QSAR Analysis of the Inhibition of Recombinant CYP 3A4 Activity by Structurally Diverse Compounds Using a Genetic Algorithm-Combined Partial Least Squares Method

Suchada Wanchana,[1] Fumiyoshi Yamashita,[1] and Mitsuru Hashida[1,2]

Received February 18, 2003; accepted May 12, 2003

Purpose. To develop a quantitative structure/activity relationship (QSAR) model for predicting drug–CYP 3A4 interactions.
Method. The inhibitory effect of 53 structurally diverse drugs on the metabolism of 7-benzyloxy-4-trifluoromethyl coumarin (BFC) by recombinant CYP 3A4 was evaluated using a rapid microtiter plate assay. For each drug, a total of 220 two-dimensional topological indices were calculated using Molconn-Z software. Using a genetic algorithm-based partial least squares (GA-PLS) method, the desired descriptors were automatically selected to maximize the predictability of the IC_{50} values.
Results. The IC_{50} values of the drugs tested ranged from 9 nM to 2 mM. Based on the GA-PLS method, five principal components derived from 20 Molconn-Z descriptors were found to be effective for QSAR modeling. Interestingly, these descriptors suggested that the molecular size would be an important factor in determining drug–CYP 3A4 interactions. In the leave-one-out prediction, the r_{pred} and the standard error of prediction (s) were 0.754 and 0.787, respectively. Even in an external validation, the predictions were in good agreement with experimental values ($r_{pred} = 0.744$, $s = 0.769$, $n = 9$).
Conclusions. The proposed model, in which two-dimensional topological descriptors were used as molecular descriptors, was able to predict drug–CYP 3A4 interactions with reasonable accuracy.

KEY WORDS: genetic algorithm; partial least squares; CYP 3A4; quantitative structure/activity relationship; Molconn-Z.

INTRODUCTION

The cytochrome P450s (CYPs) are a superfamily of heme-containing mixed-function oxygenases that catalyze the regio- and stereoselective oxidation of a wide variety of xenobiotics, including drugs. CYP 3A4 is the most abundant human hepatic CYP isoform and is responsible for the metabolism of almost 50% of known drugs by humans (1). CYP 3A4 is also known to be functionally active in the intestinal epithelium, which restricts the oral absorption of xenobiotics (2). In addition, inhibition of CYP 3A4 by coadministered drugs has been shown to result in adverse clinical drug–drug interactions, some of which may be fatal (3), because of a reduction in the body clearance of the drugs and a rapid and unexpected rise in their blood concentrations. Early identification of potential CYP 3A4 inhibitors is therefore needed to minimize the risk of clinically relevant interactions.

In addition to *in vitro* high-throughput screening techniques (4,5), *in silico* prediction of xenobiotic metabolism is now attracting increasing attention (6,7). The advantage of *in silico* technologies is that the properties of molecules can be assessed from a knowledge of their chemical structures alone, either two- or three-dimensionally. *In silico* filtering is expected to help identify and screen out compounds that are unlikely to become useful drugs, thereby maximizing the output of the drug discovery process.

In parallel with the application of homology models, where the active sites of human CYPs are predicted from crystallized structures of bacterial soluble CYP enzymes (8,9), three-dimensional quantitative structure/activity relationship (3D-QSAR) methods are being used for modeling the CYP enzymes (7,10–12). 3D-QSAR methods allow us to interpret and understand enzyme active sites and receptors through the superimposition of a set of different compounds; in other words, these methods can provide graphic representations of the binding sites even when no crystal structure is available. In the case of CYP 3A4, Ekins *et al.* (10) built a 3D-QSAR model using the program Catalyst® (Molecular Simulations, San Diego, CA) from the Michaelis-Menten constants of 38 structurally diverse substrates. However, a 3D-QSAR model assumes that the binding modes of the compounds are the same. Because of the relatively large binding pocket of CYP 3A4 (9), it is clear that substrates can have a relatively large conformational degree of freedom within the active site. The binding pocket can also accept two molecules simultaneously (13). Such a variety of binding modes for CYP 3A4 substrates and inhibitors appears to limit the predictive power of a 3D-QSAR model.

These limitations of 3D-QSAR techniques prompted us to apply a 2D-QSAR model to predicting the interaction with CYP 3A4. Topologic indices are attractive descriptors because they can be calculated easily and rapidly from any two-dimensional structural formula. Various physicochemical characteristics, e.g., partition coefficients (14) and boiling points (15), can be described by topological indices. Recently, we developed a genetic algorithm combined with the partial least squares (GA-PLS) approach for feature selection in QSAR (16,17), demonstrating that the model obtained performed well as far as predicting the aqueous solubility and membrane permeability of unknown compounds was concerned.

In the present study, we applied the GA-PLS approach to develop a 2D-QSAR model for predicting the interaction with CYP 3A4. Molconn-Z-derived topological descriptors (18) were used as structural descriptors for QSAR modeling, and these included connectivity indices, shape indices, electrotopological state (E-state) indices, and atom-type E-state indices. The inhibitory effect of 53 structurally diverse drugs on the metabolism of 7-benzyloxy-4-trifluoromethyl coumarin (BFC) by recombinant CYP 3A4 was evaluated using a rapid microtiter plate assay, and then QSAR analysis was performed.

[1] Department of Drug Delivery Research, Graduate School of Pharmaceutical Sciences, Kyoto University, Kyoto 606-8501, Japan.
[2] To whom correspondence should be addressed. (e-mail: hashidam@pharm.kyoto-u.ac.jp)

0724-8741/03/0900-1401/0 © 2003 Plenum Publishing Corporation

MATERIALS AND METHODS

Materials

Midazolam was kindly supplied by Yamanouchi Pharmaceutical Co., Ltd. (Japan). Alprazolam, amiodarone, astemizole, β-estradiol, carbamazepine, clonazepam, clotrimazole, clozapine, cortisone, buspirone, dapsone, dexamethasone, diltiazem, flutamide, hydrocortisone, ketoconazole, miconazole, mifepristone, nicardipine, nimodipine, paclitaxol, quinidine, quinine, sterigmatocystin, terfenadine, testosterone, triazolam, and warfarin were purchased from Sigma Chemicals (St. Louis, MO). Caffeine, 4-androstene-3,17,dione, benzo(a) pyrene, chlorpheniramine, lidocaine, quercetin and verapamil were from Nacalai Tesque (Kyoto, Japan). Progesterone was from ICN Biomedical Inc. (Ohio, USA). Acetaminophen, benzaldoxime, colchicine, cyclophosphamide, cyclosporin A, dextromethorphan, diazepam, erythromycin, estriol, ethynylestradiol, haloperidol, imipramine, nifedipine, tamoxifen, troleandomycin, and vinblastine were obtained from Wako Pure Chemicals Industries, Ltd. (Osaka, Japan). Baculovirus/insect cell cDNA-expressed CYP 3A4 was purchased from Gentest (Woburn, MA). Glucose-6-phosphate disodium salt, glucose-6-phosphate dehydrogenase, β-nicotinamide adenine dinucleotide phosphate sodium (β-NADP), and 7-hydroxy-4-trifluoromethylcoumarin were purchased from Sigma Chemicals (St. Louis, MO).

Synthesis of 7-Benzyloxy-4-(Trifluoromethyl)-Coumarin (BFC)

BFC was synthesized by the method of Bridge *et al.* (19). Briefly, 7-hydroxy-4-trifluoromethylcoumarin (2.47 mmole) was refluxed with benzyl bromide (4 mmole) and potassium carbonate (5.06 mmole) in 10 ml acetone for 24 h. The crude product was purified by recrystallization three times with diethyl ether and stored at −20°C until use. ^{1}H-NMR ($CDCl_3$) δ: 7.61–7.66 (m, 5H), 7.33–7.45 (m, $-C_6H_5$), 6.98–7.02 (q, 6H), 6.94–6.95 (d, 8H), 6.62 (s, 3H), 5.16 (s, $-OCH_2-$); FAB-MS m/z 321 (M+). The recovery of BFC was 21.7%. Elemental analysis gave C = 63.68% (63.75%) H = 3.33% (3.44%), F = 17.61% (17.81%), and O = 15.38% (15.00%).

Inhibition Studies

Fifty-three structurally diverse compounds that were known or suspected to interact with CYP 3A4, as substrates or inhibitors (20), were subjected to the assay. Their inhibitory effect on the metabolism of BFC by recombinant CYP 3A4 was evaluated using the same rapid microtiter plate assay as Stresser *et al.* reported (4). Incubation was conducted in a volume of 200 μl in 96-well microtiter plates (Catalog 353072, Becton Dickinson, NJ) according to a routine protocol. Serial dilutions were performed using a multichannel pipetter. A cofactor/serial dilution (C/SD) buffer was prepared in 50 mM potassium phosphate, pH 7.4. This buffer contained 2.6 mM β-NADP, 6.6 mM glucose-6-phosphate, and 0.8 U/ml glucose-6-phosphate dehydrogenase. Then 144 μl C/SD buffer that lacked test compound was added to the first well in each row, and 100 μl of the same C/SD buffer to the second and all remaining wells. Six microliters of the test compounds was added to the well in the first column. All test compounds were dissolved in acetonitrile. Fifty microliters of the solution from the first well in each row was then transferred to the second well and serially diluted 1:3 through the eighth well. Wells 9 and 10 contained no test compound, and wells 11 and 12 were used as controls for background fluorescence (enzyme and substrate were added after the reaction was terminated). The final concentration of the test compounds in the first well varied between 1 nM and 10 mM, depending on their solubility characteristics or potency. The plate was then preincubated at 37°C for 20 min, and the reaction was initiated by the addition of 100 μl 350 mM potassium phosphate buffer, pH 7.4, containing 10 pmole/ml insect cell-expressed CYP 3A4 and 100 μM BFC. The substrates were initially prepared in acetonitrile. Here, the final concentration of BFC was set to be 50 μM, in part because the metabolism of BFC is linear with respect to BFC concentration up to 100 μM (4). The reaction was terminated after 30 min by the addition of 75 μl 4:1 acetonitrile : 0.5 M Tris base solution. The fluorescence of the BFC metabolite, 7-hydroxy-4-trifluoromethyl coumarin, in each cell was measured using a Wallac multilabel counter model 1420 fluorescence plate reader (Perkin Elmer, Finland) at an excitation wavelength of 405 nm and an emission wavelength of 535 nm. The IC_{50} values of each compound were calculated through curve fitting of the Hill equation.

Data Analysis

Calculated Molecular Descriptors

The topologic descriptors were calculated by Molconn-Z software (Hall Associated Consulting, Quincy, MA) on the basis of two-dimensional structures. A total of 220 connectivity, shape, and atom-type E-state indices were calculated from the two-dimensional geometry. Molecular connectivity indices are nonempirical structure descriptions that contain information on intermolecular accessibility (21), whereas E-state indices contain information reflecting intermolecular accessibility of atoms and groups in a molecule, specifically electron accessibility (22).

Genetic Algorithm-Driven Optimization

A population of 100 random subsets of the structural descriptors was generated. Each subset was encoded as a binary string of digits, with a length corresponding to the total number of descriptors. A value of "1" implied that the descriptor was regarded as being important, whereas a value of "0" implied that the descriptor could be disregarded. The predictive q^2 (r_{pred}^2) value was used as a fitness function in the genetic algorithm optimization:

$$r_{pred}^2 = 1 - \frac{\sum (y_i - y_{pred})^2}{\sum (y_i - \bar{y})^2} \quad (1)$$

where y_i and $\bar{y}$ are the observed dependent variables and their average, and y_{pred} is the value predicted by the QSAR model. The predictability of the model was evaluated using the "leave-one-out" procedure. This method systematically removed one data point at a time from the data set. A model equation was then constructed on the basis of the reduced data set and subsequently used to predict the removed data point. This procedure was repeated until a complete set of predicted values was obtained.

In the genetic algorithm, two "parent" strings were se-

lected randomly by a roulette wheel selection method according to the fitness values. A two-point crossover of the "parent" strings was performed at a predefined probability (p) of 0.8. One of the new strings was taken, subjected to random mutation ($p = 0.01$), and stored as an "offspring" in the next generation. For each generation, the series of steps was repeated until the predefined population number (n = 100) was obtained, provided that the five best strings were kept for the next generation (Elite = 5). When the generation number reached 500, the calculation was stopped, and the best string in the generation was taken. A series of the computations were performed with an in-house program written in Microsoft Visual C++ 6.0 running on the Microsoft Windows platform.

RESULTS

Inhibition Study Results

In this study BFC was selected as a model substrate because the metabolism of BFC was linear with respect to its high concentration (~100 μM), and the IC_{50} value for BFC was in a good correlation with that for traditional compounds (4). Figure 1 shows typical examples of the inhibitory effects of eight concentrations of test compounds on recombinant CYP 3A4 activity in the microtiter plate assay. Table I summarizes the IC_{50} values of 53 structurally diverse compounds for CYP 3A4. These IC_{50} values ranged from 9 nM to 2 mM, covering the most potent inhibitors ($IC_{50} < 1$ μM, n = 6), potent inhibitors (IC_{50} = 1–10 μM, n = 11), fairly potent inhibitors (IC_{50} = 10–100 μM, n = 13) and poor inhibitors ($IC_{50} > 100$ μM, n = 14). The IC_{50} value was not determined when 50% inhibition did not occur at the highest concentration tested. The drugs that are known to cause significant drug–drug interactions, such as ketoconazole (23) and erythromycin (1), tend to inhibit the metabolism of BFC.

Genetic Algorithm-Driven Optimization in QSAR Modeling

With the drugs that have limited solubility [benzo(a)pyrene, caffeine, cortisone, dexamethasone, flutamide, hydrocortisone, taxol, and estriol] or that activate the metabolism of BFC (testosterone) omitted, the log IC_{50} values of 44 compounds were subjected to QSAR analysis. Thirty-five compounds were selected randomly as training data, and the remaining nine compounds were used for external validation of a prediction model. For all of these compounds, topological indices were calculated using Molconn-Z software. The indices with a path length of seven or higher were not used for modeling because their information is hard to interpret (18). In addition, the descriptors representing a heavily skewed distribution (skewness greater than 3) were removed. This method reduced the entire descriptor pool to 62 members.

Figure 2 shows the trajectory of the genetic algorithm-driven optimization, where the predictive q^2 obtained in a "leave-one-out" procedure for 35 compounds was used as a fitness function (Eq. 1). The average of the fitness values in the population tends to increase with increasing generation number and reaches a plateau at approximately 200 generations. The best solution at the 500th generation was taken for modeling the interaction of compounds with CYP 3A4.

Table II summarizes 22 Molconn-Z descriptors selected by genetic algorithm and their scaled PLS regression coefficients. Because the scaled PLS regression coefficients for $^4\chi_{pc}$ and SdsCH were negligibly small, these descriptors could be omitted. When the "leave-one-out" prediction was conducted using the remaining 20 descriptors, the optimal number of PLS principal components was found to be 5. The "leave-one-out" prediction gave an r_{pred} of 0.754 and a standard error of prediction (s) of 0.787 (Fig. 3A). By taking five PLS principal components of 20 descriptors from the entire data on the 35 compounds, the following linear equation for estimating the IC_{50} values was obtained:

$$\begin{aligned}\log IC_{50} = {} & 3.3681 - 0.0160\ \mathrm{nvx} + 0.1627\ \mathrm{nrings} - 0.0013\ \mathrm{fw} \\ & - 0.0203\ {}^0\chi - 0.0358\ {}^1\chi - 0.0301\ {}^2\chi + 0.0676\ {}^5\chi_p \\ & - 0.0225\ {}^0\chi_v + 0.1826\ {}^5\chi_{vp} + 0.2724\ {}^6\chi_{vp} \\ & - 6.6124\ {}^5\chi_{ch} - 5.7457\ {}^6\chi_{ch} - 6.0471\ {}^5\chi_{vch} \\ & - 0.0181\ {}^1\kappa_\alpha - 0.0695\ \mathrm{SHsOH} + 0.1848\ \mathrm{SaaN} \\ & - 0.0136\ \mathrm{SsOH} - 0.0959\ \mathrm{SHvin} + 0.6432\ \mathrm{NHBint5} \\ & + 0.6136\ \mathrm{NHBint6}\ (r = 0.88, s = 0.57, n = 35)\end{aligned} \quad (2)$$

As summarized in Table III, the five PLS principal compo-

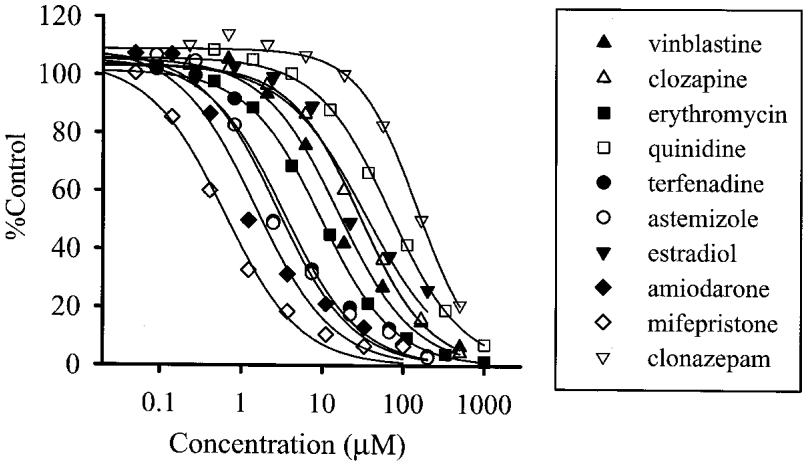

Fig. 1. Inhibitory effects of eight concentrations of the representative test compounds on the metabolism of 7-benzyloxy-4-trifluoromethyl coumarin (BFC) by recombinant CYP 3A4.

Table I. The IC_{50} Values of Test Compounds for CYP 3A4 in the Microtiter Plate Assay Calculated Using Three Parameters (the Hill Equation)

Drug	IC_{50} (μM)	Drug	IC_{50} (μM)
4-Androstene-3,17,dione	497	Flutamide	>2000
Acetaminophen	1496	Haloperidol	58.9
Alprazolam	423	Hydrocortisone	>500
Amiodarone	1.79	Imipramine	109
Astemizole	3.23	Ketoconazole	0.009
Benzaldoxime	361	Lidocaine	1451
Benzo(a)pyrene	>10	Miconazole	0.477
Buspirone	22.2	Midazolam	3.54
Caffeine	>2500	Mifepristone	0.682
Carbamazepine	376	Nicardipine	0.387
Chlorpheniramine maleate	709	Nifedipine	16.6
Clonazepam	167	Nimodipine	2.26
Clotrimazole	0.0341	Progesterone	46.0
Clozapine	30.5	Quercetin	19.9
Colchicines	194	Quinidine	76.0
Cortisone	>300	Quinine	57.3
Cyclophosphamide	1976	Sterigmatocystin	1.92
Cyclosporin A	4.19	Tamoxifen citrate	7.58
Dapsone	80.3	Taxol	>100
Dexamethasone	>1000	Terfenadine	3.70
Dextromethophan	99.2	Testosterone	Activation
Diazepam	115	Triazolam	163
Diltiazem	78.5	Troleandomycin	0.930
Erythromycin	9.82	Verapamil	2.77
β-Estradiol	37.0	Vinblastine	17.5
Estriol	>50	Warfarin	314
Ethynylestradiol	2.31		

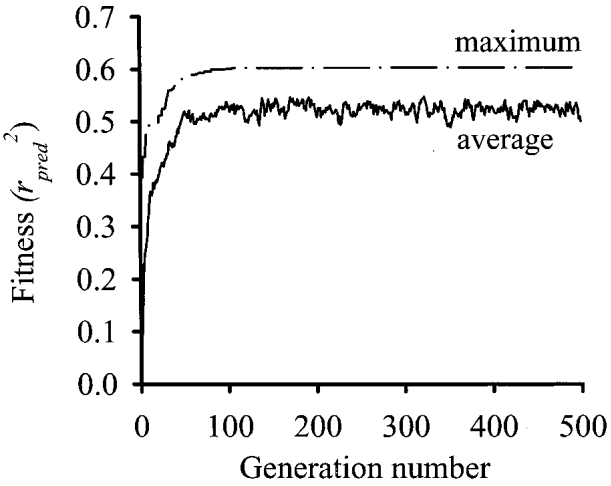

Fig. 2. Relationship between fitness and generation number in GA-PLS QSAR modeling of the interaction with CYP 3A4. The fitness function is defined as

$$r_{\text{pred}}^2 = 1 - \frac{\sum(y_i - y_{\text{pred}})^2}{\sum(y_i - \bar{y})^2}$$

where y_i and $\bar{y}$ were the observed dependent variables and their average; y_{pred} was the value predicted by the QSAR model. Dash-dot (—·—) and solid (—) lines represent the maximum and average of fitness values of individuals in the population. Each generation includes 100 individuals.

nents account for 85.7% of the variation in the factors (molecular descriptors) and 77.4% of the variation in the response (log IC_{50}). The relationship between the observed log IC_{50} values and those calculated from Eq. (1) is shown in Fig. 3B. In addition, an external validation of Eq. (2) was conducted with a testing data set (Fig. 4). The predictions were in good agreement with the observed values with an r_{pred} of 0.744 and an s of 0.769. Thus, it was demonstrated that the GA-PLS analysis gave a QSAR model of drug–CYP 3A4 interactions with a reasonable accuracy of prediction.

DISCUSSION

In the present study, a 2D-QSAR model for predicting drug–CYP 3A4 interactions was constructed using the GA-PLS method. In the PLS analysis (24), the matrix of explanatory variables is orthogonally decomposed with the inner relation between the explanatory and response variables being adhered to. Therefore, unlike MLR analysis, PLS analysis avoids any multicollinearity problems in the explanatory variables. The PLS regression is sufficiently noise-free because a minimal number of principal components are used for modeling. Thus, the PLS analysis is widely used in many fields of chemistry. However, it should be noted that the incorporation of unnecessary explanatory variables adversely affects PLS modeling. The use of a genetic algorithm is one way to eliminate this problem (16,17,25,26). Genetic algorithms (27) are search algorithms based on the mechanics of natural selection and natural genetics, which exploit the idea of the survival of the fittest and an interbreeding population. Genetic algorithms differ from more traditional optimization techniques in

Table II. Scaled PLS Regression Coefficients of the Subset of Molconn-Z Descriptors Selected by a Genetic Algorithm in Combination with PLS Regression

Descriptor	Symbol	Scaled PLS regression coefficient[a]
Significant descriptors		
Number of nonhydrogen atoms in molecule	nvx	–0.222
Number of rings in graph	nrings	0.251
Molecular weight	fw	–0.273
Path 0 simple connectivity index	$^{0}\chi$	–0.221
Path 1 simple connectivity index	$^{1}\chi$	–0.229
Path 2 simple connectivity index	$^{2}\chi$	–0.185
Path 5 simple connectivity index	$^{5}\chi_{p}$	0.251
Path 0 valence connectivity index	$^{0}\chi_{v}$	–0.210
Path 5 valence connectivity index	$^{5}\chi_{vp}$	0.357
Path 6 valence connectivity index	$^{6}\chi_{vp}$	0.389
Chain 5 simple connectivity index	$^{5}\chi_{ch}$	–0.464
Chain 6 simple connectivity index	$^{6}\chi_{ch}$	–0.414
Chain 5 valence connectivity index	$^{5}\chi_{vch}$	–0.250
Kappa simple index	$^{1}\kappa_{\alpha}$	–0.242
Hydrogen E-state index value for atom type –OH	SHsOH	–0.235
E-state index value for atom type --N--	SaaN	0.378
E-state index value for atom type –OH	SsOH	–0.175
E-state of C-atom in the vinyl group, =CH–	SHvin	–0.150
Count of potential internal hydrogen bonders	NHBint5	0.604
Count of potential internal hydrogen bonders	NHBint6	0.362
Descriptors that can be removed		
Path-cluster 4 simple connectivity index	$^{4}\chi_{pc}$	–0.072
E-state index value for atom type =CH–	SdsCH	–0.033

[a] Scaled PLS coefficients were calculated by multiplying the PLS regression coefficient and the standard deviation of the descriptors.

that they involve a search from a "population" of solutions, not from a single point. Therefore, genetic algorithms are viewed as a global optimization method. It should be noted from Fig. 2 that genetic algorithms are able to find an optimal solution very efficiently, taking into account the facts that the total number of combinations of molecular descriptors for modeling was 2^{62} (~4.6×10^{18}), and the number of individuals for each population was set to be only 100.

The optimal number of PLS principal components was determined by evaluating the predictability of the model in a

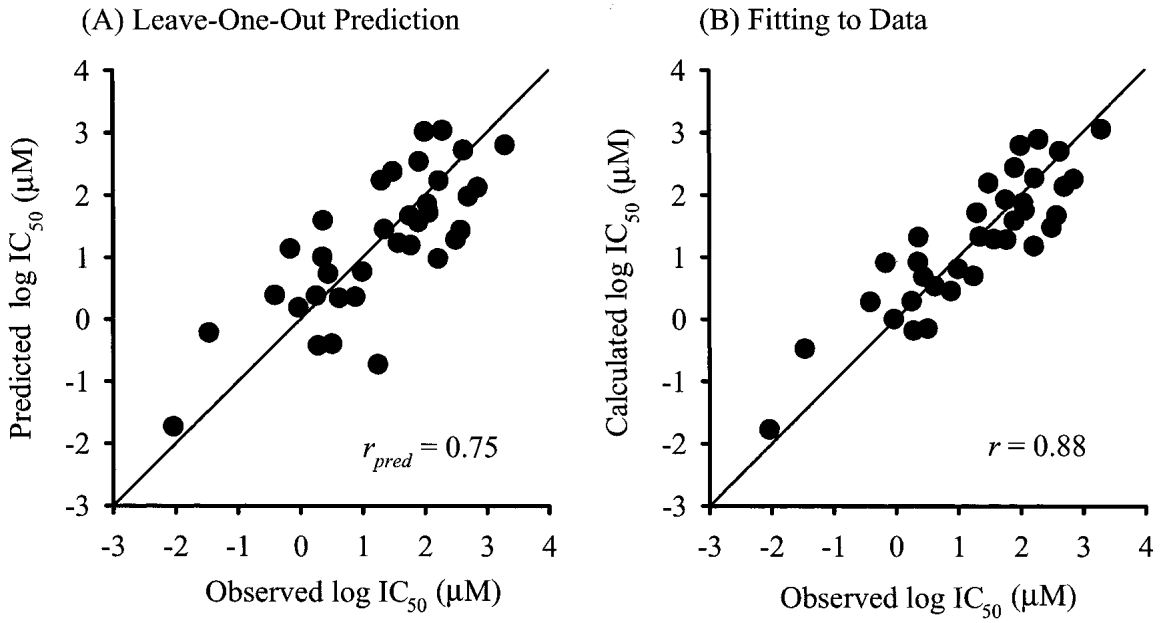

Fig. 3. Leave-one-out cross-validated predictability (A) and goodness of fit (B) of the QSAR model obtained. The model comprises the five PLS principal components of 20 descriptors listed in Table II. The correlation coefficient (r) between the experimental and calculated log IC_{50} values is also given.

Table III. Contribution of PLS Principal Components to the Variations of Response Variable (Log IC_{50}) and Explanatory Variables (Molecular Descriptors)

	Response variable		Explanatory variables	
Number of PC[a]	Variation (%)	Cumulative variation (%)	Variation (%)	Cumulative variation (%)
1	25.5	25.5	46.4	46.4
2	21.9	47.5	17.4	63.8
3	23.1	70.6	8.2	72.0
4	4.3	74.9	8.8	80.8
5	2.5	77.4	5.0	85.7

[a] PC, PLS principal components.

leave-one-out cross-validation procedure because extraction of too many principal components resulted in overfitting to the data. In this model, only five principal components were used for modeling, a number that was small enough as compared with the sample number (n = 35). In addition, this model could predict the IC_{50} values for external data as well.

The model obtained gave a reasonably accurate prediction, with an r_{pred} of 0.754 and 0.744 for the "leave-one-out" internal validation and external validation, respectively. It would be interesting to compare the predictability of the present 2D-QSAR approach with that of the 3D-QSAR approach. Because Catalyst, a representative 3D-QSAR program, was not available, we analyzed the data set of Ekins *et al.* (10) using our approach and compared the results. For the K_m values of 38 substrates against human liver microsomal CYP 3A4, the pharmacophore model built by Catalyst had an *r* value of 0.67 (n = 38), and the prediction for 12 testing data was within 1 log unit for the residual (s = 0.55, n = 12). On the other hand, our 2D-QSAR analysis gave an *r* value of 0.87 for the five PLS principal components of 14 descriptors, and the prediction for the same testing data was also within 1 log unit (s = 0.57, n = 12). Thus, 2D-QSAR models appear to be comparable with 3D-QSAR models in terms of predictability.

It is difficult to understand structural features involving CYP 3A4 interaction directly from Eq. (2) because correlations between the descriptors were relatively high. Because the PLS score, which is a linear combination of the explanatory variables, is used for the regression with the response variable, relationships of the PLS score in each principal component with physicochemical parameters such as molecular weight, hydrophobicity, and hydrogen-bonding ability were investigated. As a result, the first principal component of the 20 descriptors selected correlated well with the molecular weight, having an *r* value of –0.895 (Fig. 5). Taken together with the finding that the first PLS principal component positively correlated with the IC_{50} value (data not shown), this showed that the IC_{50} value of compounds in the CYP 3A4 interaction is smaller for larger compounds. In a CYP 3A4 homology model, erythromycin contacts more active site residues than progesterone because of its larger molecular size, where both compounds are stabilized mainly through hydrophobic interactions (9). It has been reported that the hydrophobicity of compounds plays an important role in oxidation by CYPs (28) and binding to liver microsomes (29). Because hydrophobicity is highly dependent on molecular volume (30), the molecular size would be an important factor determining any drug–CYP 3A4 interaction. However, it should be noted that the contribution ratio of the first principal component was 25.5%, which was not as high as the direct correlation between log IC_{50} and molecular weight, was detectable. In addition to the first principal component vs. molecular weight relationship, unfortunately, no other significant relationships were observed.

Riley *et al.* (31) measured inhibition of CYP 3A4 activity, i.e., N-demethylation of erythromycin, by 30 compounds and found that their IC_{50} values inversely correlated with lipo-

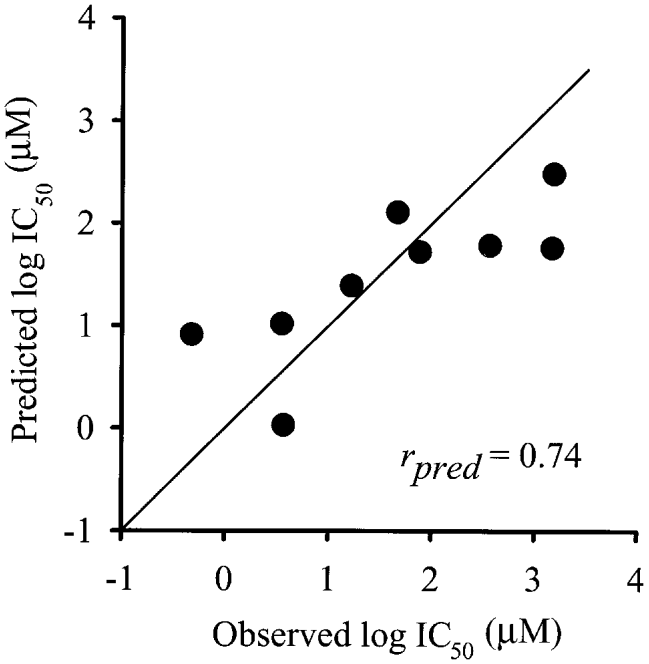

Fig. 4. External validation of the QSAR model obtained. IC_{50} values for nine compounds were predicted by using Eq. (2).

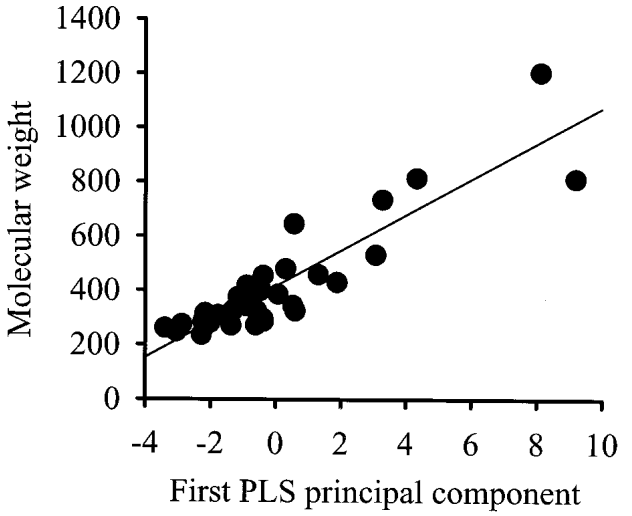

Fig. 5. Relationship between the molecular weight and the first PLS principal component of the 20 descriptors selected. The score values corresponding to the first PLS principal component were calculated based on a loading vector (b_1): b_1 = (0.335, 0.177, 0.326, 0.327, 0.338, 0.340, 0.351, 0.327, 0.313, –0.284, 0.159, –0.0168, 0.150, 0.310, 0.0940, –0.0521, 0.112, 0.127, 0.227, 0.0416). The elements in the vector are coefficients of the molecular descriptors (Table II).

philicity (log $D_{7.4}$), especially when N-containing heterocyclic compounds and others are analyzed individually. We analyzed our data in the same manner for 31 compounds, of which log $D_{7.4}$ values were available from the literature (31–34). Weak correlations were observed, with the r values of 0.67 and 0.65 for N-containing heterocyclic compounds (n = 13) and others (n = 18), respectively. However, these r values were not so high as those from our model (r = 0.88, n = 35).

An advantage of QSAR models based on two dimensional topological descriptors is that they eliminate the conformational and alignment ambiguities inherent within a 3D-QSAR process. Additionally, the two-dimensional topological descriptors are less computationally intensive, practically completely automated, and have been used to produce highly predictive models that are comparable to, or better than, those obtained using 3D-QSAR approaches (26). The most limiting feature of any 2D-QSAR approach is its insensitivity to the stereochemistry of the members of the training and prediction data sets and the lack of easily interpretable information useful for the design of new highly active drugs. In contrast, three-dimensional approaches provide graphic representations of pharmacophores (35) or putative receptor sites (36) and indicate the best directions for rational design. In view of this and the fact that two-dimensional approaches would be very helpful in screening a large number of virtual compounds, it appears that 2D- and 3D-QSAR analyses complement each other according to the purposes of the different screening stages.

In conclusion, the proposed model, in which two-dimensional topological descriptors are used as molecular descriptors, is able to predict drug–CYP 3A4 interactions with reasonable accuracy. Genetic algorithm-based approaches would be useful in selecting a set of effective descriptors for QSAR modeling.

ACKNOWLEDGMENTS

This research was supported in part by a Grant-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

REFERENCES

1. F. P. Guengerich. Role of cytochrome P450 enzymes in drug-drug interactions. *Adv. Pharmacol.* **43**:7–35 (1997).
2. M. F. Paine, M. Khalighi, J. M. Fisher, D. D. Shen, K. L. Kunze, C. L. Marsh, J. D. Perkins, and K. E. Thummel. Characterization of interintestinal and intraintestinal variations in human CYP 3A-dependent metabolism. *J. Pharmacol. Exp. Ther.* **283**:1552–1562 (1997).
3. P. K. Honig, D. C. Wortham, K. Zamani, D. Conner, J. C. Mullin, and L. R. Cantilena. Terfenadine–ketonazole interaction. *JAMA* **269**:1513–1518 (1993).
4. D. M. Stresser, A. P. Blanchard, S. D. Turner, J. C. L. Erve, A. A. Dandeneau, V. P. Miller, and C. L. Crespi. Substrate-dependent modulation of CYP 3A4 catalytic activity: analysis of 27 test compounds with four fluorometric substrates. *Drug Metab. Dispos.* **28**:1440–1448 (2000).
5. C. M. Masimirembwa, R. Thompson, and T. B. Andersson. *In vitro* high throughput screening of compounds for favorable metabolic properties in drug discovery. *Comb. Chem. High Throughput Screen.* **4**:245–263 (2001).
6. J. Langowski and A. Long. Computer systems for the prediction of xenobiotic metabolism. *Adv. Drug Deliv. Rev.* **54**:407–415 (2002).
7. S. Ekins, M. J. de Groot, and J. P. Jones. Pharmacophore and three-dimensional quantitative structure activity relationship methods for modeling cytochrome p450 active sites. *Drug Metab. Dispos.* **29**:936–944 (2001).
8. D. F. V. Lewis and B. G Lake. Molecular modelling of CYP 1A subfamily members based on an alignment with CYP 102: Rationalization of CYP 1A substrate specificity in terms of active site amino acid residues. *Xenobiotica* **26**:723–753 (1996).
9. G. D. Szklarz and J. R. Halpert. Molecular modeling of cytochrome P450 3A4. *J. Comp. Aided. Mol. Des.* **11**:265–272 (1997).
10. S. Ekins, G. Bravi, J. H. Wikel, and S. A. Wrighton. Three-dimensional quantitative structure activity relationship analysis of cytochrome P-450 3A4 substrates. *J. Pharmacol. Exp. Ther.* **291**:424–433 (1999).
11. S. Ekins, G. Bravi, S. Binkley, J. S. Gillespie, B. J. Ring, J. H. Wikel, and S. A. Wrighton. Three- and four-dimensional quantitative structure activity relationship analyses of cytochrome P-450 3A4 inhibitors. *J. Pharmacol. Exp. Ther.* **290**:429–438 (1999).
12. D. F. Lewis, S. Modi, and M. Dickins. Structure-activity relationship for human cytochrome P450 substrates and inhibitors. *Drug Metab. Rev.* **34**:69–82 (2002).
13. K. R. Korzekwa, N. Krishnamachary, M. Shou, A. Ogai, R. A. Parise, A. E. Rettie, F. J. Gonzalez, and T. S. Tracy. Evaluation of atypical cytochrome P450 kinetics with two-substrate models: evidence that multiple substrates can simultaneously bind to cytochrome P450 active sites. *Biochemistry* **37**:4137–4147 (1998).
14. J. J. Huuskonen, A. E. P. Villa, and I. V. Tetko. Prediction of partition coefficient based on atom-type electrotopological state indices. *J. Pharm. Sci.* **88**:229–233 (1999).
15. L. H. Hall and C. T. Story. Boiling point and critical temperature of a heterogeneous data set: QSAR with atom type electrotopological state indices using artificial neural networks. *J. Chem. Inf. Comp. Sci.* **36**:1004–1014 (1996).
16. S. Wanchana, F. Yamashita, and M. Hashida. Quantitative structure/property relationship analysis on aqueous solubility using genetic algorithm-combined partial least squares method. *Pharmazie* **57**:127–129 (2002).
17. F. Yamashita, S. Wanchana, and M. Hashida. Quantitative structure/property relationship analysis of Caco-2 permeability using a genetic algorithm-based partial least squares method. *J. Pharm. Sci.* **90**:2230–2239 (2002).
18. L. B. Kier and L. H. Hall. Topological information. In L. B. Kier and L. H. Hall (eds.), *Molecular Connectivity in Structure-Activity Analysis*. Research Studies, Hertfordshire, 1986, pp. 43–68.
19. W. Bridge, A. J. Crocker, T. Cubin, and A. Robertson. Experiments on the synthesis of rotenone and its derivatives. Part XIII. *J Chem. Soc.* 1530–1535 (1937).
20. A. P. Li, D. L. Kaminski, and A. Rasmussen. Substrates of human hepatic cytochrome P450 3A4. *Toxicology* **104**:1–8 (1995).
21. L. B. Kier and L. H. Hall. Intermolecular accessibility: the meaning of molecular connectivity. *J. Chem. Inf. Comput. Sci.* **40**:792–795 (2000).
22. L. H. Hall and L. B. Kier. The E-state as the basis for molecular structure space definition and structure similarity. *J. Chem. Inf. Comput. Sci.* **40**:784–791 (2000).
23. A. Varhe, K. T. Olkkola, and P. J. Neuvonen. Oral triazolam is potentially hazardous to patients receiving systemic antimycotics ketoconazole or itraconazole. *Clin. Pharmacol. Ther.* **56**:601–607 (1994).
24. P. Geladi and B. R. Kowalski. Partial least-squares regression: a tutorial. *Anal. Chim. Acta* **185**:1–17 (1986).
25. K. Hasegawa, Y. Miyashita, and K. Funatsu. GA strategy for variable selection in QSAR studies: GA-based PLS analysis of calcium channel antagonists. *J. Chem. Inf. Comput. Sci.* **37**:306–310 (1997).
26. B. Hoffman, S. J. Cho, W. Zheng, S. Wyrick, D. E. Nichols, R. B. Mailman, and A. Tropsha. Quantitative structure activity relationship modeling of dopamine D1 antagonists using comparative molecular field analysis, genetic algorithms–partial least squares, and K nearest neighbor methods. *J. Med. Chem.* **42**:3217–3226 (1999).
27. D. E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning.* Addison-Wesley, New York, 1989.
28. C. Hansch. Quantitative relationships between lipophilic character and drug metabolism. *Drug Metab. Rev.* **1**:1–14 (1972).
29. R. P. Austin, P. Barton, S. L. Cockroft, and M. C. Wenlock. The

influence of nonspecific microsomal binding on apparent intrinsic clearance, and its prediction from physicochemical properties. *Drug Metab. Dispos.* **30**:1497–1503 (2002).

30. A. Leo, C. Hansch, and P. Y. C. Jow. Dependence of hydrophobicity of apolar molecules on their molecular volume. *J. Med. Chem.* **19**:611–615 (1976).
31. R. J. Riley, A. J. Parker, S. Trigg, and C. N. Manners. Development of a generalized, quantitative physicochemical model of CYP 3A4 inhibition for use in early drug discovery. *Pharm. Res.* **18**:652–655 (2001).
32. F. Yoshida and J. G. Topliss. QSAR model for drug human oral bioavailability. *J. Med. Chem.* **43**:2575–2585 (2000).
33. F. Lombardo, M. Y. Shalaeva, K. A. Tupper, and F. Gao. Elog-D_{oct}: a toot for lipophilicity determination in drug discovery. 2. Basic and neutral compounds. *J. Med. Chem.* **44**:2490–2497 (2001).
34. C. Zhu, L. Jiang, T. M. Chen, and K. K. Hwang. A comparative study of artificial membrane permeability assay for high throughput profiling of drug absorption potential. *Eur. J. Med. Chem.* **37**:399–407 (2002).
35. J. H. Van Drie, D. Weininger, and Y. C. Martin. ALADDIN: an integrated tool for computer-assisted molecular design and pharmacophore recognition from geometric, steric, and substructure searching of three-dimensional molecular structures. *J. Comput. Aided Mol. Des.* **3**:225–251 (1989).
36. R. D. Cramer, D. E. Patterson, and J. D. Bunce. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **110**:5959–5967 (1988).