The cytochrome P450 (P450) superfamily plays a fundamental role in the metabolism of xenobiotics and endogenous compounds. Extensive research toward understanding the mechanism of P450-catalyzed reactions and of P450 substrate or inhibitor interactions has been done. The complexity of the reaction cycle and the lack of crystal structure information on membrane-bound P450 limited the process. Crystal structure data on mammalian cytochrome P450s started to emerge a few years ago. Before that, homology models were based on structure coordinates of bacterial isoforms (Sevrioukova et al., 1999). The bacterial P450s have low sequence similarity to mammalian P450s. The great structural conservation within the P450 superfamily (Graham-Lorence and Peterson, 1996) was, however, exploited toward the generation of many homology models (Lewis et al., 1998; Payne et al., 1999; Dai et al., 2000).

In January 2000 the structure of the first mammalian cytochrome P450, a construct of the rabbit isoform CYP2C5, 2C5/3LVdH, was deposited in the Brookhaven Protein Data Bank (Brookhaven PDB, 1dt6 (Williams et al., 2000). Except for the deletion of the membrane-bound N terminus, additional mutations based on the corresponding amino acids in the rabbit CYP2C3 sequence increased solubility and facilitated the crystallization of the protein. This was a landmark in homology modeling for the human isoforms due to the high sequence similarity of the CYP2C5 with human CYP2Cs. Numerous homology models are now based on this structure and have been used to evaluate substrate specificity and interactions, site of metabolism, and pharmacophore models (Lewis, 2000; Afzelius et al., 2001; Ridderstrom et al., 2001; de Groot et al., 2002; Lewis, 2002).

The next mammalian P450 structure published was from the same 2C5/3LVdH construct but with a dimethyl derivative of sulfaphenazole cocrystallized (1n6b) (Wester et al., 2003a). This was done at an improved resolution, from 3.0 Å to 2.3 Å, and the previously unresolved F-G loop was now captured and revealed two short helices denoted F’ and G’ within the loop. The B-C region was also better resolved and uncovered a B’-helix that is also present within many bacterial species (Peterson and Graham, 1998). Using modeling techniques, it was shown that two substrates could fill the electron density in the active site in two different binding modes. The construct was later also solved with diclofenac as a cocomplex (Wester et al., 2003b) (1mr6). In this complex, the site of metabolism of diclofenac was found at a reasonable distance (4.4 Å) from the iron to facilitate hydroxylation. At the same time, the first human cytochrome P450s were crystallized, a mutated form of the CYP2C9 in the substrate free

**ABBREVIATIONS:** P450, cytochrome P450; PDB, Protein Data Bank; CPCA, consensus principal component analysis; MIF, molecular interaction field; RMSD, root mean square deviation; 3D, three-dimensional; PCA, principal component analysis; MD, molecular dynamics.
form (1og2) and in complex with the anticoagulant warfarin (1og5) (Williams et al., 2003). In this structure an unexpected binding mode was seen where the site of oxidation was positioned 10 Å from the heme. Other crystal structures that have been solved include the substrate-free form of 2B4 (1PO5) that was resolved to 1.6 Å resolution. The structure is captured as a reversible homodimer in an open conformer where His 226 forms an intermolecular coordinate bond to the heme iron. An ∼15-Å-wide cleft leads down to the heme. It will be of greatest interest to study changes upon substrate binding for 2B4 and the plausible closure of the active site around the substrate to elucidate the conformational freedom available to these proteins. Structures that have been solved also include the human 2C8 to 2.7 Å (1PQ2) and the recent release (June 15th, 2004) of the wild-type human 2C9 cocryosylized with flubiprofen to 2.0 Å (1R90), in which no mutations except those in the terminal ends have been made.

The crystal structures provide a significant amount of data that can be explored by computational methods to improve our understanding of P450 enzymes. The information can be used to validate homology models and predictors of inhibition/metabolic stability/activation, but also to build new hypothesis.

Nevertheless, a tool for analysis is required to put all this information in concrete form. In this work we present the novel application of consensus principal component analysis (CPCA) to explore modeling success, which is often evaluated based on docking and stereochemical parameters (Szklarz and Halpert, 1997; Kirkton et al., 2002). The work also includes, to our knowledge, the pioneer use of CPCA as a tool for evaluating molecular dynamics simulations. Selectivity analysis between isoforms is also performed successfully with this methodology, which has been reported previously (Kastenholz et al., 2000; Ridderstrom et al., 2001). The CPCA was used for a comparative analysis of available crystal structures to homology models based on single or multiple templates and snapshots from molecular dynamics simulations for CYP2C9 and CYP2C5, respectively. The analysis is restricted to the active site of the proteins, which is described by molecular interaction fields (MIFs) calculated by the program GRID (Goodford, 1985). Multivariate data analysis is applied to these descriptors to identify selective regions of the MIFs. The selective MIF highlight regions and type of interactions in the binding site that reflect differences between the structures.

This analysis gave an increased general understanding of the structural characteristics of each isoform; selective regions were identified and changes induced by binding were traced. The study also validated the performance of computational techniques such as homology modeling and molecular dynamics simulations.

### Materials and Methods

**Equipment and Software.** Molecular interaction fields were calculated in an Irix environment on a Silicon Graphics O2 workstation (Silicon Graphics Inc., Mountain View, CA) and in Linux environment on a 32MB personal computer. The software utilized in the computational analysis was GRID v21 (Molecular Discovery Ltd., http://www.moldiscovery.com), GOLPE (Baroni et al., 1993), MetaSite (Zamora et al., 2003), MODELLER v6.1 (http://salilab.org/modeller/modeller.html), SYBYL 6.5.3 (Tripos Associates Inc., St Louis, MO), PROCHECK v.3.4.3 (Laskowski et al., 1993), ClustalX (ftp://ftp.genome.ad.jp/pub/ClustalX/clustalx1.83.alpha.tar.gz), and AMBER 7 (Case et al., 2002).

**Overview.** The analysis was performed in different steps. In the first step, all available crystal structures of CYP2C9 and CYP2C5 were compared to show whether high sequence similarity proteins share the same geometry and interaction features in the binding site and to recognize selective regions that correspond to isoform-specific reactions.

Second, the interaction maps of the active site of homology models of CYP2C9 and CYP2C5, based on different templates, were compared with the crystal structures. This enables the evaluation of techniques available for work on proteins with a high degree of similarity.

In the third step, snapshots from molecular dynamics simulations of CYP2C9 and CYP2C5 crystals in explicit water were analyzed to determine whether they could capture changes upon substrate binding and to determine which parts of the cavity were more flexible and could participate in substrate recognition and access. Finally, all structures, crystals, homology models, and molecular dynamics conformational maps for both CYP2C5 and CYP2C9 were compared in a CPCA to see whether they overlap in chemical space and how they intracorrelate dependent on the modeling technique used.

**Protein Homology Modeling.** Comparative modeling techniques were used to prepare homology models for CYP2C9 and CYP2C5. This process requires one or several homologous crystal structures referred to as template structures. The amino acid sequence for the desired protein is referred to as the target. The crystal template structures were selected from a PSI-BLAST search against the Brookhaven PDB to identify suitable template structures for comparative modeling (Kirton et al., 2002). The following templates (Table 1) were downloaded from the Brookhaven PDB (http://www.rcsb.org/pdb/): P450BM-3 (1bu7; Sevrioukova et al., 1999), P450tcp (1cp; Hasemann et al., 1994), P450tryF (1eup; Cupp-Vickery et al., 2000) and P450cam (1phc; Poulos et al., 1986), CYP2C9 (1og2; Williams et al., 2003), CYP2C9 with warfarin cocystalized (1og5; Williams et al., 2003), CYP2C9/3LVdH with diclofenac (1nr6; Wester et al., 2003b). The target and template structures were selected from a PSI-BLAST search against the Brookhaven PDB to identify suitable template structures for comparative modeling (Kirton et al., 2002). The following templates (Table 1) were downloaded from the Brookhaven PDB (http://www.rcsb.org/pdb/): P450BM-3 (1bu7; Sevrioukova et al., 1999), P450tcp (1cp; Hasemann et al., 1994), P450tryF (1eup; Cupp-Vickery et al., 2000) and P450cam (1phc; Poulos et al., 1986), CYP2C9 (1og2; Williams et al., 2003), CYP2C9 with warfarin cocystalized (1og5; Williams et al., 2003), CYP2C9/3LVdH with diclofenac (1nr6; Wester et al., 2003b). The target and template structures were downloaded from the SWISS-PROT data bank (http://us.expasy.org/spiro/).

A total of 30 homology models, 15 for CYP2C9 and 15 for CYP2C5, were built using single or multiple templates (Table 2). The exact same procedures were used to generate models for CYP2C9 and CYP2C5 independently, the only difference being that the crystal structure of CYP2C9 was the template for the CYP2C5 models and the crystal structure of CYP2C5 was the template for

### Table 1

<table>
<thead>
<tr>
<th>Crystal Structures</th>
<th>PDB Code</th>
<th>RMSD 1 vs. CYP2C9</th>
<th>RMSD 2 vs. CYP2C5</th>
<th>Ramachandran (%) *&lt;br&gt; (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CYP2C9</td>
<td>1og2</td>
<td>0</td>
<td>3.5475</td>
<td>85.8</td>
</tr>
<tr>
<td>CYP2C9-warfarin</td>
<td>1og5</td>
<td>0.3598</td>
<td>1.9591</td>
<td>87.2</td>
</tr>
<tr>
<td>CYP2C5</td>
<td>1df6</td>
<td>1.9764</td>
<td>0</td>
<td>70.6</td>
</tr>
<tr>
<td>CYP2C5-DMZ</td>
<td>1nfb</td>
<td>1.3367</td>
<td>1.559</td>
<td>85.7</td>
</tr>
<tr>
<td>CYP2C5-diclofenac</td>
<td>1nrf</td>
<td>1.46</td>
<td>1.687</td>
<td>86.9</td>
</tr>
<tr>
<td>BM-3</td>
<td>1bu7</td>
<td>*</td>
<td>*</td>
<td>93.1</td>
</tr>
<tr>
<td>P450cam</td>
<td>1phc</td>
<td>*</td>
<td>*</td>
<td>89.7</td>
</tr>
<tr>
<td>P450terp</td>
<td>1eup</td>
<td>*</td>
<td>*</td>
<td>89</td>
</tr>
<tr>
<td>P450coryF</td>
<td>1eup</td>
<td>*</td>
<td>*</td>
<td>90.8</td>
</tr>
</tbody>
</table>

DMZ, 4-methyl-N-methyl-N-(2-phenyl-2H-pyrazol-3-yl)benzenesulfonamide.  
* Sequence similarity <30%; alignment not to be relied on.  
** Percentage of residues with psi and phi conformations in the “most favored regions” of the Ramachandran plot.
Align structure using homology" mode (Tables 1 and 2). The stereochemical parameters were checked in Ramachandran plots calculated in PROCHECK. The backbone RMSD toward template and target was calculated in SYBYL using the "most favored regions" of the Ramachandran plot.

The changes in the intracorrelation of secondary structure elements during the two runs of molecular dynamics simulations for CYP2C9 were examined

<table>
<thead>
<tr>
<th>Homology Models</th>
<th>1D.</th>
<th>RMSD 1 vs CYP2C9</th>
<th>RMSD 2 vs CYP2C5</th>
<th>Ramachandran (%)b</th>
</tr>
</thead>
<tbody>
<tr>
<td>CYP2C9 2D alignment to CYP2C5</td>
<td>CYP2C9_2D_1</td>
<td>2.19</td>
<td>0.26</td>
<td>85.9</td>
</tr>
<tr>
<td></td>
<td>CYP2C9_2D_2</td>
<td>2.19</td>
<td>0.19</td>
<td>84.9</td>
</tr>
<tr>
<td></td>
<td>CYP2C9_2D_3</td>
<td>2.18</td>
<td>0.19</td>
<td>85.6</td>
</tr>
<tr>
<td></td>
<td>CYP2C9_2D_4</td>
<td>2.20</td>
<td>0.19</td>
<td>86.1</td>
</tr>
<tr>
<td></td>
<td>CYP2C9_2D_5</td>
<td>2.23</td>
<td>0.22</td>
<td>85.6</td>
</tr>
<tr>
<td>3D alignment to CYP2C5 and BM3</td>
<td>CYP2C9_3D_A_1</td>
<td>2.24</td>
<td>0.61</td>
<td>83.1</td>
</tr>
<tr>
<td></td>
<td>CYP2C9_3D_A_2</td>
<td>2.23</td>
<td>0.69</td>
<td>83.4</td>
</tr>
<tr>
<td></td>
<td>CYP2C9_3D_A_3</td>
<td>2.23</td>
<td>0.78</td>
<td>83.9</td>
</tr>
<tr>
<td></td>
<td>CYP2C9_3D_A_4</td>
<td>2.23</td>
<td>0.78</td>
<td>83.9</td>
</tr>
<tr>
<td></td>
<td>CYP2C9_3D_A_5</td>
<td>2.24</td>
<td>0.77</td>
<td>83.6</td>
</tr>
<tr>
<td>3D alignment to all crystal templates</td>
<td>CYP2C9_3D_B_1</td>
<td>2.23</td>
<td>0.86</td>
<td>80.6</td>
</tr>
<tr>
<td></td>
<td>CYP2C9_3D_B_2</td>
<td>2.26</td>
<td>0.84</td>
<td>80.4</td>
</tr>
<tr>
<td></td>
<td>CYP2C9_3D_B_3</td>
<td>2.23</td>
<td>0.64</td>
<td>81.1</td>
</tr>
<tr>
<td></td>
<td>CYP2C9_3D_B_4</td>
<td>2.23</td>
<td>0.68</td>
<td>81.6</td>
</tr>
<tr>
<td></td>
<td>CYP2C9_3D_B_5</td>
<td>2.24</td>
<td>0.59</td>
<td>80.4</td>
</tr>
<tr>
<td>CYP2C5 2D alignment to CYP2C9</td>
<td>CYP2C5_2D_1</td>
<td>0.34</td>
<td>2.18</td>
<td>91.6</td>
</tr>
<tr>
<td></td>
<td>CYP2C5_2D_2</td>
<td>0.35</td>
<td>2.17</td>
<td>92.1</td>
</tr>
<tr>
<td></td>
<td>CYP2C5_2D_3</td>
<td>0.35</td>
<td>2.19</td>
<td>92.3</td>
</tr>
<tr>
<td></td>
<td>CYP2C5_2D_4</td>
<td>0.36</td>
<td>2.17</td>
<td>93.1</td>
</tr>
<tr>
<td></td>
<td>CYP2C5_2D_5</td>
<td>0.36</td>
<td>2.17</td>
<td>91.6</td>
</tr>
<tr>
<td>3D alignment to CYP2C9 and BM3</td>
<td>CYP2C5_3D_A_1</td>
<td>0.50</td>
<td>2.13</td>
<td>89.6</td>
</tr>
<tr>
<td></td>
<td>CYP2C5_3D_A_2</td>
<td>0.50</td>
<td>2.12</td>
<td>91.3</td>
</tr>
<tr>
<td></td>
<td>CYP2C5_3D_A_3</td>
<td>0.57</td>
<td>2.13</td>
<td>89.9</td>
</tr>
<tr>
<td></td>
<td>CYP2C5_3D_A_4</td>
<td>0.64</td>
<td>2.14</td>
<td>90.8</td>
</tr>
<tr>
<td></td>
<td>CYP2C5_3D_A_5</td>
<td>0.53</td>
<td>2.13</td>
<td>90.3</td>
</tr>
<tr>
<td>3D alignment to all crystal templates</td>
<td>CYP2C5_3D_B_1</td>
<td>2.94</td>
<td>3.61</td>
<td>84</td>
</tr>
<tr>
<td></td>
<td>CYP2C5_3D_B_2</td>
<td>2.93</td>
<td>3.60</td>
<td>85.9</td>
</tr>
<tr>
<td></td>
<td>CYP2C5_3D_B_3</td>
<td>2.57</td>
<td>3.29</td>
<td>86.2</td>
</tr>
<tr>
<td></td>
<td>CYP2C5_3D_B_4</td>
<td>2.95</td>
<td>3.64</td>
<td>87.4</td>
</tr>
<tr>
<td></td>
<td>CYP2C5_3D_B_5</td>
<td>3.08</td>
<td>3.66</td>
<td>85.4</td>
</tr>
</tbody>
</table>

*Percentage of residues with psi and phi conformation in the “most favored regions” of the Ramachandran plot.

The CYP2C9 models. First, models were built based only on the mammalian CYP2C templates (CYP2C5 or CYP2C9). The two-dimensional alignment was made in ClustalX and, due to the great similarity, the deviations were small. Second, the crystal structures of BM3 were aligned to either of the human templates based on the three-dimensional structure in the MALIGN module in MODELLER with different gap penalties (0–2, 2–4, and 4–6). These alignments were considered as a profile in ClustalX and the target sequence was aligned toward it. Finally, the same procedure was repeated for the template CYP2C and multiple bacterial templates (P450BM-3, P450terp, P450eryF, and P450cam).

When this preliminary alignment was done in MALIGN and ClustalX, a detailed manual alignment process continued. The secondary structures of the sequences were predicted using PSIPRED (http://bioinf.cs.ucl.ac.uk/psipred/). In the alignment, gaps within predicted helices and sheets were removed to preserve the secondary structure since it is known to be well conserved throughout the superfAMILY of P450s (Graham-Lorence and Peterson, 1996). Thereafter, MODELLER was set to generate five models for each of the three alignments, giving 15 different models each for CYP2C9 and CYP2C5 (Table 2). To ensure quality, in terms of geometry, the stereochemical parameters were checked in Ramachandran plots calculated in PROCHECK. The backbone RMSD toward template and target was calculated in SYBYL using the “Align structure using homology” mode (Tables 1 and 2).

**Molecular Dynamics Simulations.** Setup and equilibration of CYP2C9 and CYP2C5 in solution. The molecular dynamics simulations were made using the SANDER module in AMBER 7 (Case et al., 2002). The force field used was AMBER99 (Cornell et al., 1995). The force field parameters for the heme group were adopted from the parameterization available from the parameter database of the AMBER force field (Giammona, 1984). Using the crystal structures of CYP2C9 (1og2) and CYP2C5 (1dt6), hydrogen atoms were added and crystallographic ambiguous side chain orientations were optimized using Reduce (Word et al., 1999). The hydrogen atoms were minimized for 500 optimization steps in vacuum, keeping the heavy atoms of the protein fixed. For an initial relaxation of the protein, this structure was minimized for 150 steps using a position restrain on the backbone. Two different starting structures were used in the molecular dynamics simulations, the first with the oxygen bound to the iron (the oxyferryl species) and the second without an oxygen bound to the iron (Ortiz de Montellano and De Voss, 2002).

To further relax the structure in an aqueous solution, the protein was immersed in a cubic box of TIP3 water. Electroneutrality was achieved by adding sodium ions. The water molecules were minimized and then equilibrated for 25 ps with a molecular dynamics simulation at constant volume and a temperature ramp from 5°K to 300°K, keeping the protein rigid. The position restraint on the protein was gradually removed in subsequent minimization steps. Next, the entire system (protein, water molecules, and counter ions) was heated from 5°K to 300°K over 20 ps and equilibrated for 200 ps in an additional constant pressure-constant temperature molecular dynamics simulation. The system was then run for a subsequent trajectory of 800 ps used for structural analysis.

A time step of 2 fs was used for the Particle–Mesh–Ewald molecular dynamics simulation with a nonbonded list update every 10 steps. All bonds involving hydrogens were constrained with the SHAKE (Ryckaert et al., 1977) algorithm. A cut-off of 9 Å was applied. The temperature/pressure was maintained by the Berendsen weak-coupling scheme (Berendsen et al., 1984).
by measuring distances between α-carbons in the backbone for certain selected amino acids (see Table 4). This was done to obtain indications of possible substrate access channels. For Lys72, the terminal nitrogen position was also monitored, since this flexible side chain had been suggested to gate an access channel selective for anionic compounds (Williams et al., 2003).

Molecular Interaction Field Calculations using GRID. All protein structures (see Tables 1 and 2) were 3D aligned to the CYP2C9 crystal structure (1og2) in SYBYL based on the smallest RMSD to the backbone atoms. The proteins were imported into the GRID interface called Greater. In GRID, the carboxy terminus of a protein and the carboxy groups of Asp and Glu are treated by default as anionic. Similarly, the N-terminal nitrogen and Arg and Lys side chains are treated as cationic, and the overall net charge of the protein is then calculated by summation. This is an arbitrary calculation that gives an overall net charge. If the charge differs between the proteins that are studied, the electrostatic effect can dominate the predicted interactions of charged probes. Therefore, the proteins were made neutral by positioning movable Na⁺ or Cl⁻ ions at minima energy positions calculated by GRID. Since these ions are movable they will not compete with a probe (see the GRID manual http://www.moldiscovery.com/docs/grid21/index.html). Secondly, a box was defined to include the heme, the active site, and access channels. The exact same box size, 34 × 30 × 40 Å, was used for all calculations to enable the subsequent comparison in the statistical software GOLPE. The MOVE directive, which alters flexibility of the target, was set to 0 (rigid mode) and the subsequent comparison in the statistical software GOLPE. The MOVE directive, which alters flexibility of the target, was set to 0 (rigid mode) and the subsequent comparison in the statistical software GOLPE. The MOVIE directive, which alters flexibility of the target, was set to 0 (rigid mode) and the LIST directive, which defines the file output format, to −2. The default values were used for the other parameters. The following probes were included; OH², DRY, C3, N1, N1⁺, O⁻, and O (see Table 3).

Active Site Cut-Out File Generated in MetaSite. Many interaction points that are calculated in GRID within the defined box are not accessible to the substrate. Therefore, the nonaccessible points were deleted in the subsequent CPCA to remove noise from the analysis. Cut-out files were prepared for each of the crystal structures and then merged into a single file. This was done in MetaSite, which is a program used to predict the site of metabolism of the most common cytochromes (Zamora et al., 2003) (CYP3A4, CYP2C9, CYP2D6, CYP2C19 and CYP1A2). The program automatically detects the protein cavity that are intrinsically organized in blocks in the descriptor matrix; i.e., the results for each probe generate a block in the matrix. The CPCA can be considered as a PCA on two levels. First, the principal components are derived for the entire descriptor matrix at a superlevel which is identical to the usual PCA. Second, the principal components are derived on a block level (= probe). The second derivation of principal components is not exactly a standard PCA procedure since the principal components are rotated to reproduce the scores from the superlevel PCA. The minimization criteria are not just to get the lowest residual values, but also to reproduce the values obtained in the PCA level. Consequently, the first principal component in the block level does not need to explain the maximum variance as in a normal PCA; instead, it reflects the importance of a specific probe in the superlevel analysis.

All models that are discussed were calculated using eight probes, OH², DRY, C3, N1, N1⁺, O⁻, and O. The different steps in the analysis are visualized in Fig. 2. The connection between the first and second level e.g., the influence of each block in the description of the score plot for the entire set, is analyzed by the CPCA superweight plot (Fig. 2a). From this level, the relative importance of each probe is distinguished as the plot highlights the influence of the different blocks into the principal components. For each probe the results are visualized in a score plot (Fig. 2b) in which the objects are positioned based the loadings at the probe active site cut-out file (Table 3). The CPCA is an algorithm developed (Kastenholz et al., 2000) from the principal component analysis (PCA).

A principal component analysis identifies “underlying” data structure or variables that best summarize the information of the original descriptors by explaining the variance in the data. PCA is a technique that reduces the dimensionality of a data matrix to a smaller number of underlying variables called principal component or latent variables, which are a linear combination of the variables. To a greater or lesser extent, all descriptors contribute to the component extraction. The first principal component is a line through the data space that explains the data with the least squares residual. This type of fitting ensures that the first component explains the maximum variance, the second component is orthogonal to the first component and explains as much as possible of the variance that was not explained by the first component and is derived from the residuals obtained from the first component. The first two components generate a plane to which all objects can be projected. This plane is called the score plot and describes the position of the observations based on the latent variables. Observations that are close in the score plot have a comparable variance and are similar. In addition, the methodology provides the loadings plot, in which the importance of each of the original descriptors in defining the latent variable is described. The score plots (observations; e.g., the proteins) and loading plots (descriptors; e.g., molecular interaction energies for each grid point) are related. Variables that are positively correlated to an observation are positioned at the same place in the loadings plot as the observation in the score plot. Since we are dealing with only negative energies here, where more negative interactions are more favorable, the interpretation is reversed, so that superimposed scores and loadings are negatively correlated and an observation in the score plot is positively correlated to the loadings at the same coordinates, but with opposed signs. In the PCA it can be difficult to distinguish the relative importance of different probes if more than two objects are studied. In these cases, the analysis benefits from a CPCA, to enable the recognition of single amino acids responsible for binding by studying each probe both individually and under the influence of all probes.

In the CPCA (Kastenholz et al., 2000), the data that originate from several probes are intrinsically organized in blocks in the descriptor matrix; i.e., the results for each probe generate a block in the matrix. The CPCA can be considered as a PCA on two levels. First, the principal components are derived for the entire descriptor matrix at a superlevel which is identical to the usual PCA. Second, the principal components are derived on a block level (= probe). The second derivation of principal components is not exactly a standard PCA procedure since the principal components are rotated to reproduce the scores from the superlevel PCA. The minimization criteria are not just to get the lowest residual values, but also to reproduce the values obtained in the PCA level. Consequently, the first principal component in the block level does not need to explain the maximum variance as in a normal PCA; instead, it reflects the importance of a specific probe in the superlevel analysis.

Many interaction points that are calculated in GRID within the defined box are not accessible to the substrate. Therefore, the nonaccessible points were deleted in the subsequent CPCA to remove noise from the analysis. Cut-out files were prepared for each of the crystal structures and then merged into a single file. This was done in MetaSite, which is a program used to predict the site of metabolism of the most common cytochromes (Zamora et al., 2003) (CYP3A4, CYP2C9, CYP2D6, CYP2C19 and CYP1A2). The program automatically detects the protein cavity starting from the protein reactive center, the oxygen bound to the iron, as implemented in MetaSite (Fig. 1).

Consensus Principal Component Analysis (CPCA). The molecular interaction fields generated in GRID were imported into GOLPE. Pretreatment was applied to focus on the areas of interest and increase the signal-to-noise ratio: 1) grid points within a radius of 4 Å from the active site cut-out file were selected using the cut-out tool (Fig. 1), 2) all positive energies, corresponding to close contacts between the probe and the protein, were excluded, together with all variables showing a variance of less than 0.01 standard deviation, and 3) block unscaled weights were applied. This is a scaling factor that depends on the variance for each block. All blocks are given the same variance to normalize the interaction energies between the probes. The data set was suitable for the subsequent multivariate analysis, the CPCA. The CPCA is an algorithm developed (Kastenholz et al., 2000) from the principal component analysis (PCA).

A principal component analysis identifies “underlying” data structure or variables that best summarize the information of the original descriptors by explaining the variance in the data. PCA is a technique that reduces the dimensionality of a data matrix to a smaller number of underlying variables called principal component or latent variables, which are a linear combination of the variables. To a greater or lesser extent, all descriptors contribute to the component extraction. The first principal component is a line through the data space that explains the data with the least squares residual. This type of fitting ensures that the first component explains the maximum variance, the second component is orthogonal to the first component and explains as much as possible of the variance that was not explained by the first component and is derived from the residuals obtained from the first component. The first two components generate a plane to which all objects can be projected. This plane is called the score plot and describes the position of the observations based on the latent variables. Observations that are close in the score plot have a comparable variance and are similar. In addition, the methodology provides the loadings plot, in which the importance of each of the original descriptors in defining the latent variable is described. The score plots (observations; e.g., the proteins) and loading plots (descriptors; e.g., molecular interaction energies for each grid point) are related. Variables that are positively correlated to an observation are positioned at the same place in the loadings plot as the observation in the score plot. Since we are dealing with only negative energies here, where more negative interactions are more favorable, the interpretation is reversed, so that superimposed scores and loadings are negatively correlated and an observation in the score plot is positively correlated to the loadings at the same coordinates, but with opposed signs. In the PCA it can be difficult to distinguish the relative importance of different probes if more than two objects are studied. In these cases, the analysis benefits from a CPCA, to enable the recognition of single amino acids responsible for binding by studying each probe both individually and under the influence of all probes.

In this work the massive amount of data, experimental and calculated, made it necessary to base the analysis on a multivariate analysis technique such as the CPCA. This type of analysis enables the reduction of the data dimensionality and the recognition of patterns in data. Each pattern can then be correlated to the actual discriminative structural characteristics. The protein structures analyzed can thereby be grouped together with other similar structures. If there is no matching pattern for a certain structure, it is nontypical compared with the rest of the data set. In some of the cases described below, the structural characteristics are explored on an amino acid level. In other cases, general relationships between structures are more informative.

<table>
<thead>
<tr>
<th>Probe</th>
<th>Chemical Group</th>
<th>No. of H bonds</th>
</tr>
</thead>
<tbody>
<tr>
<td>OH₂</td>
<td>Water</td>
<td>2</td>
</tr>
<tr>
<td>DRY</td>
<td>Hydrophobic</td>
<td>0</td>
</tr>
<tr>
<td>C3</td>
<td>Methyl group</td>
<td>0</td>
</tr>
<tr>
<td>N1</td>
<td>Neutral flat NH (e.g., amide)</td>
<td>0</td>
</tr>
<tr>
<td>N¹</td>
<td>sp² amine N¹ cation (+1)</td>
<td>0</td>
</tr>
<tr>
<td>O⁻</td>
<td>sp² phenolate oxygen</td>
<td>2</td>
</tr>
<tr>
<td>O</td>
<td>sp² carbonyl oxygen</td>
<td>0</td>
</tr>
</tbody>
</table>
Crystal Structures of CYP2C9 versus CYP2C5. First, the differences between the crystal structures of CYP2C9 and CYP2C5 are highlighted since they are also highly influential in the latter analysis of the homology models. The comparison between the crystal structures with and without the bound substrate has been extensively described in their original publications (Wester et al., 2003a,b; Williams et al., 2003) and will therefore not be described in depth here. Visual inspection and distance measurements between the crystal structures of CYP2C9 and CYP2C5 (without substrates bound) (Table 4; Fig. 3) reveal that the greatest changes occur in the merging region of the B-C loop, the F-G loop, and the N terminus. This is also reflected in the B-factors of the structures (Fig. 4a). In CYP2C9, the C-terminal loop (Val473, Asn474, Gly475, and Phe476) and the N-terminal loop (Ile47, Lys48) are closer to each other, making the distance between the C-terminal and the F-G loops (Ile207, Leu208, Ser209, Ser210, Pro211, and Trp212) larger than in the CYP2C5...
structure. In the CYP2C5 structure, the C terminus (Val470, Asn471, Gly472, and Phe473) and the F-G loops (Leu207, Leu208, Gly209, Thr210, Pro211, and Trp212) are closer to each other and the active site is smaller, since they are positioned nearer to the heme than in the CYP2C9 structure. The F-G loop and its F' and G' helices seem to form a lid to the active site. In the CYP2C5 structure, the lid restricts the active site, which also gives rise to a displacement of the B-C loop. However, the low resolution of the B-C loop for the CYP2C5 crystal structure and the possible influence of mutations in the F-G loop of the CYP2C9 structure have to be kept in mind. In comparison, the

### TABLE 4

<table>
<thead>
<tr>
<th>Point 1</th>
<th>Point 2</th>
<th>Molecular dynamics</th>
<th>log2</th>
<th>1dt6</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>First run</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N-terminal</td>
<td>Val38</td>
<td>β1-1</td>
<td>Gly70</td>
<td>4.37</td>
</tr>
<tr>
<td>Gly70</td>
<td>8.07</td>
<td>10.67</td>
<td>2.60</td>
<td>9.15</td>
</tr>
<tr>
<td>Asn218</td>
<td>3.90</td>
<td>7.69</td>
<td>3.79</td>
<td>4.47</td>
</tr>
<tr>
<td>N-terminal</td>
<td>Lys48</td>
<td>C-termin</td>
<td>Phe476</td>
<td>8.11</td>
</tr>
<tr>
<td>β1-1</td>
<td>5.29</td>
<td>13.25</td>
<td>7.96</td>
<td>5.60</td>
</tr>
<tr>
<td>β1-2</td>
<td>10.32</td>
<td>13.68</td>
<td>3.36</td>
<td>11.12</td>
</tr>
<tr>
<td>B-C loop</td>
<td>Pro101</td>
<td>F-G loop</td>
<td>Pro229</td>
<td>5.35</td>
</tr>
<tr>
<td>B-C loop</td>
<td>Ala106</td>
<td>G-helix</td>
<td>Lys241</td>
<td>12.55</td>
</tr>
<tr>
<td>F-helix</td>
<td>Glu206</td>
<td>I-helix</td>
<td>Glu300</td>
<td>9.68</td>
</tr>
<tr>
<td>F-G loop</td>
<td>Pro211</td>
<td>N-terminal</td>
<td>Lys48</td>
<td>7.51</td>
</tr>
<tr>
<td>F-G loop</td>
<td>Pro211</td>
<td>B-C loop</td>
<td>Pro101</td>
<td>13.67</td>
</tr>
<tr>
<td>F-G loop</td>
<td>Pro211</td>
<td>C-terminal</td>
<td>Phe476</td>
<td>7.34</td>
</tr>
<tr>
<td>Heme</td>
<td>Fe</td>
<td>F-G loop</td>
<td>Asn218</td>
<td>22.75</td>
</tr>
<tr>
<td><strong>Second run</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N-terminal</td>
<td>Val38</td>
<td>β1-1</td>
<td>Gly70</td>
<td>3.64</td>
</tr>
<tr>
<td>Gly70</td>
<td>7.73</td>
<td>12.49</td>
<td>4.75</td>
<td>9.15</td>
</tr>
<tr>
<td>Asn218</td>
<td>3.66</td>
<td>6.51</td>
<td>2.84</td>
<td>4.47</td>
</tr>
<tr>
<td>Lys72N</td>
<td>4.34</td>
<td>6.59</td>
<td>2.26</td>
<td>5.60</td>
</tr>
<tr>
<td>B-C loop</td>
<td>Pro101</td>
<td>F-G loop</td>
<td>Pro220</td>
<td>5.41</td>
</tr>
<tr>
<td>F-G loop</td>
<td>Pro211</td>
<td>N-terminal</td>
<td>Lys48</td>
<td>10.11</td>
</tr>
<tr>
<td>F-G loop</td>
<td>Pro211</td>
<td>B-C loop</td>
<td>Pro101</td>
<td>12.50</td>
</tr>
<tr>
<td>F-G loop</td>
<td>Pro211</td>
<td>C-terminal</td>
<td>Phe476</td>
<td>6.31</td>
</tr>
<tr>
<td>Heme</td>
<td>Fe</td>
<td>F-G loop</td>
<td>Asn218</td>
<td>24.68</td>
</tr>
</tbody>
</table>

aa, amino acid.

Corresponding aa in CYP2C5: * Ile38, † Phe473, ‡ Glu297, ¶ Ser99.

**Fig. 3.** Visualization of the active-site regions of the crystal structure of CYP2C9 (1og5) in magenta versus the crystal structure of CYP2C5 (1dt6) in cyan. This representation has been used consistently throughout the paper.
B-C loop in the CYP2C9 structure is well determined and more ordered with the presence of a B'-helix. The results of a CPCA, including all crystal structures, with and without substrates bound, available from the PDB for CYP2C9 and CYP2C5 are shown in Fig. 2. The superweights plot (Fig. 2a) describes the influence of each probe in the analysis. In this plot, all probes but the hydrophobic (denoted 2) have a similar contribution in the first component. The DRY probe, however, is of similar importance in the second component. The water probe (Fig. 2e) shows selectivity for CYP2C9 in a corner of the active site where Asn107, Gly109, and the backbone of Arg108, all in the B-C loop, could interact possibly via a water molecule, with Asp293 and Asn289 in the I-helix. Asp293 has previously been suggested from mutagenesis studies to have a structural role in substrate recognition and catalytic activity. This is further supported by the conservation of the Asp residue over other P450 families (Flanagan et al., 2003). Arg108, which has also been suggested to be of functional importance from mutagenesis data (Ridderstrom et al., 2000), is pointing away from the active site. This might be explained by the conformation of the B-C loop in the crystal structure, since this region is highly flexible and a charge-charge interaction would be favorable. This interaction is not possible in the CYP2C5 structure since the B-C loop is oriented differently (Fig. 3). In the region of the active site where the F-G loop (Ile208, Ser209, Ser210, and Gln214) approaches the region of the C terminus and the backbone carbonyl of Asn474, another favorable hydrogen bonding possibility is found. The DRY probe has three distinct regions in the active site that are also known to be important from mutagenesis studies (Melet et al., 2003): one close to Val113 and Phe114, a second one in the proximity of Phe476 and Phe100, and a third one near Leu102 and Ala103. The C3 probe is important in describing steric changes, and in this case, the most important loadings identify a pocket near the heme close to Thr301, Thr304, Ala477, and Leu361. The N1 probe is a mimic of an amide nitrogen and has a possibility to donate one hydrogen in a favorable direction to any available lone pair of electrons. The carbonyl of the peptide backbone between Ser478 and Val479 is highly involved, but the hydroxyl group of Thr301 and the peptide bond carbonyl between Leu361 and Leu362 are also directionally well positioned to interact with a hydrogen bond-donating group. Another selective patch is due to the carbonyl side chain of Gln214, and the backbone carbonyl groups between Leu208-Arg209 and Ser210-Pro211 and on the opposite side of the pocket between Asn474 and Gly475. Asn217 is another amino acid of importance in selectivity. The last selective patch seen for this probe is due to Asp293 and the backbone carbonyls between Arg108 and Gly109. The N1" probe is the cation of an sp3-hybridized amine and should therefore emphasize interactions with negatively charged amino acids, e.g., the carboxylic acids of aspartic and glutamic acid. The selectivity is mainly confined to Asp293, but with minor contributions from Asp360 and Asp49. These two amino acids are positioned on the protein surface and are therefore of little or no importance for the active site binding. The last two probes included in the analysis describe a phenolate and a carbonyl oxygen, respectively, and both have the possibility to accept two hydrogen bonds. The phenolate is negatively charged and will therefore interact with positively charged amino acids such as lysine and arginine. The loadings analysis showed that the exact same positions were selective for the O2- probe and the O probe and, thus, only the interactions for the O probe are further described. The carbonyl probe finds several selective regions, mainly close to the heme and Thr301. This highly conserved amino acid is of great importance to the proton transfer path (Schlichting et al., 2000) in the catalysis of substrates.

Homology Models versus Crystal Structures. In Fig. 5, a and b, the PCA score plots for homology models versus their targets are visualized for CYP2C9 and CYP2C5 separately. In Fig. 5c, the information included in Fig. 5, a and b, is combined in a single analysis. It is obvious from this simple statistical analysis of the active site that the models are close mimics of their templates. That is, the CYP2C9 homology models closely resemble the CYP2C5 crystal structure from which they were built and not the target, the CYP2C9 crystal structure. Models built from multiple targets are in one case (CYP2C5_3D:B_1-5; Fig. 5b) more similar to their target than to their templates. The idea of retaining conserved regions from multiple templates can be rewarding if the alignment is successful, as in the case mentioned. This can be a difficult task in cases where the sequence identity is low; i.e., identity lower than 30%. Then, the alignment has to be performed by initially introducing a low gap penalty to increase the weight of the low-identity structure. In the next step, gaps must then be removed to retain the secondary structure. In
this procedure, it can be difficult to determine, for example, which is the first and last amino acid in a secondary element, since secondary structure predictors give slightly different results.

Even though spatial restraints force the geometry of the homology models toward the template, the influence of the different amino acid sequences makes the homology model resemble the target. The influence of different amino acid sequences on a ligand can be explored in different ways and is a measurement of the quality of the model. An excellent choice is to dock compounds that are substrates of both 2C9 and 2C5 but with different sites of metabolism and see how well this outcome is predicted. However, such specific information is difficult to obtain. Also, due to the low specificity and selectivity among isoforms, commercial docking algorithms most often give rise to multiple binding modes. Scoring functions are seldom able to pick the preferred orientation, in which the site of metabolism points toward the heme at a reasonable distance for a reaction to occur. In this work, we have explored the active sites of crystal versus homology models with different probes and then focused on regions that have major differences defined by a CPCA differential plot, which signals lower pattern recognition in AMBER did not induce fundamental changes as compared to the first run: a horseshoe shape. On the second run, the two molecular dynamics simulations were believed to generate a similar plot in the CPCA space. This analysis showed that the minimization in AMBER did not induce fundamental changes as compared to the unsolvated crystal structures, and the structures were placed in the same quadrant as the initial crystal structures (1og2 and 1og5). The exception was the finally minimized structure in AMBER that approached the first conformations of the simulations. Because the simulations should explore the conformational freedom of the protein, the hypothesis was that the crystal structure with and without bound substrate would be found within the conformational space covered by the molecular dynamics simulations. Although the starting structures were slightly different (see Materials and Methods), the two molecular dynamics simulations were believed to generate a similar plot in which at least part of the structures would mix in the CPCA space. The snapshots from the second molecular dynamics run were analyzed in the exact same way as for the first run. The shape of the distribution was very similar to the first run: a horseshoe shape. On the
basis of the distribution in the CPCA score plot, the same four conformations, 1, 50, 150, and 200, were chosen as representatives for the entire run. In the next step, the selected conformers from both simulations were analyzed together with the starting structures; the results are shown in Fig. 7a. The first component discriminated between the first MD run and the second MD run together with the starting structures. The second component discriminated between the second MD run and the starting structure. Each of the runs and the starting structures were distributed in one quadrant each and did not overlap at any simulation time.

To gain a structural understanding of the movements during the molecular dynamics simulations, a number of distances between the secondary elements were measured for each of the snapshots for the two separate runs and compared with the crystal structure (Table 4). The overall RMSD and the average range for all distances measured correspond well between the two runs, but the individual distances differ significantly. The most apparent change is that the nitrogen of the Lys72 in the first run moves over an 8-Å range compared with its original position, whereas it moves only 2.3 Å in the second run (see Discussion). With regard to the other movements, the same regions are flexible through both runs, but the internal correlations of movements differ. The main conclusion made from these MD runs is that the protein is highly flexible. The parts of the protein that have high B-factors in the crystal structure also show great flexibility in the dynamics (Fig. 4b). As a next step, the homology models were added to the analysis to find out how these efforts correlate to the molecular dynamics simulations.

It must be kept in mind throughout the analysis that the crystal structures of CYP2C9 are solved from a construct including seven amino acid substitutions based on CYP2C sequences to improve properties for crystallization (Williams et al., 2003). All of them are located in the F-G loop that is found throughout the analysis to be of great importance for substrate recognition, binding, active-site volume, and probably membrane association. Only a wild-type CYP2C9 crystal structure will show the actual influence of these mutations. A wild-type crystal structure of CYP2C9 (1R9O) was released on the June 15, 2004, but seven amino acids, Gly214-Ser220, of the F-G loop were not located in the experiment. Due to the recent release, the structure was not included in the calculations but was considered in the interpretation. Apart from that, the protein has been released from the membrane by cutting the N terminus, and this could also induce structural changes. The CYP2C5 structure also lacks the N-terminal part, and five amino acid positions have been changed for the corresponding residues in CYP2C3 (Williams et al., 2000). In the first

Discussion

This work had two separate aims: to evaluate different modeling techniques and to make a detailed structural characterization of CYP2C9. To achieve these goals, the CPCA technique and distance measurements were used to explore the available crystal structures, the newly built homology models, and repeated molecular dynamics simulations.

FIG. 6. Selectivity between the homology models CYP2C9_3D:A_2 (gray) and CYP2C9_3D:B_3 (cyan) for the water probe. Regions of selectivity are marked by the amino acids responsible for the interactions. The Arg108 is highly interesting since it points in opposite directions in the discussed homology models. These two modes are consistent with the contradictory findings in the two crystal structures of human 2C9 available to date.
crystal structure, without substrate bound, the F-G loop was not resolved and the loop coordinates were described on the basis of calculations. During the process of structural determination, errors can be introduced at several stages. Apart from pure experimental measurement errors, loops can be trapped in unphysiological conformations and errors can be introduced in the process of modeling atoms into the electron density. A crystallographic model should therefore be evaluated with a degree of uncertainty.

The CPCA is based on possibilities for interacting with different probes, and it is very useful to elucidate structure characteristics with regard to steric and electrostatic properties. It is thus highly dependent on how the possible sites of interactions, e.g., amino acid side chains, are oriented. In some cases these results will be biased by how the side chains were built into the electron density and the template for the refinement process. A difference between a carbon and an oxygen is not distinguishable in the electron density at this resolution due to the similar number of electrons.

The homology models are of good quality from a stereochemical point of view, where the percentage within the most favored regions in a Ramachandran plot is higher in the models than in the templates. Nevertheless, this work clearly shows that a model typically resembles the underlying target. The homology modeling algorithms are per se trained to base the models on similarities. In one case, the introduction of multiple templates improved the results significantly, and this approach is therefore recommended, although it requires careful manual adjustment of the alignment. Initially, lower gap penalties have to be introduced in the alignment to facilitate an influence from templates of lower similarity, but then gaps corrupting evolutionary conserved regions must be removed manually. If several templates of high amino acid identity are present, great improvements can be expected, which can be the case in the near future when several, although far from all, isoforms have been crystallized. CYP2C8 or CYP2C19 will be interesting targets since they represent two templates of similar high sequence identity, like CYP2C9 and CYP2C5. The resulting data can then be assessed in a manner similar to that done here.

The reliability of homology models must be evaluated on the basis of the question asked. Substrates with a low selectivity and moderate affinity could probably be predicted successfully, whereas rational design of high-affinity and high-selectivity compounds are likely to have a much lower reliability since these are dependent on strongly corresponding complementarities. Despite the drawbacks presented
here, the benefits of homology models are indisputable. The models, which are being revised over time as new experimental data emerge, give insight and understanding by the mutual forming and discarding of hypotheses, as the models are refined. Apart from that, due to the limitations of the process of structural determination, even the most sophisticated crystal diffraction data will be afflicted with an experimental error, which makes the final result a model itself.

As a next step in the analysis, the molecular dynamics information was evaluated. The molecular dynamics simulations cover a different CPCA space from the crystal structures with and without substrate bound, independent of the different starting structures. Consequently, the simulations cannot be used to predict changes associated with substrate binding. On the other hand, it would not be probable, since the driving force of an approaching substrate is not present.

Nevertheless, the data can be used to produce mechanistically possible suggestions of where flexibility occurs and how it affects the surroundings. That was explored by measuring distances between backbone α-carbons of secondary structure elements to explore openings and closures that could correspond to access channels. These results show that there is great conformational rearrangement in the protein. The overall RMSD and range of movement are similar for both MD runs, but the most flexible parts differ between the runs. During the first run, one major conformational freedom was seen in the opening between the B-C loop and the β1-1 sheet. This entrance is guarded by a lysine, Lys72, where the terminal nitrogen moves up to 8 Å during the simulations. It has been suggested that substrate recognition and access occur at the opening between the B-C loop and the β1-1 sheet. This entrance is guarded by a lysine, Lys72, where the terminal nitrogen moves up to 8 Å during the simulations. It has been suggested that substrate recognition and access occur at the opening between the B-C loop and the F-G loop, which opens a channel toward the N terminus, where the substrate specificity is seen (Jung et al., 1998). However, recent mutagenesis data show that Lys72 has little or no effect on the interaction with the polar compounds ibuprofen and diclofenac. These findings rule out the critical role of this amino acid in determining substrate specificity (Davies et al., 2004). In the second MD run, this region shows only minor flexibility which emphasizes the difficulty of making structural assumptions on the basis of molecular dynamics in the case of the highly flexible CYP2C enzymes.

In both runs, conformational flexibility is seen between the B-C loop and the F-G loop, which opens a channel toward the N terminus, which also moves out, whereas the C terminus remains in place. In CYP2C5, an access channel is described between the B'- and the C-helices and helices G and I. This channel is closed upon substrate binding by hydrogen bonding of Lys241 with the backbone carboxyl of Val106 (Wester et al., 2003b). This opening is larger in the CYP2C9 crystal structure and increases during the simulation. A second solvent channel was seen in the CYP2C5 structure between the F- and I-helices. The distance between the conserved amino acids Glu300 (Glu297 in CYP2C5) of the I-helix and Glu206 of the F-helix was therefore measured and the backbone moved considerably during both runs. The influence of the mutations in the F-G loop must be taken into consideration when these results are interpreted, especially since the positively charged amino acid Glu206. The association to the membrane is also likely to affect the conformational freedom in this region. The F-G loop may function as a lid to the active site and the hydrophobic outer part is likely to be attached to the membrane. Recognition and access of substrates or solvent seem possible, in the region between the G'- and B'-helices, between the B-C loop and the β1-1, and between the F- and I-helices. The CYP2C5 structure has a smaller active site, where the F-G loop is positioned closer to the heme (Table 4), and it seems reasonable to believe that this is also a possible conformation for CYP2C9. The wider active site described by the CYP2C9 crystal structure could reflect the conformation in which this crystal has been captured. A half-opened structure could then also rationalize the position of warfarin as being a transition state. It seems reasonable, therefore, to believe that CYP2C9 and CYP2C5 are more similar than they seem from experiments and that differences could be a result of the experimental uncertainties introduced by low resolution and mutations.

It is our opinion that the results of molecular dynamic simulations have to be evaluated with care. At the moment, it is still difficult to simulate or validate a biological process such as substrate access and recognition. Simulations might give a picture of what is mechanistically possible for the protein, but the actual course of events also strongly depends on external factors such as the influence of an approaching substrate and the presence of the membrane.

Nevertheless, modeling attempts such as homology modeling and molecular dynamics simulations create hypotheses that can later be used to design experiments and analyze emerging experimental data.

References


Address correspondence to: Lovisa Afzelius, AstraZeneca R&D, Mölndal, S-431 83 Mölndal, Sweden. E-mail: Lovisa.afzelius@astrazeneca.com