

# Class-dependent sequence alignment strategy improves the structural and functional modeling of P450s

Jerome Baudry<sup>1,4</sup>, Sanjeewa Rupasinghe<sup>2,4</sup> and Mary A. Schuler<sup>2,3</sup>

<sup>1</sup>School of Chemical Sciences, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA and <sup>2</sup>Department of Cell and Developmental Biology, University of Illinois at Urbana-Champaign, 1201 W. Gregory Drive, 161 Edward R. Madigan Laboratory, Urbana, IL 61801, USA

<sup>3</sup>To whom Correspondence should be addressed.  
E-mail: maryschu@uiuc.edu

<sup>4</sup>These authors contributed equally to the work.

**Different procedures for obtaining homology models for P450s are investigated using various sequence alignments sharing various levels of sequence identity with available P450 crystal structures. In this analysis, we have investigated how well homology modeling can reproduce known crystal structures as well as how effectively these homology models can be used to reproduce known ligand-binding modes. Homology models obtained from sequence alignments that discriminate between Class I and Class II P450s are significantly closer to the experimental crystal structures and more closely reproduce known ligand's binding modes, than those obtained using sequence alignments that combine Class I and Class II P450s. The quality of the models is slightly improved by constructing hybrid-structure models that model three of the most variable regions of P450s independently from the rest of the protein: the B region that includes SRS1, the FG region that includes SRS2 and SRS3 and the  $\beta$ 4 region that includes SRS6.**

**Keywords:** Molecular modeling/cytochrome P450 monooxygenases/P450s

## Introduction

Cytochrome P450 monooxygenases (P450s) represent a particularly important class of proteins found in all major branches of the evolutionary tree (e.g. bacteria, fungi, plants, insects, animals). Of the over 5000 P450 gene sequences currently defined (<http://drnelson.utmem.edu/cytochromeP450.html>), those with known functions span the spectrum of life's biochemical reactions and include hydroxylations on aliphatic and aromatic carbons, hydroxylations on nitrogen and sulfur heteroatoms, dehalogenations, dealkylations, deaminations and epoxidations. Many of the targets for these types of reactions are drugs that are inactivated by P450-mediated activities, plant chemicals that are detoxified by the insects and animals that encounter them and intermediates in biosynthetic pathways that generate bioactive molecules (e.g. steroid hormones, signaling molecules, etc.) and cellular components (e.g. fatty acids, phenylpropanoids, etc.).

Despite the biochemical activities of these enzymes, relatively few crystal structures exist with 17 soluble bacterial P450s and six membrane-bound mammalian P450s now defined (Table I). One of the primary limitations in the

structural analysis of membrane-bound P450s has been the fact that these endoplasmic reticulum-localized proteins have their N-terminal signal sequence anchored in the membrane and hydrophobic residues in the FG loop (between the F- and G-helices) associated with the cytosolic side of the ER membrane. Success in the arena of crystallizing mammalian microsomal proteins has come after reengineering their sequences for expression in *Escherichia coli* by replacing the N-terminal anchor domain with a KKTSSKG/K sequence (von Wachenfeldt *et al.*, 1997; Williams *et al.*, 2000, 2003; Scott *et al.*, 2001, 2003; Wester *et al.*, 2003; Schoch *et al.*, 2004) and, for CYP2C5 and CYP2C9, mutagenizing seven amino acids in the FG region to eliminate potential interactions with the lipid bilayer (von Wachenfeldt *et al.*, 1997; Williams *et al.*, 2003). The one exception to these procedures has been CYP3A4 whose crystallization was achieved after deletion of residues 3–23 without replacement by additional sequences (Yano *et al.*, 2004).

Given the large number of P450s that exist and the limited number that can be defined at the level of crystal structure, it is increasingly important to be able to turn from primary sequences to modeled structures that are robust enough to rationalize experimental observations and make predictions on potential substrates and inhibitors accommodated in each catalytic site (deGraaf *et al.*, 2005). This task is complicated by the fact that the large P450 superfamily has evolved into a collection of P450s often sharing less than 15% amino acid identity with one another (<http://drnelson.utmem.edu/cytochromeP450.html>). Despite this diversity, P450 sequences have managed to maintain secondary and tertiary structures that generate core structures containing  $\alpha$ -helices (labeled A–K) and  $\beta$ -pleated sheets (labeled 1–4) surrounding their buried catalytic sites (Graham and Peterson, 1999; Stout, 2004). While implying that there are many primary sequence 'solutions' for forming three-dimensional P450 catalytic sites, this diversity confounds the molecular modeling process since having the right overall folds is not enough to ensure that the catalytic site will be modeled appropriately.

While the core structure maintains many secondary and tertiary structural components, some regions designated as SRS (substrate recognition sites; Gotoh, 1992) are known to be especially diverse and important in a number of mammalian P450 activities. Studies, which have highlighted the importance of these regions, have largely dealt with members of the human CYP2 family and human CYP3A4 because of their roles in a variety of drug metabolisms (Domanski and Halpert, 2001). Amino acids designated to exist in SRS include those in the loop region between the B and C helices (SRS1), the C-terminal end of the F helix (SRS2), part of the FG loop and N-terminal end of the G helix (SRS3), the I helix containing SRS4 extending over the pyrrole ring B in the active site (SRS4), the loop between the K helix and strand 4 of  $\beta$ -sheet 1 (SRS5) and the  $\beta$ -turn in  $\beta$ -sheet

**Table I.** P450 crystal structures

	P450	Organism	PDB ID	References	
Class I	<b>CYP102</b> (P450BM3)	<i>Bacillus megaterium</i>	2HPD, 1BU7, 1BVY	Ravichandran <i>et al.</i> (1993)	
	<b>CYP2C8</b>	<i>Homo sapiens</i>	<u>1PQ2</u>	Schoch <i>et al.</i> (2004)	
	<b>CYP2C9</b>	<i>H. sapiens</i>	1OG5, 1OG2, 1R9O	Williams <i>et al.</i> (2003)	
	<b>CYP3A4</b>	<i>H. sapiens</i>	<u>1TQN</u> , 1W0E, 1WOF, 1W0G	Yano <i>et al.</i> (2004)	
	<b>CYP2C5</b>	<i>Oryctolagus cuniculus</i>	<u>1N6B</u> , 1DT6	Wester <i>et al.</i> (2003)	
	<b>CYP2B4</b>	<i>O. cuniculus</i>	1SU0, 1PO5	Scott <i>et al.</i> (2004)	
	<b>CYP2A6</b>	<i>H. sapiens</i>	1Z10, 1Z11	Yano <i>et al.</i> (2005)	
	<b>CYP175A1</b>	<i>Thermus thermophilus</i>	1N97, 1WIY	Yano <i>et al.</i> (2003)	
	Class II	<b>CYP165B1</b> (OxyB)	<i>Amycolatopsis orientalis</i>	<u>1LFK</u>	Zerbe <i>et al.</i> (2002)
		<b>CYP165C1</b> (OxyC)	<i>A.orientalis</i>	1UED	Pylypenko <i>et al.</i> (2003)
		CYP51B1-Mt	<i>Mycobacterium tuberculosis</i>	1X8V, 1E9X 1EA1, 1HSZ	Podust <i>et al.</i> (2004)
<b>CYP121</b>		<i>M.tuberculosis</i>	1N40	Leys <i>et al.</i> (2003)	
<b>CYP167A1</b> (P450epok)		<i>Polyangium cellulosum</i>	1Q5D, 1PKF	Nagano <i>et al.</i> (2003)	
<b>CYP101</b> (P450cam)		<i>Pseudomonas putida</i>	<u>2CPP</u> , <u>1AKD</u> , 1PHC	Poulos <i>et al.</i> (1987)	
<b>CYP108</b> (P450terp)		<i>Pseudomonas</i> sp	<u>1CPT</u>	Hasemann <i>et al.</i> (1994)	
<b>CYP107A1</b> (P450eryf)		<i>Saccharopolyspora erythraea</i>	1OXA, 1EGY	Cupp-Vickery and Poulos (1995)	
<b>CYP154C1</b>		<i>Streptomyces coelicolor</i>	<u>1GWI</u>	Podust <i>et al.</i> (2003)	
<b>CYP154A1</b>		<i>S. coelicolor</i>	1ODO	Podust <i>et al.</i> (2004)	
<b>CYP158A2</b>		<i>S. coelicolor</i>	1S1F	Zhao <i>et al.</i> (2005)	
CYP119		<i>Sulfolobus solfataricus</i>	1I07, 1I08, 1I09, 1F4T	Park <i>et al.</i> (2000)	
P450st		<i>Sulfolobus tokodaii</i>	1UE8	Oku <i>et al.</i> (2004)	
Class III		<b>CYP152A1</b> (P450BS-β)	<i>Bacillus subtilis</i>	1IZO	Lee <i>et al.</i> (2003)
	<b>CYP55A2</b> (P450nor)	<i>Fusarium oxysporum</i>	1EHE, 1CL6, 1GED	Shimizu <i>et al.</i> (2000)	

Underlined PDB entries were used in the homology/crystal structure comparisons.

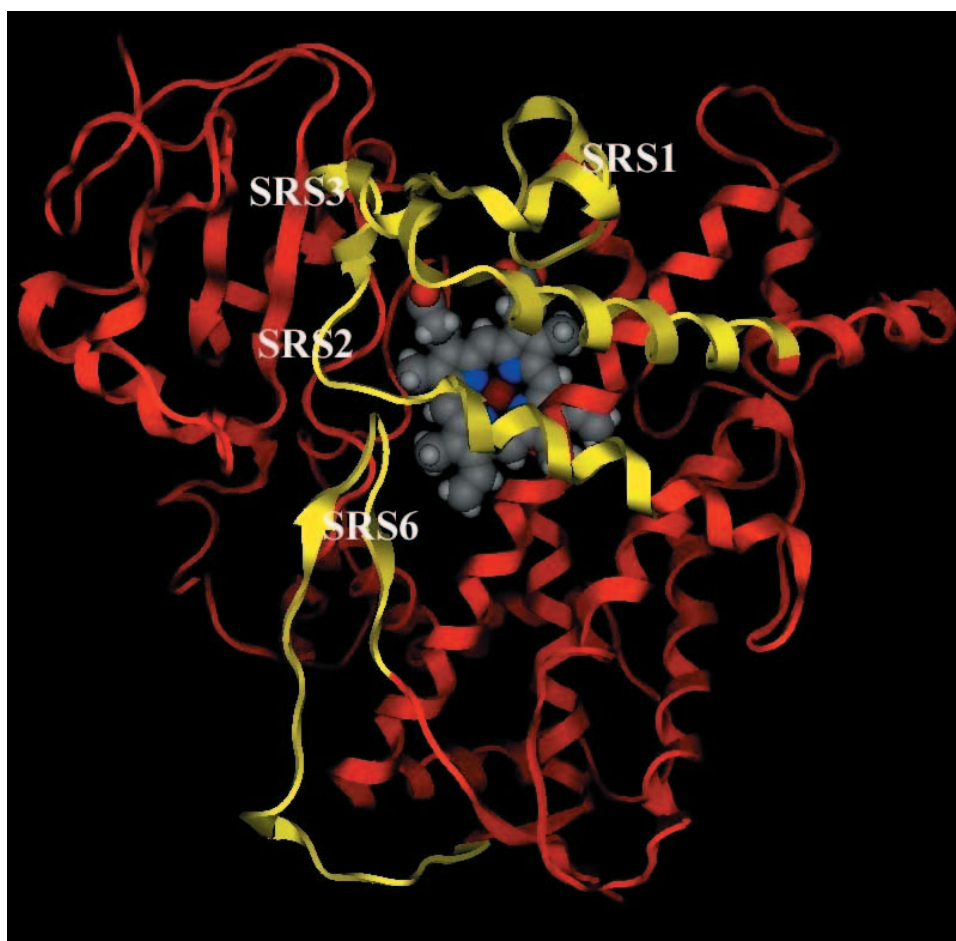
4 (SRS6). Site-directed mutagenesis within these regions which protrude into the catalytic site over the heme have identified various residues as important for defining substrate reactivities (Domanski and Halpert, 2001).

To improve the structural and functional modeling of P450s whose crystal structures may never be defined in the near future, strategies that maximize the amount of information contained in the already-known P450 structures are critical for limiting the range of predicted structures. Previous efforts aimed at designing a general strategy for robust comparative modeling all pointed to the critical importance of relevant sequence alignment of potential templates for homology modeling. The earliest systematic evaluation of sequence alignment procedures indicated that, for the modeling of CYP105C1 (Chang and Loew, 1996) and CYP3A4 (Szkларz and Halpert, 1997), it was important to align multiple sequences rather than to maximize the pairwise alignment of a target with a single template. Even with the relatively low number of available structures included in the alignments (four bacterial sequences in both studies cited above), it was suggested in Chang and Loew (1996) that a well-defined and manually checked sequence alignment (assuring that structurally conserved regions were also sequentially aligned) could serve as a ‘universal’ alignment to be used for modeling of any P450. The work by Szkларz and Halpert (1997) further suggested that some less structurally conserved parts of the target could be modeled individually from template structures that were not necessarily those used to model the SCRs (Structurally Conserved Regions) of the rest of the target. In both of these studies, the quality of the modeling strategies and structures modeled was defined by the ability of known ligands to dock in the modeled structures. This integrated approach, that combines structural and functional *in silico* approaches, is the focus of modern developments in P450 modeling (deGraaf *et al.*, 2005).

In more recent modeling studies that have systematically compared several other alignment and modeling strategies,

the use of a slightly more expanded set of template sequences (four bacterial sequences and the mammalian CYP2C5 sequence) again indicated that a multiple sequence alignment with the target led to better quality models than a single sequence alignment with the target. This was true even when, in the modeling of CYP2D6 (Kirton *et al.*, 2002a), one of the template sequences was significantly more homologous to the target than the other sequences used in the alignment. As in Szkларz and Halpert (1997), comparison of the homology model of CYP2C5 with the experimental crystal structure of the same protein (Kirton *et al.*, 2002b) suggested that there was potentially a need for a ‘fragment-based’ approach modeling SCRs and structurally diverse regions (SDRs) using different templates. The criteria used in Kirton *et al.* (2002b) to judge the structural quality of their models were the C $\alpha$  root mean square distance (RMSD) between the modeled and the experimental CYP2C5 structures. Models built using alignments of multiple sequences exhibited C $\alpha$  RMSD (in the 4.7–5.6 Å range) lower than that of models from alignments of a single sequence (RMSD of 6.3 Å). As was the case in the earlier studies, the capacities of the modeled structure to bind known ligands were used as essential criteria to assess the quality of each model.

To further enhance the predictive capabilities of P450 models, we have developed modeling strategies that accommodate the structural diversity of P450s by subdividing the 20 available P450 crystal structures into individual classes (Table I) based on their known/presumed electron transfer partners and by modeling three of the SDRs in the target sequence independently from SCRs. In initiating this study, our structural alignments of the available crystal structures showed that the C $\alpha$  atoms of three domains harboring SRS exist outside an RMSD of 2.5 Å. As depicted in Figure 1, these structurally diverse regions include: (i) the loop between strand 5 of  $\beta$ -sheet 1, B' helix, B helix and B–C loop (harbors SRS1) (collectively referred to as the B region in the following text); (ii) the C-terminus of the F helix, the FG loop



**Fig. 1.** Structurally variant regions. The backbone of the CYP2C5 structure (1N6B) is depicted with structurally conserved regions in the 20 available crystal structures falling within an RMSD of 2.5 Å shown in red and structurally diverse regions falling outside an RMSD of 2.5 Å shown in yellow. Viewed in a top-down perspective, the heme in the floor of the catalytic site is shown in grey and the elongated I helix (SRS4) diagonally spanning the right side of the structure shown in red.

and the N-terminus of the G helix (harbors SRS2 and SRS3) (referred to as the FG region); and (iii)  $\beta$ -sheet 4 (harbors SRS6) (referred to as the  $\beta$ 4 region). In this paper, homology models built using various strategies for sequence alignments and modeling of these variable regions have been assessed by comparing six target protein models to their experimentally defined structures. These comparisons have indicated that homology models obtained from sequence alignments discriminating between Class I and Class II P450s are significantly closer to experimental crystal structures than those lumping all available template sequences.

## Results

### Alignment and modeling procedures

Twenty-three unique P450 crystal structures have now been solved with six from eukaryotic (mammalian) systems and the rest from prokaryotic systems (Table I). According to the P450 classification system that differentiates these proteins based on their electron transfer partner (Kelly *et al.*, 2005), these six mammalian P450s are designated as Class II enzymes because of their need to interact with NADPH P450 reductase. Bacterial CYP102 (P450 BM3) isolated from *Bacillus megaterium* has also been designated as a Class II P450 because its original isolation demonstrated that it was translationally fused to a NADPH P450 reductase domain

(Narhi and Fulco, 1987). Although much shorter than CYP102 (378 versus 455 amino acids), bacterial CYP175A1 isolated from the thermophilic *Thermus thermophilus* has a structure that closely resembles CYP102 (Yano *et al.*, 2003; Poulos and Johnson, 2005).

Among the more unusual crystal structures of Class I bacterial P450s, two from thermophiles, CYP119 isolated from *Sulfolobus solfataricus* (Yano *et al.*, 2000) and P450st isolated from *Sulfolobus tokodaii* (Oku *et al.*, 2004), are much shorter than the mesophilic P450s (367 amino acids for both) and exhibit significant differences in their B' helix and FG loop regions (Yano *et al.*, 2000; Oku *et al.*, 2004). CYP51 isolated from *Mycobacterium tuberculosis* (CYP51B1-Mt) (Zanno *et al.*, 2005) also displays significant differences from other known P450s in that it contains a distinct bend in its I helix and an open conformation in its BC loop (Podust *et al.*, 2001). The crystal structures of the other Class I bacterial P450s listed in Table I display more overall similarity to one another than these three examples and, except for the absence of an N-terminal membrane anchor, overall similarity to the Class II P450s. The crystal structures of two other P450s, CYP55A2 (P450nor) isolated from *Fusarium oxysporum* (Shimizu *et al.*, 2000) and CYP152A1 (a peroxygenase) isolated from *Bacillus subtilis* (Lee *et al.*, 2003), designated as class III P450s since they require no redox partners are substantially more similar to the majority of the bacterial class

I P450s. Because of the unusually short length of the thermophilic CYP119 and P450st sequences and the unusually structured regions of the CYP51B1-Mt sequence, these three structures were excluded from our alignments as described below. For the remaining 20 sequences, the apo form with the highest resolution was selected for use as a template when several crystal structures were available for a single sequence.

In our comparison of various alignment procedures, two approaches were tested. In the first case designated as the 'one-class' alignment procedure, all 20 sequences shown in bold in Table I were aligned regardless of their biological and structural differences. In the other case designated as the 'two-class' alignment procedure, the sequences of 10 class I P450s (those shown in bold in Table I) and two class III P450s were aligned with one another and the sequences of eight class II P450s were aligned separately from those in Classes I and III. Bacterial CYP102 and CYP175A1 sequences were included in the Class II alignment since, as noted above, their electron transfer system and/or structure more closely resemble those of the eukaryotic P450s. Mixed structural and sequence alignments of these different collections of P450s were performed using the ALIGN facility in MOE versions 2004 and 2005 (Chemical Computing Group Inc., Montreal, Canada) and are shown in Supplementary Figures 1 and 2 available at *PEDS* online. The detailed alignment procedures are described below.

For each of these sets of alignments, two different approaches were used to homology model the target sequence against the template structures. In one approach designated as the 'one-structure' approach, the target sequence was aligned with all P450s included in the alignment procedure and the template sequence sharing the highest sequence identity was used to model the target structure. In the other approach designated as the 'hybrid-structure' approach, three regions of the target P450 sequence were aligned separately from the overall sequence. These regions corresponded to the B region (loop between strand 5 of  $\beta$ -sheet 1, B' helix, B helix and B-C loop), the FG region (the C-terminus of the F helix, the FG loop and the N-terminus of the G helix) and the  $\beta$ 4 region ( $\beta$ -sheet 4) and contain or reside in close proximity to four of the SRS regions (SRS1, SRS2, SRS3, SRS6). In this hybrid-structure modeling approach, portions from a maximum of four different sequences were used to construct a hybrid model of each target sequence: the sequence having the highest identity with the B region of the target, the sequence having the highest identity with the FG region, the sequence having the highest identity with the  $\beta$ 4 region and the sequence having the overall highest sequence identity for the rest of the target protein (not including the B, FG and  $\beta$ 4 regions). The sequences sharing highest identity with each of these regions were determined using the BLOSUM62 scoring matrix (Henikoff and Henikoff, 1992) within the ALIGN facility in MOE. The alignment was done using MOE's default procedure that involve several steps (Kelly, 1996) as follows: (i) initial pairwise build-up using tree-based approach; (ii) round-robin realignment, where each chain in succession is extracted from the global alignment and realigned against the remaining chains; and (iii) randomized iterative refinement. Penalties settings to start a gap in the sequence alignment and to extend this gap were 7 and 1, respectively, following MOE's default settings.

In addition, structural information was included to refine the initial sequence-based alignment of crystal structures, as implemented in the mixed sequence/structural alignment protocol in MOE (Kelly, 1996), and chains are realigned according to their structure. From the multiple sequence alignment, a new similarity matrix is generated using the relative alpha carbon coordinates that result from a multi-body superposition. This matrix is used to realign just these alpha carbon populated chains. This procedure is then repeated until the RMSD of the superposition failed to improve. At this point, the structured chains were reintroduced as an indivisible unit among the unstructured chains, and steps (i) through (iii) above were repeated.

Once these alignments were completed, homology modeling was performed using the HOMOLOGY module within MOE to generate 10 coarsely energy-minimized models. MOE-HOMOLOGY copies the initial geometry from regions of one or more template chains. The procedure is described in detail in MOE's internal documentation and by Kelly (1999) and is simply stated as copying all coordinates where residue identity is conserved between the template and the model and only backbone coordinates where no residue identity exists between the template and the model. For the construction of each model, 10 independent models of the target protein were built using a Boltzmann-weighted randomized modeling procedure in MOE that is adapted from Levitt (1992), and Fichteler *et al.* (1995). Each of these intermediate models was evaluated by a residue packing quality function which is sensitive to the degrees to which non-polar side-chain groups are buried and hydrogen bonding opportunities are satisfied. The model with the best packing quality function was selected in our study for further inspection and comparison with the crystal structure of the corresponding P450.

During all of the modeling processes, the atomic coordinates of the heme from the overall template structure were explicitly included. The one-structure modeling procedures are essentially those described in several of our previous papers (Baudry *et al.*, 2003; Rupasinghe *et al.*, 2003) as well as in the quasi-totality of known published models even though Szklarz and Halpert (1997) and Kirton *et al.* (2002b) described approaches similar to that of our hybrid-structure approach developed here. For the hybrid-structure modeling procedures, different parts of the structure corresponding to the sequences sharing highest identity with the overall sequence and B, FG and  $\beta$ 4 regions were combined by identifying B, FG and  $\beta$ 4 regions existing outside 2.5 Å RMSD (Figure 1) and finding the best local sequence match for each. If this match was different from the global match, the C $\alpha$  co-ordinates of the local match were used as template for that particular region. This local template was adjusted so it ends at less than a 2.0 Å RMSD with the main template. When selecting the best local match only, complete loops were selected even at a lower sequence identity and loops with the best length match were selected even though they might share lower sequence identities. With these adjustments, 10 structures were generated for each template and partially energy-minimized using the CHARMM27 force field as implemented in MOE. A non-bonded cutoff between 8 and 10 Å and a distance-dependent dielectric value for electrostatic interactions were used to approximate solvent effects. These modeled structures were then compared with their corresponding crystal structures (PDB entries) by performing

structural alignments between the modeled structures and the corresponding crystal structure (PDB entries used in these comparisons are underlined in Table I) using MOE and then comparing the RMSD calculated on all C $\alpha$  atoms from these structurally aligned structures for the entire modeled structure as well as the B, FG and  $\beta$ 4 regions.

### Alignment and modeling assessments

In our first tests of these alignment and modeling procedures, six proteins with known structures (CYP2C5, CYP2C8, CYP3A4, CYP102, CYP101, CYP158A2) were selected as test sequences since these P450s contain the most structurally diverse B, FG and  $\beta$ 4 regions. To avoid bias in our modeling procedures, the sequence of each of these test sequences was deleted from the 20 P450 alignment set prior to initiating its modeling. The particulars for each alignment are listed below with Table II detailing percent identities in the overall sequence and the three variable regions using either the one-class or two-class alignment approaches.

**CYP2C5 case:** The sequence sharing the highest sequence identity with CYP2C5 is CYP2C9 with the exception of the B region that is most identical to CYP2C8. These findings are independent of whether the one-class or two-class approach is used in the alignment process. However, using the one-class approach leads to a slightly lower percentage of overall sequence identity (68%) than using the two-class approach (75%). This means that even though the CYP2C9 sequence is consistently identified as the most identical, the weight of other sequences in the one-class approach shifts the alignment of individual residues in the target sequence toward a less identical match. This has important implications in the homology modeling process since some residues considered conserved in the ‘two-class’ approach are not conserved in the ‘one-class’ approach necessitating that their Cartesian coordinates be reconstructed during the homology steps—bringing *a priori* more serendipity to the process and increasing the potential for structural differences with that of the experimental crystal structures. For CYP2C5, these same observations apply to the FG region. In contrast to these two regions, the sequence identities defined by both one-class and two-class approaches are identical for the  $\beta$ 4 region meaning that the Cartesian coordinates of the corresponding residues would be the identically built during the homology modeling process.

It is also worth noting that the sequence of the B region of CYP2C5 is more identical to that of CYP2C8 than to CYP2C9. This difference would be predicted to translate into different coordinates for the corresponding residues when either the one-structure or hybrid-structure approaches are used in the homology modeling process.

**CYP2C8 case:** The sequence identities defined for CYP2C8 are qualitatively identical to those described for CYP2C5 above in that it shares the highest sequence identity with CYP2C9 with the exception of the B region that is most identical to CYP2C5.

**CYP3A4 case:** The sequence having the highest overall sequence identity with CYP3A4 is different depending on whether the one-class or two-class alignment approach is used. Naturally, this has significant implications during the homology modeling stage, since different templates would be used in each of these cases. The differences, however, would be limited in this case since CYP2C8 (most identical using the two-class approach) and CYP2C9 (most identical using the one-class approach) share relatively high sequence identity with each other and align similar residues with the CYP3A4 target sequence. The  $\beta$ 4 region of CYP3A4 also displays significant differences depending on the alignment approach with its sequence being most identical to CYP2B4 by the two-class approach and to CYP175A1 by the one-class approach.

**CYP102 case:** The sequence having the highest overall sequence identity with CYP102 is the same sequence (CYP3A4) whether the one-class or two-class approach is used for the alignment process. However, the two-class approach identifies different sequences sharing highest identities with the B, FG and  $\beta$ 4 regions while the one-class approach identifies the same sequence sharing highest identity with the FG and B regions. Comparison between the one-class and two-class approaches shows that only the FG region is most identical to the same sequence in both alignment procedures.

**CYP101 case:** The sequence having the highest overall sequence identity as well as B, FG and  $\beta$ 4 region identity in the two-class approach is CYP55A2. The one-class approach identifies four different sequences having highest sequence identity with the overall sequence and the B, FG and  $\beta$ 4 regions described here.

**CYP158A2 case:** As in the CYP102 case, one unique sequence is found to be most identical to the overall sequence using either the one-class or two-class approaches but

**Table II.** Sequence identities in P450 test sequences

Target	Overall	B	FG	$\beta$ 4	Type of alignment
<b>CYP2C5</b>	2C9 (75) <sup>a</sup>	2C8 (50)	2C9 (71)	2C9 (73)	2-class
	2C9 (68)	2C8 (50)	2C9 (61)	2C9 (73)	1-class
<b>CYP2C8</b>	2C9 (77)	2C5 (55)	2C9 (68)	2C9 (58)	2-class
	2C9 (67)	2C5 (46)	2C9 (74)	2C9 (56)	1-class
<b>CYP3A4</b>	2C8 (24)	2C8 (24)	2C8 (24)	2B4 (14)	2-class
	2C9 (21)	2C8 (10)	2C9 (13)	175A1 (18)	1-class
<b>CYP102</b> (P450BM3)	3A4 (22)	2A4 (24)	175A1 (12)	2B4 (26)	2-class
	3A4 (22)	175A1 (24)	175A1 (16)	108 (21)	1-class
<b>CYP101</b> (P450cam)	55A1 (24)	55A1 (24)	55A1 (24)	55A1 (24)	2-class
	108 (24)	55A1 (14)	165B1 (12)	154A1 (25)	1-class
<b>CYP158A2</b>	165B1 (31)	121 (8)	CYP55A1 (16)	165C1 (33)	2-class
	165B1 (32)	121 (12)	CYP55A1 (16)	165C1 (33)	1-class

<sup>a</sup>P450 sharing the highest amino acid identity in the designated region with percent identity shown in parentheses.

**Table III.** RMSD between crystal structures and modeled structures

	Single-structure								Hybrid-structure							
	Overall		B		FG		$\beta 4$		Overall		B		FG		$\beta 4$	
	<i>a</i>	<i>b</i>	<i>a</i>	<i>b</i>	<i>a</i>	<i>b</i>	<i>a</i>	<i>b</i>	<i>a</i>	<i>b</i>	<i>a</i>	<i>b</i>	<i>a</i>	<i>b</i>	<i>a</i>	<i>b</i>
CYP2C5	<b>2.0</b>	2.5	<b>1.8</b>	3.9	<b>2.4</b>	3.2	<b>0.9</b>	0.9	<b>2.0</b>	2.3	<b>0.8</b>	1.8	<b>2.4</b>	3.5	<b>0.9</b>	0.9
CYP2C8	<b>1.2</b>	1.2	<b>1.6</b>	1.6	<b>1.5</b>	1.5	<b>0.9</b>	0.9	<b>1.2</b>	1.2	<b>0.8</b>	0.8	<b>1.5</b>	1.5	<b>0.9</b>	0.9
CYP3A4	<b>3.9</b>	4.7	<b>2.6</b>	5.7	<b>5.2</b>	5.2	<b>3.6</b>	3.8	<b>3.8</b>	4.6	<b>3.2</b>	4.4	<b>5.2</b>	5.4	<b>3.6</b>	4.7
CYP102	<b>5.6</b>	5.8	<b>4.2</b>	3.9	<b>5.5</b>	5.2	<b>10.8</b>	11.4	<b>5.4</b>	5.3	<b>3.6</b>	2.2	<b>3.6</b>	3.0	<b>12.7</b>	11.0
CYP101	<b>3.7</b>	3.7	NA	5.1	NA	NA	NA	2.6	NA	NA	NA	2.3	NA	3.5	NA	2.1
CYP158A2	<b>3.9</b>	4.4	<b>2.2</b>	3.9	<b>3.4</b>	3.9	<b>2.0</b>	4.3	<b>3.6</b>	4.4	<b>1.5</b>	4.6	<b>3.2</b>	3.5	<b>1.9</b>	1.7

All values are in angstroms (Å).

*a*, Two-class alignment; *b*, One-class alignment.

different sequences align with each of the B, FG and  $\beta 4$  regions using these various approaches.

MOE-homology models built with these various sequence alignments are compared with the known crystal structures in Table III with the RMSD calculated comparing one-class versus two-class alignment procedures and single-structure versus hybrid-structure approaches. In this analysis, columns *a* and *b* show how closely the experimentally obtained crystal structures agree with the homology models obtained using either the one-class or two-class alignment procedures. Columns in the single-structure half of the table use the same alignment for the entire target sequence with RMSD for the individual B, FG and  $\beta 4$  regions reported. Columns in the hybrid-structure half of the table use the different B, FG and  $\beta 4$  regions designated in Table II. Comparisons between the two halves of this table allow one us to quantify how overall identity variations translate into modeling variations and affect the structural quality of the predictive model.

#### Single-structure modeling approach

The data derived from the single-structure modeling approach indicate that the overall structural quality of the homology models obtained from the two-class alignment approach are significantly better for three of the six models investigated here: CYP2C5 (2.0 Å for two-class alignment, 2.5 Å for one-class alignment), CYP3A4 (3.9 Å for two-class alignment versus 4.7 Å for one-class alignment) and CYP158A2 (3.9 Å for two-class alignment versus 4.4 Å one-class alignment). This improvement in the RMSD between predictive models and known crystal structures of  $\sim 0.5$  Å is significant and substantially lower than the range of model/experimental Ca RMSD in earlier systematic comparisons of modeling strategies (Kirton *et al.*, 2002b). In each of these three P450 models, this overall improvement results from improvements in at least one of the B, FG or  $\beta 4$  regions that contain the most highly variable SRS. In CYP2C5, the RMSD of the FG region is significantly lower in the two-class alignment model (2.4 Å) than in the one-class alignment model (3.2 Å) and also in the B region (1.8 Å for two-class alignment versus 3.9 Å for one-class alignment). In CYP3A4, the RMSD of the B region in the two-class alignment model is significantly closer to that of the crystal structure than in the one-class alignment model (2.6 Å for two-class alignment versus 5.7 Å for one-class alignment). In CYP158A2, the RMSD of all three regions are significantly closer to the crystal structure in the two-class alignment model than the one-class alignment model with the most significant differences occurring in the

B and  $\beta 4$  regions (2.2 and 2.0 Å in the two-class alignments versus 3.9 and 4.3 Å in the one-class alignments). The three remaining models (CYP2C8, CYP102, CYP101) do not exhibit significant reduction in their overall RMSD when two-class alignment models are compared with one-class alignment models.

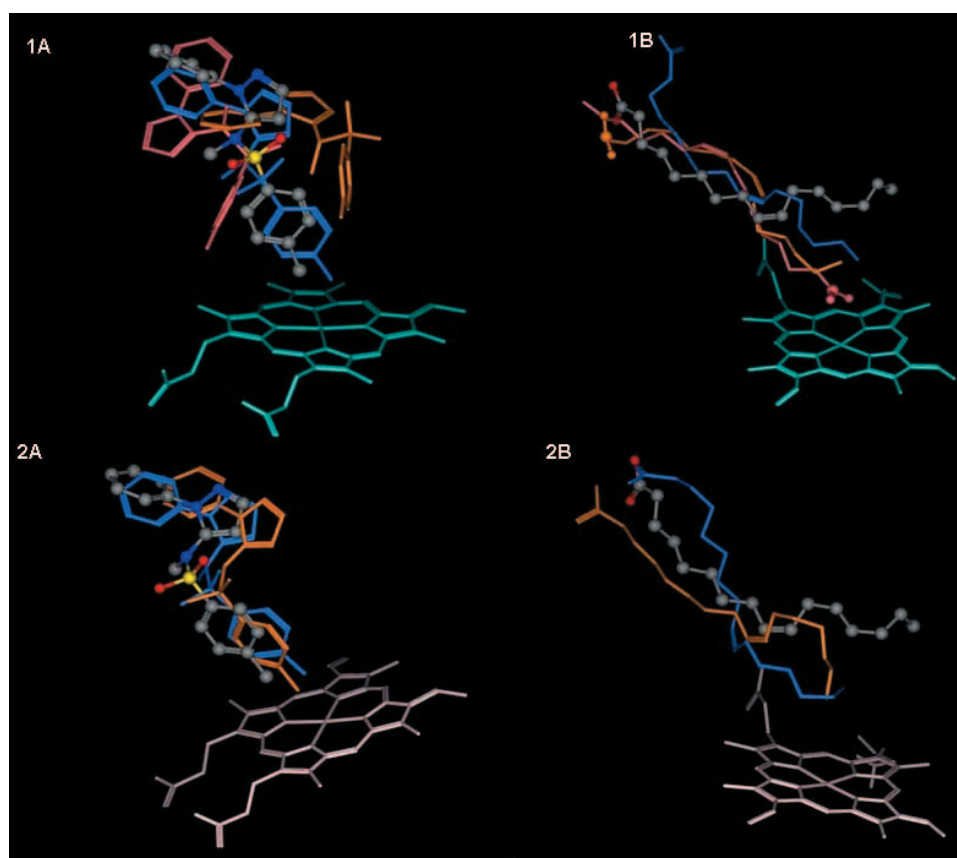
#### Hybrid-structure modeling approach

The data derived from the hybrid-structure modeling approach indicate that again for the same three P450 test sequences, the models obtained using the two-class alignment approach exhibit a better agreement with the crystal structures than the models obtained using the one-class alignment approach.

Comparisons between these two modeling approaches indicate that use of the hybrid-structure modeling approach, where the B, FG and  $\beta 4$  region models are constructed from alignments that may be different from the alignments used for the rest of the structure, does not result in significant improvement of the RMSD values calculated for the overall structure. But importantly, the RMSD values calculated for the individual B, FG and  $\beta 4$  substructures are often reduced in the hybrid-structure models compared to the RMSD values calculated for single-structure models, even in cases where the same sequence was used as template in the hybrid and single-structure approaches. This is particularly evident for the B region of CYP2C5 (0.8 Å for the hybrid-structure model versus 1.8 Å for the single-structure model), the B region of CYP2C8 (0.8 Å for the hybrid-structure model versus 1.6 Å for the single-structure model), the FG region of CYP102 (3.6 Å for the hybrid-structure model versus 5.5 Å for the single-structure model), the B region of CYP102 (3.6 Å for the hybrid-structure model versus 4.2 Å for the single-structure model) and the B region of CYP158A2 (1.5 Å for the hybrid-structure model versus 2.2 Å for the single-structure model). Interestingly, several of these reductions in the RMSDs of individual domains are found in cases where no RMSD improvements were achieved using the two-class versus one-class alignment approaches. Examples of this include the differences shown for CYP2C8 and CYP102 in Table III.

#### Ligand binding in homology models

In addition to quantifying the structural similarity of homology models to experimental crystal structures, the capacities of various homology models to reproduce experimentally described ligand-binding modes were investigated. Because



**Fig. 2.** Comparison of substrate-docked models with substrate-bound crystal structures. (1A) DMZ substrate in CYP2C5 model developed using one-class approach, (1B) palmitoleic acid substrate in CYP102 model developed using one-class approach, (2A) DMZ docked substrate in CYP2C5 model developed using two-class approach, (2B) palmitoleic acid substrate in CYP102 model developed using two-class approach. The grey ball-and-stick diagram shows the binding mode defined for each substrate in their respective crystal structures [1FAG with palmitoleic acid in CYP102 (Li and Poulos, 1997); 1N6B with DMZ in CYP2C5 (Wester *et al.*, 2003)]. The orange structures show predicted ligand orientations in models developed using the hybrid-structure modeling approach. The blue structures show predicted ligand orientations in models developed using the one-fragment modeling approach.

crystal structures of only two P450s analyzed here (CYP2C5, CYP102) co-crystallized with their respective ligands are available in the PDB databases (CYP102: 1FAG, Li and Poulos, 1997; CYP2C5: 1N6B, Wester *et al.*, 2003), homology models for only these P450s were subjected to *in silico* docking calculations. For each of these P450s, models obtained from the one-class and two-class alignment procedures as well as models from single-structure and hybrid-structure modeling procedures were carried through docking using the DOCK facility in MOE with palmitoleic acid for CYP102 and DMZ (4-methyl-*N*-methyl-*N*-(2-phenyl-2H-pyrazol-3-yl) benzenesulfonamide) for CYP2C5. The results some of which are shown in the top panels of Figure 2 indicate that the one-class alignment with single-structure modeling procedures do not systematically reproduce the experimental binding modes of their respective substrates in the CYP2C5 and CYP102 crystal structures. In the case of DMZ-bound CYP2C5 (panel 1A), several binding modes correctly reproduce the binding mode seen in the crystal structure, but there exists a diverse set of possible binding modes seen as energetically possible. One of these many possible binding modes does consistently dock in a conformation similar to that observed in the CYP2C5 crystal structure with the functional groups located in the same regions as experimentally defined (colored blue in panel 1A) but the large number of other potential binding modes (data

not shown) makes it difficult to discriminate between this and inappropriately bound conformers. In the case of palmitoleic acid-bound CYP102 (panel 1B), the one-class alignment with hybrid-structure modeling procedures generates several possible binding modes with none particularly close to the crystal structure; some even position the carboxylic acid moiety of palmitoleic acid close to heme rather than away. Like the hybrid-structure modeling, the one-class alignment with single-structure modeling procedures do consistently generate a binding mode in an orientation close to but not exactly the same as that of the crystal structure.

The results shown in the bottom panels of Figure 2 indicate that, for CYP2C5, the model obtained from the two-class alignment approach (panel 2A) consistently docks the DMZ ligand in a manner similar to that seen in the crystal structure. It is not possible at this level of accuracy, to qualitatively differentiate between the results obtained by the hybrid-structure modeling procedures and single-structure modeling procedures (orange versus blue in panel 2A). For CYP102, the binding mode of the long aliphatic chain observed in the 1FAG crystal structure could not be exactly reproduced in the two-class alignment models constructed using either the single-structure or hybrid-structure modeling procedures. However, the carboxylic acid moiety on palmitoleic acid is correctly interacting with the hydroxyl group of Tyr51 and no docking mode predicted to locate the acid close to the heme

group, unlike the predicted docking mode obtained in the one-class alignment with hybrid-structure modeling procedures (panel 1B).

## Discussion

This work aimed at identifying possible strategies to improve the quality and usefulness of P450 models has indicated that models using the two-class alignment approach consistently yielded ligand-docking modes that were in better agreement with the crystal structures than models obtained using the one-class alignment approach. From the present model comparisons, it is clear that several approaches can be used to improve the quality of P450 homology models even when amino acid identities are extremely low (<20%) when compared with sequences whose crystal structure has already been defined. Discriminating between class I and class II sequences (i.e. using two-class alignment procedures rather than one-class alignment procedures) led, in several clearly defined cases, to models that reproduce known crystal structures better than models obtained with class I and class II sequences grouped together. Additionally, modeling of SRS regions independent from the rest of the P450 target sequence (i.e. using hybrid-structure modeling procedures rather than single-structure modeling procedures) can lead to significant improvements in functionally relevant SRS regions. From the examples presented here, hybrid-structure models do not seem to significantly improve on models that have already benefited from the two-class alignment procedure. But, hybrid-structure modeling does improve the structural modeling of SRS regions that are not otherwise improved by the two-class alignment (e.g. CYP102). As more and more P450 structures are defined and their classes assigned, the number of structures that can be used in each class for homology modeling will obviously increase. This naturally has potential to change the alignments used for homology modeling and affect the final structures. The sequences used in the present study should be extended to include any new experimentally known structures as well as their class information.

It is also apparent from the two P450s analyzed for predicted binding modes that ligand docking can also be improved (with respect to how well an experimentally known ligand-binding structure is reproduced) using models obtained from the two-class alignment procedure. Analysis of this small subset of ligand-bound P450s does not demonstrate any substantial benefit to docking in hybrid-structure models compared to single-structure models (i.e. once two-class alignment procedures are used, hybrid-structure modeling does not yield models that are systematically better than single-structure modeling).

The limitations of the approaches used in this study are that the dynamics of the protein structure as well as the dynamics of the ligand-binding process (e.g. thermodynamics of ligand-binding pathway and induced-fit changes in the enzyme's catalytic site) are not included. The plasticity of the enzyme binding site is an important part of the protein's structure and function and a feature closely linked to enzymatic efficiency and organism's specialization (Li *et al.*, 2004). Using these improved modeling procedures, it will be interesting to see how the inclusion of molecular dynamics simulations in future studies can improve the quality of P450 models.

## Acknowledgements

This work was funded by NIH grant R01 GM71826 and NSF grant MCB0115068.

## References

- Baudry, J., Li, W., Pan, L., Berenbaum, M.R. and Schuler, M.A. (2003) *Protein Eng.*, **16**, 577–587.
- Chang, Y.T. and Loew, G.H. (1996) *Protein Eng.*, **9**, 755–766.
- Cupp-Vickery, J.R. and Poulos, T.L. (1995) *Nat. Struct. Biol.*, **2**, 144–153.
- deGraaf, C., Vermeulen, N.P.E. and Feenstra, K.A. (2005) *J. Med. Chem.*, **48**, 2725–2755.
- Domanski, T.L. and Halpert, J.R. (2001) *Curr. Drug Metab.*, **2**, 117–137.
- Fechteler, T., Dengler, U. and Schomberg, D. (1995) *J. Mol. Biol.*, **253**, 114–131.
- Gotoh, O. (1992) *J. Biol. Chem.*, **267**, 83–90.
- Graham, S.E. and Peterson, J.A. (1999) *Arch. Biochem. Biophys.*, **369**, 24–29.
- Hasemann, C.A., Ravichandran, K.G., Peterson, J.A. and Deisenhofer, J. (1994) *J. Mol. Biol.*, **236**, 1169–1185.
- Henikoff, S. and Henikoff, J.G. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.
- Kelly, K. (1996) <http://www.chemcomp.com/journal/align.htm>.
- Kelly, K. (1999) <http://www.chemcomp.com/journal/biol1999.htm>.
- Kelly, S.L., Kelly, D.E., Jackson, C.J., Warrilow, A.G.S. and Lamb, D.C. (2005) In Paul R. Ortiz de Montellano (ed.), *Cytochrome P450: Structure, Mechanism, and Biochemistry*, 3rd edn. Kluwer Academic/Plenum Publishers, New York, pp. 585–617.
- Kirton, S.B., Baxter, C.A. and Sutcliffe, M.J. (2002a) *Adv. Drug Deliv. Rev.*, **54**, 385–406.
- Kirton, S.B., Kemp, C.A., Tomkinson, N.P., St-Galley, S. and Sutcliffe, M.J. (2002b) *Proteins: Struct. Funct. Genet.*, **49**, 216–231.
- Lee, D.S., Yamada, A., Sugimoto, H., Matsunaga, I., Ogura, H., Ichihara, K., Adachi, S., Park, S.Y. and Shiro, Y. (2003) *J. Biol. Chem.*, **278**, 9761–9767.
- Levitt, M. (1992) *J. Mol. Biol.*, **226**, 507–533.
- Ley, D., Mowat, C.G., McLean, K.J., Richmond, A., Chapman, S.K., Walkinshaw, M.D. and Munro, A.W. (2003) *J. Biol. Chem.*, **278**, 5141–5147.
- Li, H. and Poulos, T.L. (1997) *Nat. Struct. Biol.*, **4**, 140–146.
- Li, X., Baudry, J., Berenbaum, M.R. and Schuler, M.A. (2004) *Proc. Natl Acad. Sci. USA*, **101**, 2939–2944.
- Nagano, S., Li, H., Shimizu, H., Nishida, C., Ogura, H., Ortiz de Montellano, P.R. and Poulos, T.L. (2003) *J. Biol. Chem.*, **278**, 44886–44893.
- Narhi, L.O. and Fulco, A.J. (1987) *J. Biol. Chem.*, **262**, 6683–6690.
- Oku, Y., Ohtaki, A., Kamitori, S., Nakamura, N., Yohda, M., Ohno, H. and Kawarabayashi, Y. (2004) *J. Inorg. Biochem.*, **98**, 1194–1199.
- Park, S.Y., Yamane, K., Adachi, S., Shiro, Y., Weiss, K.E. and Sligar, S.G. (2000) *Acta Crystallogr. D Biol. Crystallogr.*, **56**, 1173–1175.
- Podust, L.M., Poulos, T.L. and Waterman, M.R. (2001) *Proc. Natl Acad. Sci. USA*, **98**, 3068–3073.
- Podust, L.M., Kim, Y., Arase, M., Neely, B.A., Beck, B.J., Bach, H., Sherman, D.H., Lamb, D.C., Kelly, S.L. and Waterman, M.R. (2003) *J. Biol. Chem.*, **278**, 12214–12221.
- Podust, L.M., Bach, H., Kim, Y., Lamb, D.C., Arase, M., Sherman, D.H., Kelly, S.L. and Waterman, M.R. (2004) *Protein Sci.*, **13**, 255–268.
- Podust, L.M., Yermalitskaya, L.V., Lepesheva, G.I., Podust, V.N., Dalamasso, E.A. and Waterman, M.R. (2004) *Structure*, **12**, 1937–1945.
- Poulos, T.L. and Johnson, E.F. (2005) In Paul R. Ortiz de Montellano (ed.) *Cytochrome P450: Structure, Mechanism, and Biochemistry* 3rd edn. Kluwer Academic/Plenum Publishers, New York, pp. 87–114.
- Poulos, T.L., Finzel, B.C. and Howard, A.J. (1987) *J. Mol. Biol.*, **195**, 687–700.
- Pylypenko, O., Vitali, F., Zerbe, K., Robinson, J.A. and Schlichting, I. (2003) *J. Biol. Chem.*, **278**, 46727–46733.
- Ravichandran, K.G., Boddupalli, S.S., Hasemann, C.A., Peterson, J.A. and Deisenhofer, J. (1993) *Science*, **261**, 731–736.
- Rupasinghe, S., Baudry, J. and Schuler, M.A. (2003) *Protein Eng.*, **16**, 721–731.
- Schoch, G.A., Yano, J.K., Wester, M.R., Griffin, K.J., Stout, C.D. and Johnson, E.F. (2004) *J. Biol. Chem.*, **279**, 9497–9503.
- Scott, E.E., Spatzenegger, M. and Halpert, J.R. (2001) *Arch. Biochem. Biophys.*, **395**, 57–68.
- Scott, E.E., He, Y.A., Wester, M.R., White, M.A., Chin, C.C., Halpert, J.R., Johnson, E.F. and Stout, C.D. (2003) *Proc. Natl Acad. Sci. USA*, **100**, 13196–13201.
- Scott, E.E., White, M.A., He, Y.A., Johnson, E.F., Stout, C.D. and Halpert, J.R. (2004) *J. Biol. Chem.*, **279**, 27294–27301.
- Shimizu, H., Park, S., Lee, D., Shoun, H. and Shiro, Y. (2000) *J. Inorg. Biochem.*, **81**, 191–205.

- Stout,C.D. (2004) *Structure*, **12**, 1921–1922.
- Szklarz,G.D. and Halpert,J.R. (1997) *J. Comput. Aided Mol. Design.*, **11**, 265–272.
- vonWachenfeldt,C., Richardson,T.H., Cosme,J. and Johnson,E.F. (1997) *Arch. Biochem. Biophys.*, **339**, 107–114.
- Wester,M.R., Johnson,E.F., Marques-Soares,C., Dansette,P.M., Mansuy,D. and Stout,C.D. (2003) *Biochemistry*, **42**, 6370–6379.
- Wester,M.R., Johnson,E.F., Marques-Soares,C., Dijols,S., Dansette,P.M., Mansuy,D. and Stout,C.D. (2003) *Biochemistry*, **42**, 9335–9345.
- Williams,P.A., Cosme,J., Sridhar,V., Johnson,E.F. and McRee,D.E. (2000) *Mol. Cell*, **5**, 121–131.
- Williams,P.A., Cosme,J., Ward,A., Angove,H.C., Matak Vinkovic,D. and Jhoti,H. (2003) *Nature*, **424**, 464–468.
- Yano,J.K., Hsu,M.H., Griffin,K.J., Stout,C.D. and Johnson,E.F. (2005) *Nat. Struct. Mol. Biol.*, **12**, 822–823.
- Yano,J.K., Koo,L.S., Schuller,D.J., Li,H., Ortiz de Montellano,P.R. and Poulos,T.L. (2000) *J. Biol. Chem.*, **275**, 31086–31092.
- Yano,J.K., Blasco,F., Li,H., Schmid,R.D., Henne,A. and Poulos,T.L. (2003) *J. Biol. Chem.*, **278**, 608–616.
- Yano,J.K., Wester,M.R., Schoch,G.A., Griffin,K.J., Stout,C.D. and Johnson,E.F. (2004) *J. Biol. Chem.*, **279**, 38091–38094.
- Zanno,A., Kwiatkowski,N., Vaz,A.D.N. and Guardiola-Diaz,H.M. (2005) *Biochim. Biophys. Acta*, **1707**, 157–169.
- Zerbe,K.*et al.* (2002) *J. Biol. Chem.*, **277**, 47476–47485.
- Zhao,B.*et al.* (2005) *J. Biol. Chem.*, **280**, 11599–11607.

Received January 5, 2006; accepted March 21, 2006

Edited by Patrick Stayton