Physiol* **148**: 1295–1308.

Ouzzine M, Magdalou J, Burchell B, and Fournel-Gigleux S (1999) An internal signal sequence mediates the targeting and retention of the human UDP-glucuronosyltransferase 1A6 to the endoplasmic reticulum. *J Biol Chem* **274**:31401-31409.

Patana AS, Kurkela M, Finel M, and Goldman A. (2008) Mutation analysis in UGT1A9 suggests a relationship between substrate and catalytic residues in UDP-glucuronosyltransferases. *Protein Eng Des Sel* **9**: 537-43.

Rost B, Yachdav G, and Liu J (2004) The PredictProtein Server. Nucleic Acids Res 32: W321-W326.

Sali A and Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **234**: 779-815.

Seppen J, Bosma PJ, Goldhoorn BG, Bakker CTM, Roy-Chowdhury J, Roy-Chowdhury N, Jansen PLM, and Oude Elferink RPJ (1994) Discrimination between Crigler-Najjar Type I and II by Expression of Mutant Bilirubin Uridine Diphosphate-Glucuronosyltransferase. *J Clin Invest* **94**: 2385-2391.

Stewart JJP (2007) Optimization of Parameters for Semiempirical Methods V: Modification of NDDO Approximations and Application to 70 Elements. *J Mol Model* **13**: 1173-1213.

Thorsøe KS, Bak S, Olsen CE, Imberty A, Breton C, and Lindberg Møller B (2005) Determination of Catalytic Key Amino Acids and UDP Sugar Donor Specificity of the Cyanohydrin Glycosyltransferase UGT85B1 from *Sorghum bicolor*. Molecular Modeling Substantiated by Site-SpecificMutagenesis and Biochemical Analyses. *Plant Physiol* **139**: 664–673.

Ünligil UM and Rini JM (2000) Glycosyltransferase structure and mechanism. *Curr Opin Struct Biol* **10**: 510–517.

White SH (2009) Biophysical dissection of membrane proteins. *Nature* **459***:* 344-6.



FOOTNOTE

This work was supported by the Sigrid Juselius Foundation and the Magnus Ehrnrooth foundation.

Molecular Pharmacology Fast Forward. Published on March 9, 2010 as DOI: 10.1124/mol.109.063289
This article has not been copyedited and formatted. The final version may differ from this version.



FIGURE CAPTIONS

Fig. 1. Structure of the UGT1A1. a) One of the GT-B fold templates, UGT72B1 from *A. thaliana* (pdb code 2vce, *14*). Helices are shown as spirals and beta strands as arrows. The N-terminal domain is on the top and the C-terminal domain on the bottom., and the envelope helices on the left. The core beta strands are shaded dark. The space filling molecule between the domains is the UDP-sugar bound in the active site. The molecular images were rendered with VMD (Humphrey, 1996). b) Predicted topology diagram for UGT1A1. Beta strands are represented by triangles, core helices as spheres and envelope helices by rectangles. Well-conserved structural elements are drawn with thick lines. Helices in loop 5 are not shown individually. The N- and C-termini are marked with N and C, respectively, and the four structural domains discussed in the text are separated by dashed lines. The figure was drawn with Xfig, http://www.xfig.org/. c) The position of the secondary structure elements in the UGT1A1 sequence, based on the structural alignment of known GT-B structures and all human UGT sequences. Strands are marked with arrows and helices with dashed lines.

Fig. 2. Internal variation among the ten 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 in green, and the envelope helices and the transmembrane helix blue. The gray space-filling molecule in the middle is the UDP-

Molecular Pharmacology Fast Forward. Published on March 9, 2010 as DOI: 10.1124/mol.109.063289
This article has not been copyedited and formatted. The final version may differ from this version.



sugar. The pink stick-structures represent the N-linked sugars. The one on the right is attached to N295 of UGT1A1, and the one on left, shown in 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.

Fig. 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 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 toward loop 5. The CN-I mutation sites in the N-terminal domain are marked by red spheres.

Molecular Pharmacology Fast Forward. Published on March 9, 2010 as DOI: 10.1124/mol.109.063289
This article has not been copyedited and formatted. The final version may differ from this version.



Table1. Structural relatives of the human UGT1A1. Only unique sequence hits are shown. For the PSI-Blast search, the expectation value is given, and for the Dali searches the z-score. 2o6l was used as the query in the Dali search. First authors are given of modeling work where a structure has been used as a template.

| PDB ID, source, | PSI-Blast | Dali | Used as a template |
|---|---|---|---|
| 2ACV, *Medicago truncatula* triterpene UGT71G | 1e-51 | 21.2 | Locuson, Li, Osmani |
| 2C1X, *Vitis vinifera*, UDP-glucose flavonoid GT | 2e-50 | 20.7 | |
| 1IIR, *Amycolatopsis orientalis* GtfB | 5e-50 | 19.3 | This work, Hans, Thorsoe |
| 2PQ6, *Medicago truncatula* multifunctional GT | 7e-46 | 21.3 | |
| 2IYA, *Streptomyces antibioticus* macrolide GT | 2e-45 | 21.3 | This work |
| 2IYF, *Streptomyces antibioticus* macrolide GT | 6e-45 | 20.3 | |
| 1RRV, *Amycolatopsis orientalis* GtfD | 1e-44 | 19.6 | Locuson and Tracy |
| 2O6L, *Homo sapiens*, UGT2B7 | 1e-44 | 36.1 | This work |
| 1PN3, *Amycolatopsis orientalis* GtfA | 2e-39 | 18.5 | Thorsoe, Fujiwara |
| 2VCE, *Arabidopsis thaliana* UGT72B1 | 1e-42 | 20.3 | This work, Osmani |
| 2P6P, *Streptomyces fradiae* C-C bond forming GT | 2e-39 | 16.8 | This work for Nα3-1 |
| 3D0Q, *Micromonaspora echinospora* CALG3 | 9e-39 | 17.8 | |
| 1F0K, *Eschericia coli* MurG | 4e-36 | 14.1 | |

Molecular Pharmacology Fast Forward. Published on March 9, 2010 as DOI: 10.1124/mol.109.063289
This article has not been copyedited and formatted. The final version may differ from this version.



Table 2. CN-I point mutations, their localization in the model, and hypothetical function. Residues printed in bold face belong to the "signature sequence".

| Mutation | Localization in the model | Suggested function |
|---|---|---|
| H39D | First turn of Nα1 | Catalytic residue |
| C177R | Early in loop 5 | Possible palmitoylation site |
| G276R | In the linker, next to env1 & env2 | Conserved Gly-site |
| A292V | In Cα0, facing the solvent | Inhibits glycosylation at N295? |
| G308E | Right after Cβ1, packing to UDP | Disturbs the binding pocket |
| R336W/Q/L | Late in Cβ2, next to UDP | Interferes with the active site |
| **Q357R** | Starts Cα3, binds to uridine | Disturbs the binding pocket |
| **A368I** | Middle of Cβ4, packs to Cα6& Cβ1 | Interferes with packing |
| **S375R** | Starts Cα4, at domain interface | Interferes with packing |
| **S381R** | In Cα4, packed toCβ4 and Cβ5 | Conserved small residue site |
| **P387R/S** | Starts Cβ5, next to Cα6 & env1 | Conserved Pro-site |
| **G395V** | Before Cα5, next to UDP | Disturbs the binding pocket |
| A401P | In Cα5, packs tight to Cβ5 and Cβ6 | Interferes with packing |
| K428E | In Cα6, contacts E424 and N425 | Destabilizes domain interface |
| W461R | In env2, next to Cβ7 | Interferes with packing |

Molecular Pharmacology Fast Forward. Published on March 9, 2010 as DOI: 10.1124/mol.109.063289
This article has not been copyedited and formatted. The final version may differ from this version.



Table 3. Difference in the positions of the secondary structure elements in the N-terminal domain in three homology models of the human UGT1A1. The averages (ave) are calculated from unsigned differences, and ave' denotes the average calculated without helix Nα3.

| | Nβ1 | Nα1 | Nβ2 | Nα2 | Nβ3 | Nα3 | Nβ4 | Nα4 | Nβ5 | Nβ6 | Nα6 | Nβ7 | ave /ave' |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Li/This work | 0 | 0 | -2 | -4 | -5 | +19 | 0 | 0 | 0 | +8 | +7 | 0 | 3.75/2.36 |
| Locuson/This work | 0 | 0 | -2 | -9 | 0 | +4 | 0 | 0 | 0 | -1 | -4 | -1 | 1.75/1.54 |
| Li/Locuson | 0 | 0 | 0 | +5 | +5 | +15 | 0 | 0 | 0 | +9 | +11 | +1 | 3.83/2.82 |

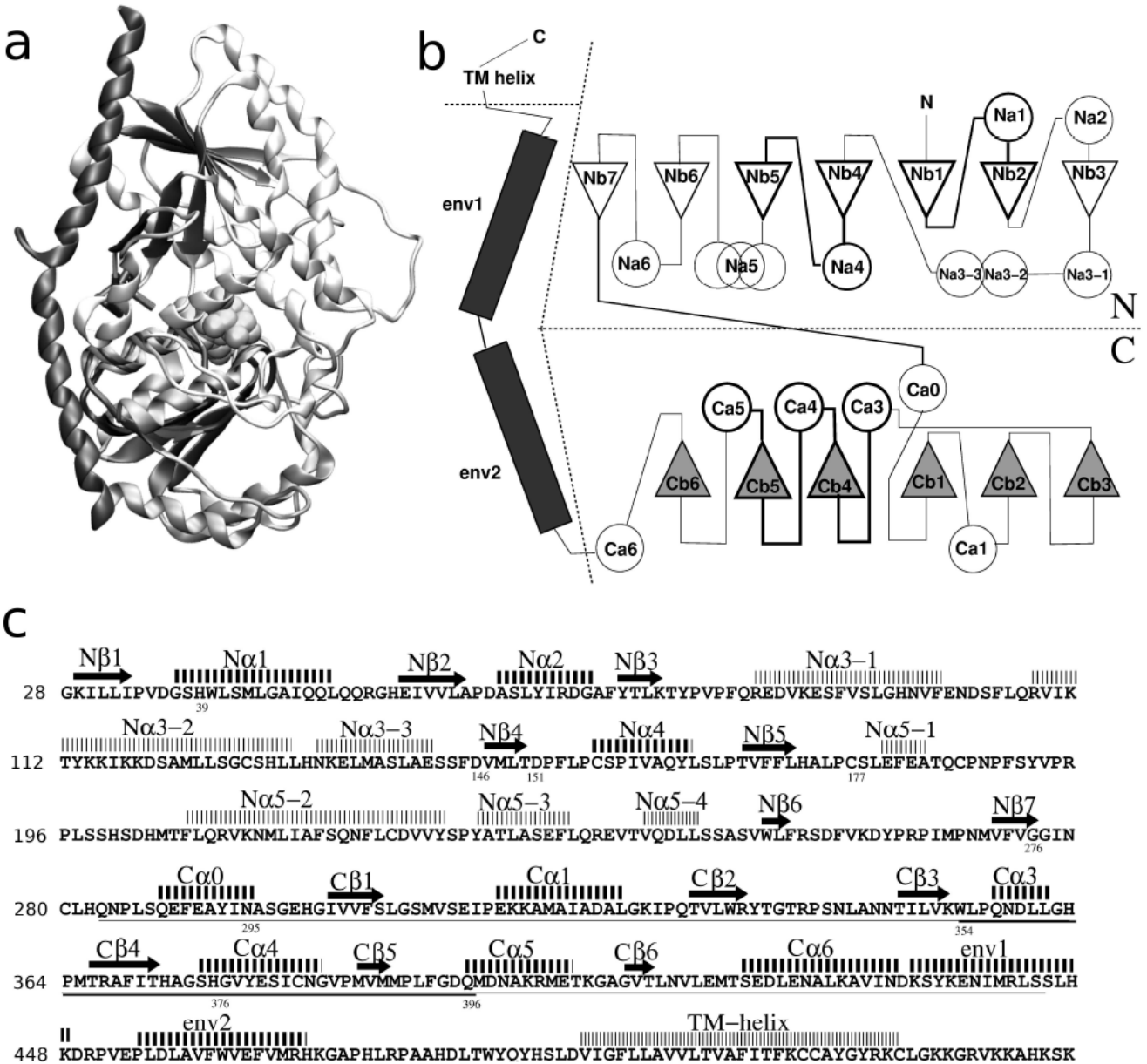


Fig. 1

Molecular Pharmacology Fast Forward. Published on March 9, 2010 as DOI: 10.1124/mol.109.063289
This article has not been copyedited and formatted. The final version may differ from this version.


Fig. 2

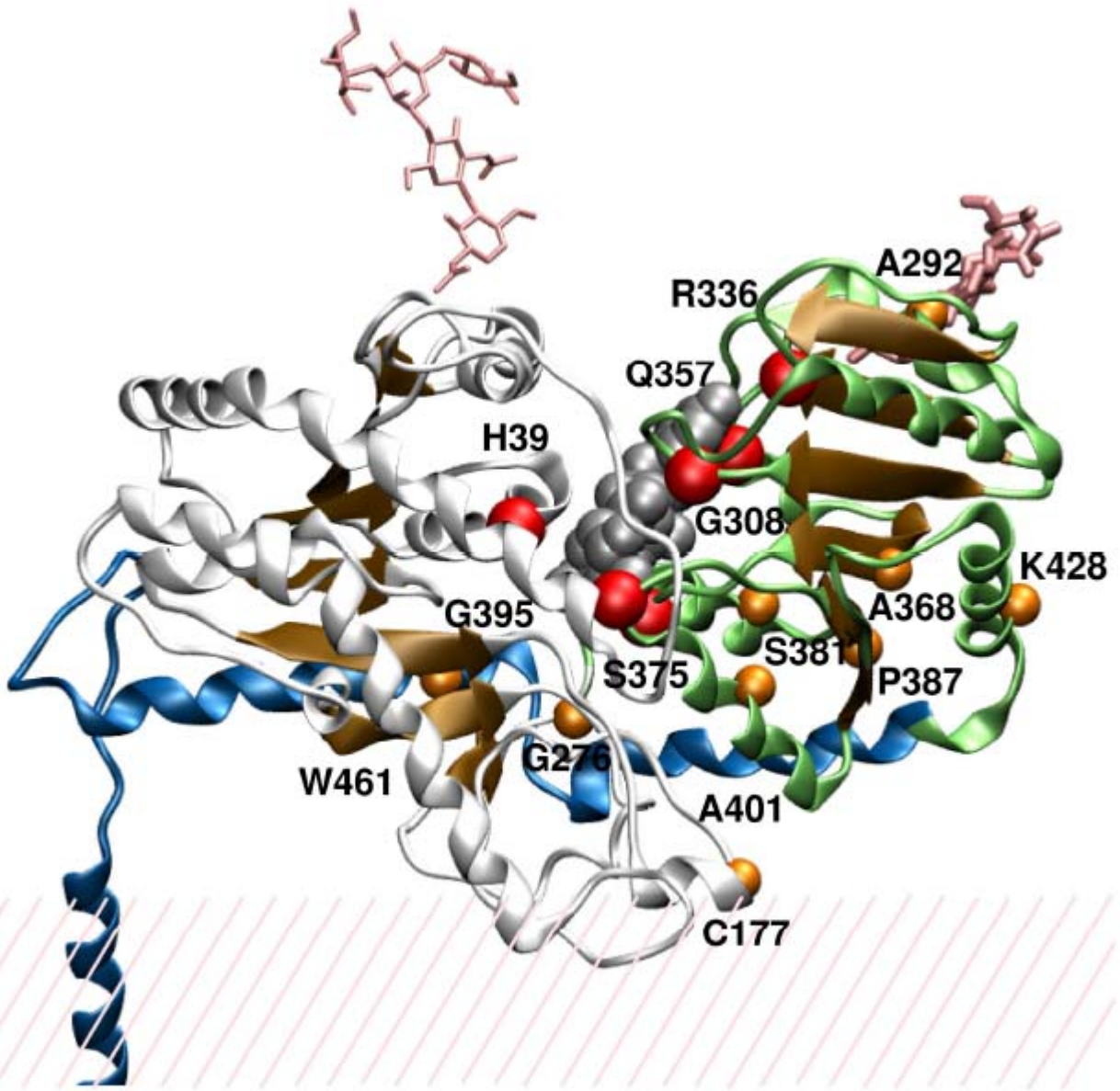


Fig. 3

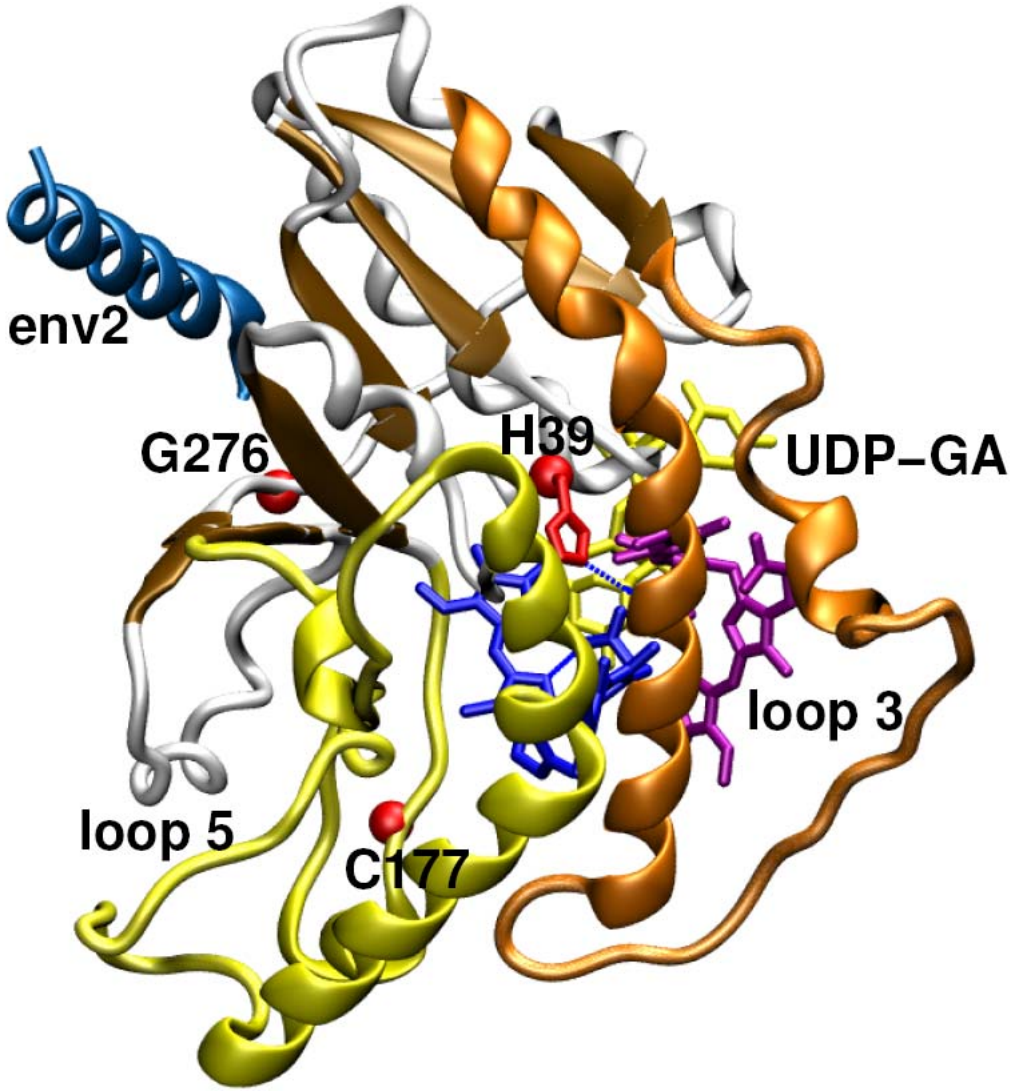


Fig. 4.