New Combined Model for the Prediction of Regioselectivity in Cytochrome P450/3A4 Mediated Metabolism

Won Seok Oh,[†] Doo Nam Kim,[‡] Jihoon Jung,[†] Kwang-Hwi Cho,[§] and Kyoung Tai No*[,†,‡]

Department of Biotechnology, Yonsei University, Seoul 120-749, Korea, Bioinformatics and Molecular Design Research Center, Seoul 120-749, Korea, and Department of Bioinformatics and CAMD Research Center, Soongsil University, Seoul 156-743, Korea

Received September 28, 2007

Cytochrome P450 3A4 metabolizes nearly 50% of the drugs currently in clinical use with a broad range of substrate specificity. Early prediction of metabolites of xenobiotic compounds is crucial for cost efficient drug discovery and development. We developed a new combined model, MLite, for the prediction of regioselectivity in the cytochrome P450 3A4 mediated metabolism. In the model, the ensemble catalyticphore-based docking method was implemented for the accessibility prediction, and the activation energy estimation method of Korzekwa et al. was used for the reactivity prediction. Four major metabolic reactions, aliphatic hydroxylation, N-dealkylation, O-dealkylation, and aromatic hydroxylation reaction, were included and the reaction data, metabolite information, were collected for 72 well-known substrates. The 47 drug molecules were used as the training set, and the 25 well-known substrates were used as the test set for the ensemble catalyticphore-based docking method. MLite predicted correctly about 76% of the first two sites in the ranking list of the test set. This predictability is comparable with that of another combined model, MetaSite, and the recently published QSAR model proposed by Sheridan et al. MLite also offers information about binding configurations of the substrate−enzyme complex. This may be useful in drug modification by the structure-based drug design.

INTRODUCTION

Members of the cytochrome P450 (CYP) superfamily are monooxygenases which metabolize a wide range of xenobiotic compounds. CYP3A4 is expressed in the human liver and the small intestine and metabolizes nearly 50% of the drugs currently in clinical use with a broad range of substrate specificity.[1] The aim of xenobiotic metabolism is detoxification and efficient excretion of potentially harmful compounds. In some cases, CYPs transform nontoxic compounds to toxic compounds leading to toxic drug side effects.[2] Early consideration of the metabolic properties of compounds of interest is one of the major requirements for efficient drug discovery.[3]

There are many parameters in drug metabolism such as regioselectivity, involved isoenzyme, reaction rate (K_m), and inhibition affinity (K_i) of metabolic substrate or inhibitor. Regioselectivity is the preference of one direction of metabolic reaction over the other possible directions. Selection of a metabolic reaction site determines the metabolite formed. The knowledge of metabolites formed from a drug candidate gives a direction for further modification of the compound. By the modification of labile metabolic sites, one can reduce the rate of metabolism or can avoid the formation of potentially toxic metabolites.

For the prediction of CYP regioselectivity, three classes of in silico methods have been proposed: rule-based methods,[4,5] quantitative structure activity relationship (QSAR) methods,[6,7] and mechanism-based methods.[8−16] Rule-based methods employ the patterns of metabolic reactions that are generated from a metabolism database. Although rule-based methods have been widely used by medicinal chemists, these methods tend to propose a relatively large number of possible metabolites. The suggestion of a large number of metabolic reactions reduces the usefulness on the point of view of prediction by the model. In QSAR methods, models were constructed using retrieved common physicochemical and/or pharmacological properties of given CYP substrates. Since CYP450s metabolize structurally diverse compounds, there is a limitation in the predictions of regioselectivity using QSAR methods. In spite of the limitations of the QSAR methods, Sheridan et al.[7] recently proposed a new QSAR model for the prediction of regioselectivity in CYP 3A4, 2D6, and 2C9. These models showed good predictability with three substructural descriptors and two physical property descriptors. Mechanism-based methods attempt to mimic the process of a metabolic reaction by the model. Mechanism-based methods include pharmacophore-based methods,[8,9] reactivity-based ab initio calculations on substrates,[10−12] and combined methods of accessibility and reactivity.[13−16] Although these methods have the limitation of long computational times, they give the physical meaning in the prediction with good predictability. Among the combined methods, MetaSite[16] is the widely distributed package for the prediction of regioselectivity.

The process of the metabolic reaction of xenobiotic consists of a series of processes, including substrate binding to the enzyme, catalytic reaction of a substrate by the

* Corresponding author phone: 82-2-393-9551; fax: 82-2-393-9554; e-mail: ktno@yonsei.ac.kr.
† Yonsei University.
‡ Bioinformatics and Molecular Design Research Center.
§ Soongsil University.

10.1021/ci7003576 CCC: $40.75
Published on Web 02/22/2008

enzyme, and release of a metabolite from the enzyme. At first, the substrate must bind in close proximity between the metabolic reaction atom within the substrate and the catalytic site of the CYP enzyme (i.e., heme oxygen). Force field based docking techniques[17,18] and molecular dynamics (MD) simulations[19] can mimic this complex formation process and the dynamic motion of the substrate−enzyme complex. Although these techniques give accurate descriptions, they are quite time-consuming. Instead of docking techniques, Cruciani et al.[15,16] compared the protein correlogram reporting the interaction energies and the distance from the heme-iron moiety of protein with a set of substrate correlograms of the possible atoms within a substrate. This technique is a very fast tool to describe the binding affinity of substrates to the enzyme.

Second, the substrate binds to the enzyme followed by the catalytic reaction of phase I metabolism. The rate determining step of phase I metabolism is the formation of a radical species. Since the catalytic reaction is an irreversible process, the fundamental form of the reaction rate, or probability, is an Arrhenius type equation. Reaction probability is determined by the activation energy of the reaction. Korzekwa et al.[10,11] developed an empirical activation energy (E_a) estimation model using the AM1 reaction enthalpy (ΔH_r) and ionization potential (IP) as descriptors to predict the preferred metabolic sites of a molecule. Singh et al.[12] used reaction enthalpy instead of activation energy for the prediction of the reactivity of each hydrogen atom within a substrate. Calculation of reaction energy is simpler and more straightforward than calculation of activation energy. Although the model of regioselectivity prediction with the estimated activation energy works well for molecules defined by the training set, it is necessary to introduce information about the binding site for more accurate predictions in wider applications.

The final process of the metabolic reaction is the release of a metabolite formed by the metabolic reaction from the enzyme. The metabolic reaction is the polarization of a substrate for an efficient elimination of xenobiotics. The metabolite is similar with the substrate in shape but more polar than the substrate. Since more polar compounds are more soluble in water, it is regarded that the metabolite will be easily released from the enzyme. Metabolites having high binding affinity with the enzyme act as inhibitors. The prediction of binding affinity of the metabolite is another field of research.

The merit of mechanism-based methods is that they give information about the process of real metabolic reaction and direction of modification. The modification of metabolically labile xenobiotic compounds can be done by preventing a compound from binding on the enzyme or reducing the reactivity of the metabolic labile site. Analysis of the binding mode or the metabolic reaction site is important in drug modification. A new combined model that is fast enough for the application in HTS and also gives a binding mode is required. In this paper, we proposed a new combined model, called MLite (metabolite prediction). In the model, the ensemble catalyticphore-based docking method and the activation energy estimation method of Korzekwa et al. were implemented for the prediction of accessibility and reactivity, respectively. By introducing both accessibility and reactivity in the model, we were able to predict the regioselectivity of the CYP3A4 substrate with high predictability.

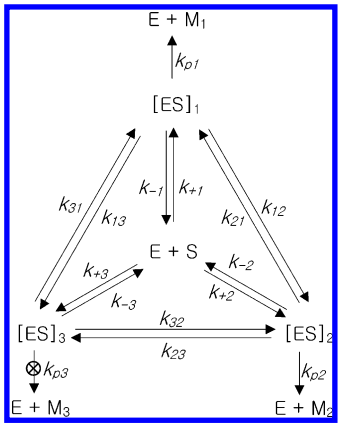


Figure 1. Kinetic model of a metabolic reaction. E, S, [ES], M, and *k* are enzyme, substrate, enzyme−substrate complex, metabolite, and rate constant, respectively.

METHOD

Proposed Model. One or more metabolites are reported for a substrate experimentally. Each metabolite has a different reaction path, such as a different binding mode and a different catalytic reaction. If we assume that the rate determining step of a metabolic reaction is the radical formation, then the metabolite formation rate is proportional to the substrate−enzyme complex concentration multiplied by the radical formation rate constant (Figure 1). Since there are not enough experimental results about interconvertibility of binding modes, we disregarded it in this study (k_{12}, k_{21}, k_{23}, k_{32}, k_{31}, $k_{13} \approx 0$). The homotropic and heterotropic cooperativity also observed for many CYP3A4 substrates and many kinetic models about them are reported.[20] CYP3A4 has multiple substrate binding sites. A binding of the substrate and the effecter molecule will change the binding affinity and/or the reaction rate of the substrate on the other binding sites. Since consideration of cooperativity for diverse substrates is hard, we also disregarded it in this study. After the substrate binds to the enzyme, a hydrogen abstraction reaction may occur. If the substrate−enzyme complex formation reaction rate (k_{+1}/k_{-1}) is slower than the hydrogen abstraction rate (k_{p1}), $k_{+1}/k_{-1} \ll k_{p1}$, then the population of the substrate−enzyme complex will determine the overall rate of the metabolic reaction. If the substrate−enzyme complex formation reaction rate (k_{+1}/k_{-1}) is faster than the radical formation reaction rate (k_{p1}), $k_{+1}/k_{-1} \gg k_{p1}$, then the radical formation reaction rate will be rate-limiting. If the radical formation reaction rate is close to zero, $k_{p3} \approx 0$, then the substrate will act as an inhibitor.

In this paper, we assumed that the reaction rate of substrate−enzyme complex formation is faster than that of radical formation if a substrate−enzyme complex is possible. In this case, the concentration of the substrate−enzyme complex is not important. The radical formation reaction rates determine the order of proposed possible reaction sites in the complex. The ensemble catalyticphore-based docking method was used to find possible binding modes of substrate−enzyme complexes, not binding probability. The metabolic reaction has broad substrate specificity and multiple substrate binding sites. It cannot be expressed by a

single catalyticphore or a binding mode. We used the multiple catalyticphore to find the binding mode of substrates to CYP3A4. In this paper, "ensemble" means that multiple catalyticphores were used. The quantum mechanical (QM) calculations on the rate-determining step at each possible reaction site determined the order of possible reaction sites.

Data Preparation. Metabolic reaction data were obtained from Rendic's review paper[21] and references therein. Four major metabolic reactions mediated by CYP3A4, aliphatic hydroxylation, N-dealkylation, O-dealkylation, and aromatic hydroxylation, were included in data set. Some minor metabolic reaction types such as epoxidation, reduction, and N-hydroxylation were not considered. The well-known CYP3A4 substrates for which special comments are annotated in the literature such as major reaction, major metabolite, major enzyme, marker, and high activity were selected for database construction. Among 72 substrates the one with drug activity belongs to the training set, and the rest of them were classified in the test set for the model of ensemble catalyticphore-based docking. In the total of 72 collected substrates, only one metabolite was reported for 45 substrates, whereas two or more metabolites were reported for 27 substrates in the CYP3A4 mediated metabolism. The first set was used as the training set and the second set as the test set for the model of ensemble catalyticphore-based docking.

Ensemble Catalyticphore. A number of human CYP3A4 structures have been solved, the ligand free-structure[22,23] and the complex structure with small[22] and large ligands.[24] William et al.[22] and Yano et al.[23] independently reported the ligand-free CYP3A4 structures. These two ligand-free structures are very similar. The main difference between the two ligand-free structures is the orientation of the Arg212 residue in the active site. The side chain of Arg212 is positioned inside the active site in the Yano et al. structure and outside in the William et al. structure. This may represent the flexibility of this residue. The role of the Arg212 residue is not clearly investigated. However, there was a report that Arg212 participates in the hydrogen bond with the substrate in the flexible docking study of diltiazem.[25] The binding of the small ligand does not show a conformational change of protein. The ligand-free structure and the complex structure with small ligands of William et al. have similar conformations. However, the complex structure in which the large ligand occurs increases in the active site volume. This represents the flexibility of the enzyme in recognizing substrates.[24]

In order to include participation of the Arg212 residue, a CYP3A4 crystal structure of Yano et al. (PDB 1TQN)[23] with 2.05 Å resolution was used for defining the catalyticphore in this study. The flexibility of the enzyme in recognizing substrates is treated with a small radius of exclusion volume and a rough tolerance of the catalyticphore. The ferric oxygen missing in the crystal structure was manually added on the distal position of the heme within 1.7 Å. There were reports that water molecules participate in the interaction between substrate and CYP(s).[25] Water molecules in the catalytic site stabilize the substrate−enzyme complex. Several mediating water molecules participate in hydrogen bonds between the substrate and the enzyme. However, the role and the position of water molecules vary as to the kind of substrate. The flexible position of water molecules is required to represent all possibilities. The use of fixed positions for the water molecules obtained from a crystal structure is not practical to treat a large number of compounds.[26] In this study, we did not consider crystallographic water molecules. All structural waters, including buried and surface waters, were removed.

All calculations of catalyticphore-based docking were done with the Structure Based Focusing (SBF) module of Cerius[2] version 4.10.[27] Two things must be prepared to perform catalyticphore-based docking: the query that is a hypothetical representation of the catalyticphore of the enzyme and the substrate DB consisting of multiple conformations. The SBF module generates the catalyticphore from the structure of the enzyme. The position of the catalytic site was determined based on the CYP3A4 crystal structure and was confirmed by visual inspection. The features of a catalyticphore were described by hydrogen bond interactions, lipophilic interactions, and exclusion volumes. It was assumed that hydrogen bond and lipophilic interaction are the major interactions in the enzyme−substrate complex. These interaction features were generated with the default option of SBF. Hydrogen bond interaction features that are directed outward from the catalyticphore and weak interaction features were deleted manually. The interaction features were clustered as 6 hydrogen-bond donors, 6 hydrogen-bond acceptors, and 6 lipophilic interaction centers. The query was generated by taking three features from the 18 clustered features of the catalyticphore, ${}_{18}C_3$ queries. The exclusion volume represents the shape of the catalytic site in the enzyme. This was generated using heavy atoms exposed to the catalytic site with a volume radius of 0.8 Å and were added to the query. Since we did not treat electron-transfer reactions in this study, all reaction sites were carbon atoms with hydrogen. To represent the carbon atoms of the reaction site, we added a carbon feature that is 3.94 Å above the ferrioxygen atom to the query. The SBF module treats ligand flexibility by generating multiple conformations. The substrate DB consisting of multiple conformations of substrates was constructed with the option that the maximum number of conformers per molecule is specified as 999.[28] The average number of conformer for all substrates is 98. The maximum number of conformers is 931 in sildenafil, and the minimum is 2 in diazepam, midazolam, and aflatoxin B1.

Since CYP3A4 metabolizes compounds with diverse physicochemical properties, i.e., size and hydrophobicity, and has multiple substrate binding sites,[29] it is necessary to introduce multiple queries to represent these characteristics of the CYP3A4 catalyticphore. Substrates with common interaction features may have a similar binding mode and can be described by the same query. We selected minimal representative queries based on the training set of substrates and confirmed by the test set of substrates. The query selection was performed by two steps. To save computational time, catalyticphore-based docking that gives a list of retrieved substrates was performed on the training set to filter the query. The top 20 queries were selected based on the number of retrieved substrates of each query. These 20 queries were reapplied on the training set by catalyticphore-based docking that gives all possible binding modes of substrates to CYP3A4. We defined that a hydrogen atom is exposed to heme when it is within 3.5 Å of the ferric oxygen atom and that the binding mode is correct when the exposed

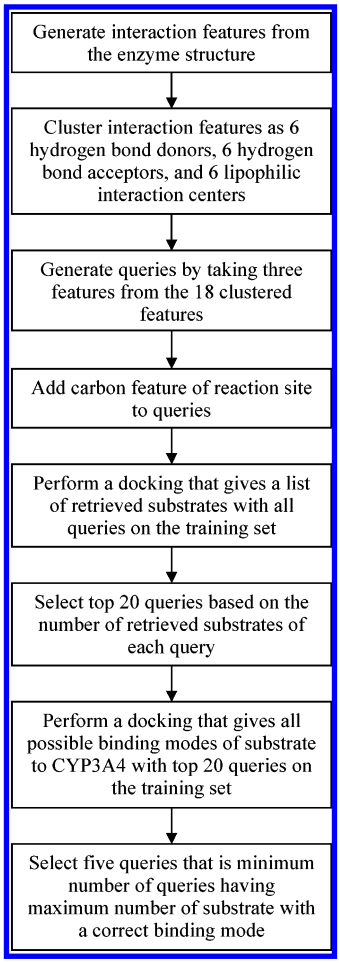


Figure 2. Flowchart for the selection of multiple queries in catalyticphore.

Table 1. Elementary Features of Queries Used in Catalyticphore-Based Docking

query	features[a]
query 694	ACC 2, LIP 4, LIP 5
query 726	ACC 3, LIP 2, LIP 6
query 728	ACC 3, LIP 3, LIP 5
query 749	ACC 4, LIP 1, LIP 5
query 806	LIP 1, LIP 5, LIP 6

[a] ACC is a hydrogen bond acceptor feature, and LIP is a lipophilic interaction feature.

site is consistent with an experimentally known metabolic reaction site. Finally, we selected 5 queries that are the minimum number of queries having the maximum number of substrates with a correct binding mode (Figure 2). We used these five queries to calculate accessibility of the substrate while bound to the enzyme. Final queries consisted of three hydrogen bond acceptors and six lipophilic interaction features (Table 1). Four queries of them consisted of one hydrogen bond acceptor and two lipophilic interaction features. The other query consisted of three lipophilic interaction features. Hydrogen bond donor features were not selected in the final queries. All hydrogen bond acceptor features are counterparts of Arg212, and lipophilic features interact with the hydrophobic cluster of CYP3A4 (Figure 3).

Activation Energy Estimation. There are two kinds of models that represent the reactivity of labile sites within a substrate: proposed by Korzekwa et al.[10,11] and Singh et al.[12]

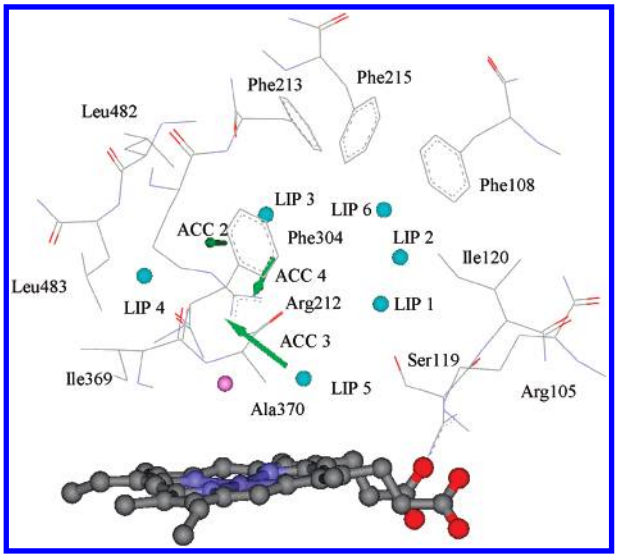


Figure 3. Elementary features in CYP3A4 catalyticphore. Shown are three hydrogen bond acceptors (arrow), six hydrophobic features (green sphere), a carbon feature (pink sphere), interacting residues (line), and a heme (ball-and-stick). All hydrogen bond acceptor features interact with Arg212. Lipophilic interaction features 2, 3, and 6 interact with the hydrophobic cluster of CYP3A4.

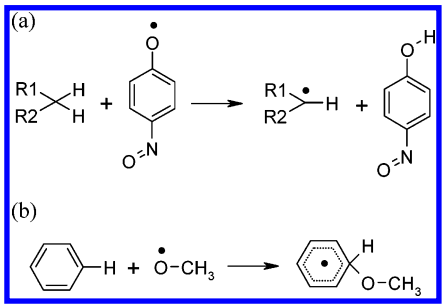


Figure 4. Representation of metabolic reaction models: (a) hydrogen abstraction reaction with p-nitrosophenoxy radical and (b) aromatic hydroxylation reaction with methoxy radical.

In the model of Korzekwa et al.,[10,11] the p-nitrosophenoxy radical (PNR) and the methoxy radical were used as surrogates of CYP for aliphatic hydroxylation and aromatic hydroxylation, respectively (Figure 4). The enthalpy of reaction of aliphatic hydroxylation, $\Delta H_{rxn(Habs)}$, and aromatic hydroxylation, $\Delta H_{rxn(Arom)}$, were obtained using AM1 molecular orbital (MO) calculations. Empirical equations for the activation energy estimation of the aliphatic hydroxylation, $\Delta H_{act(Habs)}$, and aromatic hydroxylation, $\Delta H_{act(Arom)}$, are given by

$$\Delta H_{act(Habs)} = 2.60 + 0.22\Delta H_{rxn(Habs)} + 2.38\mathrm{IP} \quad (1)$$

$$\Delta H_{act(Arom)} = 21.91 + 0.61\Delta H_{rxn(Arom)} \quad (2)$$

IP is the ionization potential of the product radical. By combining experimental data and estimated activation energies for aliphatic and aromatic hydroxylation reactions, the energy difference in activation energies for aromatic and aliphatic hydroxylation reactions ($\Delta\Delta G$) is given by

$$\Delta\Delta G = 1.22\Delta H_{act(Habs)} - 1.1\Delta H_{act(Arom)} - 17.5 \quad (3)$$

Singh et al.[12] treated the reactivity of a hydrogen atom within a substrate as the reaction energy of hydrogen abstraction. The reaction energy of hydrogen abstraction on

Table 2. Prediction Ratio for CYP3A4 Substrates Found Using the Top Two Sites for the Reactivity Models

	model of Korzekwa et al.			model of Singh et al.		
reaction types	N^a	Nc[b]	ratio (%)[c]	N^a	Nc[b]	ratio (%)[c]
aliphatic hydroxylation	16	5	31.3	20	13	65.0
N-dealkylation	37	28	75.7	38	21	55.3
O-dealkylation	11	3	27.3	10	3	30.0
aromatic hydroxylation	8	5	62.5	4	1	25.0
total	72	41	56.9	72	38	52.8

[a] N = number of substrates. [b] Nc = number of correctly predicted substrates. [c] Percentage of correct predictions.

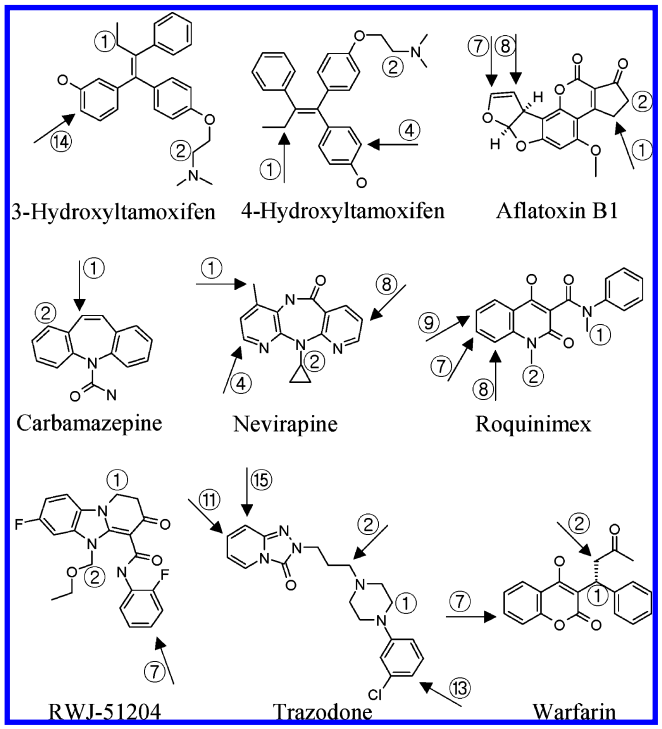


Figure 5. Prediction results of the model of Singh et al. on CYP3A4 substrates for which aromatic hydroxylation sites are known. Experimentally known major sites of metabolism are indicated with arrows. Orders of reactivity predicted with the model of Singh et al. are indicated with circled numbers.

an isolated substrate is given by

$$\Delta H_{\mathrm{rxn}} = \Delta H_2 - \Delta H_1 \qquad (4)$$

The hydrogen abstraction reaction energy (ΔH_{rxn}) is the difference between the heat of formation of the native substrate (ΔH_1) and that of its radical (ΔH_2).

We used the previously published two models for the prediction of reactivity of labile sites within a substrate. The semiempirical calculation for each hydrogen was performed using the AM1 method with Gaussian03.[30] All open shell calculations were treated with a UHF Hamiltonian. Geometry optimizations were performed with default options in Gaussian03. In hydrogen abstraction reactions, hydrogen atoms attached to the same carbon atom have the same radical species in a metabolic reaction. In aromatic hydroxylation reaction, the tetrahedral intermediates of a labile site differ in stereochemistry by the approaching direction of the activated oxygen radical. For a systematic calculation, we used the lowest energy intermediate in the prediction.

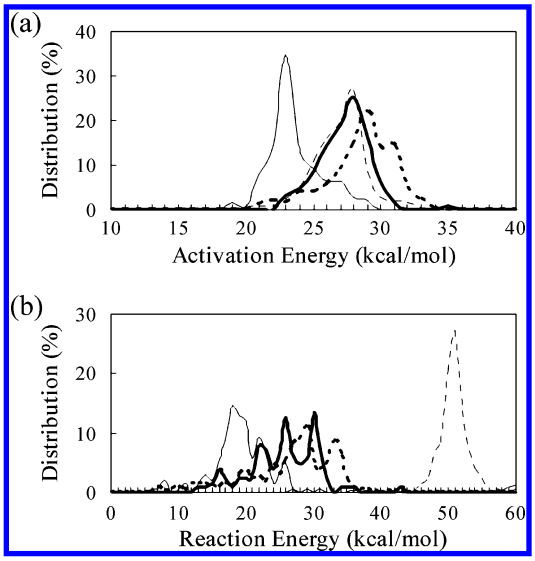
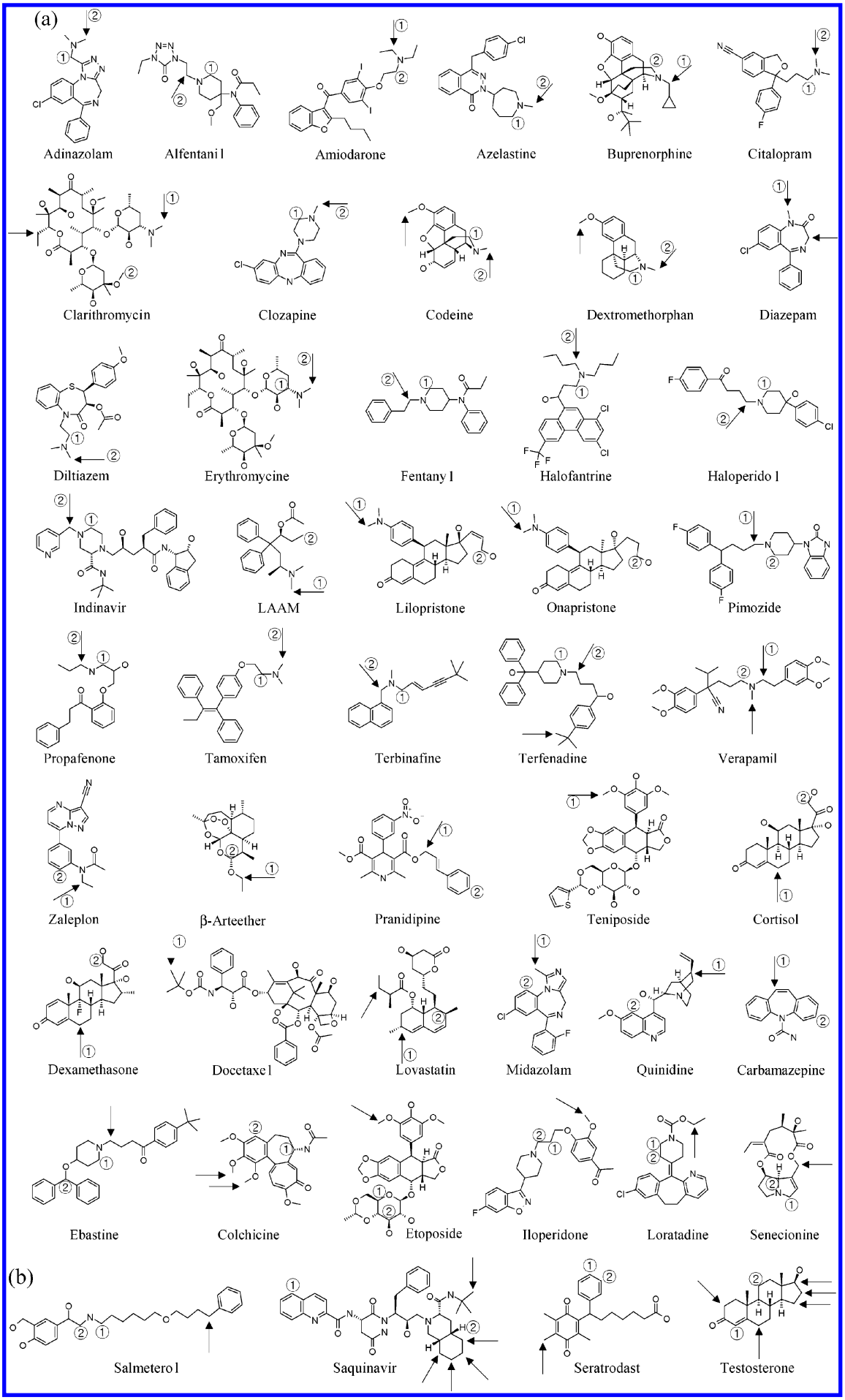


Figure 6. Energy distributions of (a) the model of Korzekwa et al. and (b) the model of Singh et al. for 72 CYP3A4 substrates. The number of data points for aromatic hydroxylation reactions (dashed line), aliphatic hydroxylation reactions (bold dashed line), N-dealkylation reactions (line), and O-dealkylation reactions (bold line) are 337, 365, 140, and 126, respectively.

Data Analysis. Metabolic labile sites within a substrate were predicted by two steps. Exposed hydrogen atoms were selected by the ensemble catalyticphore-based docking model and were ranked by the estimated activation energy of Korzekwa's model. Since the calculation method in the model contains inaccuracy, if the experimental major metabolite is coincident with one of the top two ranked metabolites, it was assumed that the regioselectivity prediction is correct. In the prediction of the substrates with multiple sites, when at least one metabolic reaction site was predicted by one of the first two ranked positions, we treated this substrate as correctly predicted. Validation of the prediction of the metabolic reaction site was performed by counting the number of substrates that is correctly predicted. The outcome of the catalyticphore-based docking is the possible binding configuration of the substrate bound to the enzyme. For the analysis of binding configuration, the PMF score[31] of the binding complex was calculated with the Ligand Fit module of Cerius[2] version 4.10.[27]

RESULTS AND DISCUSSION

Analysis of Hydrogen Atom Reactivity. Reactivity prediction models for hydrogen atoms proposed by Korzekwa et al.[10,11] and Singh et al.[12] were analyzed using CYP3A4 substrates. The total number of substrates used in the test was 72, having 1061 nonequivalent hydrogen atoms attached to carbon. The reactivity prediction models of Korzekwa et al. and Singh et al. showed similar prediction success rates of about 57% and 53% (Table 2). Although CYP3A4 has a large binding pocket and acts on diverse substrates, the model based on only the MO calculation showed low predictability of the CYP3A4 mediated metabolism. Many of the failed cases were sterically hindered atoms such as the tertiary carbon and the bridge head atom having a relatively stable radical product. Singh et al. proposed solvent accessible surface area exposure to filter out this kind of atoms.[12] However, they showed still low predictability in the prediction of CYP3A4 regioselectivity. Consideration of the protein structure may alleviate this problem.


(a)
Adinazolam
Alfentanil
Amiodarone
Azelastine
Buprenorphine
Citalopram
Clarithromycin
Clozapine
Codeine
Dextromethorphan
Diazepam
Diltiazem
Erythromycine
Fentanyl
Halofantrine
Haloperidol
Indinavir
LAAM
Lilopristone
Onapristone
Pimozide
Propafenone
Tamoxifen
Terbinafine
Terfenadine
Verapamil
Zaleplon
β-Arteether
Pranidipine
Teniposide
Cortisol
Dexamethasone
Docetaxel
Lovastatin
Midazolam
Quinidine
Carbamazepine
Ebastine
Colchicine
Etoposide
Iloperidone
Loratadine
Senecionine
(b)
Salmeterol
Saquinavir
Seratrodast
Testosterone

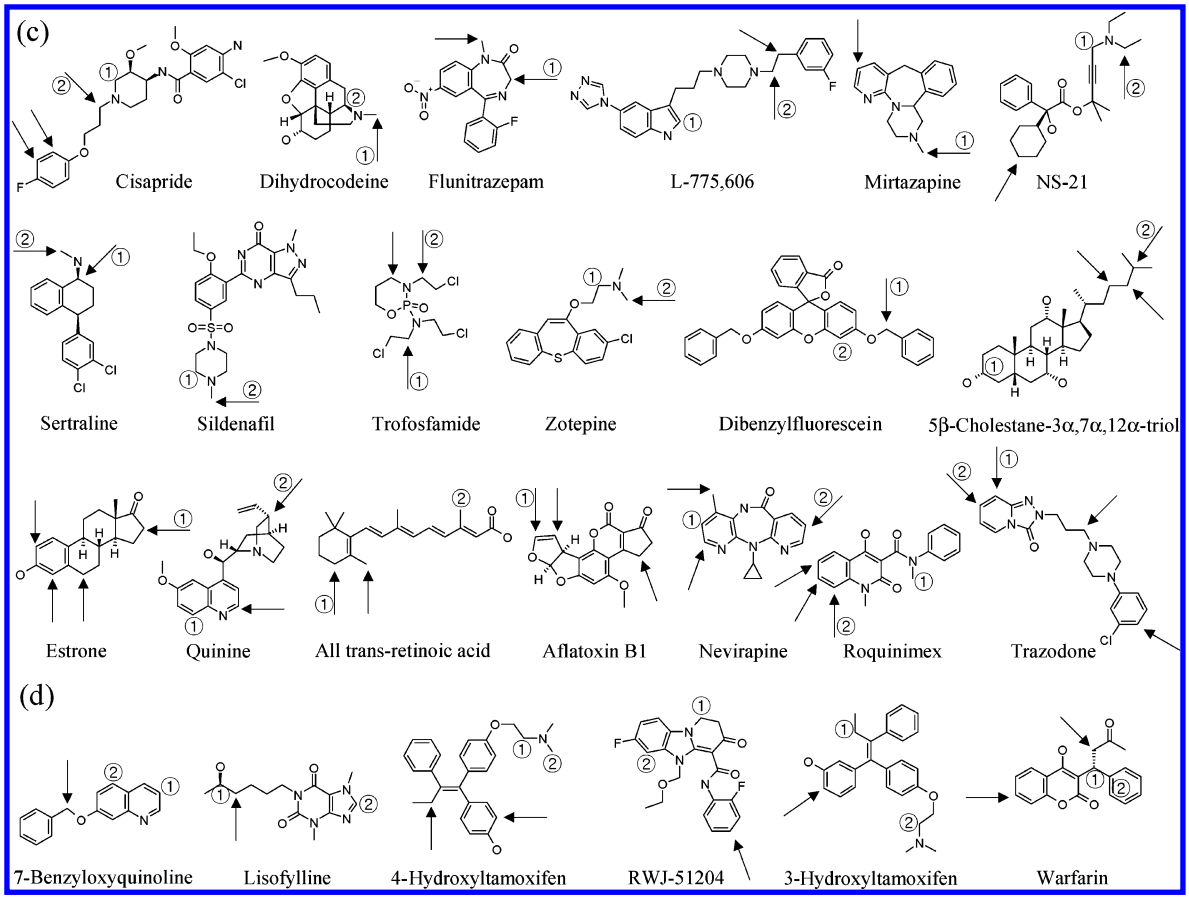


Figure 7. Chemical structures of (a) successful cases in the training set, (b) failed cases in the training set, (c) successful cases in the test set, and (d) failed cases in the test set. Experimentally known major sites of metabolism are indicated with arrows. Top 2 labile sites predicted with MLite are indicated with circled numbers.

Reactivity prediction models were analyzed according to the type of metabolic reaction (Table 2). The numbers of substrates of each reaction type differ in two models. Twenty-seven substrates have two or more metabolic reactions to CYP3A4. In the prediction of the substrates with multiple sites, we assigned the reaction type of this substrate as the reaction type that is the high ranking position (high reactivity) by the model. For example, trazodone has four metabolic sites by CYP3A4, one N-dealkylation site and three aromatic hydroxylation sites. In the model of Korzekwa et al., the aromatic hydroxylation reaction is more probable than the N-dealkylation reaction of trazodone by CYP3A4, and we assigned the reaction type of this substrate as aromatic hydroxylation. In the model of Singh et al., trazodone has more reactivity on the N-dealkylation reaction than the aromatic hydroxylation reaction, and the reaction type was assigned as N-dealkylation. The model of Singh et al. showed higher predictability in aliphatic hydroxylation reactions and lower predictability in N-dealkylation and aromatic hydroxylation reactions than the mode of Korzekwa et al. In particular, aromatic hydroxylation reaction sites within substrates with multiple sites were not detected in the model of Singh et al. (Figure 5). The aromatic hydroxylation site of carbamazepine was correctly predicted in the model of Singh et al. However aromatic hydroxylation was the only possible reaction type in the substrate. The reaction mechanisms of aliphatic hydroxylation and aromatic hydroxylation differ. It is impossible for the application of the simplified hydrogen abstraction model to the aromatic hydroxylation reaction. In Singh's model, the energy distribution of aromatic hydroxylation reactions shifted to a higher energy range than the energy distribution of the other metabolic reactions (Figure 6(b)). When aromatic hydroxylation reaction sites were compared with the other kind metabolic reaction sites in the same substrate, aromatic hydroxylation reaction sites had a lower reactivity (higher reaction energy) than the other kinds of metabolic reaction sites. The energy scaling between the different reaction types is required in the model of Singh et al. The activation energy estimation model of Korzekwa et al. showed a similar energy distribution in all metabolic reaction types (Figure 6(a)). However, this model showed low predictability in aliphatic hydroxylation reactions. This weakness seems to have originated from insufficient experimental data used to build an empirical model. This was complemented when accessibility was taken into account in our model. We combined the activation energy estimation model of Korzekwa et al. with the ensemble catalyticphore-based docking model.

Even though the result has been improved a lot by including accessibility as well as reactivity, there is a fundamental problem in embedding the model of Korzekwa et al. into the prediction system—MO calculation of substrates and metabolites remain time-consuming. More simple and fast empirical rules to estimate the activation energy of metabolic reactions are required to save computational time for the prediction. We are currently developing empirical functions to predict the activation energy with atomic descriptors instead of time-consuming MO calculation.

Application of Combined Model. MLite, a combined model with the ensemble catalyticphore model and the

Table 3. Prediction Ratio for CYP3A4 Substrates Found Using the Top Two Sites for MLite[a]

	training set			test set		
reaction type	N[b]	Nc[c]	ratio (%)[d]	N[b]	Nc[c]	ratio (%)[d]
aliphatic hydroxylation	10	6	60.0	6	4	66.7
N-dealkylation	28	27	96.4	10	10	100.0
O-dealkylation	8	3	37.5	2	1	50.0
aromatic hydroxylation	1	1	100.0	7	4	57.1
total	47	37	78.7	25	19	76.0

[a] Combined model with the ensemble catalyticphore model and the activation energy estimation model of Korzekwa et al. [b] N = number of substrates. [c] Nc = number of correctly predicted substrates. [d] Percentage of correct predictions.

Table 4. Prediction Ratio for CYP3A4 Substrates Found Using the Top Two Sites for MLite[a] and Modified MLite[b]

		percentage of correct prediction (%)	
reaction type	number of substrate	MLite	modified MLite
aliphatic hydroxylation	17	64.7	64.7
N-dealkylation	37	97.3	91.9
O-dealkylation	10	40.0	90.0
aromatic hydroxylation	8	62.5	62.5
total	72	77.8	81.9

[a] Combined model with the ensemble catalyticphore model and the activation energy estimation model of Korzekwa et al. [b] In the modified MLite, a penalty value (−4 kcal/mol) was added to the activation energy of O-dealkylation reactions neighboring a double bond.

activation energy estimation model of Korzekwa et al., was applied to predict regioselectivity in CYP3A4 substrates. Ensemble catalyticphore-based docking was used for the selection of exposed hydrogen atoms within a substrate. Catalyticphore-based docking with the final five queries generated binding modes of the substrate on the CYP3A4 enzyme. By the analysis of the binding mode, hydrogen atoms exposed to heme were selected. These hydrogen atoms are ranked by the activation energy of Korzekwa's model. This combined model, MLite, predicted the experimentally known metabolic sites within the first two sites in the ranking list in about 79% of the cases for the training set and about 76% of the cases for the test set (Table 3, Figure 7). This predictability is comparable with the MetaSite,[16] which is a widely distributed combined model with a predictability of 78% within the first three sites,[26] and a recently published QSAR model proposed by Sheridan et al.[7] with the predictability of 74% with the first two sites.

The main advantage was shown on the reaction types of aliphatic hydroxylation and N-dealkylation. Especially, MLite predicted the N-dealkylation reaction type correctly within the first two sites in the ranking list except for ebastine. Many atoms that have high reactivity but are not the reaction site cannot take a close proximity to heme oxygen atom in the CYP3A4−substrate complex. Predictability of MLite is similar with that of the model of Korzekwa et al. on the reaction types of O-dealkylation and aromatic hydroxylation. The number of data with aromatic hydroxylation is too small in the training set. When we moved some of the data from the test set to the training set, we had the similar result for aromatic hydroxylation. Many of the failed cases were O-dealkylation reactions of oxygen atoms neighboring a double bond (Figure 7). To confirm the estimated activation energy model of Korzekwa et al., transition state calculations were performed in the AM1 level. The directly calculated and estimated activation energies showed a similar trend. The activation energy of O-dealkylation reaction neighboring a double bond has a similar pattern with activation energy of the other reaction types in DFT/B3LYP level calculations.[33] However, when we introduced a penalty value (−4 kcal/mol) to the activation energy of O-dealkylation reaction neighboring a double bond, predictability increased for the O-dealkylation reaction without a detriment to the other reaction types (Table 4). It seems quite probable that the activation energy of the metabolic reaction in enzymes is different than in isolated systems. In enzymes, additional interactions between the substrate and the enzyme will interrupt the reaction path and change the activation energy. More detail analysis of reaction paths and binding modes in the enzyme system are required for more accurate predictions of regioselectivity.

To increase the predictability of our combined model, scoring functions were introduced for the representation of the binding affinity of each binding mode. Quantitative prediction of binding affinities are hard, especially with crude binding orientations like those from pharmacophore based docking.[32] Many selection schemes based on the score (such as considering the best docking solution, the top 10 docking solutions, or the top 10% docking solutions for each substrate) were tried, but there were no improvements in predictability (data not shown), considering all docking solutions for each substrate showed higher predictability than any of the scoring functions we tried.

In this study, enzyme flexibility was not considered. CYP3A4 has many flexible residues in its active site. Arg212, which was used as a source of the hydrogen bond acceptor in catalyticphore generation, is the flexible residue in the active site. The side chain of Arg212 is positioned inside of the active site in the ligand-free form and outside in some of the substrate complexes.[25] The conformation of protein dramatically changes based on the binding ligand. The active site volume of the complex structure with a large ligand is greater than that of the ligand-free structure by >80%.[24] New catalyticphore models which include enzyme flexibility are required for more accurate predictions.

Binding Mode Analysis. The merit of MLite is offering binding structures of the substrate−enzyme complex. The structure of the binding complex gives information about the metabolic reaction process and the direction of modification. Selection of the productive binding mode by the scoring function is impossible in crude docking results. However, the selection of a more reasonable binding configuration among the productive binding mode is possible. Reproduction of the 3D structure of the experimentally determined substrate−enzyme complex is the general approach to validate the docking method. However, there are limited experimental data on CYP3A4 substrate−enzyme complexes. The CYP3A4-progesterone structure[22] is far from the active site, and the CYP3A4-erythromycin structure[24] is not the productive binding mode. We compared the catalyticphore-based docking result with published force-field based docking results.[25] The PMF score[31] calculation was performed to represent the stability of the binding complex. The lowest energy binding conformations of amiodarone, diltiazem, and

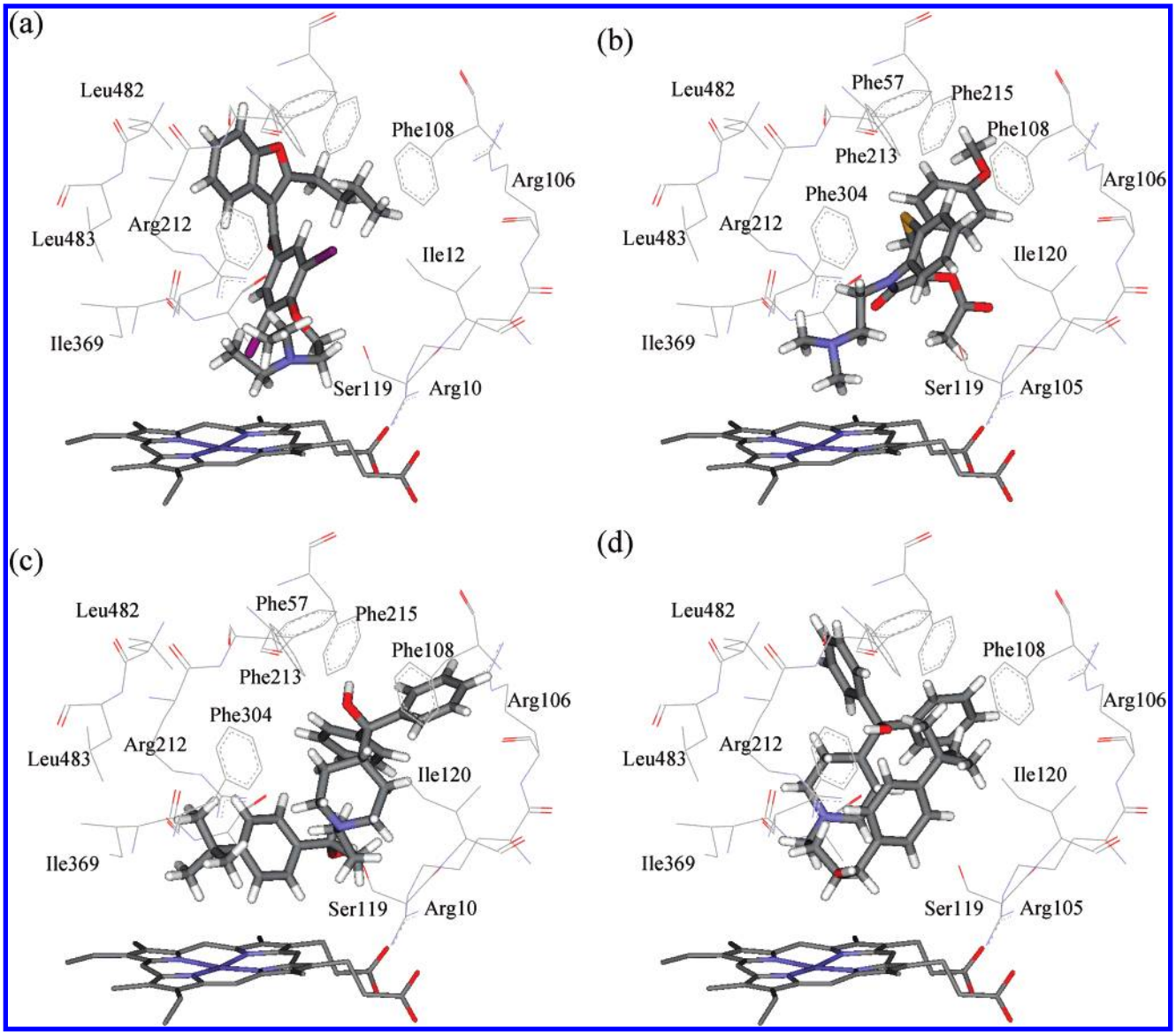


Figure 8. The lowest energy binding configurations of (a) amiodarone, (b) diltiazem, and (c) terfenadine and the second lowest energy binding configure of (d) terfenadine bound to CYP3A4. The important residues interacting with the bound substrate in the CYP3A4 complex and a heme group are shown.

terfenadine are in good agreement with force-field based docking results (Figure 8).

Deethylation is the major metabolic reaction of amiodarone by CYP3A4. The ethyl group of amiodarone is in close proximity to the heme oxygen atom of enzyme on the lowest energy complex structure of CYP3A4 with amiodarone (Figure 8(a)). The butyl group of the drug forms a lipophilic interaction with the hydrophobic cluster of CYP3A4 and the oxygen atom of the ketone group hydrogen bonded to the side chain of Arg212. This binding configuration is very similar with that of force-field based docking and consistent with experimental results. The dominant metabolic reaction of diltiazem by CYP3A4 is N-demethylation. The lowest energy binding configuration of CYP3A4 with diltiazem is suitable for that reaction (Figure 8(b)). The benzene portion of the heterocyclic ring and anisole group of diltiazem form $\pi-\pi$ interactions with the hydrophobic cluster of CYP3A4 in the diltiazem−CYP3A4 complex. Terfenadine having two metabolic reactions by CYP3A4 is metabolized to azacyclonol or terfenadine alcohol. In the lowest energy binding configuration of CYP3A4 with terfenadine, the CH_2 group neighboring the piperidyl nitrogen atom is in close proximity to the heme oxygen atom (Figure 8(c)). In the second lowest energy binding configuration, two of the tert-butyl methyl groups are located in close proximity to the heme oxygen atom (Figure 8(d)). In these two binding configurations, the phenyl groups of terfenadine form $\pi-\pi$ interactions with the hydrophobic cluster of CYP3A4. These binding configurations may be useful in drug modification by the structure-based drug design.

CONCLUSION

Regioselectivity prediction with reactivity methods, such as the models of Korzekwa et al. and Singh et al., showed similar prediction success rates of about 50%. Many of the failed cases were sterically hindered atoms such as the tertiary carbon and the bridge head atom having relatively stable radical products. It is necessary to introduce information about the binding site for more accurate prediction. A combined model, MLite, has been proposed to predict the

regioselectivity of the CYP3A4 substrate. The model is the combination of the ensemble catalyticphore model and the activation energy estimation model of Korzekwa et al. which were employed for the prediction of accessibility and reactivity, respectively. MLite predicted accurately about 76% of the first two sites in the ranking list of CYP3A4 substrates. This success rate is comparable with a MetaSite, a widely distributed package, and is high enough to be useful in drug discovery and development. MLite also offers the binding structures of the substrate−enzyme complex. These binding configurations may be useful in drug modification by the structure-based drug design.

In conclusion, this combined model predicted reasonably well labile metabolic sites of the CYP3A4 substrates. The development of empirical rules to calculate substrate reactivity will make this model more applicable in HTS. The approach used in the model can be easily expanded for the qualitative prediction of drug metabolism mediated by not only CYP3A4 but also other CYP450 family enzymes. Although CYP3A4 is the main isoenzyme which participates in drug metabolism, the other isoenzymes, CYP2D6, CYP1A2, CYP2C9, and CYP2C19, must be considered. The different local relative amount of various metabolites will arise in different tissues depending upon not only phase I metabolism but also phase II conjugation, protein binding, the route of administration, efflux mechanism of the interesting drug, and so on. Integrated consideration of CYPs is required to develop more efficient and helpful in silico packages for drug discovery and development.

ACKNOWLEDGMENT

This work was financially supported by the Ministry of Commerce, Industry and Energy (MOCIE), Korea. We thank the Korea Science and Engineering Foundation through the Hyperstructured Organic Materials Research Center in Seoul National University for the financial support.

REFERENCES AND NOTES

(1) Guengerich, F. P. Cytochrome P-450 3A4: regulation and role in drug metabolism. *Annu. Rev. Pharmacol. Toxicol.* **1999**, *39*, 1−17.
(2) Guengerich, F. P. The 1992 Bernard B. Brodie Award lecture. Bioactivation and detoxification of toxic and carcinogenic chemicals. *Drug Metab. Dispos.* **1993**, *21*, 1−6.
(3) Hou, T. J.; Xu, X. J. Recent development and application of virtual screening in drug discovery: an overview. *Curr. Pharm. Des.* **2004**, *10*, 1011−1033.
(4) Klopman, G.; Dimayuga, M.; Talafous, J. META 1. A program for the evaluation of metabolic transformation of chemicals. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1320−1325.
(5) Darvas, F. Predicting metabolic pathways by logic programming. *J. Mol. Graphics* **1988**, *6*, 80−86.
(6) Borodina, Y.; Rudik, A.; Filimonov, D.; Kharchevnikova, N.; Dmitriev, A.; Blinova, V.; Poroikov, V. A new statistical approach to predicting aromatic hydroxylation sites. Comparison with model-based approaches. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1998−2009.
(7) Sheridan, R. P.; Korzekwa, K. R.; Torres, R. A.; Walker, M. J. Empirical regioselectivity models for human cytochromes P450 3A4, 2D6, and 2C9. *J. Med. Chem.* **2007**, *50*, 3173−3184.
(8) Ekins, S.; Bravi, G.; Wikel, J. H.; Wrighton, S. A. Three-dimensional-quantitative structure activity relationship analysis of cytochrome P-450 3A4 substrates. *J. Pharmacol. Exp. Ther.* **1999**, *291*, 424−433.
(9) Ekins, S.; Bravi, G.; Ring, B. J.; Gillespie, T. A.; Gillespie, J. S.; Vandenbranden, M.; Wrighton, S. A.; Wikel, J. H. Three-dimensional quantitative structure activity relationship analyses of substrates for CYP2B6. *J. Pharmacol. Exp. Ther.* **1999**, *288*, 21−29.
(10) Korzekwa, K. R.; Jones, J. P.; Gilette, J. R. Theoretical studies on cytochrome P-450 mediated hydroxylation: A predictive model for hydrogen atom abstractions. *J. Am. Chem. Soc.* **1990**, *112*, 7042−7046.
(11) Jones, J. P.; Mysinger, M.; Korzekwa, K. R. Computational models for cytochrome P450: A predictive electronic model for aromatic oxidation and hydrogen atom abstraction. *Drug Metab. Dispos.* **2002**, *30*, 7−12.
(12) Singh, S. B.; Shen, L. Q.; Walker, M. J.; Sheridan, R. P. A model for predicting likely sites of CYP3A4-mediated metabolism on drug-like molecules. *J. Med. Chem.* **2003**, *46*, 1330−1336.
(13) de Groot, M. J.; Ackland, M. J.; Horne, V. A.; Alex, A. A.; Jones, B. C. A novel approach to predicting P450 mediated drug metabolism. CYP2D6 catalyzed N-dealkylation reactions and qualitative metabolite predictions using a combined protein and pharmacophore model for CYP2D6. *J. Med. Chem.* **1999**, *42*, 4062−4070.
(14) de Groot, M. J.; Alex, A. A.; Jones, B. C. Development of a combined protein and pharmacophore model for cytochrome P450 2C9. *J. Med. Chem.* **2002**, *45*, 1983−1993.
(15) Zamora, I.; Afzelius, L.; Cruciani, G. Predicting drug metabolism: a site of metabolism prediction tool applied to the cytochrome P450 2C9. *J. Med. Chem.* **2003**, *46*, 2313−2324.
(16) Cruciani, G.; Carosati, E.; Boeck, B. D.; Ethirajulu, K.; Mackie, C.; Howe, T.; Vianello, R. MetaSite: Understanding metabolism in human cytochromes from the perspective of the chemist. *J. Med. Chem.* **2005**, *48*, 6970−6979.
(17) de Graaf, C.; Pospisil, P.; Pos, W.; Folkers, G.; Vermeulen, N. P. E. Binding mode prediction of cytochrome P450 and thymidine kinase protein-ligand complexes by consideration of water and rescoring in automated docking. *J. Med. Chem.* **2005**, *48*, 2308−2318.
(18) De Voss, J. J.; Sibbesen, O.; Zhang, Z.; Ortiz de Montellano, P. R. Substrate docking algorithms and prediction of the substrate specificity of cytochrome $P450_{cam}$ and its L244A mutant. *J. Am. Chem. Soc.* **1997**, *119*, 5489−5498.
(19) Park, J. Y.; Harris, D. Construction and assessment of models of CYP2E1: Predictions of metabolism from docking, molecular dynamics, and density functional theoretical calculations. *J. Med. Chem.* **2003**, *46*, 1645−1660.
(20) He, Y. A.; Roussel, F.; Halpert, J. R. Analysis of homotropic and heterotropic cooperativity of diazepam oxidation by CYP3A4 using site-directed mutagenesis and kinetic modeling. *Arch. Biochem. Biophys.* **2003**, *409*, 92−101.
(21) Rendic, S. Summary of information on human CYP enzymes: human P450 metabolism data. *Drug Metab. Rev.* **2002**, *34*, 83−448.
(22) Williams, P. A.; Cosme, J.; Vinkovic, D. M.; Ward, A.; Angove, H. C.; Day, P. J.; Vonrhein, C.; Tickle, I. J.; Jhoti, H. Crystal structures of human Cytochrome P450 3A4 bound to metyrapone and progesterone. *Science* **2004**, *305*, 683−686.
(23) Yano, J. K.; Wester, M. R.; Schoch, G. A.; Griffin, K. J.; Stout, C. D.; Johnson, E. F. The structure of human microsomal cytochrome P450 3A4 determined by X-ray crystallography to 2.05-Å resolution. *J. Biol. Chem.* **2004**, *279*, 38091−38094.
(24) Ekroos, M.; Sjogren, T. Structural basis for ligand promiscuity in Cytochrome P450 3A4. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 13682−13687.
(25) Lill, M. A.; Dobler, M.; Vedani, A. Prediction of small-molecule binding to cytochrome P450 3A4: flexible docking combined with multidimensional QSAR. *ChemMedChem* **2006**, *1*, 73−81.
(26) Zhou, D.; Afzelius, L.; Grimm, S. W.; Andersson, T. B.; Zauhar, R. J.; Zamora, I. Comparison of methods for the prediction of the metabolic sites for CYP3A4-mediated metabolic reactions. *Drug Metab. Dispos.* **2006**, *34*. 976−983.
(27) *Cerius*^{2}, *Version 4.10*; Accelrys Software Inc.: San Diego, CA, 2005.
(28) A multiple conformer represents the flexibility of a molecule. All conformers within 20 kcal/mol of the lowest energy conformer are generated with a poling technique. The maximum number of conformers was specified as 999 instead of the default value of 100.
(29) Narasimhulu, S. Differential behavior of the sub-sites of cytochrome 450 active site in binding of substrates, and products (implications for coupling/uncoupling). *Biochim. Biophys. Acta* **2007**, *1770*, 360−375.
(30) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko,

A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03, Revision C.02*; Gaussian, Inc.: Wallingford, CT, 2004.

(31) Muegge, I.; Martin, Y. C. A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *J. Med. Chem.* **1999**, *42*, 791.

(32) Gohlke, H.; Klebe, G. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angew. Chem. Int. Ed.* **2002**, *41*, 2644−2676.

(33) Olsen, L.; Rydberg, P.; Rod, T. H.; Ryde, U. Prediction of activation energies for hydrogen abstraction by cytochrome p450. *J. Med. Chem.* **2006**, *49*, 6489−6499.

CI7003576