Common active site architecture and binding strategy of four phenylpropanoid P450s from *Arabidopsis thaliana* as revealed by molecular modeling

Sanjeewa Rupasinghe¹, Jerome Baudry² and Mary A. Schuler¹,³

¹Department of Cell and Structural Biology and ³School of Chemical Sciences, University of Illinois, Urbana, IL 61801, USA

To whom correspondence should be addressed.
E-mail: maryschu@uiuc.edu

Despite extensive primary sequence diversity, crystal structures of several bacterial cytochrome P450 monoxygenases (P450s) and a single eukaryotic P450 indicate that these enzymes share a structural core of α-helices and β-sheets and vary in the loop regions contacting individual substrates. To determine the extent to which individual structural features are conserved among divergent P450s existing in a single biosynthetic pathway, we have modeled the structures of four highly divergent P450s (CYP73A5, CYP84A1, CYP75B1, CYP98A3) in the *Arabidopsis* phenylpropanoid pathway synthesizing lignins, flavonoids and anthocyanins. Analysis of these models has indicated that, despite primary sequence identities as low as 13%, the structural cores and several loop regions of these P450s are highly conserved. Substrate docking indicated that all four enzymes employ a common strategy to identify their substrates in that their cinnamate-derived substrates align along helix I with their aromatic ring positioned towards the C-terminus of this helix and their aliphatic tails positioned towards the N-terminus. Further similarity was observed in the way the substrates contact the consensus P450 recognition sites (SRS). Residues predicted to contact the aromatic ring region exist in SRS5, SRS6 and the C-terminal portion of SRS4 and residues contacting the distal end of each substrate exist in SRS1, SRS2 and the N-terminal portion of SRS4. Alignments of the regions contacting the aromatic ring region indicate that SRS4, SRS5 and SRS6 share higher degrees of sequence conservation than found in SRS1, SRS2 or the full-length protein.

*Keywords*: homology modeling/P450/P450 monoxygenases/phenylpropanoid pathway/substrate docking

---

**Introduction**

Cytochrome P450 monoxygenases (P450s) are heme-dependent mixed function oxidases that utilize reducing equivalents from various sources to reductively cleave atmospheric dioxygen to produce a functionalized organic product and a molecule of water. These enzymes vary in their subcellular locations, organismal ranges and the identities of their electron transfer partners (Werck-Reichhart and Feyereisen, 2000). Class I P450s, which are soluble in bacteria and in the inner mitochondrial membrane of vertebrates, insects and nematodes, utilize a FAD-containing reductase and a separate iron–sulfur protein. Bacterial class II P450s, which are soluble in bacteria, utilize a FAD/FMN-containing reductase that is fused to the catalytic P450. Eukaryotic Class II P450s, which are found in the endoplasmic reticulum of a wide variety of eukaryotic cell types, utilize an NADPH-dependent P450 reductase and, sometimes an NADH-dependent cytochrome b₅ reductase in conjunction with cytochrome b₅. Class III P450s, which have been described in a number of different subcellular locations in animals and plants, are self-sufficient enzymes that can accept electrons directly from NADPH (i.e. they require no electron transfer partner) and/or utilize alternative oxygen donors in their reactions. Class IV P450s, which have been described in fungi, accept electrons directly from NADH (Werck-Reichhart and Feyereisen, 2000).

Since the cloning of the first plant P450 in 1990 (Bozak et al., 1990), substantial work has been done on the cloning and characterization of a large array of P450s involved in essential and non-essential plant biochemistries (Schuler, 1996; Chapple, 1998; and Werck-Reichhart et al., 2002; Schuler and Werck-Reichhart, 2003). P450s are found in all major plant biosynthetic pathways, including those for flavonoids, anthocyanins, phenylpropanoids, terpenoids, alkaloids, cyanogenic glycosides, fatty acids, hormones and signaling molecules (Werck-Reichhart, 1995). Analysis at the primary sequence level of these many P450s with perspectives on evolution (Kahn and Durst, 2000) has suggested that divergence of these enzymes allows for the acquisition of new biochemical reactivities and/or sets of reactivities that have potential for evolving new biochemical pathways. Among closely related P450s, one clear example of this is the set of four maize and four wheat CYP71C subfamily proteins that mediate consecutive steps in the synthesis of DIMBOA, a defense agent against fungal and insect pests (Frey et al., 1997; Nomura et al., 2002). Another example is the set of *Arabidopsis* CYP90 family proteins that mediate consecutive steps in brassinolide synthesis (Szkeres et al., 1996; Choe et al., 1998).

Even so, there are perhaps more examples of biochemical pathways containing P450s categorized in very different families that are capable of handling similar substrates. Examples of these are the *Arabidopsis* CYP701A3 and redundant CYP88A3/CYP88A4 proteins involved in gibberellin synthesis (Helliwell et al., 1998, 1999, 2001). Others are the *Arabidopsis* CYP86A and *Vicia* CYP94A subfamily proteins that mediate ω-hydroxylations on C₁₂ fatty acids (Benveniste et al., 1998; Tijet et al., 1998; Le Bouquin et al., 1999; Wellesen et al., 2001). Within *Arabidopsis*, additional examples exist in the CYP73A5, CYP75B1, CYP84A1 and CYP98A3 proteins that mediate various steps in phenylpropanoid biosynthesis (Figure 1), including one in the core pathway, two in lignin synthesis and one in flavonoid/anthocyanin synthesis (Mizutani et al., 1997; Urban et al., 1997; Humphreys et al., 1999; Schoenbohm et al., 2000; Schoch et al., 2001; Franke et al., 2002). In this last example,
CYP73A5 (t-cinnamic acid hydroxylase, t-CAH) catalyzes the 4-hydroxylation of t-cinnamate to p-coumarate (Figure 1). CYP84A1 catalyzes the 5-hydroxylation of the similarly sized coniferaldehyde, coniferyl alcohol and ferulic acid, CYP75B1 catalyzes the 3-hydroxylation of the larger dihydrokaempferol structure and CYP98A3, which represents the newest member of this collection of phenylpropanoid enzymes, catalyzes the 3-hydroxylation of the larger p-coumaroylshikimic and quinic acids. Given their high degree of sequence divergence, it has been unclear whether the catalytic sites of these divergent proteins maintain side chain conservations important for interactions with conserved features of their phenylpropanoid substrates.

To describe further the degree of catalytic site conservation in divergent P450s within a biochemical pathway, we have modeled and substrate docked the structures of the four phenylpropanoid pathway enzymes described above. Comparisons between these structures indicate that their predicted catalytic core structures are highly conserved in spite of extensive sequence diversity in their primary sequences and higher order structures.

Materials and methods
Homology models were built using MOE programs (Chemical Computing Group, Montreal, Canada) using Arabidopsis P450 protein sequences, which were obtained from the University of Illinois Arabidopsis P450 website (http://Arabidopsis-p450.biotec.uiuc.edu). Crystal structure information for four bacterial P450s [P450cam (CYP101), P450BM3 (CYP102), P450eryF (CYP107A1), P450terp (CYP108)] and one mammalian P450 (CYP2C5) corresponding to PDB files designated 3CPP, 2HPD, 1OXA, 1CPT and 1DT6, respectively, were obtained from the Brookhaven protein data bank (RCSB) (Berman et al., 2000). Multiple sequence alignments for the five bacterial P450s and single mammalian P450 were done with the four Arabidopsis P450s using the blosum62 substitution matrix in the MOE program. The sequence alignments of each Arabidopsis sequence with the five crystal structures were subsequently homology modeled using the MOE program.

Although the CYP2C5 sequence had a slightly higher sequence identity with the Arabidopsis P450 target sequences, evaluation of the CYP2C5 crystal structure that is described in Results suggests that it may have several regions of high energy, suggesting the possibility that the residues in these regions are in unfavorable environments. To minimize the possibility of including high-energy regions in the modeled P450s, the CYP102 structure was chosen as the template.

In the initial modeling of each P450 target sequence, 10 models were generated for each P450 with the explicit inclusion of heme coordinates in all steps of homology model generation. These models were subjected to coarse energy minimization procedures in order to remove potentially bad van der Waals contacts between atoms. In the subsequent modeling, the best model ranked by MOE’s residue packing quality function was selected and the heme coordinates were copied from the CYP102 crystal structure with a covalent bond created between the heme’s iron atom and the sulfur of the conserved cysteine axial ligand. Further energy minimizations for each protein were performed using the CHARMm22 force field (MacKerell et al., 1998) within the MOE distribution until the final energy gradient was <0.01 kcal/mol Å. A distance-dependent dielectric constant was used in the calculations with a cutoff between 6.5 and 7 Å.

Known substrates for each of the Arabidopsis P450s were docked within the catalytic site of the energy-minimized model using the Monte Carlo docking procedure of MOE after attaching a single oxygen to the heme plane (representing the iron–oxo intermediate). Parameters for the F–O bond and N–Fe–O angle were obtained using MOE’s parameter assignment facility and are listed in Table I in CHARMm notation. In this docking, each substrate was initially placed above the heme plane and allowed to vary through Monte Carlo simulations removing any bias due to manual placement. Twenty-five possible conformations were generated for each substrate while maintaining rigid side chains and these were ranked according to the sum of the ligand’s internal energy and the van der Waals and electrostatic energy terms of the potential energy function. The binding conformation with the lowest energy and appropriate hydroxylation site closest to...
the heme was selected as the optimal conformation and subjected to energy minimization using the MMFF94 force field (Halgren, 1996) in MOE while allowing full side chain relaxation. In these protein/ligand minimizations, the heme coordinates were fixed to prevent distortion of the heme plane originating from bonded parameters in the MOE’s implementation of the MMFF94 force field.

In these structures, substrate recognition sequences (SRS) first highlighted as important for substrate specificity in mammalian CYP2A proteins by Gotoh (Gotoh, 1992) were identified by alignments of each Arabidopsis P450 sequence with 10 members of the mammalian CYP2 family. The boundaries of these SRS regions were refined to correspond to the structural feature defined for each particular SRS region and subsequently aligned using the VectorNTI AlignX algorithm and VectorNTI phylogeny (InforMax, Wisconsin, MD).

The final Arabidopsis P450 models were subjected to several tests to assess their reliability. The first test was to examine the distribution of $\phi$ and $\psi$ angles using Ramachandran plots generated within the MOE program. The second test was to apply energy criteria using Prosa II (version 3.02) analysis (Center for Applied Molecular Engineering, University of Salzburg, Austria). In this program, the Prosa Z-scores and energy profiles were calculated for each model in order to assess their reliability. The third test was to calculate residue compatibility scores (3D–1D score) and the sum of residue compatibility scores for each model using Profiles 3D analysis (InsightII Homology module, MSI, San Diego, CA). The fourth test was to determine the thermodynamic and structural stability of the models during 200 ps unconstrained molecular dynamics (MD) simulations of the unbound enzyme using CHARMm22 force fields (MacKerell et al., 1998) within the MOE program. MD simulations were run in the absence of explicit water molecules but with a distance-dependent dielectric constant. The canonical ensemble (NVT) was used with a target temperature of 300 K. An integration time step of 1 fs was used and structures were saved to disk every 100 steps.

### Results

#### Molecular modeling

A number of comparisons of the crystal structures of seven bacterial P450s and a mammalian P450 have now indicated that, despite a relatively low degree of sequence identity (only 13%) and significant size heterogeneity (45–62 kDa), their protein architectures are remarkably conserved (Hasemann et al., 1995; Peterson and Graham-Lorence, 1995; Williams et al., 2000). The conserved architecture of the P450 crystal structures consists of four to five $\beta$-pleated sheets (designated 1–5) and 15 $\alpha$-helixes (designated A–L). Overall, the structure is classified into an $\alpha$-domain that contains most of the $\alpha$-helixes with three small $\beta$-pleated sheets and a $\beta$-domain that contains the larger $\beta$- and $\beta$-sheets with three $\alpha$-helixes. Structural overlays of the C\(\alpha\) backbones of all four bacterial crystal structures [P450cam (CYP101) (Poulos et al., 1987), P450BM3 (CYP102) (Ravichandran et al., 1993; Li and Poulos, 1994, 1997), P450eryF (CYP107A1) (Cupp-Vickery and Poulos, 1995), P450orp (CYP108) (Hasemann et al., 1994)] and one mammalian P450 (CYP2C5) (Williams et al., 2000) have indicated, as shown in Figure 2, that >50% of each protein’s backbone occurs within an r.m.s.d. of 2 Å and that at least 90% of the backbone occurs within an r.m.s.d. of 5 Å. The most conserved structural elements with respect to their length and relative C\(\alpha\) positions are $\beta$-pleated sheets 1, 4 and 5 and $\alpha$-helixes D, E, I, K and L. In addition, the C\(\alpha\) backbones of the loop between helix K and strand 4 of $\beta$-pleated sheet 1 ($\beta$-sheet 1–4), the heme binding loop and the loop between helices L and K’ (designated as the ‘meander region’) are highly

<table>
<thead>
<tr>
<th>Force constant</th>
<th>Equilibrium value</th>
</tr>
</thead>
<tbody>
<tr>
<td>273.472 kcal/molÅ(^2)</td>
<td>1.786 Å</td>
</tr>
<tr>
<td>58.773 kcal/mol.degree(^2)</td>
<td>90.0°</td>
</tr>
</tbody>
</table>
S.Rupasinghe, J.Baudry and M.A.Schuler

conserved. Conserved residues shared by all of these P450s are those in the P450 signature motif (F—G—R—C—G) that contains the heme binding cysteine and the (E/D)/T pair in helix I that mediates dioxygen activation through a distal charge relay system (Atkins and Siglar, 1988; Aikens and Siglar, 1994).

Comparisons within the vertebrate CYP2 family have designated six regions (SRS1—SRS6) as important for regulating substrate specificities (Gotoh, 1992). Alignment of these regions with the compiled crystal structures has indicated that SRS1 region is located between helices B and C, SRS2 is located at the C-terminal end of helix F extending into the F—G loop, SRS3 begins in the F—G loop and extends into helix G, SRS4 is located in the middle of helix I, SRS5 is located in the loop between helix K and β1-sheet strand 4 and SRS6 is located in the loop between the two strands of the β-4 sheet.

As detailed in Materials and methods, the four phenylpropanoid pathway P450s from Arabidopsis thaliana were modeled using the CYP102 structure as the template. The four Arabidopsis P450s share 15.8–17.5% sequence identity with this bacterial P450 compared with 22.5–24.3% sequence identity with the eukaryotic CYP2C5 protein. The CYP102 structure was chosen as template because evaluation of the crystal structures with Prosa II indicated some high-energy regions in CYP2C5 structure. Profiles 3D, 3D–1D self-compatibility scores (Table IV) suggest that some residues in the structure are in incompatible environments, including residues 209–236 and 342–346 (corresponding to residues 179–206 and 312–316 in Figure 4). CYP2C5’s Prosa II normalized Z-scores of (Table III) is the lowest among all crystal structures investigated here, as has also been reported recently (Kirton et al., 2002).

Using the CYP102 structural coordinates, 10 models were generated for each P450 protein with the explicit inclusion of the heme coordinates in all steps of homology model generation. After subjecting all of these models to coarse energy minimizations that eliminate bad van der Waals contacts between atoms, the best-ranked model was further minimized using the CHARMM22 force field within the MOE program (version 2002) until the final energy gradient was <0.01 kcal/mol Å.

All four of the final models display considerable similarity in the α-domain with the P450 core structure shown in Figure 2. Specifically, all contain a well-defined A-helix with 1–3 turns, a B-helix with 1–2 turns, a B′-helix with 1–3 turns, an E-helix with 4–5 turns, a G-helix with 5–8 turns, a J-helix with 4–5 turns and a K-helix with 3–4 turns. Some variability occurs between the models in the lengths of their D-helix (2–5 turns), F-helix (3–4 turns) and L-helix (2–5 turns). Additionally, the CYP73A5 and CYP75B1 models predict a kink in the middle of the I-helix at amino acids Ala306 and Gly303, respectively. At this level of comparison, all four P450 models contain a structurally conserved five-stranded β1-sheet and more variability in the remaining three β-pleated sheets of their β-domains. Specifically, the CYP75B1 and CYP98A3 models contain an intact two-stranded β2-sheet that is not present in the CYP73A5 and CYP84A1 models. The CYP73A5 model contains an intact three-stranded β3-sheet, the CYP98A1 model contains the second and third strands of a β3-sheet and neither the CYP75B1 nor the CYP98A3 model contains a β3-sheet. The CYP75B1 and CYP98A3 models contain a two-stranded β4-sheet, the CYP73A5 model contains a shortened β4-sheet and the CYP84A1 model lacks all strands of this β-sheet.

Table II. Distribution of φ and ψ angles of the residues of the generated and minimized models, within and outside the allowed region

<table>
<thead>
<tr>
<th>Crystal structure/homology model</th>
<th>No. of residues</th>
<th>Residues outside allowed regions</th>
<th>% in allowed regions</th>
</tr>
</thead>
<tbody>
<tr>
<td>CYP73A5</td>
<td>475</td>
<td>8</td>
<td>98.3</td>
</tr>
<tr>
<td>CYP75B1</td>
<td>483</td>
<td>1</td>
<td>99.7</td>
</tr>
<tr>
<td>CYP98A3</td>
<td>478</td>
<td>9</td>
<td>98.1</td>
</tr>
</tbody>
</table>

Quality of the models
To investigate the quality of these substrate-free Arabidopsis P450 models, a variety of tests described in Materials and methods were performed. The first of these tests used Ramachandran plots to analyze the φ and ψ angle distributions for each model (Table II). This analysis demonstrates that all the models display more than 97% of their residues (excluding Gly and Pro residues) in allowed areas of the Ramachandran map and indicates that all four of these models are of a quality similar to those published for other bacterial P450 models (Chang and Loew, 1996, 2000; Chang et al., 1997).

The second of these tests used the Prosa II program to define interaction energies for each residue of the model with the remainder of the protein. The Prosa II energy profile derived for the five known crystal structures as well as for the four Arabidopsis P450 models are shown in Figure 3 with the predominant area of the profile displaying negative scores and a small number of areas representing high-energy regions displaying positive scores. The Prosa II program was also used to determine the overall folding of the models by calculating the Prosa combined Z-scores for the models and the crystal structures (Table III). The normalized Z-scores for our four models all fall just slightly below 0.7 where normalized Z-scores of ≥0.7 are accepted as good models. These slightly lower Z-scores are due to some potential high-energy regions in the B-helix and/or the B′-B′ loop in all four models, the G–H loop in two models, the D–E loop in another two models and very small areas in the D and E helices. None of these high-energy regions are located in or close to any of the predicted SRS regions with the exception of the last 17 C-terminal residues in the CYP75B1 model that includes Leu487, Thr488 and Leu489 within the SRS6 region.

The third test used Profiles 3D analysis to evaluate whether each residue’s environment is consistent with the environment in known protein structures. In this analysis (Figure 4), a residue compatibility 3D–1D score is assigned to each residue in high-energy regions. The sum of the 3D–1D scores indicates the overall correctness of the protein folds. For all four Arabidopsis P450 models, the calculated scores fall within the acceptable range for appropriately folded structures (Table IV). The high-energy regions suggested by the Profiles 3D analysis were small and distributed in several locations of the protein that were not located within or close to any predicted SRS regions. Importantly, residues in the large 17-residue region in the CYP75B1 model suggested as a high-energy region by Prosa II energy analysis was not identified as being present in the Profiles 3D analysis.

The last test applied to these energy-minimized models was to run 200 ps MD simulations to determine the thermodynamic and structural stability of each ligand-free model. Figure 5 shows the time series of thermodynamic and structural...
properties for each model computed without explicit hydration and indicates that all reach equilibrium before the first 50 ps of the MD simulation. In the evaluation of the radius of gyration and r.m.s.d. during the simulation time (Figure 5, middle and bottom panels), the sizes of the proteins do not change drastically after 50 ps. At this time point, all of their dimensions stabilize with slightly expanded dimensions compared with those predicted in the original energy-minimized model. Throughout these simulations, the temperature of the systems fluctuates at ~300 ± 10 K (Figure 5, top panel).

Substrate docking
Substrates (Figure 1) were docked within the active site of each P450 using the Monte Carlo docking procedures within the MOE program and repeated cycles of protein and substrate minimization as described in Materials and methods. Of the 25 different conformations obtained in this substrate docking procedure, the lowest energy conformation having the substrate hydroxylation site positioned in closest proximity to the heme-bound oxygen was selected as the most probable binding mode. This docking mode was subjected to energy minimization using the MMFF94 force field in MOE, allowing full side chain relaxation while keeping the heme coordinates fixed to prevent distortion of the heme plane originating from bonded parameters in the MOE’s implementation of the MMFF94 force field.

Examination of the selected binding conformations of all four models indicates definite similarities in the orientations (Figure 6). In all four models, the aromatic ring of each ligand is positioned above the protoporphyrin IX heme at an angle ranging from 53 to 83° relative to the heme, two of these models have their ligand’s axis (defined by the non-reacting region) positioned above the heme rings III and I and two (CYP73A5 and CYP98A3) have the ligand positioned above heme rings III and II. Additional similarities between the binding modes are apparent in the positioning of the substrates relative to the Cα backbone of helix I, which traverses the catalytic site above heme rings I and IV (Figure 6). In all of these predicted binding modes, the aromatic ring region is oriented toward the C-terminus of helix I and the variable length ‘tail’ is oriented toward its N-terminus. In this orientation, the distance between the iron–oxo intermediate and the conserved Thr in helix I presumed, by analogy to other P450s (Atkins and Sligar, 1988; Aikens and Sligar, 1994), to be involved in oxygen activation is 4.4 Å for Thr310 in CYP73A5, 4.3 Å for Thr323 in CYP84A1, 4.6 Å for Thr303 in CYP98A3 and 5.1 Å for Thr306 in CYP75B1. As shown in Figure 7, which depicts only those amino acids lying within 4 Å of each substrate, these orientations place the aromatic rings of all substrates in proximity to amino acids in SRS4, SRS5 and SRS6 and the variable ‘tails’ in proximity to amino acids in SRS1 and SRS2 (Figure 7).

Discussion
As previously noted by Hasemann et al. (Hasemann et al., 1995), P450 protein architectures are remarkably more conserved than their primary sequences might indicate. Our structural overlays of multiple P450 crystal structures, including P450BM3, P450terp, P450cam, P450eryF and CYP2C5, indicate that all retain the P450 core proposed by Hasemann et al. (Hasemann et al., 1995) and extend the structural similarity of the Cα backbone to a mammalian P450 and another bacterial P450. The models that we have developed for four Arabidopsis P450s in the phenylpropanoid pathway retain most of the conserved α-helices in their α-domains and varying numbers of β-pleated sheets in their β-domains.

Table III. Prosa II Z-scores for the defined P450 crystal structures and Arabidopsis P450 model structures

<table>
<thead>
<tr>
<th>Crystal structure/homology model</th>
<th>Prosa II normalized Z-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>CYP102 (BM3)</td>
<td>1.08</td>
</tr>
<tr>
<td>CYP101 (cam)</td>
<td>1.10</td>
</tr>
<tr>
<td>CYP107A1 (eryF)</td>
<td>1.08</td>
</tr>
<tr>
<td>CYP108 (terp)</td>
<td>0.97</td>
</tr>
<tr>
<td>CYP2C5</td>
<td>0.82</td>
</tr>
<tr>
<td>CYP73A5</td>
<td>0.69</td>
</tr>
<tr>
<td>CYP84A1</td>
<td>0.59</td>
</tr>
<tr>
<td>CYP75B1</td>
<td>0.66</td>
</tr>
<tr>
<td>CYP98A3</td>
<td>0.64</td>
</tr>
</tbody>
</table>

Fig. 3. Prosa II energy profiles. The Prosa II energy profiles of the four bacterial P450 crystal structures (CYP101, CYP102, CYP107A1, CYP108) are shown with amino acid positions corresponding to those in the full-length protein. The Prosa II energy profiles of the eukaryotic P450 crystal structure (CYP2C5) and four Arabidopsis P450 models are shown with amino acid positions corresponding to those in proteins lacking 30 amino acids of each signal sequence.
Of the numerous methods used to analyze the quality of the predicted structures, all indicate that, overall, the models are well folded and thermodynamically stable. Some test results, such as those obtained for \( \phi \) and \( \psi \) angle distributions in Ramachandran plots, indicate a small number of residues existing outside ‘allowed’ angles. Examination of individual models indicates that these residues typically exist outside predicted SRS regions defining catalytic site specificities in various P450s. Other test results, such as those obtained for \( Z \)-scores, which fall slightly below the recommended score of 0.7 generated in the Prosa II analysis or the \( S \)-scores generated in the Profiles 3D analysis, suggest the possibility of short regions of high energy or residues being in incompatible environments but, again, all of these residues are located outside the predicted SRS regions. More telling MD simulations indicate that these substrate-free P450 models are thermodynamically stable with r.m.s.d. and radius of gyration reaching equilibrium within 50–60 ps of the start of the simulation and the temperature of the system remaining constant throughout each simulation.

Substrate docking experiments indicate that, although these four P450s share only 13% amino acid identity, similar mechanisms for substrate binding are predicted for all of these proteins. Specifically, all of these P450s display optimal binding modes that position the aromatic rings of their substrates at a 53–83\(^\circ\) angle to the heme plane with the hydroxylation site positioned at a distance of 2.4–3.6 Å to the oxygen of the iron–oxo intermediate. In all of the four lowest energy binding conformations, the axis of the substrates aligns parallel to helix I with the aromatic ring oriented towards its C-terminus and the remainder of the molecule oriented towards the N-terminal end.

In all four models, all of the residues having at least one atom within 4 Å of the reactive ring are located in the C-terminal end of the SRS4 region or in the SRS5 and SRS6 regions (Figure 7). Conversely, all of the residues having at least one atom within 4 Å from the aliphatic end of each substrate are located in the SRS1 and SRS2 regions and the N-terminal end of the SRS4 region. These results suggest that the SRS5 and SRS6 regions and the C-terminal end of the SRS4 region are important in contacting the aromatic rings of these substrates and that the SRS1 and SRS2 regions and the N-terminal end of the SRS4 region are important in contacting the aliphatic regions of these substrates. It is worth noting that on the larger substrates, including dihydrokaempferol and \( p \)-coumaroylshikimic acid, contacts with SRS4 are not more extended than on the smaller substrates, all SRS4 interactions terminate at approximately the position of the carboxylic acid group on \( t \)-cinnamate or the aldehyde group on coniferaldehyde.

Comparison with the available substrate-bound CYP102 crystal structure (Li and Poulos, 1997) indicates that the Arabidopsis P450 substrate binding modes are different from the CYP102 substrate binding mode in which palmitoleic acid is positioned nearly perpendicular to helix I and over heme rings II and III (Figure 8). Comparisons of substrate positionings with other available substrate-bound P450 crystal structures are difficult because many of these substrates bear little resemblance to the linear cinnamate derivatives examined here. Notably, our predicted substrate binding mode for \( t \)-cinnamate in the Arabidopsis CYP73A5 catalytic site is consistent with NMR data showing that \( t \)-cinnamate binds parallel to helix I in the Helianthus tuberosus CYP73A1 protein (D.Werck-Reichhart, personal communication).

Table IV. Profiles 3D, 3D–1D scores for the defined P450s and the homology models

<table>
<thead>
<tr>
<th>Crystal structure/homology model</th>
<th>Lowest score</th>
<th>Optimum score</th>
<th>Obtained score</th>
</tr>
</thead>
<tbody>
<tr>
<td>CYP102 (BM3)</td>
<td>94.4</td>
<td>209.8</td>
<td>221.5</td>
</tr>
<tr>
<td>CYP101 (cam)</td>
<td>83.1</td>
<td>184.7</td>
<td>219.1</td>
</tr>
<tr>
<td>CYP107A1 (eryF)</td>
<td>82.7</td>
<td>183.8</td>
<td>213.8</td>
</tr>
<tr>
<td>CYP108 (terp)</td>
<td>84.6</td>
<td>187.9</td>
<td>208.5</td>
</tr>
<tr>
<td>CYP2C5</td>
<td>92.2</td>
<td>204.9</td>
<td>182.2</td>
</tr>
<tr>
<td>CYP73A5</td>
<td>97.6</td>
<td>216.9</td>
<td>131.7</td>
</tr>
<tr>
<td>CYP84A1</td>
<td>100.1</td>
<td>222.4</td>
<td>142.1</td>
</tr>
<tr>
<td>CYP75B1</td>
<td>99.3</td>
<td>220.6</td>
<td>140.8</td>
</tr>
<tr>
<td>CYP98A3</td>
<td>98.2</td>
<td>218.3</td>
<td>132.3</td>
</tr>
</tbody>
</table>
Sequence comparisons among the SRS regions in these various phenylpropanoid pathway proteins indicate that the SRS5 region predicted to contact the aromatic ring contains a highly conserved HPPhPL−L−H sequence [where lower case h is a hydrophobic residue (Karplus, 1997)] in three of the four phenylpropanoid proteins but not the CYP73A5 sequence (Figure 9). When this fourth sequence is considered, all contain a conserved hPLh−H near the C-terminus of this region. The degree of sequence identity in the longer conserved SRS5 sequence is significantly higher (7 of 10) than apparent in comparisons of these full-length proteins or in comparisons of other SRS regions (Figure 9). As shown with underlines in Figure 9, within this longer sequence and in the more divergent CYP73A5 sequence, one position corresponding to a hydrophobic Ile/Thr is consistently predicted to contact the aromatic end of the substrate in each protein. In the longer sequence, the
conserved Leu and the adjacent residue (Ser in CYP75B1 and Met in CYP98A3) are predicted to contact the aromatic ring in the larger dihydrokaempferol and p-coumaroylshikimic acid substrates metabolized by the CYP75B1 and CYP98A3 proteins, respectively. Similarly in the CYP73A5 sequence, a Leu residue aligning with the CYP75B1 Ser and CYP98A3 Met is predicted to contact its t-cinnamate substrate. Although the two conserved proline residues come within the 4 Å range for the substrate in CYP98A3, the Thr364, Leu366 and Met367 which correspond to the substrate-contacting residues of CYP75B1 are located closest to the substrate in CYP98A3.

These hydrophobic Ile/Thr and Leu contact points are embedded in a Pro-rich sequence in the CYP84A1, CYP75B1 and CYP98A3 proteins and flank a Pro residue in the CYP73A5 protein. Our models predict that the kinks in the Cα backbone induced by these Pro residues project these hydrophobic residues into the catalytic site for contacts with various aromatic substrates (Figure 7c). The first of these hydrophobic Ile/Thr residues corresponds in primary sequence alignments with Phe363 in spearmint CYP71D18, Ile371 in H. tuberosus CYP73A1 and Thr364 in Arabidopsis CYP98A3 that have been defined by site-directed mutagenesis as SRS5 residues critical for defining reactivity against limonene, t-cinnamic acid and p-coumaroylshikimic acid, their respective aromatic substrates (D.Werck-Reichhart, personal communication). In the CYP71D18 protein, mutation of Phe363 to Ile363 alters the C-6 hydroxylation of limonene to C-3 hydroxylation (Schalk and Croteau, 2000) and emphasizes the importance of SRS5 contacts with the unsaturated ring region of the substrates as suggested by our present models.

By comparison, other SRS regions of these P450s display significantly fewer absolute sequence identities (Figure 9). The SRS4 region contains a wide variety of residues that are not conserved in simple primary sequence alignments of these four divergent P450s, in contrast to comparisons of P450s done with the same subfamily. Despite these variations, there are clear conservations in the length and geometry of helix I in all four proteins and in their pattern of hydrophobic, charged and specific residues within the conserved h−c−hAGhaThA—sequence (where lower case c is a charged residue and lower case a is an acidic residue). Within this sequence, several closely aligned residues are predicted to contact the substrate in all four proteins. These include the conserved (Asp/Glu)Thr pair of residues that align with Asp251 and Thr252 in bacterial P450cam that mediate dioxygen activation (Atkins and Sligar, 1988; Aikens and Sligar, 1994), the charged Asp/Asn residue near the N-terminus of helix I and the (Ala/Gly)(Ala/Gly) pair in the center of helix I.

The SRS6 region contains an M−a−hGLh sequence that is conserved in the CYP84A1, CYP75B1 and CYP98A3 proteins and divergent in the CYP73A5 protein. Another residue N-terminal to the conserved Leu is also predicted to contact the substrate in three models except in CYP98A3 where a residue C-terminal to the Leu is predicted to make contacts. Among these, the hydrophobic Leu aligns with Phe494 in Vicia CYP94A2 that is critical for ω-fatty acid hydroxylation (Kahn et al., 2001).

Fig. 7. Predicted substrate contacts in SRS regions. The Cα backbone traces of each Arabidopsis P450 model are shown with stick diagrams depicting all amino acids containing at least one atom within 4 Å radius of the docked substrate. (a) SRS1 and SRS2 regions; (b) SRS4 regions; (c) SRS5 and SRS6 regions.
Relative to these other SRS regions, the SRS2 region lacks sequence conservations and the SRS1 region has relatively few. This is understandable in the light of their predicted roles in contacting the diversified tails on these substrates that vary significantly in their size. Even the CYP73A5 and CYP84A1 proteins that have similarly sized groups contacting this region exhibit little conservation in the identity of their contact residues apart from the aromatic residue in SRS2 that contacts both substrates.

In conclusion, these results indicate a common substrate-recognition mechanism among these four P450 proteins that involves similarities in (i) the orientation of the substrate relative to the Cα backbone and heme group of the models, (ii) the localization of residues in SRS regions contacting the substrate, and (iii) the sequence similarities of the predicted SRS regions. Future site-directed mutagenesis on these proteins will clarify the roles of individual amino acids in this substrate recognition process and the flexibility of each catalytic site in accepting alternative substrates.

Acknowledgements
The authors thank Dr Stephen Sligar and Dr Daniele Werck-Reichhart for discussions on P450 biochemistries. This work was supported by National Institutes of Health grant R01-GM50007 and National Science Foundation grant MCB0115068.

References