Prediction of Cytochrome P450 3A4, 2D6, and 2C9 Inhibitors and Substrates by Using Support Vector Machines

C. W. Yap[†] and Y. Z. Chen*[,†,‡]

Department of Computational Science, National University of Singapore, Blk SOC1, Level 7, 3 Science Drive 2, Singapore 117543, and Shanghai Center for Bioinformation Technology, 100 Qinzhou Road, Shanghai 200235, P.R. China

Received February 13, 2005

Statistical learning methods have been used in developing filters for predicting inhibitors of two P450 isoenzymes, CYP3A4 and CYP2D6. This work explores the use of different statistical learning methods for predicting inhibitors of these enzymes and an additional P450 enzyme, CYP2C9, and the substrates of the three P450 isoenzymes. Two consensus support vector machine (CSVM) methods, "positive majority" (PM-CSVM) and "positive probability" (PP-CSVM), were used in this work. These methods were first tested for the prediction of inhibitors of CYP3A4 and CYP2D6 by using a significantly higher number of inhibitors and noninhibitors than that used in earlier studies. They were then applied to the prediction of inhibitors of CYP2C9 and substrates of the three enzymes. Both methods predict inhibitors of CYP3A4 and CYP2D6 at a similar level of accuracy as those of earlier studies. For classification of inhibitors of CYP2C9, the best CSVM method gives an accuracy of 88.9% for inhibitors and 96.3% for noninhibitors. The accuracies for classification of substrates and nonsubstrates of CYP3A4, CYP2D6, and CYP2C9 are 98.2 and 90.9%, 96.6 and 94.4%, and 85.7 and 98.8%, respectively. Both CSVM methods are potentially useful as filters for predicting inhibitors and substrates of P450 isoenzymes. These methods generally give better accuracies than single SVM classification systems, and the performance of the PP-CSVM method is slightly better than that of the PM-CSVM method.

INTRODUCTION

Drug metabolism is a process whereby a drug is modified by a metabolizing enzyme, and these processes play important roles in pharmacokinetics and therapeutic actions of drugs.[1] For instance, lipophilic drugs need to be metabolized to hydrophilic metabolites so that they can be readily excreted.[2] Although the primary site of drug metabolism is in the liver, metabolism can also occur in the intestines, blood, and other tissues.

Profiles of drug metabolism have increasingly become an important consideration in early stages of drug development because of the profound effect of metabolism on such important drug properties as metabolic stability, drug−drug interactions, and drug toxicity.[1,3] Lower metabolic stability of a drug generally reduces its efficacy as it becomes more difficult to reach an adequate therapeutic concentration at a target site, whereas higher metabolic stability of a drug may lead to harmful effects because of the prolonged half-life.[4] A significant portion of adverse drug reactions have been attributed to drug−drug interactions that involve the interference of the normal metabolism of a drug as a result of the inhibition or induction of its metabolic enzyme by another drug.[5,6] Drug metabolism is also known to produce metabolites more toxic than their parent compound.[7]

There are mainly two phases in drug metabolism processes. The first involves phase I enzymes responsible for drug oxidation, reduction, or hydrolysis. The second involves phase II enzymes responsible for drug conjugation of the phase I metabolite with a water-solubilizing endogenous moiety.[8] The cytochrome P450 isoenzymes are responsible for most of the phase I metabolism processes,[2,9] with CYP3A4, CYP2D6, and CYP2C9 mediating the metabolism of nearly 70% of all phase I metabolism.[10] CYP3A4 is responsible for the metabolism of over 50% of drugs,[2,11,12] and its ability to metabolize a wide variety of drugs of varying molecular weights and physicochemical properties is attributed to its relatively large active site that facilitates weak hydrophobic interactions with its substrates.[2,8,11] CYP2D6 is a polymorphic enzyme primarily responsible for the metabolism of substrates containing a basic nitrogen,[13] which includes antiarrhythmics, antidepressants, and beta blockers.[14] Its metabolism activity is, in many cases, facilitated by an ion pair interaction between an aspartic acid residue at the active site and a protonated nitrogen atom of the substrate.[13] CYP2C9 is primarily involved in the metabolism of many polar drugs that are ionized at physiological pH, such as ibuprofen, naproxen, diclofenac, and sulphaphenazole.[11,15] Most of the substrates of CYP2C9 contain an aromatic group, and drug−enzyme interaction has been attributed to the $\pi-\pi$ interactions between the aromatic groups of the substrate and the specific residue at the binding site[13] and to hydrogen bonding.[9] Therefore, the prediction of inhibitors, substrates, and inducers of these P450 isoenzymes is important for the analysis of drug metabolism and for developing efficient tools for screening drugs of appropriate metabolism profiles.

Several computer prediction systems have been developed by using statistical learning methods for the identification of inhibitors of specific P450 isoenzymes. Zuegge et al.[12] developed a filter for predicting CYP3A4 inhibition by using

* Corresponding author tel.: 65-6874-6877; fax: 65-6774-6756; e-mail: yzchen@cz3.nus.edu.sg.
[†] National University of Singapore.
[‡] Shanghai Center for Bioinformation Technology.

10.1021/ci0500536 CCC: $30.25
Published on Web 06/04/2005

a linear partial least-squares-based approach, which gives an accuracy of 93% for 29 inhibitors and 86% for 21 noninhibitors. Another filter for the prediction of CYP3A4 inhibition was developed by Molnar and Keseru[6] by using neural networks, which gives an accuracy of 91.7% for 36 inhibitors and 88.9% for 36 noninhibitors. A consensus filter for predicting CYP2D6 inhibitors was developed by Susnow and Dixon[14] using recursive partitioning, which gives an accuracy of 100% for 10 inhibitors and 76% for 41 noninhibitors. Ekin et al.[16] also used recursive partitioning to develop filters for predicting CYP3A4 and CYP2D6 inhibitors, which gives Spearman's ρ values of 0.48 and 0.61 for a test set of 98 compounds, respectively. The success of these methods raises an interest in the exploration of other statistical learning methods that have been used in a variety of drug studies.[17−22]

The aim of this work is to explore the use of support vector machine (SVM) methods for facilitating the prediction of substrates and nonsubstrates and inhibitors and noninhibitors of P450 isoenzymes. SVM has been successfully used in a wide range of problems including p-glycoprotein substrates,[21] blood−brain barrier penetration,[17,18] human intestinal absorption,[20] torsade de pointes prediction,[22] and protein function prediction.[19] The main advantage of SVM over other statistical learning methods is its relatively low sensitivity to data overfitting, even with the use of a large number of redundant and overlapping molecular descriptors. This is because SVM is based on the structural risk minimization principle. However, as with other statistical learning methods, SVM requires a sufficient number of samples to develop a classification system, and irrelevant molecular descriptors may reduce the prediction accuracies of the SVM classification systems. Thus, in this work, a larger number of inhibitors and noninhibitors of P450 isoenzymes were used to train the SVM classification systems than in previous studies. For the same reason, a larger number of substrates and nonsubstrates were used to train the respective SVM systems.

A genetic-algorithm-based descriptor selection method[23,24] is used to select relevant molecular descriptors for SVM classification of the substrates and inhibitors of P450 isoenzymes. Because of the high number of redundant and overlapping descriptors, many sets of descriptors, which describe similar overall physicochemical properties but are derived from slightly different algorithms and parameters, can be selected by this genetic algorithm with a different random seed. The consensus modeling strategy has been introduced for developing prediction systems based on multiple descriptor sets.[25] In this work, this strategy was applied to the development of consensus SVM (CSVM) classification systems for the prediction of inhibitors and substrates of P450 isoenzymes by using multiple descriptor sets generated from genetic algorithms of different seeds.

Our method was first applied to the prediction of the inhibitors of CYP3A4 and CYP2D6 by using a substantially higher number of inhibitors and noninhibitors than in earlier studies,[6,12,14] which serves as a test of the capability of our method. It was then used for the prediction of the inhibitors of CYP2C9 and the substrates of CYP3A4, CYP2D6, and CYP2C9. The relevance of the selected descriptors by the CSVM methods to drug interactions with P450 isoenzymes is discussed.

METHODS

Datasets. Inhibitors and substrates of CYP3A4, CYP2D6, and CYP2C9 P450 isoenzymes were collected from various sources.[26−29] To ensure that interlaboratory variations in experimental protocols do not significantly affect the quality of the data sets, the most common range of Ki values for the compounds investigated in more than one source was used to select compounds as inhibitors or substrates.[14] The generated datasets are composed of 241 inhibitors and 368 substrates for CYP3A4, 180 inhibitors and 198 substrates for CYP2D6, and 167 inhibitors and 144 substrates for CYP2C9. Noninhibitors and nonsubstrates are seldom described in the literature, and few of these compounds are specified in a known chemical database. For instance, a comprehensive search of the literature sources[26−29] identified only seven noninhibitors and six nonsubstrates for CYP3A4, nine noninhibitors and eight nonsubstrates for CYP2D6, and eight noninhibitors and seven nonsubstrates for CYP2C9. In an earlier study of the prediction of CYP3A4 inhibitors,[6] noninhibitors of the enzyme were selected from those well-studied agents that are known inhibitors/substrates/agonists of proteins other than that enzyme, and there is no report that any of these is an inhibitor of that enzyme.[26−29] Such a method is based on the assumption that, as they have been well studied, if these compounds have not been reported to be inhibitors or substrates of a specific enzyme, it is highly likely that they are not. In this work, this method was used to generate noninhibitors or nonsubstrates of the P450 isoenzymes. From this procedure, 461 noninhibitors and 334 nonsubstrates for CYP3A4, 522 noninhibitors and 504 nonsubstrates for CYP2D6, and 535 noninhibitors and 558 nonsubstrates for CYP2C9 were generated. Substrates and inhibitors of an isoenzyme were denoted as belonging to the positive class ($P+$) of the isoenzyme, and nonsubstrates and noninhibitors of the isoenzyme were denoted as belonging to the negative class ($P-$) of the isoenzyme.

Representative training and validation sets were constructed from the datasets according to their distribution in the chemical space. Here, chemical space is defined by the 1607 structural and chemical descriptors used to represent a compound. Each compound occupies a particular location in this chemical space. All possible pairs of these compounds were generated, and a similarity score was computed for each pair. These pairs were then ranked in terms of their similarity scores, based on which compounds of similar structural and chemical features were evenly assigned into separate datasets. For those compounds without enough structurally and chemically similar counterparts, they were assigned to the training set. The training and validation sets for the inhibitors or substrates of each of these enzymes are given in Table 1. The list of compounds in these six datasets and their allocation into the training and validation sets is provided in the Supporting Information.

Prediction accuracy of statistical learning systems is known to be strongly affected by the diversity of samples used in the training set.[30,31] Independent validation sets have frequently been used for evaluating the predictive performance of these classification systems, and these need also to be diverse and representative of the samples studied in order to accurately assess the capabilities of the prediction systems.[30,31] The diversity of these datasets can be determined

Table 1. Number of Compounds in the Training, Independent Validation, Modeling Training, and Modeling Testing Sets for the Inhibitors/Substrates of Different Cytochrome P450 Isoenzymes

		training set		validation set		modeling training set		modeling testing set	
dataset	CYP	$P+$[a]	P−[b]	$P+$[a]	$P-$[b]	$P+$[a]	$P-$[b]	$P+$[a]	$P-$[b]
inhibitors/ noninhibitors	3A4	216	386	25	75	196	306	20	80
	2D6	160	442	20	80	143	359	17	83
	2C9	149	453	18	82	134	368	15	85
substrates/ nonsubstrates	3A4	312	290	56	44	256	246	56	44
	2D6	169	433	29	71	149	353	20	80
	2C9	130	472	14	86	121	381	9	91

[a] Inhibitors or substrates. [b] Noninhibitors or nonsubstrates.

by calculating the diversity index (DI), which is the average value of the similarity between all the pairs of compounds in a dataset:[32]

$$\mathrm{DI} = \frac{\sum_{i=1}^{N}\sum_{j=1,i\neq j}^{N} \mathrm{sim}(i,j)}{N(N-1)} \quad (1)$$

where sim(i,j) is a measure of the similarity between compounds i and j and N is the number of compounds in the dataset. The diversity of a dataset increases with decreasing DI. The similarity between two compounds i and j is commonly described by the Tanimoto coefficient:[6,33,34]

$$\mathrm{sim}(i,j) = \frac{\sum_{d=1}^{l} x_{di}x_{dj}}{\sum_{d=1}^{l}(x_{di})^2 + \sum_{d=1}^{l}(x_{dj})^2 - \sum_{d=1}^{l} x_{di}x_{dj}} \quad (2)$$

where l is the number of descriptors of the compounds in the dataset. The mean Tanimoto coefficient of the compounds in datasets A and B can be used as a representativeness index (RI) to measure the level of representativeness of dataset A by dataset B. Dataset B is more representative of dataset A if the RI value between datasets A and B is higher. The DIs of the six training sets and the six validation sets are in the ranges between 0.001 and 0.005 and between 0.002 and 0.020, respectively. The low DI values of the $P+$ and $P-$ compounds for all of the training and validation sets suggest that these datasets are sufficiently diverse. The RI value between each of the training sets and its corresponding validation set is in the range between 0.446 and 0.511, which suggests that these validation sets are representative of their corresponding training sets and these validation sets are suitable for assessing the systems developed in this work.

Molecular Structures and Descriptors. The 2D structures of each of the compounds studied were generated by using DS ViewerPro 5.0[35] and were subsequently converted into 3D structure by using CONCORD.[36] The 3D structure of each compound was manually inspected to ensure that the chirality of each chiral agent was properly represented. By using DRAGON Web version 3.0,[37] we derived a total of 1497 1D, 2D, and 3D molecular descriptors from the 3D structure of each compound. These descriptors can be divided into 18 classes including 47 constitutional descriptors, 70 geometrical descriptors, 266 topological descriptors, 150 RDF descriptors,[38] 21 molecular walk counts,[39] 160 3D-MoRSE descriptors,[40] 64 BCUT descriptors,[41] 99 WHIM descriptors,[42] 21 Galvez topological charge indices,[43] 197 GETAWAY descriptors,[44] 96 2D autocorrelations, 121 functional groups, 14 charge descriptors, 120 atom-centered descriptors, 4 aromaticity indices,[45] 3 empirical descriptors, 41 Randic molecular profiles,[46] and 3 molecular properties. Moreover, an additional set of 105 electrotopological state descriptors[47] and 5 linear solvation energy relationship descriptors[48] were computed by using our own developed code. Our code has been tested on a number of compounds used in earlier studies to ensure the accuracy of the computed descriptors.

Descriptor Selection. A genetic algorithm (GA)[49] was used to remove descriptors irrelevant to the prediction of CYP450 inhibitors and substrates. The retained descriptors from this process were used for representing the compounds studied in this work. All of the descriptors in the training set were first normalized in the range of −1 to +1 by using the following formula to ensure that none of them had a biased influence on a classification system by virtue of its absolute value:[50]

$$X'_{ij} = \frac{2(X_{ij} - X_{j,\min})}{X_{j,\max} - X_{j,\min}} - 1 \quad (3)$$

where X'_{ij} is the scaled value for descriptor j of compound i and $X_{j,\min}$ and $X_{j,\max}$ are the minimum and maximum values of descriptor j, respectively. An initial population of 50 randomly selected descriptor subsets was generated and screened for 100 generations. In each generation, the descriptor subsets were first ranked by their fitness value. The higher ranked descriptor subsets were given a higher probability of being chosen for reproduction. The top 40 selected descriptor subsets were then used to replace the 40 lowest ranking descriptor subsets in the population. These 40 new descriptor subsets, together with the 10 highest ranked descriptor subsets in the current generation, form a new generation of descriptor subsets. The 40 new descriptor subsets were subsequently subjected to a one-point crossover and mutation to increase the diversity of the population. In the mutation process, descriptors might be randomly added to or deleted from a descriptor subset. At the end of 100 generations, the highest ranked descriptor subset was used to construct the final SVM classification system.

In the descriptor selection process, ranking of the different descriptor subsets can be determined by using either 10-fold cross-validation, 5-fold cross-validation, or a modeling testing set. Our analysis of the 30 P450 isoenzyme SVM classification systems derived from each of these cross validation

methods showed that the modeling testing method gives the best performance, and thus, this validation method was used in all of the descriptor selection processes in this study. The modeling testing set was derived by dividing the original training set into a modeling training set and modeling testing set of 502 and 100 compounds, respectively, using the same procedure as that for dividing a dataset into the training and validation sets described in the previous section. The modeling training and modeling testing sets for the inhibitors or substrates of each of these enzymes are given in Table 1. The modeling training set was used for constructing the SVM classification systems in the genetic algorithm. The Matthews correlation coefficient (C)[51] was used as the fitness function for genetic algorithm optimization:

$$C = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FN} \times \mathrm{FP}}{\sqrt{(\mathrm{TP} + \mathrm{FN})(\mathrm{TP} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})}} \quad (4)$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

SVM Algorithm. The theory of SVM has been extensively described.[52−54] Thus, only a brief description is given here. SVM is based on the structural risk minimization principle from statistical learning theory.[52] SVM constructs a hyperplane, which separates the two classes of vectors with a maximum margin. Each instance is represented by a vector $\mathbf{x}_i$, which is its molecular descriptors. The hyperplane can be represented by

$$f(\mathbf{x}) = \mathrm{sign}[\sum_{i=1}^{l} \alpha_i^0 y_i K(\mathbf{x},\mathbf{x}_i) + b] \quad (5)$$

where y_i is the class index, $K(\mathbf{x}_i, \mathbf{x}_j)$ is a kernel function that maps the vectors into a high dimensional feature space, and the coefficients α_i^0 and b are determined by maximizing the following Langrangian expression:

$$\sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i,\mathbf{x}_j) \quad (6)$$

under the following conditions:

$$0 \le a_i \le C \quad (7)$$

where C is a penalty for training errors. A positive or

$$\sum_{i=1}^{l} \alpha_i y_i = 0 \quad (8)$$

negative value from eq 5 indicates that the vector $\mathbf{x}$ belongs to the positive or negative class, respectively.

CSVM Methods. Two types of CSVM methods were used. The first is a "positive majority" CSVM classification system (PM-CSVM), which classifies a compound as $P+$ if the majority of its SVM classification systems classify the compound as $P+$.[55] A PM-CSVM requires an odd number of SVM classification systems to prevent ambiguity in its prediction. The second is a "positive probability" CSVM classification system (PP-CSVM), which explicitly computes the probability for a compound to be $P+$ using the following formulas:[56]

$$\Pr(S_i^+|P_i) = \frac{\Pr(S_{i-1}^+|P_{i-1})\alpha_i^+}{(1 - \alpha_i^-) + (\alpha_i^+ + \alpha_i^- - 1) \times \Pr(S_{i-1}^+|P_{i-1})} \quad (9)$$

where $\Pr(S_i^+|P_i)$ is the posterior probability that a compound is $P+$ given the classification result from SVM classification system i and where α_i^+ and α_i^- are the sensitivity and specificity of SVM classification system i, respectively. Sensitivity and specificity represent the prediction accuracies of $P+$ and $P-$, respectively. Equation 9 or 10 was used when SVM classification system i classified the compound as $P+$ or $P-$, respectively. In the absence of the knowledge about the ratio of $P+$ to $P-$ compounds in the population, the prior probability of a compound to be $P+$ was tentatively set at 0.5. Sensitivity and specificity of SVM classification system i were estimated by using the validation method of the descriptor selection process.

$$\Pr(S_i^+|P_i) = \frac{\Pr(S_{i-1}^+|P_{i-1})(1 - \alpha_i^+)}{\alpha_i^- - (\alpha_i^+ + \alpha_i^- - 1) \times \Pr(S_{i-1}^+|P_{i-1})} \quad (10)$$

To determine an appropriate number of SVM classification systems for the CSVM methods, the descriptor selection process was repeated 101 times, producing a pool of SVM classification systems. SVM classification systems were randomly selected, with replacement, from the pool of SVM classification systems to form nine classes of CSVMs, each containing 11, 21, 31, 41, 51, 61, 71, 81, or 91 SVM classification systems. This random selection of SVM classification systems from the pool of SVM classification systems and construction of CSVMs were repeated 1000 times. Our analysis of these nine CSVM classes showed that the best accuracies for the two types of CSVM methods were obtained when at least 81 SVM classification systems were used to develop CSVMs, and the accuracies roughly level off at higher numbers of SVM classification systems. Thus, 81 SVM classification systems appear to be the optimum number of systems for constructing CSVMs, which are used for developing CSVMs for all of the datasets in this work.

RESULTS

The SVM classification system with the best cross-validation accuracies was selected from the 81 SVM classification systems as the "best-trained" single SVM classification system. This selection method has been used by several studies that used GA as the descriptor selection method.[57,58] The prediction accuracies of this system were determined by using the independent validation set described in the Methods section. A PM-CSVM and a PP-CSVM were constructed by using the 81 SVM classification systems. The prediction accuracies of these three systems were determined by using the independent validation set and are given in Table 2. It is found that both CSVM methods give better accuracies than the "best-trained" single SVM classification system. Moreover, PP-CSVM gives similar sensitivities and slightly better specificities and PM-CSVM gives slightly lower sensitivities and slightly better specificities than those of earlier classification systems for the prediction of inhibitors CYP3A4[6,12] and CYP2D6.[14] Thus, PP-CSVM appears to be more useful than PM-CSVM for predicting inhibitors and substrates of P450 isoenzymes.

Table 2. Accuracies of the "Best-Trained" Single SVM Classification Systems, PM-CSVM, and PP-CSVM for the Prediction of CYP3A4 and CYP2D6 Inhibitors/Noninhibitors by Using the Independent Validation Sets

CYP	classification system	TP	FN	TN	FP	sensitivity (%)	specificity (%)	concordance (%)	Matthews correlation coefficient
3A4	"best-trained" single SVM classification system	20	5	72	3	80.0	96.0	92.0	0.782
	PM-CSVM	21	4	75	0	84.0	100.0	96.0	0.893
	PP-CSVM	23	2	73	2	92.0	97.3	96.0	0.893
2D6	"best-trained" single SVM classification system	15	5	77	3	75.0	96.3	92.0	0.742
	PM-CSVM	16	4	78	2	80.0	97.5	94.0	0.807
	PP-CSVM	18	2	76	4	90.0	95.0	94.0	0.821

Table 3. Accuracies of PP-CSVM for the Prediction of CYP2C9 Inhibitors/Noninhibitors and CYP3A4, CYP2D6, and CYP2C9 Substrates/Nonsubstrates by Using the Independent Validation Sets

dataset	CYP	TP	FN	TN	FP	sensitivity (%)	specificity (%)	concordance (%)	Matthews correlation coefficient
inhibitors/noninhibitors	2C9	16	2	79	3	88.9	96.3	95.0	0.835
substrates/nonsubstrates	3A4	55	1	40	4	98.2	90.9	95.0	0.899
	2D6	28	1	67	4	96.6	94.4	95.0	0.884
	2C9	12	2	85	1	85.7	98.8	97.0	0.872

The accuracies of PP-CSVM for the prediction of inhibitors of CYP2C9 and substrates of CYP3A4, CYP2D6, and CYP2C9 are given in Table 3. The prediction accuracies of these CSVMs are at a similar level as those of the inhibitors of CYP3A4 and CYP2D6, which suggests that these CSVM methods, particularly PP-CSVM, are generally useful for predicting both the inhibitors and the substrates of different P450 isoenzymes.

DISCUSSION

Overall Prediction Accuracies. The difference between the specificities of the current CSVMs and those of classification systems from earlier studies may be due to the difference in the number of $P-$ compounds used for training the classification systems. In our work, the number of $P-$ compounds in the training set ranges from 290 to 472, whereas earlier classification systems were developed by using 41−145 $P-$ compounds. Statistical learning methods require a large number of compounds for the development of classification systems. Therefore, it is not surprising that the methods of the current work, which uses a larger number of $P-$ compounds, give higher specificities than those of earlier studies. Another possible reason for the improved specificities is the use of SVM, which has been found to be consistently superior to other classification methods in most classification problems.[59−61]

For all of the datasets, with the exception of the CYP3A4 substrates/nonsubstrates dataset, the number of $P-$ compounds is always higher than the number of $P+$ compounds. This may create a bias of the SVM classification systems to predict unknown compounds as $P-$, resulting in a higher number of false negatives. However, previous studies suggest that SVMs are not significantly affected by unbalanced datasets,[19,62] especially if there are more than 80−100 compounds of each class in the training set.[63] All of the datasets used in this work contain at least 130 compounds of each class in the training set, and thus, the unbalanced dataset is not expected to significantly affect the predictive ability of the SVM classification systems.

Evaluation of Prediction Performance. The results of our SVM systems were compared with those of several statistical learning methods including multiple linear regression, partial least squares, logistic regression, C4.5 decision tree, and *k*-nearest neighbors. GA was used to determine the optimum descriptor subsets for each of these classification methods by using 30 different random seeds, from which 30 separate classification models were generated for each method. The prediction accuracies of these classification models were determined by using the independent validation set. Table 4 gives the results for CYP3A4 substrates/nonsubstrates. The accuracies for the other P450 isoenzymes datasets are similar and, thus, are not given here. It is found that the SVM classification systems give the highest prediction accuracies when compared with those of other methods.

To determine whether the selected descriptors of the SVM classification systems include those irrelevant for the prediction of the inhibitors or substrates of the respective enzymes, 10 groups of classification systems were generated by using the GA-based descriptor selection method. These groups are SVM_{100}, SVM_{200}, SVM_{300}, SVM_{400}, SVM_{500}, SVM_{600}, SVM_{700}, SVM_{800}, SVM_{900}, and SVM_{1000}, in which the subscript denotes the number of descriptors used. Each group contains 30 SVM classification systems. The prediction accuracies of these SVM classification systems were determined by using the independent validation sets. Table 5 gives the results for the CYP3A4 substrates/nonsubstrates, which shows that prediction accuracies begin to decrease when more than 400 descriptors are used in a SVM classification system. This suggests that the maximum number of relevant descriptors for the CYP3A4 substrates/nonsubstrates dataset

Table 4. Average Accuracies of Different Statistical Learning Classification Systems for the Prediction of CYP3A4 Substrates/Nonsubstrates by Using Independent Validation Sets

classification method	sensitivity (%)[a]	specificity (%)[a]	concordance (%)[a]	Matthews correlation coefficient[a]
multiple linear regression	86.1 (3.9)	71.4 (4.4)	79.6 (2.9)	0.586 (0.060)
logistic regression	83.8 (3.9)	71.0 (5.1)	78.1 (3.0)	0.555 (0.063)
partial least squares	79.9 (5.8)	72.5 (5.2)	76.7 (3.7)	0.528 (0.073)
C4.5 decision tree	75.5 (6.8)	66.4 (6.7)	71.5 (4.3)	0.423 (0.087)
k-nearest neighbor	92.4 (2.0)	82.6 (3.4)	88.1 (1.7)	0.759 (0.034)
SVM	98.0 (1.4)	85.3 (3.1)	92.4 (1.2)	0.849 (0.024)

[a] Numbers in parentheses are the standard deviations.

Table 5. Average Accuracies of 10 Groups of SVM Classification Systems for the Prediction of CYP3A4 Substrates/Nonsubstrates by Using Independent Validation Sets

number of descriptors	sensitivity (%)[a]	specificity (%)[a]	concordance (%)[a]	Matthews correlation coefficient
100	93.0 (3.1)	80.4 (4.4)	87.5 (2.7)	0.747 (0.054)
200	96.7 (2.0)	83.0 (3.3)	90.7 (1.9)	0.814 (0.039)
300	98.0 (1.6)	85.6 (3.6)	92.6 (1.9)	0.853 (0.037)
400	98.0 (1.3)	82.4 (3.4)	91.1 (1.6)	0.825 (0.032)
500	98.2 (1.0)	80.9 (3.1)	90.6 (1.4)	0.815 (0.028)
600	98.6 (0.8)	74.5 (3.3)	88.0 (1.5)	0.769 (0.028)
700	99.3 (0.9)	66.4 (5.4)	84.8 (2.3)	0.715 (0.040)
800	100.0 (0.0)	51.5 (3.1)	78.7 (1.4)	0.611 (0.024)
900	99.9 (0.3)	45.7 (2.4)	76.1 (1.0)	0.565 (0.017)
1000	100.0 (0.0)	37.3 (3.2)	72.4 (1.4)	0.500 (0.026)

[a] Numbers in parentheses are the standard deviations.

Table 6. Comparison of the Average Accuracies of SVM Classification Systems for the Prediction of Inhibitors/Substrates of Different P450 Isoenzymes by Using Modeling Testing Sets and Independent Validation Sets

		modeling testing set[a]				independent validation set[a]			
dataset	CYP	sensitivity (%)	specificity (%)	concordance (%)	Matthews correlation coefficient	sensitivity (%)	specificity (%)	concordance (%)	Matthews correlation coefficient
inhibitors/ noninhibitors	3A4	76.5 (6.2)	98.8 (1.3)	94.3 (0.8)	0.817 (0.026)	82.1 (4.5)	97.9 (1.5)	93.9 (1.3)	0.835 (0.036)
	2D6	79.1 (7.3)	98.5 (1.4)	95.2 (0.8)	0.828 (0.028)	79.3 (5.4)	96.7 (1.6)	93.2 (1.7)	0.783 (0.054)
	2C9	81.9 (4.7)	98.8 (1.0)	96.3 (0.6)	0.851 (0.025)	86.4 (5.0)	97.3 (1.3)	95.3 (1.1)	0.842 (0.039)
substrates/ nonsubstrates	3A4	96.3 (1.5)	86.7 (2.7)	92.1 (0.8)	0.841 (0.015)	98.0 (1.3)	85.2 (3.0)	92.4 (1.3)	0.849 (0.026)
	2D6	84.6 (5.0)	98.9 (1.3)	96.0 (0.6)	0.874 (0.018)	86.9 (4.7)	96.9 (1.5)	94.0 (1.7)	0.852 (0.043)
	2C9	77.0 (8.2)	98.9 (1.0)	97.0 (0.8)	0.810 (0.047)	72.3 (7.9)	99.2 (0.9)	95.4 (1.1)	0.801 (0.051)

[a] Values in parentheses are the standard deviations.

is around 400. Because the original 81 SVM classification systems for the CYP3A4 substrates/nonsubstrates dataset contains 214−402 descriptors, our results seem to suggest that the original 81 SVM classification systems are unlikely to contain irrelevant descriptors. Similar conclusions are also made for the rest of the P450 isoenzymes datasets on the basis of our computational studies.

It has been shown that chance correlations may occur during descriptor selection, especially if the number of descriptors available for selection is large.[64,65] y randomization has been frequently used to determine the probability of chance correlation during descriptor selection processes.[66,67] In y randomization, a portion of $P+$ compounds in the training set was randomly selected and converted to $P-$ compounds. Another portion of $P-$ compounds was also randomly selected and converted to $P+$ compounds. The ratios of $P+$ to $P-$ compounds were kept unchanged during y randomization. The scrambled training set was then used for the descriptor selection process. The process of scrambling the training set and descriptor selection process was repeated 81 times. The average Matthews correlation coefficient of these scrambled SVM classification systems derived by using the independent validation sets was found to be in the range between 0.189 and 0.288, which is significantly lower than those of the original SVM classification systems, which are in the range between 0.783 and 0.852. This suggests that the original SVM classification systems are relevant and unlikely to arise as a result of chance correlation.

A frequently used method for checking whether a prediction system is overfitted is to compare the prediction accuracies determined by using cross-validation methods with those determined by using independent validation sets.[68] Because descriptor selection was performed by using the modeling testing sets as the cross-validation method, an overfitted classification system is expected to have a much higher prediction accuracy for the modeling testing sets than for the independent validation sets. As shown in Table 6, the prediction accuracies of the SVM systems based on the modeling testing sets and those based on independent validation sets are similar. This suggests that the SVM classification systems in this work are unlikely to overfit.

Table 7. Important Descriptor Classes Selected for the Prediction of Inhibitors/Substrates of Different P450 Isoenzymes

dataset	CYP	electrostatic (%)	hydrogen bond acceptors (%)	hydrogen bond donors (%)	hydrophobic (%)	shape (%)	size (%)
inhibitors/ noninhibitors	3A4	20.4	3.6	3.3	8.8	56.8	7.1
	2D6	20.5	2.4	2.5	10.0	57.1	7.5
	2C9	20.1	2.0	2.9	8.8	59.0	7.2
substrates/ nonsubstrates	3A4	21.0	2.8	1.9	9.5	57.2	7.5
	2D6	18.9	3.1	3.5	8.5	59.7	6.3
	2C9	19.1	3.5	3.0	9.4	58.2	6.8

The Selected Descriptors. The majority of the selected descriptors in our SVM classification systems are composite descriptors, which can be divided into three groups: 3D-MoRSE, RDF, and Randic molecular profiles. 3D-MoRSE descriptors, which are representations of the 3D structure of a molecule and encode features such as molecular weight, van der Waals volume, electronegativities, and polarizabilities, have been used for the classification of dopamine D1 and D2 agonists and modeling the binding of steroids to corticosteroid binding globulin.[40] RDF descriptors provide information about bond lengths, ring types, planar and nonplanar systems, atom types, and molecular weight and have been used for pharmacokinetic studies.[69] Randic molecular profiles measure interactions between atoms in a molecule and encode information on molecular shape, which is an important factor in ligand−enzyme interactions. Because shape and chemical complementarity between a ligand and an enzyme are important for ligand−enzyme binding, it is not surprising that these three classes of 3D descriptors, which provide information on hydrophobicity, electronegativities, polarizabilities, and the shape of a molecule, are frequently selected by the descriptor selection process.

Because composite descriptors encode multiple physicochemical and structural aspects of the molecule, it is difficult to extract from these descriptors information about which specific molecular characteristics are important for the inhibitors and substrates of these P450 isoenzymes. Nonetheless, it is possible to infer some information from noncomposite descriptors. As many descriptors are overlapping and some of them are redundant, it is more appropriate to group them into classes of descriptors of similar properties and discuss their contribution to the inhibitors/substrates predictions at the class level. Table 7 gives the classes of noncomposite descriptors selected by our computations. It is found that shape is the dominant factor involved in ligand−P450 isoenzyme interaction. This is not surprising because shape complementarity is important for ligand−protein interactions. In addition to the shape descriptors, electrostatic and hydrophobic interactions are found to be the dominant forces involved in ligand−P450 isoenzyme interactions. Descriptors that describe hydrogen bonding also appear to be important for the ligand−P450 isoenzyme interactions, which is consistent with the findings that hydrogen bonds are involved in the ligand−P450 isoenzyme interactions.[9]

It is also possible to roughly distinguish between *P*+ and *P*− compounds and to roughly distinguish between inhibitors and substrates from the values of six selected descriptors, S, nHAcc, nHDon, MLOGP, MW, and SPH. These descriptors are representative of the four dominant interaction forces, electrostatic, hydrogen bond acceptor, hydrogen bond donor and hydrophobicity, and size and shape of the compounds, respectively. S is the combined dipolarity/polarizability, nHAcc and nHDon are the number of acceptor and donor atoms for hydrogen bonds, respectively, MLOGP is the Moriguchi Log P,[70] MW is the molecular weight, and SPH is the spherosity. The average values of these four descriptors for *P*+ and *P*− compounds of all of the various datasets are given in Table 8. Substrates of CYP3A4 are generally larger in size, less spherical in shape, more hydrophobic, and have more hydrogen bonding sites than nonsubstrates. Inhibitors of CYP3A4 are generally less hydrophobic than substrates but are larger in size and contain more hydrogen bond donors and acceptors. Substrates of CYP2D6 are generally smaller in size, more hydrophobic than nonsubstrates, and contain one hydrogen bond donor. There are only minor differences between inhibitors and substrates of CYP2D6, which suggests that there is considerable overlap between the inhibitors

Table 8. Differences in the Values of Descriptors Important for Distinguishing between *P*+ and *P*− Compounds

			average value[a]	
dataset	CYP	descriptor	*P*+	*P*−
inhibitors/ noninhibitors	3A4	S	2.56 (1.24)	2.36 (1.12)
		nHAcc	6.47 (4.05)	4.59 (2.64)
		nHDon	2.27 (2.44)	1.23 (1.40)
		MLogP	1.83 (2.02)	1.96 (2.06)
		MW	417 (185)	313 (116)
		SPH	0.77 (0.13)	0.77 (0.13)
	2D6	S	2.17 (1.00)	2.52 (1.20)
		nHAcc	4.57 (2.70)	5.47 (3.48)
		nHDon	1.57 (1.81)	1.59 (1.92)
		MLogP	2.54 (1.76)	1.70 (2.09)
		MW	355 (125)	346 (159)
		SPH	0.78 (0.13)	0.77 (0.13)
	2C9	S	2.56 (1.21)	2.39 (1.15)
		nHAcc	5.31 (2.65)	5.21 (3.50)
		nHDon	1.49 (1.52)	1.62 (1.99)
		MLogP	1.78 (2.11)	1.96 (2.02)
		MW	351 (123)	348 (159)
		SPH	0.76 (0.13)	0.78 (0.13)
substrates/ nonsubstrates	3A4	S	2.56 (1.15)	2.29 (1.17)
		nHAcc	5.53 (3.45)	4.91 (3.14)
		nHDon	1.72 (1.99)	1.44 (1.75)
		MLogP	2.20 (1.99)	1.60 (2.06)
		MW	379 (157)	315 (137)
		SPH	0.76 (0.13)	0.78 (0.13)
	2D6	S	2.19 (1.08)	2.53 (1.18)
		nHAcc	4.10 (2.13)	5.68 (3.58)
		nHDon	1.15 (1.22)	1.76 (2.07)
		MLogP	2.51 (1.74)	1.68 (2.11)
		MW	320 (100)	360 (166)
		SPH	0.78 (0.14)	0.77 (0.13)
	2C9	S	2.52 (1.26)	2.41 (1.14)
		nHAcc	4.69 (2.52)	5.38 (3.48)
		nHDon	1.03 (1.14)	1.73 (2.01)
		MLogP	2.05 (2.04)	1.88 (2.05)
		MW	326 (112)	354 (160)
		SPH	0.75 (0.14)	0.78 (0.13)

[a] Values in parentheses are the standard deviations.

and substrates of CYP2D6. Substrates of CYP2C9 generally are more hydrophobic than inhibitors of CYP2C9 but are smaller in size and have lesser hydrogen bonding capacity.

CYP3A4 has a relatively large active site that facilitates weak hydrophobic interactions with its substrates.[2,8,11] A pharmacophoric model of the substrates suggests that there are four important features: two hydrogen bond acceptors, one hydrogen bond donor, and one hydrophobic region.[71] Some of the descriptor classes frequently selected by the SVM classification systems for the prediction of substrates and nonsubstrates of CYP3A4 are related to the hydrophobicity and hydrogen bonding ability of the molecule. Examples of descriptors in these classes include ARR, which is the aromatic ratio; aaCH and aasC, which are electrotopological descriptors for carbons in aromatic rings; nHAcc; and nHDon. The differences in the distribution of intermolecular forces between inhibitors and substrates of CYP3A4 suggest that the inhibitors have less electrostatic and hydrophobic interactions and more hydrogen bonding at the binding site than the substrates.

The pharmacophoric model for substrates of CYP2D6 consists of a basic nitrogen atom and a flat hydrophobic region.[5,13] Some of the frequently selected descriptor classes by SVM classification systems for predicting substrates and nonsubstrates of CYP2D6 match this model. Examples of descriptors in these classes include MAXDP, which is the maximal electrotopological positive variation topological descriptor and is related to the electrophilicity of the molecule; nN, which is the number of nitrogen atoms; and BLI, which is the Kier benzene-likeness index. These descriptor classes are also selected by the SVM classification systems for predicting inhibitors and noninhibitors of CYP2D6. However, differences in the distribution of intermolecular forces between inhibitors of CYP2D6 suggest that the inhibitors may have increased electrostatic and hydrophobic interactions at the active site. This is consistent with the findings from pharmacophoric studies of inhibitors of CYP2D6 which suggests that the inhibitors have an additional region in which functional groups with lone pairs enhance inhibitory potency and a region for hydrophobic groups.[5]

Descriptors encoding aromaticity, polarity, and hydrogen bond donors are frequently selected by SVM classification systems for predicting substrates and nonsubstrates of CYP2C9. These include aasC, which is the electrotopological state atom index for aromatic carbons; MAXDN, which is the maximal electrotopological negative variation topological descriptor and is related to the nucleophilicity of the molecule; and nHDon. These selected descriptors are consistent with the findings that the substrates of CYP2C9 are primarily polar compounds that contain an aromatic group and that drug−CYP2C9 interaction is mediated by both hydrogen bonding[9] and $\pi-\pi$ interactions at the binding site.[13] The differences in the distribution of intermolecular forces between inhibitors and substrates of CYP2C9 suggest that the inhibitors have fewer hydrogen bonds but increased electrostatic interactions at the active site compared to the substrates.

Potential Training Errors and Misclassified Compounds. In this work, noninhibitors and nonsubstrates were selected from those compounds without a report identifying them as an inhibitor or a substrate. There is also a certain level of overlapping between noninhibitors of different CYP subtypes, between noninhibitors and nonsubstrates of a specific CYP subtype, and between noninhibitors and substrates of a particular CYP subtype. A potential problem with this method is that a small number of true inhibitors or substrates may be selected as noninhibitors or nonsubstrates (false negatives). The extent of training errors caused by false negatives can be roughly estimated by using experimentally confirmed noninhibitors/nonsubstrates. However, there is only a limited number of experimentally confirmed noninhibitors/nonsubstrates. In the CYP3A4 substrate/nonsubstrate validation set, only irbesartan is a known nonsubstrate.[26] In the CYP2C9 inhibitor/noninhibitor validation set, only reboxetine is experimentally determined to be a noninhibitor.[26] In the CYP2D6 substrate/nonsubstrate validation set, only nilvadipine is a known nonsubstrate.[26] In the CYP2D6 inhibitor/noninhibitor validation set, only gatifloxacin is a known noninhibitor.[26] All of these compounds, except irbesartan, were correctly predicted by the CSVMs to be noninhibitors/nonsubstrates. These results, together with the reported high accuracies of the SVM classification systems for other systems,[21,72] suggest that by using soft-margin SVM,[52] the training errors caused by false negatives can be kept at a minimum.

Table 9 gives the list of compounds misclassified by more than 50% of the SVM classification systems for each dataset. A possible reason for the misclassification of some of these compounds is that some descriptor subsets may be inadequate to properly describe these compounds. Examples of these compounds are carbamazepine; chlorphenamine; cinnarizine; doxepin; methadone; olanzapine and zuclopenthixol, which contain two aromatic rings separated by an atom; and irbesartan and lorsartan, which contain a highly polar

Table 9. List of Misclassified Compounds in This Work[a]

dataset	CYP	misclassified compounds
inhibitors/ noninhibitors	3A4	pilocarpine (*P*+)
		stiripentol (*P*+)
		olanzapine (*P*+)
		cyclophosphamide (*P*+)
	2D6	lobeline (*P*+)
		propafenone (*P*+)
		reboxetine (*P*+)
		sulconazole (*P*+)
		doxepin (*P*+)
		isoconazole (*P*−)
	2C9	stiripentol (*P*−)
		sulconazole (*P*+)
		isoconazole (*P*−)
substrates/ nonsubstrates	3A4	chlorphenamine (*P*+)
		flurithromycin (*P*−)
		irbesartan (*P*−)
		oxomemazine (*P*−)
		pargyline (*P*−)
		pentazocine (*P*−)
		sulindac (*P*−)
	2D6	carbamazepine (*P*+)
		cinnarizine (*P*+)
		zuclopenthixol (*P*+)
		domperidone (*P*−)
		emedastine (*P*−)
	2C9	cinnarizine (*P*+)
		losartan (*P*+)
		methadone (*P*+)

[a] All of the compounds misclassified by more than 50% of the 81 classification systems are included.

tetrazole ring. Among the misclassified noninhibitors or nonsubstrates, only irbesartan is a known nonsubstrate.[26] Oxomemazine is a known inducer and flurithromycin is a known inhibitor of CYP3A4.[27] Thus, it may be possible that both oxomemazine and flurithromycin are actually false negatives, as more than 60% of the CYP3A4 inhibitors in the dataset are both CYP3A4 inhibitors and substrates. Similarly, doxepin, which is a known CYP2D6 substrate,[27] may also be a false negative, as nearly 50% of the CYP2D6 substrates are both CYP2D6 substrates and inhibitors.

Comparison of the Two CSVM Systems. The results from our studies show that PP-CSVM gives slightly better accuracies than PM-CSVM. This is because individual SVM classification systems in PP-CSVM are ranked according to their accuracies and SVM classification systems with better accuracies have more influence on the final classification of a compound. This is different from PM-CSVM where all individual SVM classification systems, regardless of their accuracies, contribute equally to the final classification of a compound. Thus, it is expected that PP-CSVM, by reducing the contribution from SVM classification systems with lower accuracies, gives better or at least equal accuracies to those of PM-CSVM.

There are two potential problems with PP-CSVM. The first is that the prior probability, which was tentatively set at 0.5, may not always be the most appropriate value for representing the ratio of $P+$ to $P-$ compounds in the population. This problem can be partially solved by using a large number of individual SVM classification systems to construct a CSVM so that the influence of prior probability on the final classification result is reduced. In this study, we have found that the same classification results were obtained even when the prior probability was varied from 0.05 to 0.95 when 81 SVM classification systems were used to construct the CSVM. The second problem is the difficulty in determining the true sensitivities and specificities of the individual SVM classification systems, which are required by eqs 9 and 10. In the present study, sensitivities and specificities of the SVM classification systems were estimated by using the modeling testing set and have a mean absolute difference of 2.0% and 3.4%, respectively, from those derived by using the independent validation set. If sensitivities and specificities of the individual SVM classification systems derived from the independent validation set are used in PP-CSVM, the resultant CSVMs are found to give slightly higher accuracies, suggesting a possible need for a more accurate estimate of the performance of some SVM classification systems.

CONCLUSION

Results from this work are consistent with earlier studies that suggest that consensus classification systems give better predictive performance than single classification systems. All of the PP-CSVMs for predicting inhibitors/substrates of the three P450 isoenzymes, CYP3A4, CYP2D6, and CYP2C9, show high prediction accuracies, with improved specificities compared to those of earlier studies. A potential problem of this work is that the selection criteria for noninhibitors and nonsubstrates may result in a small number of false negatives. However, the use of soft-margin SVMs in this work can help to achieve a balance between training errors and prediction accuracies. The accuracies of the SVM classification systems may also be improved by the addition of a correction factor to the SVM decision function. The present CSVMs are only suitable for distinguishing between inhibitors and noninhibitors or substrates and nonsubstrates. With the availability of more detailed experimental data, it is possible to use multiclass SVMs[73] for classification of noninhibitors, weak inhibitors, and strong inhibitors or SVM regression[74] for quantitative prediction of the Ki values of inhibitors. Our computational results suggest that PP-CSVM is better than PM-CSVM for constructing CSVMs for classifying inhibitors and substrates of various P450 isoenzymes. Thus, CSVMs, particularly PP-CSVM, are potentially useful for developing filters for the prediction of inhibitors and substrates of P450 isoenzymes.

ACKNOWLEDGMENT

This work was supported in part by grants from Shanghai Commission for Science and Technology (04DZ19850, 04QMX1450, 04DZ14005), and the "973" National Key Basic Research Program of China (2004CB720103, 2004CB715901).

Supporting Information Available: A list of the compounds in the six datasets and their allocation into the training and validation sets. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) van de Waterbeemd, H.; Gifford, E. ADMET in silico modelling: towards prediction paradise? *Nat. Rev. Drug Discovery* **2003**, *2* (3), 192−204.

(2) Smith, D. A.; Ackland, M. J.; Jones, B. C. Properties of cytochrome P450 isoenzymes and their substrates Part 1: active site characteristics. *Drug Discovery Today* **1997**, *2* (10), 406−414.

(3) Li, A. P. Screening for human ADME/Tox drug properties in drug discovery. *Drug Discovery Today* **2001**, *6* (7), 357−366.

(4) Keseru, G. M. A virtual high throughput screen for high affinity cytochrome P450cam substrates. Implications for in silico prediction of drug metabolism. *J. Comput.-Aided Mol. Des.* **2001**, *15* (7), 649−657.

(5) Ekins, S.; de Groot, M. J.; Jones, J. P. Pharmacophore and three-dimensional quantitative structure−activity relationship methods for modeling cytochrome P450 active sites. *Drug Metab. Dispos.* **2001**, *29* (7), 936−944.

(6) Molnar, L.; Keseru, G. M., A neural network based virtual screening of cytochrome P450 3A4 inhibitors. *Bioorg. Med. Chem. Lett.* **2002**, *12* (3), 419−421.

(7) Li, A. P.; Kaminski, D. L.; Rasmussen, A. Substrates of human hepatic cytochrome P450 3A4. *Toxicology* **1995**, *104* (1−3), 1−8.

(8) Long, A.; Walker, J. D. Quantitative structure−activity relationships for predicting metabolism and modeling cytochrome P450 enzyme activities. *Environ. Toxicol. Chem.* **2003**, *22* (8), 1894−1899.

(9) de Groot, M. J.; Ekins, S. Pharmacophore modeling of cytochromes P450. *Adv. Drug Delivery Rev.* **2002**, *54* (3), 367−383.

(10) Lewis, D. F.; Modi, S.; Dickins, M. Structure−activity relationship for human cytochrome P450 substrates and inhibitors. *Drug Metab. Rev.* **2002**, *34* (1−2), 69−82.

(11) Smith, D. A.; Acklanda, M. J.; Jones, B. C. Properties of cytochrome P450 isoenzymes and their substrates Part 2: properties of cytochrome P450 substrates. *Drug Discovery Today* **1997**, *2* (11), 479−486.

(12) Zuegge, J.; Fechner, U.; Roche, O.; Parrott, N. J.; Engkvist, O.; Schneider, G. A fast virtual screening filter for cytochrome P450 3A4 inhibition liability of compound libraries. *Quant. Struct.-Act. Relat.* **2002**, *21* (3), 249−256.

(13) Langowski, J.; Long, A. Computer systems for the prediction of xenobiotic metabolism. *Adv. Drug Delivery Rev.* **2002**, *54* (3), 407−415.

(14) Susnow, R. G.; Dixon, S. L. Use of robust classification techniques for the prediction of human cytochrome P450 2D6 inhibition. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (4), 1308−1315.

(15) Ekins, S.; Bravi, G.; Binkley, S.; Gillespie, J. S.; Ring, B. J.; Wikel, J. H.; Wrighton, S. A. Three- and four-dimensional-quantitative structure activity relationship (3D/4D-QSAR) analyses of CYP2C9 inhibitors. *Drug Metab. Dispos.* **2000**, *28* (8), 994−1002.
(16) Ekins, S.; Berbaum, J.; Harrison, R. K. Generation and validation of rapid computational filters for CYP2D6 and CYP3A4. *Drug Metab. Dispos.* **2003**, *31* (9), 1077−1080.
(17) Doniger, S.; Hofmann, T.; Yeh, J. Predicting CNS permeability of drug molecules: comparison of neural network and support vector machine algorithms. *J. Comput. Biol.* **2002**, *9* (6), 849−864.
(18) Trotter, M. W. B.; Buxton, B. F.; Holden, S. B. Support vector machines in combinatorial chemistry. *Meas. Control* **2001**, *34* (8), 235−239.
(19) Cai, C. Z.; Han, L. Y.; Ji, Z. L.; Chen, X.; Chen, Y. Z. SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucl. Acids Res.* **2003**, *31* (13), 3692−3697.
(20) Xue, Y.; Li, Z. R.; Yap, C. W.; Sun, L. Z.; Chen, X.; Chen, Y. Z. Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (5), 1630−1638.
(21) Xue, Y.; Yap, C. W.; Sun, L. Z.; Cao, Z. W.; Wang, J. F.; Chen, Y. Z. Prediction of p-glycoprotein substrates by support vector machine approach. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (4), 1497−1505.
(22) Yap, C. W.; Cai, C. Z.; Xue, Y.; Chen, Y. Z. Prediction of torsade-causing potential of drugs by support vector machine approach. *Toxicol. Sci.* **2004**, *79* (1), 170−177.
(23) Gao, H.; Lajiness, M. S.; Van Drie, J. Enhancement of binary QSAR analysis by a GA-based variable selection method. *J. Mol. Graphics Modell.* **2002**, *20* (4), 259−268.
(24) Frohlich, H.; Chapelle, O.; Scholkopf, B. Feature selection for support vector machines by means of genetic algorithm. In *Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence*, 2003; pp 142−148.
(25) Gramatica, P.; Pilutti, P.; Papa, E. Validated QSAR prediction of OH tropospheric degradation of VOCs: Splitting into training-test sets and consensus modeling. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (5), 1794−1802.
(26) *DRUGDEX System*; MICROMEDEX: Greenwood Village, CO, 2003.
(27) Rendic, S. Summary of information on human CYP enzymes: human P450 metabolism data. *Drug Metab. Rev.* **2002**, *34* (1−2), 83−448.
(28) Lacy, C. F. *Drug information handbook*, 10th ed.; Lexi-Comp, Inc.: Hudson, OH, 2002.
(29) Cytochrome P450 drug-interaction table. http://medicine.iupui.edu/flockhart/table.htm (November 2003).
(30) Schultz, T. W.; Netzeva, T. I.; Cronin, M. T. D. Selection of data sets for QSARs: analyses of Tetrahymena toxicity from aromatic compounds. *SAR QSAR Environ. Res.* **2003**, *14* (1), 59−81.
(31) Rajer-Kanduc, K.; Zupan, J. M., N. Separation of data on the training and test set for modelling: a case study for modelling of five colour properties of a white pigment. *Chemom. Intell. Lab. Syst.* **2003**, *65* (2), 221−229.
(32) Perez, J. J. Managing molecular diversity. *Chem. Soc. Rev.* **2005**, *34* (2), 143−152.
(33) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (6), 983−996.
(34) Potter, T.; Matter, H. Random or rational design? Evaluation of diverse compound subsets from chemical structure databases. *J. Med. Chem.* **1998**, *41* (4), 478−488.
(35) *Accelrys DS ViewerPro*, version 5.0; Accelrys: San Diego, CA.
(36) Pearlman, R. S. *CONCORD User's Manual*; Tripos Inc.: St. Louis, MO.
(37) Todeschini, R.; Consonni, V.; Mauri, A.; Pavan, M. *DRAGON Web*, version 3.0; Talete SRL: Milan, 2003.
(38) Hemmer, M. C.; Steinhauer, V.; Gasteiger, J. Deriving the 3D structure of organic molecules from their infrared spectra. *Vib. Spectrosc.* **1999**, *19* (1), 151−164.
(39) Rücker, G.; Rücker, C. Counts of all walks as atomic and molecular descriptors. *J. Chem. Inf. Comput. Sci.* **1993**, *33* (5), 683−695.
(40) Schuur, J. H.; Setzer, P.; Gasteiger, J. The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure−spectra correlations and studies of biological activity. *J. Chem. Inf. Comput. Sci.* **1996**, *36* (2), 334−344.
(41) Pearlman, R. S.; Smith, K. M. Metric validation and the receptor-relevant subspace concept. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (1), 28−35.
(42) Bravi, G.; Gancia, E.; Mascagni, P.; Pegna, M.; Todeschini, R.; Zaliani, A. MS−WHIM, new 3D theoretical descriptors derived from molecular surface properties: A comparative 3D QSAR study in a series of steroids. *J. Comput.-Aided Mol. Des.* **1997**, *11* (1), 79−92.
(43) Galvez, J.; Garcia, R.; Salabert, M. T.; Soler, R. Charge indexes. New topological descriptors. *J. Chem. Inf. Comput. Sci.* **1994**, *34* (3), 520−525.
(44) Consonni, V.; Todeschini, R.; Pavan, M. Structure/Response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (3), 682−692.
(45) Randic, M. Graph theoretical approach to local and overall aromaticity of benzenoid hydrocarbons. *Tetrahedron* **1975**, *31* (11−12), 1477−1481.
(46) Randic, M. Molecular profiles. Novel geometry-dependent molecular descriptors. *New J. Chem.***1995**, *19*, 781−791.
(47) Kier, L. B.; Hall, L. H. *Molecular structure description: The electrotopological state*; Academic Press: San Diego, CA, 1999.
(48) Platts, J. A.; Butina, D.; Abraham, M. H.; Hersey, A. Estimation of molecular free energy relation descriptors using a group contribution approach. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (5), 835−845.
(49) Lucasius, C. B.; Kateman, G. Understanding and using genetic algorithms Part 1. Concepts, properties and context. *Chemom. Intell. Lab. Syst.* **1993**, *19* (1), 1−33.
(50) Livingstone, D. J. Data pretreatment. In *Data analysis for chemists: Applications to QSAR and chemical product design*; Oxford University Press: Oxford, 1995; pp 48−64.
(51) Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **1975**, *405* (2), 442−451.
(52) Vapnik, V. N. *The nature of statistical learning theory*. Springer: New York, 1995.
(53) Burges, C. J. C. A tutorial on support vector machines for pattern recognition. *Data Min. Knowledge Discovery* **1998**, *2* (2), 127−167.
(54) Evgeniou, T.; Pontil, M. Support vector machines: theory and applications. In *Machine learning and its applications. Advanced lectures*; Paliouras, G., Karkaletsis, V., Spyropoulos, C. D., Eds.; Springer: New York, 2001; pp 249−257.
(55) Eriksson, L.; Jaworska, J.; Cronin, M.; Worth, A.; Gramatica, P.; McDowell, R. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environ. Health Perspect.* **2003**, *111* (10), 1361−1375.
(56) McDowell, R.; Jaworska, J. Bayesian analysis and inference from QSAR predictive model results. *SAR QSAR Environ. Res.* **2002**, *13*, 111−125.
(57) Sutherland, J. J.; Weaver, D. F. Development of quantitative structure−activity relationships and classification models for anticonvulsant activity of hydantoin analogues. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (3), 1028−1036.
(58) Yap, C. W.; Chen, Y. Z. Quantitative structure-pharmacokinetic relationships for drug distribution properties by using general regression neural network. *J. Pharm. Sci.* **2005**, *94* (1), 153−168.
(59) Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Comput. Chem.* **2001**, *26* (1), 5−14.
(60) Czerminski, R.; Yasri, A.; Hartsough, D. Use of support vector machine in pattern classification: Application to QSAR studies. *Quant. Struct.-Act. Relat.* **2001**, *20* (3), 227−240.
(61) Meyer, D.; Leischa, F.; Hornik, K. The support vector machine under test. *Neurocomputing* **2003**, *55* (1−2), 169−186.
(62) Lessmann, S. Solving unbalanced classification problems with support vector machines. *Proceedings of the International Conference on Artificial Intelligence*, IC-AI'04 2004; pp 214−220.
(63) Han, L. Y.; Cai, C. Z.; Lo, S. L.; Chung, M. C. M.; Chen, Y. Z. Prediction of RNA-binding proteins from primary sequence by support vector machine approach. *RNA* **2004**, *10* (3), 355−368.
(64) Topliss, J. G.; Edwards, R. P. Chance factors in studies of quantitative structure−activity relationships. *J. Med. Chem.* **1979**, *22* (10), 1238−1244.
(65) Jouan-Rimbaud, D.; Massart, D. L.; de Noord, O. E. Random correlation in variable selection for multivariate calibration with a genetic algorithm. *Chemom. Intell. Lab. Syst.* **1996**, *35* (2), 213−220.
(66) Manly, B. F. J. *Randomization bootstrap and Monte Carlo methods in biology*, 2nd ed.; Chapman and Hall: London, 1997.
(67) Leardia, R.; González, A. L. Genetic algorithms applied to feature selection in PLS regression: How and when to use them. *Chemom. Intell. Lab. Syst.* **1998**, *41* (2), 195−207.
(68) Hawkins, D. M. The problem of overfitting. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (1), 1−12.
(69) Wegner, J. K.; Fröhlich, H.; Zell, A. Feature selection for descriptor based classification models. 2. Human intestinal absorption (HIA). *J. Chem. Inf. Comput. Sci.* **2004**, *44* (3), 931−939.
(70) Moriguchi, I.; Hirono, S.; Liu, Q.; Nakagome, I.; Matsushita, Y. Simple method of calculating octanol/water partition coefficient. *Chem. Pharm. Bull.* **1992**, *40* (1), 127−130.

(71) 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* (1), 424−433.
(72) Sorich, M. J.; Miners, J. O.; McKinnon, R. A.; Winkler, D. A.; Burden, F. R.; Smith, P. A. Comparison of linear and nonlinear classification algorithms for the prediction of drug and chemical metabolism by human UDP-glucuronosyltransferase isoforms. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (6), 2019−2024.
(73) Angulo, C.; Parra, X.; Catala, A. K. SVCR. A support vector machine for multi-class classification. *Neurocomputing* **2003**, *55* (1−2), 57−77.
(74) Smola, A. J.; Scholkopf, B. In *A tutorial on support vector regression*, NeuroCOLT2 Technical Report Series.

CI0500536