



## 3D-QSAR methods on the basis of ligand–receptor complexes. Application of COMBINE and GRID/GOLPE methodologies to a series of CYP1A2 ligands

Juan José Lozano<sup>a,b</sup>, Manuel Pastor<sup>c</sup>, Gabriele Cruciani<sup>c</sup>, Katrin Gaedt<sup>a</sup>, Nuria B. Centeno<sup>a,b</sup>, Federico Gago<sup>d</sup> & Ferran Sanz<sup>a,e,\*</sup>

<sup>a</sup>Research Group on Medical Informatics, IMIM, C/ Dr. Aiguader, 80, E-08003 Barcelona, Spain; <sup>b</sup>School of Medicine, Universitat Autònoma de Barcelona, C/ Dr. Aiguader, 80, E-08003 Barcelona, Spain; <sup>c</sup>Laboratory on Chemometrics, Department of Chemistry, Università di Perugia, I-06123 Perugia, Italy; <sup>d</sup>Department of Pharmacology, Universidad de Alcalá, E-28871 Alcalá de Henares, Madrid, Spain; <sup>e</sup>School of Health and Life Sciences, Universitat Pompeu Fabra, C/ Dr. Aiguader, 80, E-08003 Barcelona, Spain

Received 3 March 1999; Accepted 9 November 1999

**Key words:** AUTODOCK, COMBINE, cooked food heterocyclic amines, cytochrome P450 1A2, docking simulation, GRID/GOLPE, 3D-QSAR

### Summary

Many heterocyclic amines (HCA) present in cooked food exert a genotoxic activity when they are metabolised (N-oxidated) by the human cytochrome P450 1A2 (CYP1A2h). In order to rationalize the observed differences in activity of this enzyme on a series of 12 HCA, 3D-QSAR methods were applied on the basis of models of HCA–CYP1A2h complexes. The CYP1A2h enzyme model has been previously reported and was built by homology modeling based on cytochrome P450 BM3. The complexes were automatically generated applying the AUTODOCK software and refined using AMBER. A COMBINE analysis on the complexes identified the most important enzyme–ligand interactions that account for the differences in activity within the series. A GRID/GOLPE analysis was then performed on just the ligands, in the conformations and orientations found in the modeled complexes. The results from both methods were concordant and confirmed the advantages of incorporating structural information from series of ligand–receptor complexes into 3D-QSAR methodologies.

### Introduction

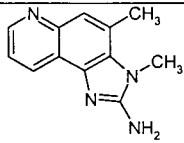
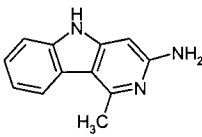
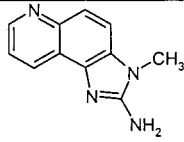
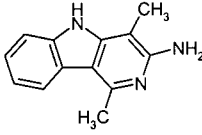
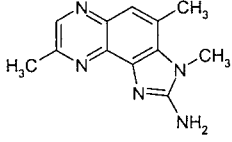
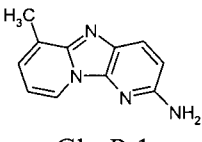
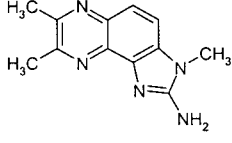
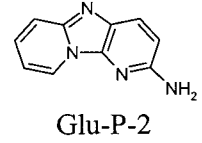
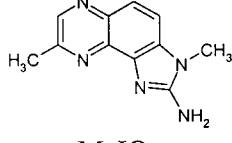
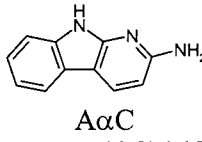
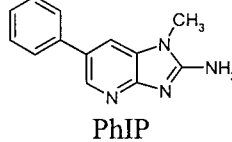
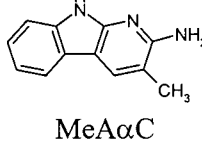
The cytochromes P450 (CYP) group of monooxygenases is a superfamily of enzymes present in both prokaryotic and eukaryotic organisms, responsible for the oxidative metabolism of many endogenous and exogenous compounds [1, 2]. CYP enzymes play a key role in the metabolic transformations required for the excretion of such compounds, but in some circumstances they participate in the bioactivation of some compounds to toxins or carcinogens [3]. CYP1A2 is a member of the CYP1 family involved in the metabolism of conjugated planar compounds, such

as xanthines [4] and some antibacterial quinolones [5]. CYP1A2 is also responsible for the bioactivation of heterocyclic aromatic amines (HCA), formed during the cooking of meat and meat-derived products, to hydroxylamines that exhibit different degrees of mutagenic activity [6, 7] (Table 1).

The absence of three-dimensional structures of the enzyme, either free or complexed with the substrates, hampers the rationalization of the observed differences in activity among series of homologous CYP1A2 substrates. The experimental determination of 3D structures of mammalian cytochromes P450 has proved to be complex because they are membrane-bound proteins. However, some 3D models of mammalian cytochromes P450 [8–10], built by homology

\*To whom correspondence should be addressed. E-mail: ferran@imim.es

Table 1. Mutagenic heterocyclic amines

Formulae and name	TA98 <sup>a</sup>	CYP1A2 act <sup>b</sup>	Formulae and name	TA98 <sup>a</sup>	CYP1A2 act <sup>b</sup>
 <p><b>MeIQ</b> (2-amino-3-methylimidazo- [4,5-f]quinoline)</p>	5.82	4.53	 <p><b>Trp-P-2</b> (3-amino-1-methyl-5H-pyrido- [4,3-b]indole)</p>	5.00	2.87
 <p><b>IQ</b> (2-amino-3,4-dimethylimidazo- [4,5-f]quinoline)</p>	5.62	4.47	 <p><b>Trp-P-1</b> (3-amino-1,4-dimethyl-5H- pyrido[4,3-b]indole)</p>	4.50	3.06
 <p><b>4,8-DiMeIQx</b> (2-amino-3,4,8-trimethylimidazo- [4,5-f]quinoxaline)</p>	5.26	n.a	 <p><b>Glu-P-1</b> (2-amino-6-methyldipyrido- [1,2-a:3',2'-d]imidazole)</p>	4.50	3.69
 <p><b>7,8-DiMeIQx</b> (2-amino-3,7,8-trimethylimidazo- [4,5-f]quinoxaline)</p>	5.21	n.a	 <p><b>Glu-P-2</b> (2-aminodipyrido- [1,2-a:3',2'-d]imidazole)</p>	3.27	1.95
 <p><b>MeIQx</b> (2-amino-3,8-dimethylimidazo- [4,5-f]quinoxaline)</p>	5.16	3.73	 <p><b>AαC</b> (2-amino-9H-pyrido[2,3-b]indole)</p>	2.47	n.a
 <p><b>PhIP</b> (2-amino-1-methyl- 6-phenylimidazo[4,5-b]pyridine)</p>	3.20	n.a	 <p><b>MeAαC</b> (2-amino-3-methyl-9H-pyrido- [2,3-b]indole)</p>	2.30	n.a

<sup>a</sup>Base-10 logarithm of mutagenicity estimated using *S. typhimurium* TA98 [7].

<sup>b</sup>Activation of procarcinogens in the presence of purified human CYP1A2 measured as the base-10 logarithm of umu gene expression [6].

techniques from known 3D structures of prokaryotic cytochromes P450 [11–14], have been published.

Recently, a model of CYP1A2h was built in our group with the aim of improving our understanding of this important enzyme [10]. In that work, the docking of some substrates was studied and the identity of the most important residues for the enzyme–substrate interaction was suggested. The present work attempts to gain further insight by means of the automated docking of a larger series of compounds into the binding site of CYP1A2h, and the subsequent analysis of the quantitative relationships between the molecular structures and the biological activities.

The free energy of interaction between a ligand and its receptor provides a measure of the strength of the association between the two molecules. Although simple estimates of binding energies can sometimes successfully correlate with experimental differences in enzyme activity, as shown for a series of mutants of the P450 2a4/2a5 system [15], more accurate theoretical estimations of binding free energies are often necessary. For this purpose, thermodynamic integration and free energy perturbation calculations [16, 18] or semi-empirical approaches based on linear response theory [17] can yield absolute free energies of binding. These methods, however, are still time-demanding and beyond average computational resources.

3D-QSAR approaches are commonly viewed as less costly and more convenient alternatives to the calculation of free energy differences. Provided that the compounds are similar enough, these methods can supply reliable models relating the molecular differences within the series with the differences in biological activity. In such cases, many of the terms that are difficult to estimate in free energy studies cancel out, because they take approximately the same values for all the compounds in the series. A 3D-QSAR approach closely related to docking studies is the COMBINE approach [19–22], which relies on the use of a series of structures of ligand–receptor complexes (modeled or experimentally determined) to quantify interaction energies by molecular mechanics computations. Then, partitioning of each ligand–receptor interaction energy into van der Waals and electrostatic contributions per residue is followed by a partial least squares (PLS) regression analysis in order to derive a 3D-QSAR model. The interpretation of this model can help identify the residues that are more relevant for the differences in activity and quantify their relative importance.

On the basis of 3D distributions of interaction energies between selected probes and the studied compounds, 3D-QSAR models are being developed. Interesting approaches are the one developed by Goodford [23], CoMFA [24] or GRID/GOLPE [25, 26] methods. In these, although the structure of the ligand–receptor complexes is not strictly required, it is necessary to define a relevant alignment for the ligands. Instead of computing the ligand–receptor interaction energies, interaction energy fields are calculated on a lattice of points around the ligands by means of ‘probes’ that represent chemical groups potentially present on the receptor. The subsequent PLS analysis, using the energy fields as descriptors and the biological activity as the dependent variable, can highlight the relative importance for activity of certain types of interaction and certain regions around the compounds. In the present study, both methodologies (COMBINE and GRID/GOLPE) were applied to a series of HCA-CYP1A2h complexes. Application to the same problem of these two independent methodological approaches, one based on the structures of the ligands alone and the other based on the structures of the ligand–protein complexes, should enable us to compare the performance of the two methods and to increase the feasibility of the results.

## Materials and methods

### *Biological activities of the heterocyclic amines*

Table 1 shows the chemical structures and biological activities of the HCA studied. These 12 CYP1A2h substrates were chosen for our study because they are frequently found in cooked foods [27, 28]. Directly measured binding affinities of such amines for the CYP1A2h are not available at present. In the absence of a direct measure, HCA activation in the presence of purified CYP1A2h would be a relevant index, but unfortunately this index is only available for seven of the compounds studied [6]. An alternative biological index that is available for more compounds is the mutagenicity of the metabolites expressed as revertants/nmol in *Salmonella typhimurium* TA98 [7]. The metabolic path driving the bioactivation of HCA to mutagens involves two steps: HCA metabolism by CYP1A2, and N-acetylation mostly undertaken by the human arylamine-N-acetyltransferase isoenzyme NAT2. Since the first process has been described as the limiting step [30], a close correlation should exist

between the susceptibility of a certain HCA to activation by CYP1A2h and its mutagenicity. Indeed, for the seven amines with known values for both activities, the correlation coefficient is  $r^2 = 0.91$  (Figure 1). The linear relation between mutagenicity and CYP1A2 activity is confirmed by experimental information on PhiP, which has both low mutagenicity and low activity [31] (this information is not included in Figure 1 because it originates from a different experiment).

#### *HCA docking in human CYP1A2*

In order to apply the COMBINE method to the series under study, feasible 3D models of the amines docked within the previously reported CYP1A2h structure [10] had to be generated. With this aim, the docking program AUTODOCK 2.4 [32] was used to perform an automatic exploration of possible docking orientations for different conformations of the ligands. The original force field parameters were used, with the exception of those defining the interaction of the nitrogen atom of the exocyclic amino group of the HCA with the iron of the heme group. For this pair of atoms an equilibrium distance of 4.2 Å and a force constant of 50 kcal Å<sup>-2</sup> were used in order to force the ligands to approximately keep the experimentally known distance between the oxidation site and the iron atom [33]. The AUTODOCK exploration was carried out within a 30 Å cube, and a 0.5 Å grid spacing was used. The simulated annealing protocol used in the AUTODOCK simulations consisted of 20 runs of 50 cycles, each cycle including 25000 accepted or 25000 rejected relative positions. The annealing temperature, RT, was set to 616 cal mol<sup>-1</sup> during the first cycle, and then linearly reduced at the end of each cycle, following the protocol used by Morris et al. [32]. The clustering tolerance was set to a root-mean-square (RMS) deviation of 1.5 Å. Ligands were considered conformationally flexible by defining the torsional angles about which rotation was allowed. From the set of docking positions suggested by AUTODOCK, the lowest energy conformations from the most populated clusters were selected. In order to take into account the conformational flexibility of the protein, the resulting ligand–enzyme complexes were refined using the AMBER 4.1 program [34, 35], as explained below.

Solvation of the ligands in the binding site was considered explicitly because water molecules are known to play an important role in enzymatic catalysis, as seen in hydrophilic P450 substrates like 6-desoxyerythronolide B and phenylimidazoles [14, 36].

On the other hand, from available crystallographic structures of bacterial P450 cam [11] it is known that, in substrate-free structures, approximately six water molecules are present in the active site. When a ligand binds the active site of the enzyme, some or all of these solvent molecules are displaced. Solvent molecules remaining after the binding process can be justified as filling empty space or helping ligands to establish water-bridged hydrogen bonds. In the active site of the crystal structure of the complex of P450 cam with 2-phenyl-imidazole two water molecules were found [36]. One acts as the sixth ligand to the heme iron whereas the other one is a well-defined solvent molecule that participates in hydrogen bonds between the heterocyclic nitrogen of 2-phenyl-imidazole and adjacent protein residues. In the present work the same solvation scheme based on the electrostatic potential distributions of the ligands, as described and applied to MeIQ in our previous study [10], was extended to the remaining 11 HCA.

#### *AMBER calculations*

In order to allow for force-field differences between the present work and the receptor modeling work [10] a mild relaxation of the docked complexes was required. Otherwise, the ligand–enzyme interactions could be affected by differences in van der Waals radii, charges and force constants. In a first step, only the hydrogen atoms of each complex were allowed to reorientate. Then, the geometry of the ligand was optimized, as well as the water molecules present in the binding site. The water molecules were restrained to their positions by applying an harmonic force constant of 5 kcal mol<sup>-1</sup> Å<sup>-2</sup> on the water oxygens. Finally, the geometry of the whole complex was optimized, while restraining the protein atoms to their original positions by a force constant of 32 kcal mol<sup>-1</sup> Å<sup>-2</sup>. Each of the energy minimizations was carried out using the steepest descent algorithm for 2000 steps or until the RMS value of the potential energy gradient was lower than 0.01 kcal mol<sup>-1</sup> Å<sup>-1</sup>. A cut-off of 10 Å for the non-bonded interactions and a distance-dependent dielectric constant equal to 4r were used.

Atom centered charges for the ligands were derived by projecting the quantum-mechanical molecular electrostatic potential distributions, calculated at the HF/6-31G\*\* level of theory using GAUSSIAN 94 [37]. In order to obtain a maximum consistence with the rest of the charges included in AMBER 4.1 (force field parm91), the RESP program [38] was used. The

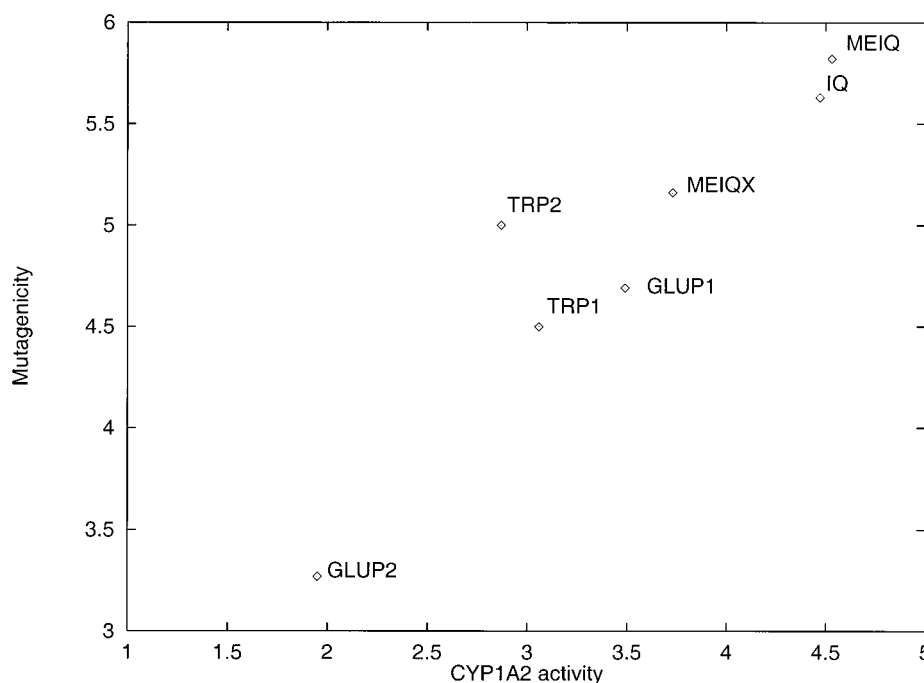


Figure 1. Mutagenicity [7] versus human CYP1A2 activity [6].

rest of the parameters for the ligands were derived by analogy from those already present in the AMBER database. The parameters proposed by Bayly, and implemented in AMBER 4.1, were used for the heme group [34].

#### COMBINE analysis

The ligand–enzyme interaction energies of the refined complexes were calculated and partitioned on a per-residue basis using the ANAL module of the AMBER software. Each solvated ligand or residue was regarded as a single fragment and no intramolecular energy terms were considered. Some of the ligands studied bind two molecules of water, whereas others bind only one. Thus, we decided to incorporate into the ligand only the water molecule that is present in all the complexes. The fine-tuning hydrogen bonding term present in the AMBER parameterization used [35] was considered separately from the electrostatic interaction because the compounds of the series have potential hydrogen bond acceptor sites, which play an important role in ligand orientation within the binding site. Since the 479 residues of CYP1A2 plus the heme group were independently considered in the partitioning scheme, and three energy contributions (van der Waals, electrostatic and hydrogen bonding) were

calculated for each residue, 1440 energy variables ( $480 \times 3$ ) were used to characterize each complex.

The resulting energy matrix was pre-treated by zeroing all the variables with absolute values lower than  $0.05 \text{ kcal mol}^{-1}$  and by removing those with a standard deviation below  $0.05 \text{ kcal mol}^{-1}$ . In order to equalize the importance of the blocks of variables the Block Unscaled Weight (BUW) method was used, as implemented in the GOLPE program [25]. Finally, three consecutive fractional factorial design (FFD) variable selections were performed over the matrix obtained after pre-treatment. The optimal dimensionality of the PLS models was determined by monitoring the cross-validation indexes as a function of the number of latent variables extracted. For cross-validation, the compounds were assigned randomly to any of five groups of approximately the same size, and the whole procedure was repeated 20 times. The predictive ability of the resulting models is reported by both the cross-validated correlation coefficient ( $q^2$ ) and the standard deviation of error of predictions (SDEP).

#### GRID/GOLPE analysis

Using a superimposition of the enzyme atoms as a common frame of reference, the previously described complexes provided the alignment of the HCA that

was used in the GRID/GOLPE analysis. The molecular interaction fields were calculated for each ligand using a phenolic OH probe as implemented in the program GRID [29]. This group is able to donate and accept hydrogen bonds and it has an electronic environment that corresponds to the interaction with the  $\pi$  system of a phenyl ring, resulting in a hydrogen bonding pattern different from that of an aliphatic hydroxyl probe. All GRID calculations were performed in a box with dimensions equal to  $14 \times 17 \times 16 \text{ \AA}^3$  using a grid spacing of  $1 \text{ \AA}$ . The resulting 3808 probe–target interactions for each compound were unfolded to produce a one-dimensional vector of variables for each compound, which were assembled into the so-called X matrix. This matrix was pre-treated by first using a cut-off of  $+5 \text{ kcal mol}^{-1}$  to produce a more symmetrical distribution of energy values, then zeroing those values with absolute values smaller than  $0.1 \text{ kcal mol}^{-1}$ , and finally removing any variables with a standard deviation below  $0.1 \text{ \AA}$ . In addition, variables taking only two or three values and presenting a skewed distribution were also removed [25].

For the variable selection procedure, the Smart Region Definition (SRD/GOLPE) method [39] was used defining a critical distance of  $2.0 \text{ \AA}$  and a collapsing cut-off of  $27.2 \text{ \AA}$ .

## Results and discussion

### Docking simulations

Three representative examples of HCA in their complexes with CYP1A2h (MeIQ, Trp-P-1 and PhIP) are shown in Figures 2 to 4. These HCA have been chosen because MeIQ is the most active compound, Trp-P-1 exhibits an intermediate activity and shows a particular docking geometry, and PhIP is the only low-activity compound that does not possess a fused tricyclic core.

The MeIQ complex (Figure 2) is stabilized by three hydrogen bonds involving the two water molecules considered in the simulation. One hydrogen is donated from the exocyclic amine to the oxygen of the heme-coordinated water. The heterocyclic nitrogen vicinal to the exocyclic amine accepts a hydrogen bond from the same water molecule, forming the second hydrogen bond. The third hydrogen bond in the MeIQ complex connects the pyridine-like nitrogen, which acts as a hydrogen bond acceptor, with the second water molecule, which in turn accepts another hydrogen bond from the aliphatic hydroxyl group of Thr223.

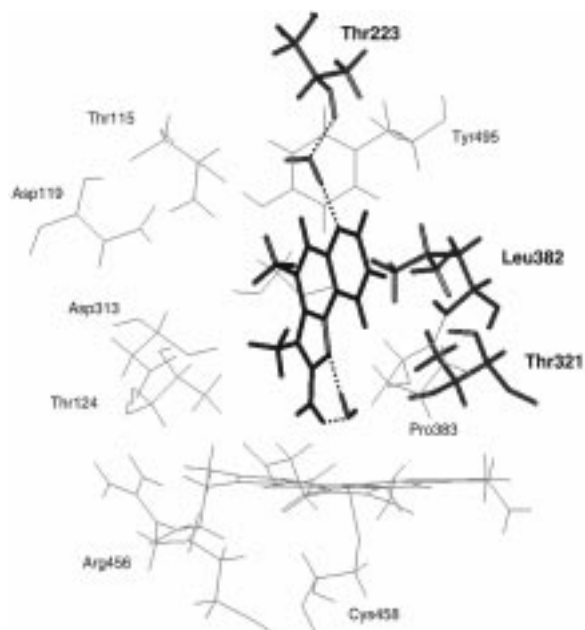


Figure 2. Model of MeIQ (black) docked into the binding site of CYP1A2h (gray). Important residues are shown as sticks.

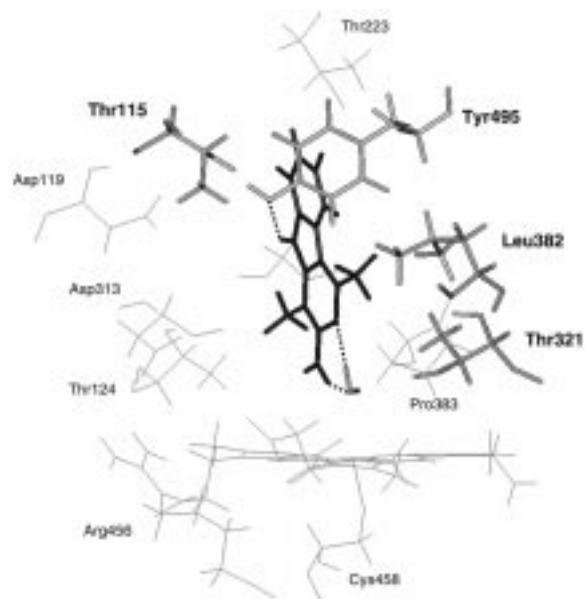


Figure 3. Model of Trp-P-1 (black) docked into the binding site of CYP1A2h (gray). Important residues are shown as sticks.

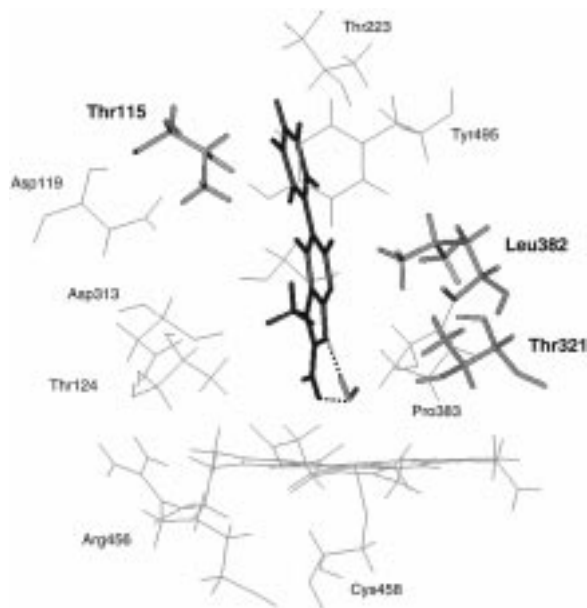


Figure 4. Model of PhIP (black) docked into the binding site of CYP1A2h (gray). Important residues are shown as sticks.

The Trp-P-1 complex (Figure 3) is also stabilized by three hydrogen bonds, the first and the second being the same as in the MeIQ complex, but the third one is formed between the hydrogen of the nitrogen located in the central pyrrole-like ring and the hydroxyl group of Tyr495. Finally, in the PhIP complex (Figure 4) only two hydrogen bonds were present. Instead of the third hydrogen bond, a favorable but weaker aromatic-aromatic interaction between the phenyl groups of the substrate and Tyr495 was observed.

On the basis of their complexes with CYP1A2h, the alignment used for the 12 HCAs was that shown in Figure 5. It can be observed that the docking simulation and subsequent energy refinement protocols yielded two clusters of orientations corresponding to the more active (dark gray) and the less active (light gray) HCA, respectively. The more active compounds (MeIQ, IQ, 4,8-DiMeIQx, 7,8-DiMeIQx, MeIQx and Glu-P-1) orientate their third ring towards a zone of the active site that is unoccupied in five of the six complexes with the less active compounds (Trp-P-1, Trp-P-2, PhIP, A $\alpha$ C and MeA $\alpha$ C), and the planes in which the heterocyclic systems of both kinds of compounds are located appear slightly rotated. The only exception of the mentioned clustering is Glu-P-2, which adopts the orientation of the most active compounds although it is a low-activity compound. The results of the COMBINE and GRID/GOLPE analyses

should be expected to be influenced by these different orientations of the more active and the less active compounds.

#### COMBINE analysis

In order to assess the optimal number of latent variables (LV) to include in the PLS model, indexes of the fitting ( $r^2$ ) and predictive power ( $q^2$ ) of the models were plotted against the model dimensionality (Figure 6a). The maximum quality was obtained for a model with three LV, but the slight improvement in comparison to a model with only two LV and the small sample size ( $n = 12$ ) recommend the use of the model with only two LV. This model includes 75 original variables in the PLS model and exhibits a  $r^2 = 0.90$ , a  $q^2 = 0.74$  and a SDEP = 0.60. The activities calculated by the COMBINE model (Figure 6b) were, in general, quite similar to the experimental values, with the exception of Glu-P-2 ( $Act_{exp} = 3.27$  and  $Act_{pred} = 4.58$ ). This phenomenon is related to the fact that the orientation obtained for this compound in the active site was similar to the orientation of the most active compounds, as mentioned above.

In a COMBINE model, the values of the weighted PLS pseudo-coefficients (herein referred to as PLS coefficients) for the different ligand-residue interactions can be analyzed in order to know which are the most relevant ligand-residue interactions [20, 21]. The signs of those coefficients express if favorable interactions (negative energy values) correlate with an increase (negative coefficient) or a decrease (positive coefficient) in activity. Attention was focused on the PLS coefficients with absolute values larger than 0.1. The most important residues in terms of steric, electrostatic or hydrogen bonding interactions are shown in Table 2. These residues are located in four regions (herein referred to as regions A, B, C and D) which encompass the substrate recognition site proposed by Gotoh [40] and the heme region.

#### Region A

This region is located in a homologous region that comprises helix B' and turn B'-C and also residues of helix F. Because of the presence of gaps, these helices are the most variable and problematic in sequence alignments and modeling. From the comparison of the four crystallized cytochromes P450, it is inferred that the B' helices and their immediate environments can adopt completely different conformations. Therefore, this zone may be highly influential on substrate speci-

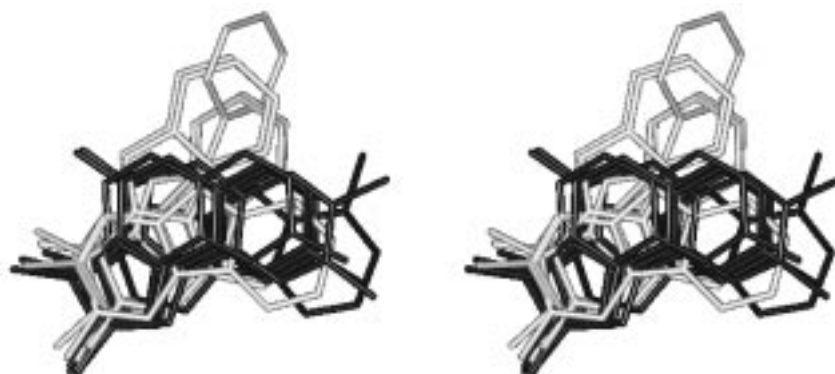


Figure 5. Stereoview of HCA superposition resulting from AUTODOCK 2.4 and AMBER refinement. The more active compounds (MeIQ, IQ, 4,8-DiMeIQx, 7,8-DiMeIQx, MeIQx, and Glu-P-1) are shown in black, and the less active compounds (Glu-P-2, Trp-P-1, Trp-P-2, PhIP, A $\alpha$ C, and MeA $\alpha$ C) are shown in gray.

Table 2. Main COMBINE PLS coefficients for CYP1A2h/HCA complexes

Region	Residue <sup>a</sup>	Location	Steric <sup>b</sup>	Electrostatic <sup>c</sup>	H-bond <sup>d</sup>
A	Thr115	Helix B'	0.762	1.023	–
	Asp119	Helix B'–C	–	0.305	–
	Thr124	Helix B'–C	–	0.429	–
	Thr223	Helix F	–	–0.377	–
B	Asp313	Helix I	–	0.313	–
	Gly316	Helix I	–	–	0.638
	Thr321	Helix I	–0.504	0.504	–
	Leu382	Near strand $\beta_{1-4}$	–0.631	–	–
	Pro383	Near strand $\beta_{1-4}$	–	–0.203	–
C	Tyr495	Near strand $\beta_{4-1}$	–	–0.372	–0.901
D	Arg456	Near heme	–	0.638	–

<sup>a</sup>Residue of human CYP1A2 sequence.

<sup>b</sup>Steric field PLS coefficients.

<sup>c</sup>Electrostatic field PLS coefficients.

<sup>d</sup>Hydrogen bond PLS coefficients.

ficity, but could also be a major source of modeling errors. In contrast, the B'–C turn shows a reasonable homology among cytochromes P450 and it is structurally similar in all crystallized structures.

To a large extent, the coefficients in this region arise from the different orientation of the third ring for the more active and the less active compounds. Different residues of this region make favorable contacts with the third ring of the less active compounds and therefore these interactions give rise to positive PLS coefficients in the model (Asp119, Thr124, and especially Thr115). Remarkably, the negative PLS coefficient for Thr223 corresponds to the water-bridged interactions observed in the complexes of the most active compounds, between polar groups present in the ligands and residue Thr223.

### Region B

This region is located in a zone that comprises helix I and a loop between helix K and  $\beta$ -strand 1–4. The results of site-directed mutagenesis experiments on residues Asp320 and Thr321 support the participation of these amino acids in catalysis, i.e. this region is important for the enzymatic reaction [41–44]. During the catalysis process this distal helix experiences a local distortion that leads to the loss of several helical hydrogen bonds. The missing hydrogen bond to the peptide carbonyl oxygen of Gly248 in P450cam (Ala317 in CYP1A2) is replaced by a hydrogen bond that is donated from the hydroxyl group of Thr252 (Thr321 in CYP1A2). Thr252 has been postulated to be the direct proton donor, part of a proton shut-



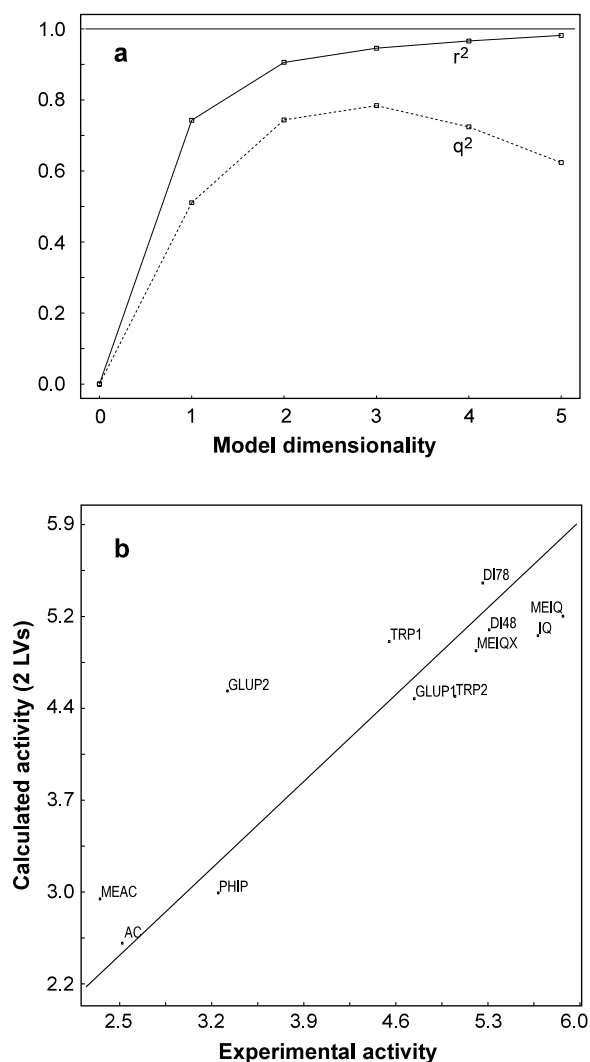


Figure 6. (a) Squared correlation coefficient ( $r^2$ ) and cross-validated squared correlation coefficient ( $q^2$ ) versus different dimensionalities of the COMBINE model. (b) Scatter plot showing the calculated versus experimental activity values.

tle or part of the pocket that helps to stabilize the oxy-complex [45].

Therefore it can be assumed that compounds that place bulky groups in this zone may facilitate the deformation of helix I, and as a consequence, activate the catalysis process. Coefficients given by the model to residue Thr321 and also to the neighboring Leu382 can be seen as an expression of this fact. The most active compounds tend to place their third ring in this region, i.e. all steric PLS coefficients in this zone have negative coefficients. In contrast, less active compounds place in this position a nitrogen able

to make electrostatic or hydrogen bond interactions with residues in helix I (like Asp320 or Thr321), resulting in a decrease of activity. Positive electrostatic field coefficients at Thr321 can be taken as a measure of repulsion with helix I. The less active compounds A $\alpha$ C and MeA $\alpha$ C tend to make hydrogen bonding interactions with Gly316, which the model reflects in the positive hydrogen bond PLS coefficients.

#### Region C

This region comprises a loop between two antiparallel  $\beta$ -strands. In this region, the electrostatic and hydrogen bonding interactions with Tyr495 display a significant negative PLS coefficient. This can be interpreted as a consequence of the polar and hydrogen bonding interactions observed between compounds with intermediate activity (Trp-P-1 and Trp-P-2) and this residue. For the more active compounds no interaction was found in this region.

#### Region D

This region comprises the cationic residue Arg456 that coordinates the propionate groups of the heme group. The importance of this conserved arginine has been reported recently [46]. It has been suggested that ligand binding induces a conformational change in Arg456 that appears to be important for the change in the oxidation state of the heme iron and, consequently, for ligand oxidation. The positive coefficients given to this residue can be seen as the consequence of the slightly different orientations adopted by the more active and the less active compounds near the heme group, exerting different electrostatic effects.

As a summary of the COMBINE results, we can assert that region A residues make better contacts (energetically more negative) with the less active compounds than with the more active compounds. Contacts with Thr223 are an exception, and the more active compounds are characterized by their ability to make electrostatic interactions with this residue. On the other hand, region B contains negative steric PLS coefficients because residues in this region interact better (more negative energy values) with the more active compounds. Region C contains Tyr495, a residue that modulates the activity of moderate active compounds like Trp-P-1 and Trp-P-2, by forming hydrogen bonds. Finally, region D modulates the activity through electrostatic interactions involving the charged residue Arg456.

Table 3. Influence of the SRD selection procedure on the PLS modeling results (GRID/GOLPE)

var. sel.	ar <sup>a</sup>	dimens <sup>b</sup>	r <sup>2c</sup>	q <sup>2d</sup>	SDEP <sup>e</sup>
Before SRD algorithm	1204	2	0.94	0.52	0.82
SRD + 2 FFD	533	2	0.96	0.79	0.54

<sup>a</sup> Number of variables used in the PLS model.

<sup>b</sup> Dimensionality of the model.

<sup>c</sup> Squared correlation coefficient.

<sup>d</sup> Cross-validated squared correlation coefficient.

<sup>e</sup> Standard deviation of error of predictions.

### GRID/GOLPE analysis

On the basis of the HCA alignment reported above a GRID/GOLPE analysis was carried out on the same set of compounds. Since both methods are quite different, concordant results can be seen as a further confirmation of the validity of the COMBINE analysis. However, the different positions of the third ring and the tilting of the molecular plane, which were shown to discriminate between the more active compounds and the less active ones (Figure 5), are expected to affect also the GRID/GOLPE results, as they affected the COMBINE analysis. Both methods should account for the same underlying reality in different ways.

As was the case in the COMBINE analysis, the optimal number of latent variables (LV) was chosen by monitoring the changes in the fitting ( $r^2$ ) and predictive ability ( $q^2$ ) indexes of the model upon addition of new LV (Figure 7a). The maximum quality was obtained for a model with two LV. Table 3 shows the characteristics of this model and stresses the advantage of using the Smart Region Definition (SRD) algorithm. In cross-validation, the predictive indices of the resulting GRID/GOLPE model are slightly better than those obtained in the COMBINE analysis ( $q^2 = 0.79$  vs.  $q^2 = 0.74$ ). The calculated activities given by the GRID/GOLPE model (Figure 7b) are generally quite similar to the experimental values, with the exception of Glu-P-2, as was the case in the COMBINE analysis. Figure 8 shows the four regions where the most important GRID/GOLPE PLS coefficients are located. Since they are located roughly in the same areas of the binding site described in the COMBINE analysis, these were also named regions A, B, C and D.

#### Region A

An important region of negative PLS coefficients is located opposite the heme group in the binding site. As seen in the COMBINE analysis, it can be interpreted

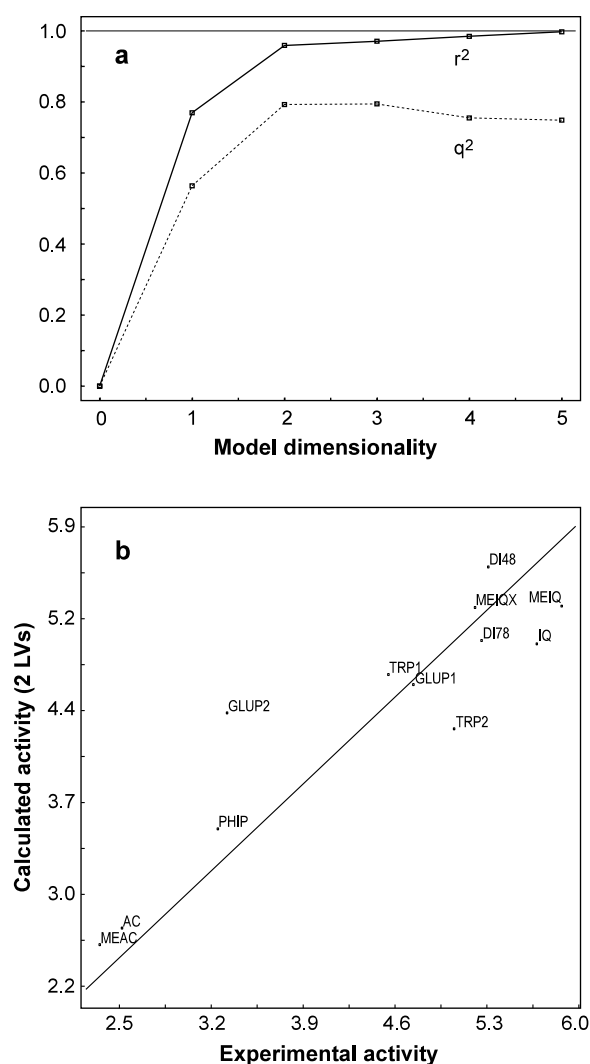


Figure 7. (a) Squared correlation coefficient ( $r^2$ ) and cross-validated squared correlation coefficient ( $q^2$ ) versus different dimensionalities of the GRID/GOLPE model. (b) Scatter plot showing the calculated versus experimental activity values.

as reflecting the possibility of the ligand to make water-bridged hydrogen bonds with Thr223. The PLS coefficients from both methods account for the high activities of MeIQ, IQ, 4,8-DiMeIQx, 7,8-DiMeIQx, MeIQx and Glu-P-1 because they are able to form the mentioned hydrogen bond, and therefore they give rise to negative interaction energies or field values. In contrast, less active compounds such as Trp-P-1, Trp-P-2, A $\alpha$ C, MeA $\alpha$ C and PhIP place bulky groups in this region, generating positive field values there. Glu-P-2 behaves rather like an outlier: this compound shares relative orientation, shape and electronic fea-

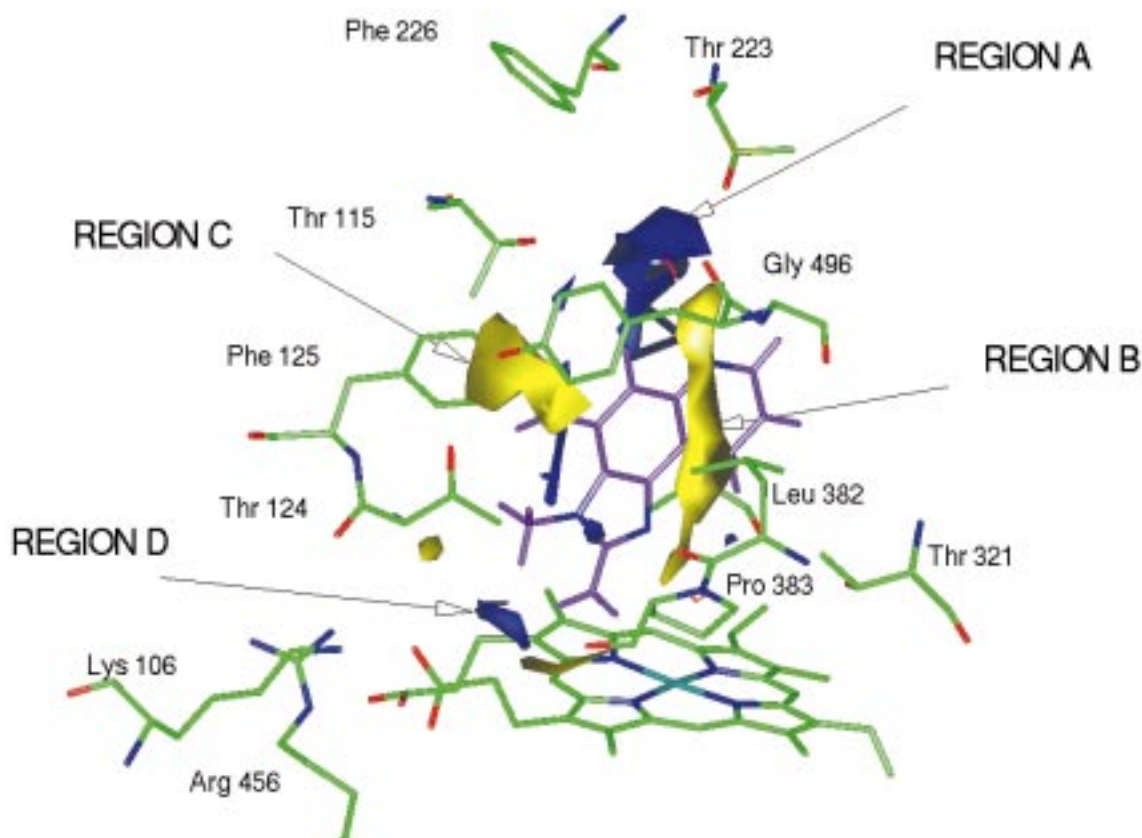


Figure 8. Contour map of the PLS coefficients for the GRID/GOLPE model. Positive PLS coefficients (contour level 0.005) are colored yellow, and negative PLS coefficients (contour level  $-0.005$ ) are colored blue.

tures with the active compounds, with the exception of a methyl group, but it is much less active. Since this is a peculiarity of this compound, its behavior is not correctly explained by either the COMBINE model or the GRID/GOLPE model.

#### Region B

This second important region is located adjacent to region A and near the catalytic center. In this case, the positive coefficients indicate that the more active compounds generate positive field values in this region while the less active compounds tend to generate negative field values. It seems that energetically favorable interactions in this zone have opposite effects on the activity.

The PLS coefficients of this region affect the rank of activity of the compounds. The most active compounds MeIQ and IQ do not generate any negative field in this region. In contrast, MeIQx, 4,8-DiMeIQx and 7,8-DiMeIQx share an heterocyclic nitrogen that generates negative fields in this region. Trp-P-1 and

Trp-P-2 place a methyl group in this region which generates the relevant field in order to increase the activities. The less active compounds, PhIP, A $\alpha$ C and MeA $\alpha$ C, show negative fields in this region due to a heterocyclic nitrogen (PhIP) or a polar hydrogen (A $\alpha$ C and MeA $\alpha$ C).

#### Region C

This is another region showing high positive PLS coefficients within the positions where the most active compounds (MeIQ or 4,8-DiMeIQx) place their methyl groups. Therefore, the presence of these bulky groups in this position produces in the model an increase of the activity. Besides, there is a small region of negative PLS coefficients located near Tyr495 and Thr115, mainly devoted to explaining the medium activity of Trp-P-1 and Trp-P-2. These HCAs are not able to form hydrogen bonds with the above mentioned water molecules but, due to hydrogen bonds being formed with Tyr495, they are probably better

orientated towards the heme group than other less active compounds.

#### *Region D*

Finally, there is a small zone with negative coefficients near the exocyclic amino group that account for the small differences in the location of the amino group and its orientation towards the heme group.

#### *Comparison of COMBINE and GRID/GOLPE models*

As was to be expected, there is a clear coincidence of the results from the GRID/GOLPE method and the COMBINE analysis. Figure 8 shows the residues closest to the regions with the largest GRID/GOLPE PLS coefficients, which correspond with the most important ligand-residue interactions disclosed in the COMBINE model. Also, the predicted vs. experimental activity plots exhibit the same pattern in both the GRID/GOLPE and the COMBINE analyses (Figures 6b and 7b).

Since the models express to a large extent the different orientations of the ligands inside the binding site, the GRID/GOLPE model uses mainly the positive part of the molecular interaction field. Most of the coefficients represent the effect of the presence or absence of a certain part of the ligand more than the interaction of this part with the receptor. This can be observed for regions B and C. The sign of the steric PLS coefficients for region A (COMBINE: positive and GRID/GOLPE: negative) reflects quite well the presence of hydrogen bonds for the more active compounds and also the trend of the less active compounds to place bulky groups in this region. In region B, the sign reversal of the steric PLS coefficients in COMBINE (negative) and GRID/GOLPE (positive) is brought about by the fact that the more active compounds place hydrophobic or bulky groups in the vicinity of the catalytic center.

With respect to region C, both methods identify the surroundings of Tyr495 as the region responsible for the increase in activity of MeIQ, 4,8-DiMeIQx, Trp-P-1 and Trp-P-2, although the interpretation made by both methods is different. The GRID/GOLPE method simply displays positive PLS coefficients in this region whereas COMBINE highlights the importance of the hydrogen bond between Tyr495 and both Trp-P-1 and Trp-P-2.

Finally, GRID/GOLPE identifies a common region D near the heme group. Its positive PLS coefficients can be seen as a consequence of the slightly different

locations of the amino group generating more negative fields for the more active compounds. COMBINE goes one step further and identifies Arg456 as the residue responsible for modulating the activity of the HCA.

## Conclusions

When using an appropriate alignment of the compounds considered, both the COMBINE and GRID/GOLPE methods offer concordant results that are useful to quantitatively rationalize the differences in activity within the series. Furthermore, the combined use of both approaches acts as a mutual validation procedure and allows a more reliable and detailed interpretation of the results.

The quality of the particular models reported here is obviously limited by the quality of the CYP2H model and the suitability of the activity data. They should be recomputed and reanalyzed when an experimental structure of human CYP1A2 or more relevant experimental activities become available. Likewise, the role of the water molecules in the active site will be of great interest for further studies. In any case, we think that the models reported in this work can be very useful for the study of regio- and stereospecificity of cytochromes P450 and for the interpretation of the increasing number of site-directed mutagenesis data.

## References

1. Black, S.D. and Coon, M.J., *Adv. Enzymol. Relat. Areas Mol. Biol.*, 60 (1987) 35.
2. Peterson, J.A. and Graham, S.E., *Structure*, 6 (1998) 1079.
3. Sugimura T., *Mutat. Res.*, 376 (1997) 211.
4. Segura, J., Roberts, D.J. and Tarrús, E.J., *J. Pharm. Pharmacol.*, 41 (1989) 129.
5. Fuhr, U., Strobl, G., Manaut, F., Anders, E., Sörgel, F., López de Brinñas, E., Chu, D.T.W., Pernet, A.G., Mahr, G., Sanz, F. and Staib, H., *Mol. Pharmacol.*, 43 (1993) 191.
6. Shimada, T., Iwasaki, M., Martín, M.V. and Guengerich, F.P., *Cancer Res.*, 49 (1989) 3218.
7. Wakabayashi, K., Nagao, M., Esumi, H. and Sugimura, T., *Cancer Res. (suppl.)*, 52 (1992) 2092.
8. Chang, Y.T., Stiffelman, O.B., Vakser, I.A., Loew, G.H., Bridges, A. and Waskell, L., *Protein Eng.*, 10 (1997) 119.
9. Szklarz, G.D. and Halpert, J.R., *Life Sci.*, 61 (1997) 2507.
10. Lozano, J.J., Lopez-de-Brinas, E., Centeno, N.B., Guido, R. and Sanz, F., *J. Comput.-Aided Mol. Design*, 11 (1997) 395.
11. Poulos, T.L., Finzel, B.C. and Howard, A.J., *Biochemistry*, 25 (1986) 5314.
12. Ravichandran, K.G., Boddupalli, S.S., Hasemann, C.A., Peterson, J.A. and Deisenhofer, J., *Science*, 261 (1993) 731.

13. Hasemann, C.A., Ravichandran, K.G., Peterson, J.A. and Deisenhofer, J., *J. Mol. Biol.*, 236 (1994) 1169.
14. Cupp-Vickery, J.R. and Poulos, T.L., *Nat. Struct. Biol.*, 2 (1995) 144.
15. Uno, T., Michell, E., Aida, K., Lambert, M.H., Darden, T.A., Pedersen, L.G. and Negishi, M., *Biochemistry*, 36 (1997) 3193.
16. Helms, V. and Wade, R.C., *J. Am. Chem. Soc.*, 120 (1998) 2710.
17. Paulsen, M.D. and Ornstein, R.L., *Protein Eng.*, 9 (1996) 567.
18. Kollman, P., *Chem. Rev.*, 93 (1993) 2395.
19. Ortiz, A.R., Pisabarro, M.T., Gago, F. and Wade, R.C., *J. Med. Chem.*, 36 (1995) 2681.
20. Ortiz, A.R., Pastor, M., Palomer, A., Cruciani, G., Gago, F. and Wade, R.C., *J. Med. Chem.*, 40 (1997) 1136.
21. Perez, C., Pastor, M., Ortiz, A.R. and Gago, F., *J. Med. Chem.*, 41 (1998) 836.
22. Wade, R.C., Ortiz, A.R. and Gago, F., In Kubinyi, H., Folkers, G. and Martin, Y.C. (Eds.), *3D QSAR in Drug Design*, Vol. 2, Kluwer/Escom, Dordrecht, 1998, p. 19.
23. Goodford, P., *J. Chemometr.*, 10 (1996) 107.
24. Cramer, R.D., Patterson, D.E. and Bunce, J.D., *J. Am. Chem. Soc.*, 110 (1988) 5959.
25. GOLPE 4.0, Multivariate Infometric Analysis, Perugia, Italy, 1998.
26. Cruciani, G. and Watson, K.A., *J. Med. Chem.*, 37 (1994) 2589.
27. Sugimura, T., *Mutat. Res.*, 150 (1985) 33.
28. Eisenbrand, G. and Tang, W., *Toxicology*, 84 (1993) 1.
29. GRID v. 15, Molecular Discovery Ltd., Oxford, 1997.
30. Wild, D., Feser, W., Michel, S., Lord, H.L. and Josephy, P.D., *Carcinogenesis*, 16 (1995) 643.
31. Shimada, T. and Guengerich, F.P., *Cancer Res.*, 51 (1991) 5284.
32. Morris, G.M., Goodsell, D.S., Huey, R. and Olson, A.J., *J. Comput.-Aided Mol. Design*, 10 (1996) 293.
33. Raag, R., Li, H., Jones, B.C. and Poulos, T.L., *Biochemistry*, 32 (1993) 4571.
34. AMBER (UCSF): Assisted Model Building with Energy Refinement, version 4.1, 1995. Department of Pharmaceutical Chemistry, University of California, San Francisco, CA.
35. Weiner, S.J., Kollman, P.A., Nguyen, D.T. and Case, D.A., *J. Comput. Chem.*, 7 (1986) 230.
36. Poulos, T.L. and Howard, A.J., *Biochemistry*, 26 (1987) 8165.
37. Gaussian 94, Revision E.1, Frisch, M.J., Trucks, G.W., Schlegel, H.B., Gill, P.M.W., Johnson, B.G., Robb, M.A., Cheeseman, J.R., Keith, T., Petersson, G.A., Montgomery, J.A., Raghavachari, K., Al-Laham, M.A., Zakrzewski, V.G., Ortiz, J.V., Foresman, J.B., Cioslowski, J., Stefanov, B.B., Nanayakkara, A., Challacombe, M., Peng, C.Y., Ayala, P.Y., Chen, W., Wong, M.W., Andres, J.L., Replogle, E.S., Gomperts, R., Martin, R.L., Fox, D.J., Binkley, J.S., Defrees, D.J., Baker, J., Stewart, J.P., Head-Gordon, M., Gonzalez, C. and Pople, J.A., Gaussian, Inc., Pittsburgh, PA, 1995.
38. Bayly, C.I., Cieplak, P., Cornell, W.D. and Kollman, P.A., *J. Phys. Chem.*, 97 (1993) 10269.
39. Pastor, M., Cruciani, G. and Clementi, S., *J. Med. Chem.*, 40 (1997) 1455.
40. Gotoh, O., *J. Biol. Chem.*, 267 (1993) 83.
41. Furuya, H., Shimizu, T., Hirano, K., Hatano, M. and Fujii-Kuriyama, Y., *Biochemistry*, 28 (1989) 6848.
42. Shimizu, T., Sadeque, A.J.M., Sadeque, G.N., Hatano, M. and Fujii-Kuriyama, Y., *Biochemistry*, 30 (1991) 1490.
43. Tuck, F.S., Hiroya, K., Shimizu, T., Hatano, M. and Ortiz de Montellano, P.R., *Biochemistry*, 32 (1993) 2548.
44. Krainev, A.G., Shimizu, T., Ishigooka, M., Hiroya, K. and Hatano, M., *Biochemistry*, 32 (1993) 1951.
45. Poulos, T.L., *Curr. Opin. Struct. Biol.*, 5 (1995) 767.
46. Oprea, T.I., Hummer, G. and Garcia, A.E., *Proc. Natl. Acad. Sci. USA*, 94 (1997) 2133.