0090-9556/09/3703-658–664$20.00
Drug Metabolism and Disposition
Copyright © 2009 by The American Society for Pharmacology and Experimental Therapeutics
DMD 37:658–664, 2009

Vol. 37, No. 3
23507/3436840
Printed in U.S.A.

Classification of Cytochrome P450 1A2 Inhibitors and Noninhibitors by Machine Learning Techniques

Poongavanam Vasanthanathan, Olivier Taboureau, Chris Oostenbrink, Nico P. E. Vermeulen, Lars Olsen, and Flemming Steen Jørgensen

Biostructural Research, Department of Medicinal Chemistry, Faculty of Pharmaceutical Sciences, University of Copenhagen, Copenhagen, Denmark (P.V., L.O., F.S.J.); Leiden-Amsterdam Center for Drug Research, Section of Molecular Toxicology, Department of Chemistry and Pharmacochemistry, Vrije Universiteit, Amsterdam, The Netherlands (P.V., C.O., N.P.E.V.); and Chemoinformatics, BioCentrum-DTU, Technical University of Denmark, Lyngby, Denmark (O.T.)

Received July 25, 2008; accepted December 1, 2008

ABSTRACT:

The cytochrome P450 (P450) superfamily plays an important role in the metabolism of drug compounds, and it is therefore highly desirable to have models that can predict whether a compound interacts with a specific isoform of the P450s. In this work, we provide in silico models for classification of CYP1A2 inhibitors and noninhibitors. Training and test sets consisted of approximately 400 and 7000 compounds, respectively. Various machine learning techniques, such as binary quantitative structure activity relationship, support vector machine (SVM), random forest, kappa nearest neighbor (kNN), and decision tree methods were used to develop in silico models, based on Volsurf and Molecular Operating Environment descriptors. The best models were obtained using the SVM, random forest, and kNN methods in combination with the BestFirst variable selection method, resulting in models with 73 to 76% of accuracy on the test set prediction (Matthews correlation coefficients of 0.51 and 0.52). Finally, a decision tree model based on Lipinski's Rule-of-Five descriptors was also developed. This model predicts 67% of the compounds correctly and gives a simple and interesting insight into the issue of classification. All of the models developed in this work are fast and precise enough to be applicable for virtual screening of CYP1A2 inhibitors or noninhibitors or can be used as simple filters in the drug discovery process.

Cytochromes P450 (P450s) are heme-containing enzymes found in both prokaryotes and eukaryotes, and they are involved in a wide range of cellular biotransformation functions. From a pharmaceutical perspective, the most important function is the degradation of drugs (Nebert and Russell, 2002). In general, hydrophobic compounds are converted into more hydrophilic species to facilitate excretion.

The most important P450 isoforms involved in metabolism of drugs in humans are CYP1A2, CYP2A6, CYP2C9, CYP2C19, CYP2D6, CYP2E1, and CYP3A4. CYP1A2 constitutes 12% of the total P450 content in the liver and plays an important role in the metabolic clearance of ~5% of currently marketed drugs. The substrates for the CYP1A subfamily are generally characterized as neutral, flat, aromatic, and lipophilic (two to four aromatic rings) with at least one putative hydrogen bond donor (Smith et al., 1997), in agreement with the observed contacts in the recent crystal structure of CYP1A2 (Sansen et al., 2007). Examples of drugs that are CYP1A2 substrates are acetaminophen, caffeine, clozapine, haloperidol, olanzapine, propranolol, tacrine, theophylline, and zolmitriptan (drug interactions: cytochrome P450 drug interaction table, Indiana University School of Medicine, http://medicine.iupui.edu/flockhart/table.htm).

In silico approaches are attractive because they can be used in an early stage of the drug discovery process and thereby reduce the number of experimental studies and improve the success rates. For this purpose, various traditional in silico modeling methods and more recently developed nonlinear machine learning methods have been used (Chohan et al., 2005; de Graaf et al., 2005; Kriegl et al., 2005a; Yap and Chen, 2005; Fox and Kriegl, 2006; Yap et al., 2006; Eitrich et al., 2007; Terfloth et al., 2007; Zhou et al., 2007). Machine learning methods are particularly useful for data mining of large databases to discover patterns or rules to derive models for problems for which the underlying mechanism is not clear. For example, support vector machine (SVM) methods have been applied to classify inhibitors of the CYP3A4 enzyme with a success rate of approximately 70% for the test set (807/538 compounds in training/test sets) (Kriegl et al., 2005a) and to predict isoform specificity of CYP3A4, CYP2D6, and CYP2C9 substrates with approximately 80% of the test set correctly predicted (146/233 compounds in training/test sets) (Terfloth et al., 2007).

Thus, in silico methods seem promising for making reliable models for sets of a large number of compounds. In this study, we used approximately 400 compounds to construct models for CYP1A2 inhibition and to explore the accuracy of various machine learning

The work was supported in part by the Drug Research Academy; the Nanoscience Biotechnology and IT program [Grant 26-04-0296]; the Danish Medical Research Council [Grant 271-07-0783]; and the Carlsberg Foundation [Grant 04-0290/20].

Article, publication date, and citation information can be found at http://dmd.aspetjournals.org.

doi:10.1124/dmd.108.023507.

ABBREVIATIONS: P450, cytochrome P450; SVM, support vector machine; kNN, kappa nearest neighbor; QSAR, quantitative structure activity relationship; 3D, three-dimensional; 2D, two-dimensional; MOE, molecular operating environment; DOOD, D-optimal onion design; PCA, principle component analysis; PC, principle component; MCC, Mathews correlation coefficient.

methods such as SVM, random forest, decision tree, kappa nearest neighbor (kNN), and binary QSAR methods for a test set containing 7000 compounds. This is, to our knowledge, the first time that the information from such a large number of compounds has been used to generate and validate in silico models for classification of CYP1A2 inhibitors and noninhibitors.

Materials and Methods

Biological Activity. The structures and biological activity of 8342 compounds were collected from the PubChem BioAssay database (http://www.ncbi.nlm.nih.gov). In brief, the inhibition of human CYP1A2 catalyzing the demethylation of luciferin 6′-methyl ether to luciferin was measured. The luciferin formation is monitored by luminescence after the addition of a luciferase detection reagent. The luciferin 6′-methyl ether concentration in the assay was equal to its K_M for CYP1A2. Of the 8342 compounds downloaded from PubChem, 4173 were assigned as active (inhibitors), 3514 as inactive inhibitors (noninhibitors), and 655 as inconclusive with respect to the activity. The inconclusive compounds were not considered in this study to avoid uncertainty in model building. In addition to the PubChem data set, we also used an external test set of 89 drug compounds, kindly provided to us by Vertex Pharmaceuticals (Cambridge, MA) for validation of the models developed (Zlokarnik et al., 2005).

3D Structure Generation. The 2D structures were converted into 3D structures by using CONCORD software (version 6.1.2, accessed via SYBYL from Tripos, St. Louis, MO). Of the 7687 structures, 218 could not be converted to 3D because of restrictions in the CONCORD program (number of rotatable bonds >25, ring size >10, number of atoms >80, or the presence of atoms such as mercury, arsenic, or antimony) or incorrect or insufficient stereochemical information in the 2D notation from PubChem. For the remaining 7469 structures various descriptors were calculated.

Molecular Operating Environment 2D Descriptors. All 3D structures were imported into the molecular modeling software Molecular Operating Environment (MOE) (version 2006.08; Chemical Computing Group Inc., Montreal, QC, Canada). They were preprocessed by removing all counter ions, solvent molecules, and salt in the structures. Subsequently, the structures were geometry-optimized using the MMFF94s force field, and 214 2D descriptors implemented in MOE were computed (representing atom and bond counts, Kier and Hall connectivity, kappa shape indices, adjacency, distance matrix descriptors, pharmacophore feature descriptors, partial charge descriptors, potential energy descriptors, surface area, volume, and shape descriptors) (Labute, 2000).

Volsurf Descriptors. The geometry-optimized compounds were imported into Volsurf software (version 4.1.4.1; Molecular Discovery Ltd., Middlesex, UK). Volsurf is a computational procedure to transform 3D molecular interaction fields from a GRID calculation into quantitative numerical descriptors. For the present study, we used the following probes: water (H_2O), hydrophobic (DRY), H-bonding carbonyl (O), and N1 amide (N).

Selection of Training and Test Set. All 214 2D MOE descriptors were merged with the 110 Volsurf descriptors for the total set of 7469 compounds. The total data set was divided into a training and a test set. The training set was selected by use of D-optimal onion design (DOOD). DOOD is a method for selecting a training set of reasonable size, which is representative for the chemical property space defined by the molecular structures (Olsson et al., 2004; Kriegl et al., 2005b). In brief, the general idea of DOOD is to score vectors calculated by either principal component analysis (PCA) or partial least-squares and then to split the data set into different layers (or subsets) enabling the selection of training set compounds from each layer. This ensures that representative molecules are selected from the inner layers of the onion design and that the training set contains a diverse range of chemical structures in addition to covering the range of chemical properties. The total number of compounds to be selected can be controlled by the number of layers and the type of regression model targeted within each layer. The performance of DOOD has already been successfully assessed in drug discovery. Compared with a random selection, which may introduce bias and variability if the selected training set is not balanced, the DOOD method has shown better stability for the prediction models (Olsson et al., 2004).

In this work, we used the scores from the PCA obtained by the SIMCA-P program (version 11.0; Umetrics, Umeå, Sweden) [correlation coefficient (R^2) = 0.81; cross-validated correlation coefficient (Q^2) = 0.66; 25 principal components (PCs)]. The layers were determined according to the Hotelling's *T*2 parameter, which measures the distance between the projection of a compound and the center of the model. The compounds were then allocated into the individual layers according to their Hotelling's *T*2 values for the corresponding model. The software package MODDE (version 7.0; Umetrics) was used to perform the DOOD-based training set selection.

We also divided the full set of compounds into a training and a test set using another approach (the kNN method, see below for a description of the method). The quality of the final models was almost identical to that of the models developed with the training set selected with DOOD (data not shown).

Machine Learning Methods. The selected methods are based on different concepts and are representative of the common approaches considered for classification. The machine learning methods applied in this work are presented in brief below.

SVM. SVM is a nonlinear model, developed by Vapnik (2000). The SVM method constructs a hyperplane, which discriminates between data points of distinct classes (binary SVM) such that the margin between both classes is maximized. This margin models the linear decision hyperplane. The final position and orientation of the hyperplane are defined by a subset of training vectors, the so-called support vectors. The SVM approaches were used in association with a radial bias function as the kernel function. The kernel exponent was set to 1.0 [SVM^D (default)] and the polynomial kernel to 2.0 (SVM^E) for linear and nonlinear SVM classifiers, respectively.

kNN algorithm. This method was developed by Zheng and Tropsha (2000). The *k*-nearest neighbor models are based on the assumption that similar compounds have similar physicochemical and biological activity profiles and can define a class membership of its nearest neighbors. Compounds are represented by their position vectors defined in the physicochemical space and then biological activity is assigned by a majority vote of its neighbors. The neighbors are taken from a set of compounds for which the correct classification is known. Finally, properties or activities of new compounds are assigned on the basis of the majority vote of neighbors defined previously. For the kappa nearest neighbor algorithm, we used Euclidian distances and five kappa nearest neighbors to avoid tied votes.

Random forest method. With the random forest method (Witten and Frank, 2005), multiple classification trees are constructed from an input vector. The input vector is then placed down of the classification trees in the forest, and each tree gives a classification, or votes, for that class. Finally, the random forest chooses the classification having the most votes. The forest error rate depends on two factors: 1) the correlation between any two trees in the forest and 2) the strength of each individual tree in the forest. A tree with a low error rate is a strong classifier. The main advantage of this approach is the possibility of handling thousands of input variables without overfitting in a very fast way. The random forest method was used with 10 trees and 1 seed.

Decision tree (C4.5/J48). This method is a divide-and-conquer approach to the problem of learning from a set of independent instances that naturally leads to a style of representation. First, an attribute is selected to place at the root node and make one branch for each possible value. This splits up the data set into subsets, one for every value of the attributes. This process can be repeated recursively for each branch, using only those instances that actually reach the branch (Witten and Frank, 2005).

Binary QSAR. This method is based on the Bayesian inference technique, which is used to classify whether a compound is active or inactive on the basis of its associated molecular descriptors. A probabilistic distribution of active and inactive compounds in a training set is determined using a partial least-squares method. The binary QSAR model derived can subsequently be used to predict the probability of new compounds to be active against given targets. The binary QSAR methodology has been described in detail in the literature (Gao et al., 1999; Labute, 1999; Gao, 2001).

The influence of various factors on the accuracy of the predictions was investigated. A smoothing factor was used to minimize the sensitivity of the derived model to the selection of binary boundaries. We analyzed different smoothing factors (Gao et al., 1999; Labute, 1999; Gao, 2001) ranging from 0.08 to 0.25 with different numbers of principal components ranging from 1 to maximum (data not shown). For descriptor selection, different variable combinations were evaluated. Finally, we selected 110 (MOE/Volsurf) descriptors for model development. The descriptors include MOE 2D descriptors such as subdivided surface area descriptors (SlogP descriptors), number of aromatic atoms, number of hydrogen atoms, atomic valence connectivity index, carbon valence connectivity index, molecular weight, and Balaban's connectivity topological index and Volsurf descriptors using the DRY and N1 probes.

Software. Weka data mining software (version 3.2; Waikato Environment for Knowledge Analysis, University of Waikato, Hamilton, NZ, http://www.cs.waikato.ac.nz/~ml/Weka/) (Witten and Frank, 2005) was used for the inhibitor/noninhibitor classification. The software provides a set of classification and regression methods, variable selection methods, and clustering methods (SVM^D, SVM^E, kNN, random forest, and decision tree). The binary QSAR was performed using the MOE program.

Attribute Selection. For attribute selection we used the automatic variable selection procedure (CfsSubsetEval) in Weka software. CfsSubsetEval was combined with either the BestFirst or a genetic algorithm (Witten and Frank, 2005).

Matthews Correlation Coefficient. Matthews correlation coefficient (MCC) is a measure of the quality of a binary classification. It takes into account true positives and negatives and is generally regarded as a balanced measure that can also be used if the classes are of very different sizes. It returns a value between −1 and +1. A coefficient of +1 represents a perfect prediction, 0 an average random prediction, and −1 the worst possible prediction. In general, MCC values greater than 0.4 are considered to be predictive in machine learning methods (Chohan et al., 2005).

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

where TP are true positives, TN are true negatives, FP are false positives, and FN are false negatives.

Cross-Validation. All models developed were cross-validated with *K*-fold cross-validation. In *K*-fold cross-validation, the original sample is divided into *K* subsamples. Of the *K* subsamples, a single subsample is retained as the validation data for testing the model, and the remaining *K* − 1 subsamples are used as training data. The cross-validation process is then repeated *K* times, with each of the *K* subsamples used exactly once as the validation data. Here, we performed 5-fold cross-validation. All cross-validations were carried out with the procedures in Weka software, and 5-fold cross-validation was chosen.

Results and Discussion

Description of Training and Test Sets. The 7469 compounds extracted from PubChem and considered in this work constitute the largest set of compounds used for constructing in silico models of CYP1A2. Visual inspection of randomly selected compounds revealed compounds resembling drug compounds as well as compounds being clearly nondrug-like. Although Lipinski's Rule-of-Five (Lipinski et al., 2001) is not a direct measure of drug-likeness, we calculated the violations of each of the four rules for the CYP1A2 data set: 544 compounds had molecular weight greater than 500; 132 compounds had more than 10 hydrogen bond acceptors; 293 compounds had more than 5 hydrogen bond donors, and 551 compounds had a partition coefficient, logP value, greater than 5. In total, only approximately 15% of the compounds in the data set violated one or several of the Lipinski's rules.

From DOOD, a training set consisting of 411 compounds (192 inhibitors and 219 noninhibitors) was constructed, and the remaining 7058 compounds were used as a test set. DOOD was used because it has previously been successfully applied for the selection of suitable training sets (Gavaghan et al., 2007). The score plot from the PCA shows that the diversity of the whole data set is satisfactorily reflected in the training set (Figure 1). Only a minority of the compounds are outside the 95% confidence interval, and, of those, most form a cluster located in the lower right part of the score plot, corresponding to the compounds violating Lipinski's Rule-of-Five. Inspection of these structures revealed that there is no structural similarity within this group, but that 95% of these compounds were noninhibitors.

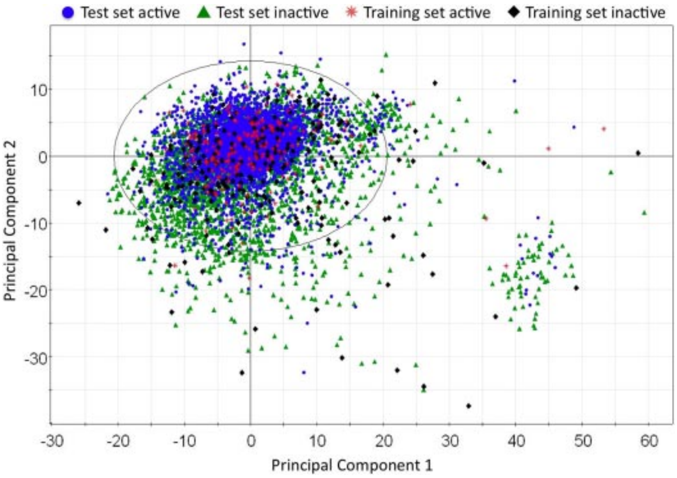

FIG. 1. Score plot from principal component analysis (two first components shown) of the PubChem data set. The compounds are colored as follows: training set and inactive as black diamonds, training set and active as red asterisks, test set and inactive as green triangles, and test set and active as blue circles.

Construction of Binary QSAR Models. Binary QSAR models for classification of active (inhibitors) and inactive (noninhibitors) compounds were developed using the 214 calculated MOE and 110 Volsurf descriptors. We selected these descriptors because previously they have been applied successfully for classification or IC$_{50}$ predictions for P450 ligands (Eriksson et al., 2004; Kriegl et al., 2005b). Most of these descriptors cover properties such as lipophilicity [e.g., LogP-derived descriptors or Volsurf descriptors based on the hydrophobic interactions the (DRY probe)], the hydrophilicity [e.g., polar surface area-derived descriptors or Volsurf descriptors based on the hydrophilic interactions (N1 probe)], or the number of various types of atoms (e.g., number of aromatic atoms or hydrogen atoms).

The influence of various factors such as the number of descriptors, the number of principle components (PCA), and smoothing factors on the accuracy of the predictions was investigated (see *Materials and Methods*). Finally, we selected 110 (MOE and Volsurf) descriptors for model development, including those describing the lipophilicity, hydrophilicity, various atom counts, and connectivity. Use of default settings in MOE for smoothing factors and number of PCs gave a binary QSAR model with a significant difference in predictive power between the training and test set, in particular for the CYP1A2 inhibitors, for which 96% are correctly predicted in the training set, but only 55% in the test set. Therefore, a model that is based on only three principal components was considered. This leads to a correctly predicted number of compounds on 70 and 66% for the training and test sets, respectively, but at the same time it yields a more balanced model with respect to the predictivity of inhibitors and noninhibitors (Table 1).

Construction of Models by Other Machine Learning Methods. We developed models with other methods to investigate whether the quality of the models could be improved. For that purpose, classification models of active and inactive compounds were generated using the SVM, kNN, random forest, and decision tree methods.

As for the binary QSAR model described in the last section, care was taken with respect to what descriptors were included in the model building to avoid overfitting and to reduce the noise. Therefore, we generated models using all MOE and Volsurf variables, as well as a variable selection by the so-called BestFirst and a genetic algorithm (Tables 2 and 3; Fig. 2). In general, the quality of the models improved when the BestFirst variable selection was used, whereas variables selected by the genetic algorithm did not improve the model

TABLE 1

Binary QSAR models based on MOE and Volsurf descriptors

Set	Predicted Compounds				Correctly Predicted	MCC
	TP	FP	TN	FN		
				%		
Training[a]	73	27	67	33	70	0.40
Training[b]	96	4	90	10	93	0.89
Test[a]	74	26	57	43	66	0.31
Test[b]	55	45	80	20	66	0.36

TP, number of true positives; FP, number of false positives; TN, number of true negatives; FN, number of false negatives.

[a] Three principal components with a smoothing factor of 0.08.

[b] Maximum principal components with a smoothing factor of 0.08 (default value in MOE).

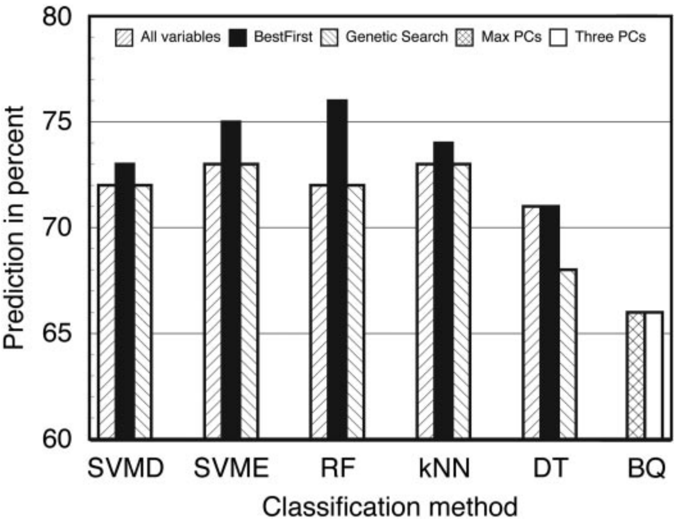

FIG. 2. Correctly classified compounds from the test set in all methods. All variables, predicted by all variables; BestFirst, predicted by BestFirst variables; Genetic Search, predicted by GeneticSearch variables; Max PCs, binary model using maximum number of principle components; Three PCs, binary model using first three principle components; SVMD, support vector machine, linear model; SVME, support vector machine, nonlinear model; RF, random forest method; kNN, kappa nearest neighbor method; DT, decision tree method; BQ, binary QSAR.

compared with use of all variables. For example, the total number of correctly predicted compounds in the test set is largest using the BestFirst method (SVM^D: 73%: SVM^E: 75%; random forest: 76%; kNN: 74%; and decision tree: 71%) (Fig. 2), and its ability to predict both active and inactive compounds is more balanced compared with the models generated with all descriptors and descriptors selected by the genetic algorithm (Table 3). Thus, in the remaining text we will only comment on the models generated by the BestFirst method for variable selection.

Overall, the different methods yield models of similar quality. Of the 7058 compounds in the test set, 4984 to 5375 (71–76%) are correctly predicted. Thus, the different methods are quite similar in performance, with SVM^E and the random forest methods being the best methods and the decision tree method being a reasonable method. This finding is also reflected in the MCC values that are 0.51 to 0.52 for the SVM^E and the random forest methods and 0.41 for the decision tree method. However, it should be noted that the smallest discrep-

TABLE 2

Summary of results using all descriptors in training and test sets

Set	Method	Predicted Compounds				Correctly Predicted	MCC	5-Fold CV
		TP	FP	TN	FN			
				%				
Training set	SVM^D	80	20	85	14	83	0.66	71
Training set	SVM^E	100	0	100	0	100	1.00	70
Training set	RF	100	0	100	0	100	1.00	73
Training set	kNN	84	16	72	28	78	0.55	68
Training set	C4.5/J48	96	4	97	3	97	0.93	67
Test set	SVM^D	80	20	61	39	72	0.42	
Test set	SVM^E	77	23	67	33	73	0.45	
Test set	kNN	79	21	65	35	73	0.44	
Test set	C4.5/J48	71	29	72	28	71	0.43	

TP, number of true positives; FP, number of false positives; TN, number of true negatives; FN, number of false negatives; CV, cross-validation; SVM^D, support vector machine, linear model; SVM^E, support vector machine, nonlinear model; RF, random forest method; kNN, kappa nearest neighbors method; C4.5/J48, decision tree method.

TABLE 3

Summary of results using BestFirst descriptors in training and test sets

Set	Method	Predicted Compounds				Correctly Predicted	MCC	5-Fold CV
		TP	FP	TN	FN			
				%				
Training set	SVM^D	74	26	79	21	77	0.54	
Training set	SVM^E	79	21	84	16	82	0.63	
Training set	RF	100	0	100	0	100	1.00	
Training set	kNN	86	14	80	20	83	0.66	
Training set	C4.5/J48	98	02	97	02	97	0.95	
Test set	SVM^D	73	27	73	27	73	0.46	
Test set	SVM^E	73	27	78	22	75	0.51	
Test set	RF	78	22	74	26	76	0.52	
Test set	kNN	79	21	68	32	74	0.47	
Test set	C4.5/J48	71	29	70	30	71	0.41	

TP, number of true positives; FP, number of false positives; TN, number of true negatives; FN, number of false negatives; MCC, Matthews correlation coefficient; CV, cross-validation; SVM^D, support vector machine, linear model; SVM^E, support vector machine, nonlinear model; RF, random forest method; kNN, kappa nearest neighbors method; C4.5/J48, decision tree method.

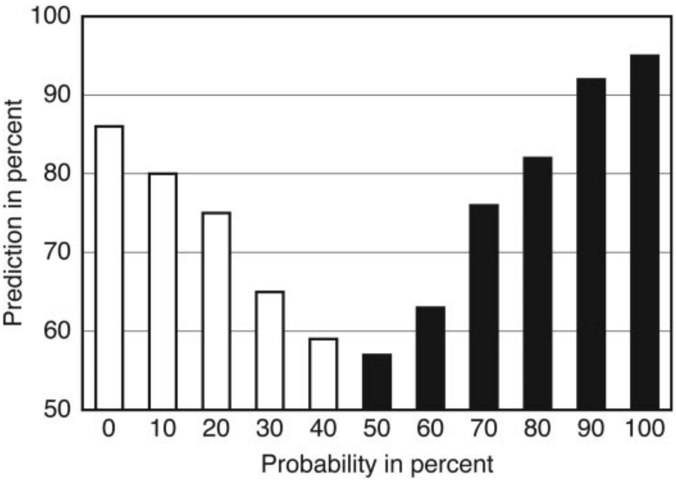

FIG. 3. True predictions for probabilities of belonging to a certain class (data shown for random forest predictions of compounds in the test set). In Weka software, compounds with a probability smaller or larger than 50% are classified as inactive or active, respectively. White/black, true predictions for inactive/active compounds.

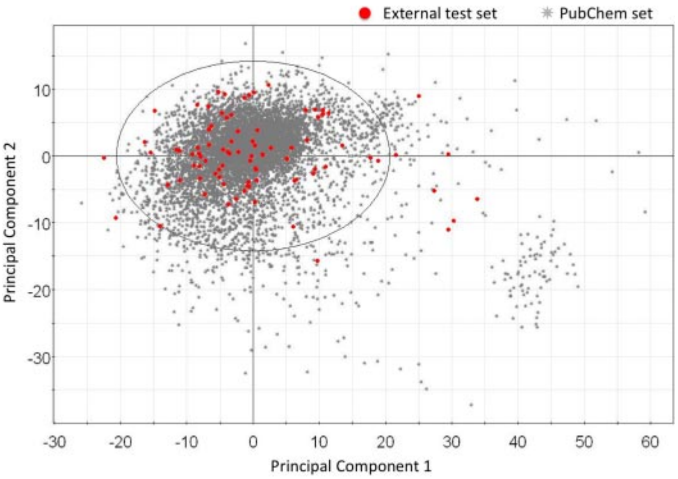

FIG. 4. The 89 drug molecules in the external test set projected on the score plot from principal component analysis of the PubChem data set.

ancy in prediction between training and tests is observed for the model developed by the SVM^D method. This result indicates that it is a well balanced and not overfitted model. As seen in Fig. 2, the binary QSAR model is clearly worse than the other methods, and, again, this result is also reflected in the MCC values of 0.31 and 0.36 (Table 1).

A probability of belonging to a certain class, active or inactive, is given for each compound with the random forest method. Compounds with a probability <50% are classified as inactive compounds, and compounds with a probability ≥50% are classified as active compounds. Figure 3 shows the percentage of true predictions for different ranges of the probabilities using the random forest model. It is interesting to note that there is a larger chance of finding either a true positive or a true negative for probabilities closest to 100 or 0%, respectively. For example, those compounds with a probability of 100% have the highest prediction rate of 95%, which is significantly larger than if one is using lower values of the probability. Likewise, the compounds with a probability of 0% also have a larger prediction rate of 86% compared with higher values of the probability. This is an important result, because it gives a way of ranking the database and therefore also a way of selecting compounds from a database search.

TABLE 4

Prediction of external test set of 89 drug compounds using the machine learning methods

Method	Predicted Compounds				Correctly Predicted	MCC
	TP	FP	TN	FN		
				%		
SVM^D	56	44	70	30	64	0.27
SVM^E	62	38	72	28	67	0.34
RF	56	44	60	40	58	0.16
kNN	77	33	44	56	58	0.19
C4.5/J48	54	46	52	48	53	0.05

TP, number of true positives; FP, number of false positives; TN, number of true negatives; FN, number of false negatives; CV, cross-validation; SVM^D, support vector machine, linear model; SVM^E, support vector machine, nonlinear model; RF, random forest method; kNN, κ nearest neighbors method; C4.5/J48, decision tree method.

External Test Set of Drug Compounds. In the following, we have focused on the best machine learning methods (SVM^D, SVM^E, random forest, kNN, and decision tree) to classify a small external set of 89 drugs (39 active and 50 inactive). Of these compounds, 24 violated Lipinski's Rule-of-five, primarily because of lipophilicity. Figure 4 shows that this small external set of compounds span the same chemical space as the large PubChem dataset; thus, it should be possible to predict their 1A2 activity with the developed models. The best method was again the SVM^E method, which classifies 60 of the 89 compounds correctly (Table 4). It is interesting to note that the performance seems to be slightly poorer for this set of compounds compared with the PubChem data set, which might be attributed to the smaller size of the set of compounds. Another explanation may arise from the fact that a different experimental assay was used to determine the inhibitory constants of the external test set compared with the PubChem data set. This may lead to different experimental classifications for which the models were not trained.

Structural Characteristics of Inhibitors, Noninhibitors, and Misclassified Compounds. It is important to understand why compounds are classified as they are. In general, it turned out that differences in the hydrophobic and hydrophilic natures of the CYP1A2 compounds are important for discriminating inhibitors from noninhibitors. In Volsurf, the hydrophobic and hydrophilic properties of a compound may be visualized with the so-called DRY (hydrophobic properties) and N1 (hydrogen donor properties) probes, as shown in Fig. 5. Compound A, an inhibitor of CYP1A2, has a large hydrophobic region and a small polar region, which is common for most of the inhibitors. Noninhibitors, such as compound B, are characterized by slightly smaller hydrophobic regions and much larger polar regions. These characteristics fit well with the general idea of CYP1A2 ligands being hydrophobic with one hydrogen bond acceptor (Smith et al., 1997; Sansen et al., 2007). A compound with characteristics such as those of compound C is likely to be predicted as a noninhibitor, because it has a large polar region and therefore fails in the prediction. Likewise, compound D, which is a noninhibitor, is mainly hydrophobic, and this is probably why this prediction also fails.

Lipinski-Decision Tree Model. Although the models derived by the machine learning methods perform satisfactorily, it would be useful to develop a simple "back-of-the-envelope calculation" such as Lipinski's Rule-of-Five. Such a method would be extremely fast and therefore potentially useful for high-throughput screening and at the same time allow easy interpretation. Thus, a decision tree model was subsequently developed on the basis of the four Lipinski parameters: molecular weight, numbers of hydrogen bond acceptors and donors, and lipophilicity. The so-called Lipinski-decision tree model performed surprisingly well (Table 5): 71 and 67% of the compounds in the training and test sets, respectively, were correctly predicted, even

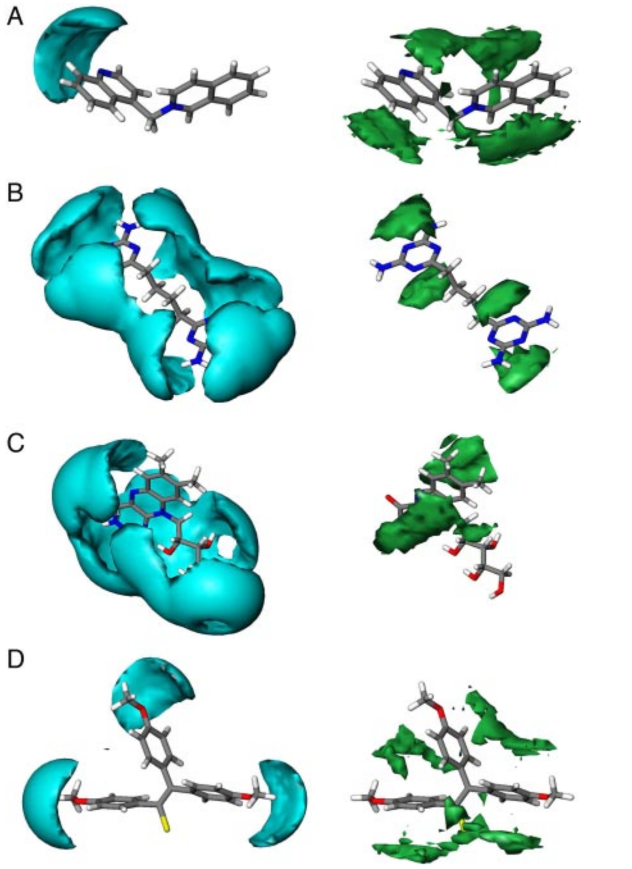

FIG. 5. GRID 3D molecular interaction fields (MIFs) of four representative compounds. The cyan contours (left column) correspond to MIFs calculated with an amide probe and contoured at −2 kcal/mol. The green contours (right column) correspond to MIFs calculated with a DRY probe and contoured at −1 kcal/mol. From top to bottom the MIFs illustrate (A) active compound classified as active, (B) inactive compound classified as inactive, (C) active compound classified as inactive, and (D) inactive compound classified as active.

TABLE 5

Result of Lipinski-decision tree model

Set	Predicted Compounds				Correctly Predicted	MCC
	TP	FP	TN	FN		
				%		
Training set	91	09	53	47	71	0.43
Test set	87	13	43	57	67	0.31

TP, number of true positives; FP, number of false positives; TN, number of true negatives; FN, number of false negatives.

though the balance between TP and TN is relatively far off. As illustrated in Fig. 6, compounds classified as inhibitors are characterized as having either 1) more than two hydrogen bond donors, logP ≥3.7 and less than six hydrogen bond donors or 2) ≤3 hydrogen bond donors and a molecular weight between 198 and 516. Compounds that fall outside these ranges are considered to be noninhibitors. This decision model is in good agreement with the conclusion derived from inspection of the molecular interaction fields (Fig. 4) discussed in the previous section. Finally, we tested the Lipinski-decision tree model on a set of known 1A2 substrates and inhibitors (Table 6) (drug interactions: cytochrome P450 drug interaction table, Indiana University School of Medicine, http://medicine.iupui.edu/flockhart/table.htm). Of these, 34 compounds corresponding to 79% were correctly predicted as 1A2 ligands.

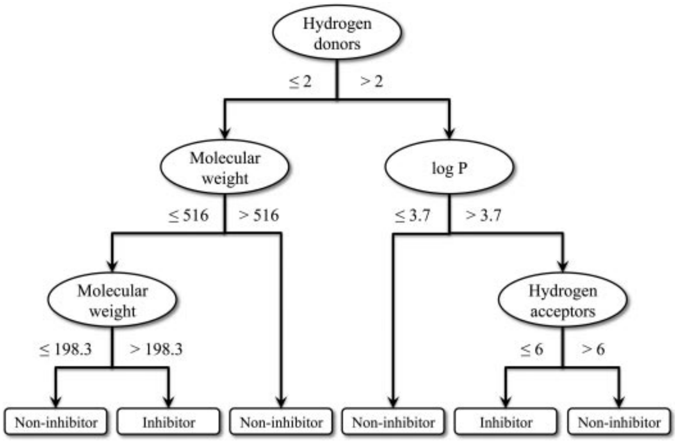

FIG. 6. Decision tree model from Lipinski Rule-of-Five descriptors (see text for details).

TABLE 6

Lipinski-decision tree model predictions of marketed CYP1A2 substrates/inhibitors

Data from Flockhart (2007).

Prediction	1A2 Substrates/Inhibitors	Predictions
		%
Correct	Amitriptyline, ciprofloxacin, clomipramine, clozapine, cyclobenzaprine, estradiol, fluvoxamine, furafylline, haloperidol, imipramine, levofloxacin, methoxsalen, mibefradil, α-naphthoflavone, naproxen, norfloxacin, olanzapine, ondansetron, propranolol, riluzole, ropivacaine, tacrine, tizanidine, verapamil, warfarin, zileuton, zolmitriptan	79
Wrong	Acetaminophen, amiodarone, caffeine, cimetidine, mexiletine, phenacetin, theophylline	21

Conclusion

In this article, we have reported the application of different machine learning classification methods such as support vector machine (linear and nonlinear), random forest, kNN, decision tree, and binary QSAR for the classification of cytochrome P450 1A2 ligands. Models that are based on the support vector machine method are superior to binary QSAR models. Use of the SVM, random forest, and kNN methods and the BestFirst variable selection method resulted in models with 73 to 76% of the test set correctly predicted. Finally, a decision tree model based on Lipinski's Rule-of-Five descriptors was developed, classifying 67% of the compounds in the test set correctly. This model is easy to interpret and offers structural insight into the classification of new CYP1A2 inhibitors.

Inspection of the molecular interaction fields illustrates the importance of hydrophobicity and hydrogen bond donors and acceptors. These features are in agreement with the Lipinski-decision tree model and agree with previously reported pharmacophore models and X-ray structure of the cytochrome P450 1A2 isoenzyme.

The models developed in this work are fast and precise enough to be applicable for virtual screening of large databases for identification of 1A2 inhibitors or noninhibitors. Moreover, they can be used as filters to quickly assess the likelihood that a newly designed compound shows an interaction with CYP1A2. As such, the models can play an important role in preventing the risk of, e.g., drug-drug interactions through metabolism at an early stage of the drug development process.

Acknowledgments. We thank Professor Gabriele Cruciani (University of Perugia, Perugia, Italy) for providing the Volsurf software.

References

Chohan KK, Paine SW, Mistry J, Barton P, and Davis AM (2005) A rapid computational filter for cytochrome P450 1A2 inhibition potential of compound libraries. *J Med Chem* **48:**5154–5161.

de Graaf C, Vermeulen NP, and Feenstra KA (2005) Cytochrome P450 in silico: an integrative modeling approach. *J Med Chem* **48:**2725–2755.

Eitrich T, Kless A, Druska C, Meyer W, and Grotendorst J (2007) Classification of highly unbalanced CYP450 data of drugs using cost sensitive machine learning techniques. *J Chem Inf Model* **47:**92–103.

Eriksson L, Arnhold T, Beck B, Fox T, Johansson E, and Kriegl JM (2004) Onion design and its application to a pharmaceutical QSAR problem. *J Chemometrics* **18:**188–202.

Fox T and Kriegl JM (2006) Machine learning techniques for in silico modeling of drug metabolism. *Curr Top Med Chem* **6:**1579–1591.

Gao H (2001) Application of BCUT metrics and genetic algorithm in binary QSAR analysis. *J Chem Inf Comput Sci* **41:**402–407.

Gao H, Williams C, Labute P, and Bajorath J (1999) Binary quantitative structure-activity relationship (QSAR) analysis of estrogen receptor ligands. *J Chem Inf Comput Sci* **39:**164–168.

Gavaghan CL, Arnby CH, Blomberg N, Strandlund G, and Boyer S (2007) Development, interpretation and temporal evaluation of a global QSAR of hERG electrophysiology screening data. *J Comput Aided Mol Des* **21:**189–206.

Kriegl JM, Arnhold T, Beck B, and Fox T (2005a) Prediction of human cytochrome P450 inhibition using support vector machines. *QSAR Comb Sci* **24:**491–502.

Kriegl JM, Eriksson L, Arnhold T, Beck B, Johansson E, and Fox T (2005b) Multivariate modeling of cytochrome P450 3A4 inhibition. *Eur J Pharm Sci* **24:**451–463.

Labute P (1999) Binary QSAR: a new method for the determination of quantitative structure activity relationships. *Pac Symp Biocomput* 444–455.

Labute P (2000) A widely applicable set of descriptors. *J Mol Graph Model* **18:**464–477.

Lipinski CA, Lombardo F, Dominy BW, and Feeney PJ (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* **46:**3–26.

Nebert DW and Russell DW (2002) Clinical importance of the cytochromes P450. *Lancet* **360:**1155–1162.

Olsson IM, Gottfries J, and Wold S (2004) D-optimal onion designs in statistical molecular design. *Chemom Intell Lab Syst* **73:**37–46.

Sansen S, Yano JK, Reynald RL, Schoch GA, Griffin KJ, Stout CD, and Johnson EF (2007) Adaptations for the oxidation of polycyclic aromatic hydrocarbons exhibited by the structure of human P450 1A2. *J Biol Chem* **282:**14348–14355.

Smith DA, Ackland MJ, and Jones BC (1997) Properties of cytochrome P450 isoenzymes and their substrates part 2: properties of cytochrome P450 substrates. *Drug Discov Today* **2:**479–486.

Terfloth L, Bienfait B, and Gasteiger J (2007) Ligand-based models for the isoform specificity of cytochrome P450 3A4, 2D6, and 2C9 substrates. *J Chem Inf Model* **47:**1688–1701.

Vapnik VN (2000) *The Nature of Statistical Learning Theory*, 2nd ed, Springer, New York.

Witten IH and Frank E (2005) *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed, Morgan Kaufmann, San Francisco.

Yap CW and Chen YZ (2005) Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines. *J Chem Inf Model* **45:**982–992.

Yap CW, Xue Y, Li ZR, and Chen YZ (2006) Application of support vector machines to in silico prediction of cytochrome p450 enzyme substrates and inhibitors. *Curr Top Med Chem* **6:**1593–1607.

Zheng W and Tropsha A (2000) Novel variable selection quantitative structure–property relationship approach based on the *k*-nearest-neighbor principle. *J Chem Inf Comput Sci* **40:**185–194.

Zhou D, Liu R, Otmani SA, Grimm SW, Zauhar RJ, and Zamora I (2007) Rapid classification of CYP3A4 inhibition potential using support vector machine approach. *Lett Drug Des Disc* **4:**192–200.

Zlokarnik G, Grootenhuis PD, and Watson JB (2005) High throughput P450 inhibition screens in early drug discovery. *Drug Discov Today* **10:**1443–1450.

Address correspondence to: Dr. Flemming Steen Jørgensen, Biostructural Research, Department of Medicinal Chemistry, Faculty of Pharmaceutical Sciences, University of Copenhagen, Universitetsparken 2, DK-2100 Copenhagen, Denmark. E-mail: fsj@farma.ku.dk