Classification of Cytochrome P450 Inhibitors with Respect to Binding Free Energy and pIC_{50} Using Common Molecular Descriptors

Onur Dagliyan,[†] I. Halil Kavakli,[†] and Metin Turkay*[,‡]

College of Engineering and Center for Computational Biology and Bioinformatics, Koç University,
Rumelifeneri yolu, Sariyer, Istanbul, 34450 Turkey

Received July 14, 2009

Virtual screening of chemical libraries following experimental assays of drug candidates is a common procedure in structure based drug discovery. However, the relationship between binding free energies and biological activities (pIC_{50}) of drug candidates is still an unsolved issue that limits the efficiency and speed of drug development processes. In this study, the relationship between them is investigated based on a common molecular descriptor set for human cytochrome P450 enzymes (CYPs). CYPs play an important role in drug−drug interactions, drug metabolism, and toxicity. Therefore, *in silico* prediction of CYP inhibition by drug candidates is one of the major considerations in drug discovery. The combination of partial least-squares regression (PLSR) and a variety of classification algorithms were employed by considering this relationship as a classification problem. Our results indicate that PLSR with classification is a powerful tool to predict more than one output such as binding free energy and pIC_{50} simultaneously. PLSR with mixed-integer linear programming based hyperboxes predicts the binding free energy and pIC_{50} with a mean accuracy of 87.18% (min: 81.67% max: 97.05%) and 88.09% (min: 79.83% max: 92.90%), respectively, for the cytochrome p450 superfamily using the common 6 molecular descriptors with a 10-fold cross-validation.

INTRODUCTION

The discovery of lead compounds for a biomolecular target is the key step in structure based drug design for a specific disease. The discovery process starts with the determination of the 3-D structure of the target protein by X-ray crystallography, NMR, or homology modeling. Then, initial drug candidates are identified using computational approaches. One of the most widely used approaches is the virtual screening using docking analysis of the drug candidates on the active or regulatory site of the target protein. Virtual screening provides a score based on the steric and electrostatic interactions of the drug candidates with the target protein. The scoring function provides a computational estimate such as binding free energy (BFE), binding constant, and docking score for the activity of the drug candidate on the target site. Next, selected drug candidates giving relatively high binding affinity are tested with *in vitro* and *in vivo* assays.[1,2] Nevertheless, both virtual screening and high throughput experiments require resources and time. Although the compounds that have low binding energies and favorable ADMET (adsorption, distribution, metabolism, elimination, and toxicity) values are selected for further analysis, these compounds can give considerably high IC_{50} values in experimental assays. Therefore, it is necessary to analyze the relationship between binding free energies and experimentally tested activities of drug candidates.

When the initial discovery process results in a drug candidate having low binding affinity or improper ADMET properties, lead optimization is carried out to generate structurally similar derivatives of the lead compounds using different methods. Quantitative—structure activity relation (QSAR) is very useful and accurate in lead optimization studies when the molecules are structurally very similar. QSAR correlates structure and function within a series of molecules in terms of physicochemical parameters and steric properties. 3D QSAR methods consider three-dimensional structures and a binding form of the ligands on the target protein.[3] This draws the attention that drug activities on specific targets (outputs) can be modeled by using a wide range of molecular descriptors (inputs). Yap et al.[4] predicted the cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates with a high accuracy by using Support Vector Machines (SVM) with 6 common molecular descriptors. Classification of 1,4-dihydropyridine calcium antagonists were performed by using the Least-Square Support Vector Machines (LSSVM) method to obtain a seven descriptor model.[5] Similarly, binding free energies of drug candidates can be predicted by using molecular descriptors. When the activity is taken as the response (output) only in a feature reduction method, prediction of the activity with the reduced descriptors is not a great drawback. This is also true for the prediction of binding free energy. Nevertheless, the accurate prediction of binding free energy and biological activity with a common molecular descriptor set requires more effort. There is a trade-off in accuracy to predict more than one output with the same features. A novel iterative algorithm (based on the combination of partial least-squares regression and mixed-integer programming based hyperboxes (MILP-HB) classification which gives higher accuracies compared to other classification methods for drug candidates was presented.[6,7]

* Corresponding author e-mail: mturkay@ku.edu.tr.
[†] Department of Chemical and Biological Engineering.
[‡] Department of Industrial Engineering.

10.1021/ci900247t CCC: $40.75 © 2009 American Chemical Society
Published on Web 09/24/2009

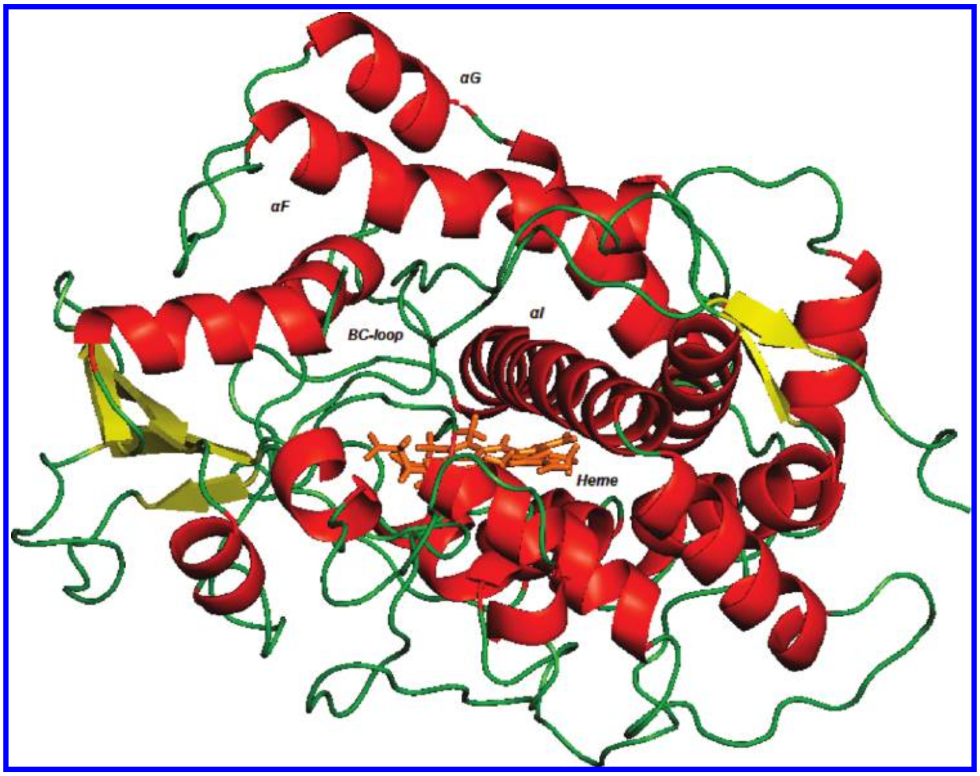

Figure 1. The structure of cytochrome P450s.

Kontijevskis et al.[8] developed a general model for the prediction of CYP enzymes with a regression approach by using various inhibition data for CYPs. We illustrate our approach to establish a relationship between binding free energy and biological activity on CYP superfamily enzymes. The structure of CYP is given in Figure 1, and it is known that the substrate binding pocket is placed in the cavity above the *heme* group. *CYP1A2, CYP2C8, CYP2C9, CYP2A6, CYP2C19, CYP2D6,* and *CYP3A4* play an important role in the metabolism of xenobiotic and endogenous substances.[9,10] In other words, inhibition or induction of *CYP* enzymes by drugs has an effect on drug metabolism resulting in adverse effects.[11] During our search of the database we found data for *CYP17* which is the enzyme responsible for the androgen synthesis, and inhibiting its activity prevents the progression of prostate cancer.[12] *CYP17* inhibition data are also included in this paper.

In this work, we have approached the binding free energy and biological activity relationship of inhibitory molecules as the elucidation of significant common descriptors that classify both BFE and pIC_{50} values with high accuracy. We present an iterative algorithm that employs partial least-squares regression, MILP-HB classification, and significance analysis to classify molecules with high or low biological activity and binding free energy. Furthermore, other classification methods were used in order to compare the accuracy of the MILP-HB classification.

METHODS

Data Sets. Members of the cytochrome superfamily enzymes including *CYP2D6* (pdb id: 2f9q), *CYP1A2* (pdb id: 2hi4), *CYP3A4* (pdb id: 1tqn), *CYP2A6* (pdb id: 1z10), *CYP2C9* (pdb id: 1og2), *CYP2C8* (pdb id: 1pq2), *CYP2C19*, and *CYP17* (pdb id: 2c17) are considered in this paper. The three-dimensional structures of these enzymes were determined by experimental methods except *CYP17*[13] and *CYP2C19* which were built by homology modeling. SWISS-MODEL[14] was used for the homology modeling of *CYP2C19* by taking the *CYP2C9* as a template having 90% sequence identity. The IC_{50} values of diverse CYP inhibitors were collected from various publications,[15−53] and 2D structures of compounds were built by MarvinSketch.[54] IC_{50} values were transformed to decimal logarithms pIC_{50}, since the range of activities is considerably large.

Molecular Dynamics. Initial structures of CYP enzymes were obtained from the Protein Data Bank. VMD was used to prepare protein structure files for MD simulation with NAMD[55] by using CHARMM[56] force field parameters. Protein and *heme* moiety structures were prepared separately due to the constraints in the *psfgen* package in VMD. Enzymes were solvated in water box with a minimum 10 Å distance from any atom of the protein to the boundary. Then, each CYP system was neutralized with the addition of Na^+ and Cl^- ions. First, 10000 steps minimization to only side chains were performed, before performing 10000 steps of minimization to all atoms without pressure control for the equilibration of the systems. Second, the system temperature was calibrated to 310 K by 10 K increment steps, and 10 ps simulation was performed after each 10 K increment. Since rmsd of proteins was converged and stabilized, molecular dynamics simulation was started to perform at 310 K. The bonded interactions, the van der Waals interaction (12 Å cutoff), and the long-range electrostatic interactions with particle-mesh Ewald (PME) were included in the calculations to define the forces acting on systems. The damping coefficient was set to 5 ps^{-1} using Langevin dynamics to handle pressure control, and 1 atm constant pressure was set with 100 fs decay period and with 50 fs damping time.

Two ns of simulation runs were performed to obtain the final structures of enzymes for docking studies.

Molecular Docking. Autodock 3.0.5[57] was used to obtain binding free energies (BFE) and bound spatial conformations of compounds. The binding site of each protein was covered by preparing a 52 × 52 × 52 size of grid box with 0.375 Å of spacing between grid points. Docking parameters were set to the following values: population size: 50, number of generations: 27000, crossover rate: 0.8, mutation rate: 0.02, number of runs: 10, number of evaluations: 1,000,000. For the preparation of charged ligand file (pdbq), grid parameter file (gpf) and grid maps, and docking parameter files (dpf) for a large number of ligands, *Python*[58] programming language scripts were written and used. After docking simulations, 3D coordinates of ligands at their lowest binding free energy were saved in MDL Mol Format (sdf), and then explicit hydrogen atoms were added with OpenBabel 2.2.0.[59]

Preprocessing and Feature Extraction. Molecular descriptors were calculated by using the E-DRAGON[60] Web server which provides more than 1600 molecular descriptors in 20 different categories. The descriptor sets used in this study are constitutional descriptors (48), walk and path counts (47), topological descriptors (119), connectivity indices (33), information indices (47), 2D autocorrelations (96), edge adjacency indices (107), Burden eigenvalues (64), topological charge indices (21), eigenvalue based indices (44), Randic molecular profiles (41), geometrical descriptors (74), RDF descriptors (150), 3D-MORSE descriptors (160), WHIM descriptors (99), GETAWAY descriptors (197), functional group counts (154), atom-centered fragments (120), charge descriptors (14), and molecular properties (29). As a preprocessing step, constant or near-constant (showing the same value for more than 90%) descriptors were eliminated. Also, Unsupervised Forward Selection[61] (UFS) was used to select the maximal linearly independent set of columns with a minimal amount of multiple correlations. UFS is an unsupervised learning method to eliminate the redundancy and to reduce the multicollinearity. UFS removes attributes with standard deviation less than user defined minimum standard deviation (*sdevmin).* It reduces the number of attributes such that the squared multiple correlation coefficients of the remaining attributes is smaller than user defined R^2_{max}. In other words, UFS is utilized to have the most informative descriptors for each data set.

Partial Least Square Regression (PLSR) is a multivariate data analysis tool to describe some predicted variables in terms of observed variables.[62] PLSR was performed using MINITAB software.[63] For the calculation of regression coefficients of descriptors, pIC_{50} values and binding free energies were taken separately as responses (Y vector), and reduced molecular descriptors by *UFS* were taken as attributes (X matrix). The 15 descriptors with the highest regression coefficient were added to the "15-most significant descriptors" set, and the first 6 of them were used as input for the initial classification, where drug molecules were classified as low active or high active regarding their experimental activity and binding free energy.

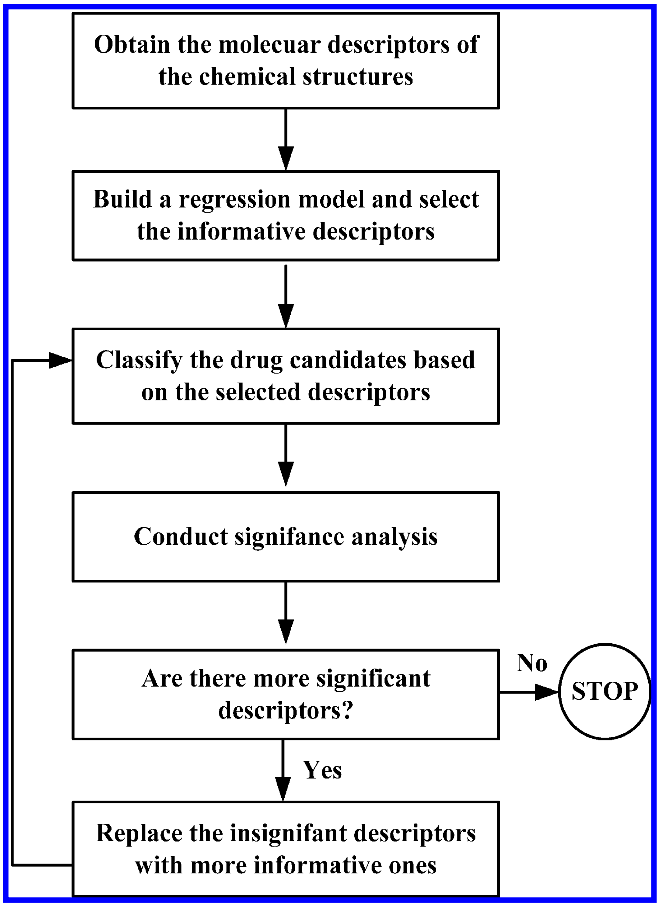
Obtain the molecuar descriptors of the chemical structures
Build a regression model and select the informative descriptors
Classify the drug candidates based on the selected descriptors
Conduct signifance analysis
Are there more significant descriptors?
No
STOP
Yes
Replace the insignifant descriptors with more informative ones

Figure 2. The outline of classification algorithm.

Significance Analysis and Iterations. After the initial classification, prediction accuracy was increased by performing significance analysis. In this analysis, whole drug set Z is divided into two classes after the classification, A and B. If the classification is successful, the variances of descriptor values should be smaller within classes A and B than it is for the whole drug set Z

$$\frac{S_{ij}^2/\sigma_i^2}{S_{ik}^2/\sigma_i^2} = S_{ij}^2/S_{ik}^2 = f_{vn} \qquad (1)$$

Equation 1 is the F distribution with degrees of freedom $v = n$-1 and $n = m$-1. S_{ij}^2 is the sample variance of values for descriptor i for drug set j, n is the number of values of descriptor i for drug set j, and m is the number of values of descriptor i for the drug set k.

Then, hypothesis testing is performed by the null hypothesis $S_{ij}^2 = S_{ik}^2$. This hypothesis proposes that the variance of drugs within the same class is equal to the variance of the whole set of drugs. Since the variance within the class should be smaller than the variance of the whole class, alternative hypothesis is defined as $H_a = S_{ij}^2 > S_{ik}^2$ where j is the member of whole data set and k is the member of the class. Accordingly, alternative hypothesis is accepted if the p value of f_{vn} in the current model was smaller than the p value of f_{vn} in the previous model. In other words, while defining the weakest descriptor to leave the model within the "6-most significant descriptors", the descriptor with the maximum p value (failed to reject H_o with the greatest error) for one of the high or low classes is selected. As a result, the weakest descriptor is replaced by the strongest one. The strongest descriptor is described as the attribute whose maximum p value for high or low classes is the minimum among other descriptors. Figure 2 summarizes the main steps of the iterative algorithm used in this study. As the weakest and

Table 1. CYP Enzymes and R^2 Values with 6 and 15 Attributes

enzyme	PDB ID	number of ligands	R^2-6 pIC_{50}	R^2-6 BFE[a]	R^2-15 pIC_{50}	R^2-15 BFE[a]
CYP2D6	2F9Q	209	0.69	0.55	0.78	0.65
CYP1A2	2HI4	160	0.74	0.91	0.85	0.96
CYP3A4	1TQN	106	0.87	0.95	0.96	0.99
CYP2A6	1Z10	79	0.86	0.98	0.96	0.99
CYP2C9	1OG2	69	0.94	0.98	0.99	0.99
CYP2C8	1PQ2	58	0.92	0.94	0.96	0.98
CYP2C19	Model	62	0.93	0.97	0.99	0.99
CYP17	2C17	54	0.91	0.98	0.98	0.99

[a] Binding free energy.

the strongest descriptors were calculated by significance analysis, the weakest descriptors are replaced by the strongest ones, and the mixed integer linear programming based hyperboxes (MILP-HB) method[64] is used for the classification at each iteration. If the classification accuracy is not improved at the end of the iteration, algorithm stops and prediction accuracies are reported.

Classification accuracies are compared with the accuracy of other classification algorithms available in the WEKA[65] data mining package. The LIBSVM algorithm was integrated into the WEKA package using WLSVM.[66]

RESULTS AND DISCUSSION

Building Initial Models by PLSR. All human CYP enzymes with their PDB IDs and number of ligands used in our data are reported in Table 1. Regression models are constructed for both pIC_{50} and BFE values as responses separately with 6 and 15 attributes (Table 1). The models having fewer than 6 variables had poor R^2 values and were not considered as descriptive models. A small number of descriptors may lead to poor models, while a large number of descriptors may lead to inefficient models due to noninformative descriptors. The aim of the regression to have a 15-attributes model is that there can be some weak descriptors in a 6-attributes model; therefore, these weak descriptors can be replaced with strong ones from a 15-attribute model after the significance analysis. We define “strong descriptors” as input descriptors that increase the classification accuracies for both pIC_{50} and BFE. The “weak descriptors” are the descriptors that decrease the classification accuracy and leave the “6-significant descriptor” set.

MILP-HB Classification and Comparison with Other Classifiers. Iterations continue until the highest classification accuracy is reached. For the assessment of effectiveness of MILP-HB algorithm, the lowest, highest, and mean accuracies are given in Table 2. All runs were performed 10 times. The maximum difference between the highest and the lowest accuracy for pIC_{50} is 8%, whereas it is 5% for BFE. This deviation indicates that MILP based hyperboxes is an efficient and reliable method for data classification. MILP-HB classification is applied in two parts: training and testing. In the training part, data points that belong to a specific class are determined, and data points that belong to different classes are differentiated. Then, hyperboxes are constructed considering the bounds of the molecular descriptors. After the defining features of classes, the effectiveness of the classification is tested. All classification runs including algorithms in WEKA and MILP-HB were performed with 10-fold cross-validation for reliability.

Table 2. Lowest, Highest, and Mean Accuracies of the MILP-Hyperbox Method

enzyme	lowest % acc.	highest % acc.	mean % acc.
pIC_{50}			
CYP2D6	96.67	98.10	97.05
CYP1A2	90.63	94.38	91.94
CYP3A4	88.18	90.91	89.27
CYP2A6	87.50	90.00	88.50
CYP2C9	81.43	87.14	83.00
CYP2C8	78.33	86.67	81.67
CYP2C19	81.67	85.00	83.17
CYP17	80.00	88.00	82.80
BFE[a]			
CYP2D6	91.90	94.29	92.90
CYP1A2	91.88	93.13	92.44
CYP3A4	87.27	90.00	87.99
CYP2A6	85.00	87.50	86.25
CYP2C9	87.14	90.00	87.86
CYP2C8	78.33	83.33	79.83
CYP2C19	83.33	88.33	84.83
CYP17	92.00	94.00	92.60

[a] Binding free energy.

Table 3. Classification Accuracies for CYP2D6[b]

	% accuracy	
classifier	pIC_{50}	BFE[a]
MILP-Hyperbox	**97.05**	**92.90**
Bayes Network	51.19	59.33
Naïve Bayes	52.15	56.94
Naïve Bayes. Updatable	52.15	56.94
Liblinear	49.76	55.98
LibSVM	60.77	59.33
RBF Network	58.37	58.85
SMO	53.11	59.33
Logistic	58.85	55.50
IBk	66.51	55.50
Bagging	67.46	58.37
Ensemble Selection	67.94	54.55
Logit Boost	64.11	56.46
LMT	66.96	59.81
NBTree	69.38	59.33
Random Forest	70.33	57.42
DTNB	50.24	59.33
OneR	64.59	58.85

[a] Binding free energy. [b] Descriptors: PW2, nR06, MW, nBM, Wap, MLOGP.

Tables from 3 to 10 report the mean accuracies of classifications for all CYPs with their selected common descriptors, and the highest results were marked in bold. *In silico* prediction of *CYP2D6* is reported in Table 3 in which MILP-HB predicts pIC_{50} and BFE with an accuracy of 97.05% and 92.90%, respectively. Although the R^2 values

Table 4. Classification Accuracies for CYP1A2[b]

	% accuracy	
classifier	pIC_{50}	BFE[a]
MILP-Hyperbox	**91.94**	**92.44**
Bayes Network	53.70	82.10
Naïve Bayes	60.49	76.54
Naïve Bayes. Updatable	60.49	76.54
Liblinear	63.58	75.93
LibSVM	53.09	67.28
RBF Network	62.96	75.93
SMO	60.49	71.60
Logistic	64.81	74.69
IBk	66.05	77.16
Bagging	64.81	82.72
Ensemble Selection	65.43	81.48
Logit Boost	66.05	80.25
LMT	66.05	77.78
NBTree	62.35	80.86
Random Forest	74.07	83.95
DTNB	53.09	80.86
OneR	57.41	69.14

[a] Binding free energy. [b] Descriptors: JGI3, PW2, X4A, PJI2, E3s, J.

Table 5. Classification Accuracies for CYP3A4[b]

	% accuracy	
classifier	pIC_{50}	BFE[a]
MILP-Hyperbox	**89.27**	**87.99**
Bayes Network	67.92	81.13
Naïve Bayes	71.70	77.36
Naïve Bayes. Updatable	71.70	77.36
Liblinear	66.04	71.70
LibSVM	70.75	65.09
RBF Network	68.88	76.42
SMO	68.87	77.36
Logistic	70.75	83.02
IBk	65.09	75.47
Bagging	70.75	83.02
Ensemble Selection	71.70	83.96
Logit Boost	72.64	83.96
LMT	70.75	81.13
NBTree	67.92	80.19
Random Forest	68.81	81.13
DTNB	67.92	83.02
OneR	64.15	83.02

[a] Binding free energy. [b] Descriptors: D/Dr05, T(N..O), MW, G (N..N), GATS7e, PJI2.

Table 6. Classification Accuracies for CYP2A6[b]

	% accuracy	
classifier	pIC_{50}	BFE[a]
MILP-Hyperbox	**88.50**	86.25
Bayes Network	74.68	81.01
Naïve Bayes	75.95	77.22
Naïve Bayes. Updatable	75.95	77.22
Liblinear	69.62	84.81
LibSVM	77.22	86.08
RBF Network	73.42	89.87
SMO	72.15	75.95
Logistic	70.89	83.55
IBk	77.22	87.34
Bagging	74.68	87.34
Ensemble Selection	74.68	83.54
Logit Boost	77.22	88.61
LMT	73.42	83.54
NBTree	69.62	83.54
Random Forest	72.15	87.34
DTNB	75.95	**91.14**
OneR	73.42	79.75

[a] Binding free energy. [b] Descriptors: MLOGP, PCR, AROM, nBM, nR06, MAXDP.

Table 7. Classification Accuracies for CYP2C9[b]

	% accuracy	
classifier	pIC_{50}	BFE[a]
MILP-Hyperbox	**83.00**	**87.86**
Bayes Network	65.22	85.51
Naïve Bayes	63.77	85.51
Naïve Bayes. Updatable	63.77	85.51
Liblinear	65.22	84.06
LibSVM	68.12	68.12
RBF Network	59.42	82.61
SMO	65.22	85.51
Logistic	65.22	81.16
IBk	60.87	81.16
Bagging	59.42	84.06
Ensemble Selection	57.97	78.26
Logit Boost	60.87	86.96
LMT	63.77	82.61
NBTree	59.42	86.96
Random Forest	60.87	86.96
DTNB	65.22	85.51
OneR	53.62	84.06

[a] Binding free energy. [b] Descriptors: MW, SPAM, PW2, nBM, PW3, nR06.

for both pIC_{50} (0.69) and BFE (0.55) are relatively low, the classification accuracies are higher than those for other enzymes. This result may be connected to the number of ligands (209) which is the largest among the data sets. MILP-HB performs better than other classifiers in *CYP1A2* data summarized in Table 4. Similar to the results of *CYP2D6*, the *Random Forest* algorithm gives the second highest accuracy for pIC_{50} and BFE, 74.07% and 83.95% respectively. As demonstrated in Table 5, MILP-HB results with 89.27% accuracy for pIC_{50} and 87.99% for BFE, and the *LogistBoost* algorithm is the second classifier with a prediction accuracy of 72.64% and 83.96% for pIC_{50} and BFE, respectively. The accuracy of MILP-HB for pIC_{50} is 88.50%, the highest score for pIC_{50} among other algorithms; however, the accuracy for BFE is 86.25% by MILP-HB, and *DTNB* predicts BFE with an accuracy of 91.14% as shown in Table 6. Also, *DTNB* is the second most successful classifier after MILP-HB that predicts the pIC_{50} and BFE with an accuracy of 83.00% and 87.86%, respectively, in Table 7 listing the accuracies for *CYP2C9*. *DTNB* is a decision table/naive Bayes hybrid classifier, and Hall et al.[67] states that this combined model performs better compared to stand-alone naïve Bayes and decision tables. Table 8 gives the classification accuracies of different classifiers for *CYP2C8*. MILP-HB, with 81.67% accuracy for pIC_{50}, again performs better than other algorithms. Nevertheless, *RBF Network* predicts BFE of *CYP2C8* with 87.93% of accuracy, whereas MILP-HB reports an accuracy of 79.83%. *RBF Network*, radial basis function network, is an artificial neural network that uses the k-means clustering algorithm for the basis functions. *RBF Network* performs better to model the complex mappings, since it has efficient nonlinear approximation properties.[68] In Table 9, the accuracy of MILP-HB is 83.17% for pIC_{50} and 84.83% for BFE classification; however, the accuracy of *Bayesian Network* for BFE

Table 8. Classification Accuracies for CYP2C8[b]

classifier	% accuracy	
	pIC_{50}	BFE[a]
MILP-Hyperbox	**81.67**	79.83
Bayes Network	58.62	77.59
Naïve Bayes	70.69	82.76
Naïve Bayes. Updatable	70.69	82.76
Liblinear	60.34	75.86
LibSVM	60.34	51.72
RBF Network	63.79	**87.93**
SMO	60.49	71.60
Logistic	70.69	77.59
IBk	62.07	74.14
Bagging	65.52	72.41
Ensemble Selection	67.24	68.97
Logit Boost	60.34	81.03
LMT	72.41	77.59
NBTree	63.79	74.14
Random Forest	63.79	81.03
DTNB	58.62	79.31
OneR	60.34	72.41

[a] Binding free energy. [b] Descriptors: MW, MLOGP, Wap, nR06, ESpm01d, JGI8.

Table 9. Classification Accuracies for CYP2C19[b]

classifier	% accuracy	
	pIC_{50}	BFE[a]
MILP-Hyperbox	83.17	84.83
Bayes Network	78.69	**88.52**
Naïve Bayes	77.05	83.61
Naïve Bayes. Updatable	77.05	83.61
Liblinear	49.18	77.05
LibSVM	72.13	59.02
RBF Network	81.97	86.89
SMO	80.33	83.61
Logistic	80.33	86.89
IBk	72.13	80.33
Bagging	78.69	86.89
Ensemble Selection	75.41	80.33
Logit Boost	78.69	85.25
LMT	**83.61**	80.33
NBTree	75.41	86.89
Random Forest	75.41	83.61
DTNB	77.05	85.25
OneR	77.05	78.69

[a] Binding free energy. [b] Descriptors: PW2, J, Wap, Mor28u, PCR, SPAM.

classification is 88.52%. *Bayesian Networks* are directed acyclic graphs that represent the joint probability distribution over a set of random variables. Each vertex represents a random attribute, and each arc between nodes represents the probabilistic correlation.[69] Although *Bayesian Networks* is a consistent, smooth (robust), flexible (the same Bayesian network model can be used for different classification tasks), variable selection and casual relations is not trivial in *Bayesian Networks*. Table 10 demonstrates the classification accuracies of drugs for *CYP17*. MILP-HB gives the highest accuracy with 82.80% for pIC_{50} classification and gives 92.60% accuracy for BFE classification. However, the result of BFE classification with the *Bagging* method and the *LogitBoost* algorithm is 96.30%. The boosting procedure is the combination of weak classifiers to have strong classifiers. *LogitBoost* uses the *AdaBoost* procedure that is an additive logistic regression model. The *AdaBoost* procedure gives weights for the training samples and gives higher weights for misclassified samples. At each stage the linear combination of the classifiers is defined. An adaptive Newton-like algorithm is used in this algorithm.[70] Bagging (Bootstrap aggregating), like the boosting algorithm, uses an independent bootstrap procedure on the learning set to get an aggregated predictor with plurality voting. In Don et al.[71] the *Bagging* approach is compared with *SVMs* and *LogitBoost*. Although SVM classifier is characterized as less overfitting and uses hyperplanes to separate data optimally, it is quite hard to choose proper parameters. *LogitBoost* focuses on adjusting the misclassified data more; hence the accuracy on classifying other samples can decrease, and it can cause overfitting problem for some data sets. Bagging focuses on global accuracy and it is less overfitting. Due to the results in the article of Don et al.[71] *Bagging* and *LogitBoost* give very similar accuracies, like the similarity in our CYP17 results.

MILP based hyperboxes is not only an efficient algorithm for binary classification but also for multigroup classification since it employs hyperboxes to define the boundaries of the classes including all or some of the points in the set. When these boundaries of hyperboxes overlap (a region of attribute space is assigned to more than one class), there is a misclassification possibility of new data point. To eliminate this possibility, more than one hyperbox must be used to include all of the data points that belong to same class. It is seen that the MILP-HB approach gives higher accuracies compared to other classification algorithms almost for all data sets except for BFE predictions of *CYP2A6, CYP2C8, CYP2C19*, and *CYP17* and pIC_{50} prediction of *CYP2C19*. In addition, MILP-HB is simpler since it is free from parametric assumptions and no weights to be adjusted. The effectiveness of other classification methods usually depends on the parametric adjustment that is a big constraint in classification and regression problems. Nevertheless, MILP-HB needs more computational effort compared to other classification methods used in this study. We obtained results from WEKA in several seconds, whereas it takes 60−90 seconds on average to obtain accuracies with MILP-HB. In other words, a large data set is computationally challenging for MILP-HB, but the computational times are reasonable for all data sets considered in this paper. Therefore, preprocessing or feature reduction before MILP-HB classification is necessary to obtain quick results.

Interpretation of Selected Descriptors. Table 11 lists the selected molecular descriptors for all CYP enzymes. Todeschini et al.[72] states that good molecular descriptors should have a structural interpretation, a good correlation with at least one property, no trivial correlation with other molecular descriptors, and not restricted to a too small class of molecule. Therefore, different types of descriptors are preferred to build proper models in this study, and the classification accuracy is the main consideration for the selection of descriptors. *MW, nBM, nDB,* and *nR06* are selected as "constitutional descriptors" that are dependent basically on the composition of molecule rather than on geometry and topology. Since the active site of cytochrome P450s is the cavity region above the heme group, the size of substrates or inhibitors play an important role in penetration into the target site of the proteins. "Topological descriptors" identify the physicochemical properties and

Table 10. Classification Accuracies for CYP17[b]

classifier	% accuracy	
	pIC_{50}	BFE[a]
MILP-Hyperbox	**82.80**	92.60
Bayes Network	72.22	90.74
Naïve Bayes	70.37	90.74
Naïve Bayes. Updatable	70.37	90.74
Liblinear	74.07	92.59
LibSVM	66.67	88.89
RBF Network	74.07	90.74
SMO	70.37	94.44
Logistic	81.48	87.04
IBk	72.22	90.74
Bagging	72.22	**96.30**
Ensemble Selection	68.52	90.74
Logit Boost	66.67	**96.30**
LMT	74.07	94.44
NBTree	68.52	90.74
Random Forest	75.93	90.74
DTNB	64.81	92.59
OneR	59.26	92.59

[a] Binding free energy. [b] Descriptors: D/Dr05, nBM, JGI8, nCrs, PCR, nDB.

compound similarities in a quantitative manner. The values of this group of descriptors do not increase substantially with molecule size or number of rings. Topological descriptors are also one of the most widely used descriptors in QSAR studies. *Wap, J, MAXDP, PW2, PW3, PJI2, D/Dr05,* and *T(N..O)* are the topological descriptors used in this article. *WAP* is a Wiener index that is the sum of the number of edges in the shortest paths in a chemical graph between all pairs of non-hydrogen atoms in a molecule.[73] Balaban index *J* describes the size, composition, and branching of a molecule by representing the extended connectivity.[74] *MAXDP*[75] is the maximum positive Kier-Hall intrinsic state difference and can be considered as a measure of electrophilicity of the molecule. Since the shape and conformation has an effect on the binding of a ligand to an enzyme, descriptors such as *PW2, PW3,* and *PCR* ("walk and path count") and *PJI2 (*represents the topological anisometry) are responsible for enzyme-ligand interaction. Another descriptor *X4A* was selected as a *"connectivity index"*. It is known that molecular connectivity indices reflect the relative accessibility of each bond to encounter other bonds of the same molecule and have a role in biomolecular interactions.[76] *GATS7e* is the only descriptor from 2D "autocorrelation indices" that depends on the constitution and connectivity of the molecule but independent from the conformation. *Espm01d* is spectral moment 01 from "edge adjacency matrix" weighed by dipole moments. *JGI8* and *JGI3* are selected as "topological charge indices" identifying the charge transfer between atom pairs and global charge transfer in molecule. *SPAM* (statistical properties estimation of polymer chains considering short-range interactions), *AROM* (aromaticity), and *G(N..N)* were selected as "geometrical descriptors" reflecting the 3D structures of cytochrome inhibitors and substrates. *E3s* was selected from the "WHIM descriptor" block independent from molecular rotations and translations. WHIM descriptors use Cartesian coordinates to define 3-D chemical information in terms of size, shape, symmetry, atom distribution, and electrical properties. From the "molecular properties" block *MLOGP* which is the Moriguchi octanol−water partition coefficient and considered as a measure of lipophilicity of a drug was selected. *Mor28u* is selected as a "3D-Morse Descriptor" that is the distribution of different properties in the molecule and obtained by summing products of atomic properties weighed by different angular scattering functions but unweighted in this case. *nCrs* is a "functional group",

Table 11. Brief Explanation of the Most Significant Descriptors

symbol	description
PW2	path/walk 2 - Randic shape index
nR06	number of 6-membered rings
MW	molecular weight
Wap	all-path Wiener index
MLOGP	Moriguchi octanol−water partition coefficient
JGI3	mean topological charge index of order3
X4A	average connectivity index chi-4
PJI2	2D Petitjean shape index
E3s	third component accessibility directional WHIM index/weighted by atomic electrotopological states
J	Balaban distance connectivity index
D/Dr05	distance/detour ring index of order 5
T(N..O)	sum of topological distances between N..O
G (N..N)	sum of geometrical distances between N..N
GATS7e	Geary autocorrelation - lag 7/weighted by atomic Sanderson electronegativities
PCR	ratio of multiple path count over path count
AROM	aromaticity (trial)
nBM	number of multiple bonds
MAXDP	maximal electrotopological positive variation
SPAM	average span R
PW3	path/walk 3 - Randic shape index
ESpm01d	spectral moment 01 from edge adj. matrix weighted by dipole moments
JGI8	mean topological charge index of order8
Mor28u	3D-MoRSE - signal 28/unweighted
nCrs	number of ring secondary C(sp3)
nDB	number of double bonds

and it is number of ring secondary C(sp3) as a common variable for the classification of *CYP17* inhibitors.

CONCLUSIONS

The relationship between binding free energy obtained from simulations and experimental activity of drugs is one of the most important concerns in drug discovery and design. In this study, such a relationship is built by finding the significant common descriptors that is effective in predicting both pIC_{50} and binding free energy with a high accuracy. It is shown that these descriptors can be efficiently obtained by combining the partial least-squares regression with different type of classifiers. Among these classifiers, the MILP based hyperboxes method gives very accurate predictions for all data sets used in this work. Our novel method is applied to diverse chemical compounds bound to human P450 cytochrome enzymes that play an important role in drug metabolism, drug−drug interactions, and hormone synthesis. It is shown that our approach can be a practical and reliable tool in drug design for the direct prediction of the activity and binding free energy of drugs rather than generating models and correlations. This approach will not only enable to us reduce time but also will increase reliability of processes during the drug discovery. Besides, it can give insights into what structural characteristics are related to the inhibition of CYP enzymes which must be one of the main considerations in any new drug design of any biomolecular targets.

ACKNOWLEDGMENT

The computing hardware support from the IBM SUR Award is gratefully acknowledged. I.H.K is thankful for the support of the Turkish National Academy of Science of Turkey for the young investigator program (TUBA-GEBIP).

REFERENCES AND NOTES

(1) Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **2004**, *303* (5665), 1813–1818.

(2) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nat. Rev. Drug Discovery* **2004**, *3* (11), 935–949.

(3) Kubinyi, H. QSAR and 3D QSAR in drug design 0.1. methodology. *Drug Discovery Today* **1997**, *2* (11), 457–467.

(4) Yap, C. W.; Chen, Y. Z. Prediction of cytochrome p450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines. *J. Chem. Inf. Model.* **2005**, *45* (4), 982–992.

(5) Yao, X. J.; Liu, H. X.; Zhang, R. S.; Liu, M. C.; Hu, Z. D.; Panaye, A.; Doucet, J. P.; Fan, B. T. QSAR and classification study of 1,4-dihydropyridine calcium channel antagonists based on least squares support vector machines. *Mol. Pharmaceutics* **2005**, *2* (5), 348–356.

(6) Kahraman, P.; Turkay, M. Classification of 1,4-dihydropyridine calcium channel antagonists using the hyperbox approach. *Ind. Eng. Chem. Res.* **2007**, *46* (14), 4921–4929.

(7) Armutlu, P.; Ozdemir, M. E.; Uney-Yuksektepe, F.; Kavakli, I. H.; Turkay, M. Classification of drug molecules considering their IC50 values using mixed-integer linear programming based hyper-boxes method. *BMC Bioinf.* **2008**, *9*, -.

(8) Kontijevskis, A.; Komorowski, J.; Wikberg, J. E. S. Generalized proteochemometric model of multiple cytochrome P450 enzymes and their inhibitors. *J. Chem. Inf. Model.* **2008**, *48* (9), 1840–1850.

(9) Robertson, G. R.; Field, J.; Goodwin, B.; Bierach, S.; Tran, M.; Lehnert, A.; Liddle, C. Transgenic mouse models of human CYP3A4 gene regulation. *Mol. Pharmacol.* **2003**, *64* (1), 42–50.

(10) Arimoto, R. Computational models for predicting interactions with cytochrome p450 enzyme. *Curr. Top. Med. Chem.* **2006**, *6* (15), 1609–1618.

(11) Lynch, T.; Price, A. The effect of cytochrome P450 metabolism on drug response, interactions, and adverse effects. *Am. Fam. Physicia.* **2007**, *76* (3), 391–396.

(12) Liu, Y.; Yao, Z. X.; Papadopoulos, V. Cytochrome P450 17 alpha hydroxylase/17,20 lyase (CYP17) function in cholesterol biosynthesis: Identification of squalene monooxygenase (epoxidase) activity associated with CYP17 in Leydig cells. *Mol. Endocrinol.* **2005**, *19* (7), 1918–1931.

(13) Auchus, R. J.; Miller, W. L. Molecular modeling of human P450c17 (17 alpha-hydroxylase/17,20-lyase): Insights into reaction mechanisms and effects of mutations. *Mol. Endocrinol.* **1999**, *13* (7), 1169–1182.

(14) Schwede, T.; Kopp, J.; Guex, N.; Peitsch, M. C. SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res.* **2003**, *31* (13), 3381–3385.

(15) Prakash, C.; Kamel, A.; Cui, D.; Whalen, R. D.; Miceli, J. J.; Tweedie, D. Identification of the major human liver cytochrome P450 isoform(s) responsible for the formation of the primary metabolites of ziprasidone and prediction of possible drug interactions. *Br. J. Clin. Pharmacol.* **2000**, *49*, 35s–42s.

(16) Rahnasto, M.; Raunio, H.; Poso, A.; Wittekindt, C.; Juvonen, R. O. Quantitative structure-activity relationship analysis of inhibitors of the nicotine metabolizing CYP2A6 enzyme. *J. Med. Chem.* **2005**, *48* (2), 440–449.

(17) Korhonen, L. E.; Rahnasto, M.; Mahonen, N. J.; Wittekindt, C.; Poso, A.; Juvonen, R. O.; Raunio, H. Predictive three-dimensional quantitative structure-activity relationship of cytochrome P450 1A2 inhibitors. *J. Med. Chem.* **2005**, *48* (11), 3808–15.

(18) Chohan, K. K.; Paine, S. W.; Mistry, J.; Barton, P.; Davis, A. M. A rapid computational filter for cytochrome P450 1A2 inhibition potential of compound libraries. *J. Med. Chem.* **2005**, *48* (16), 5154–61.

(19) Asikainen, A. H.; Ruuskanen, J.; Tuppurainen, K. A. Alternative QSAR models for selected estradiol and cytochrome P450 ligands: comparison between classical, spectroscopic, CoMFA and GRID/GOLPE methods. *SAR QSAR Environ. Res.* **2005**, *16* (6), 555–65.

(20) Moon, T.; Chi, M. H.; Kim, D. H.; Yoon, C. N.; Choi, Y. S. Quantitative structure-activity relationships (QSAR) study of flavonoid derivatives for inhibition of cytochrome P450 1A2. *Quant. Struct-Act. Rel.* **2000**, *19* (3), 257–263.

(21) Fischer, V.; Johanson, L.; Heitz, F.; Tullman, R.; Graham, E.; Baldeck, J. P.; Robinson, W. T. The 3-hydroxy-3-methylglutaryl coenzyme A reductase inhibitor fluvastatin: Effect on human cytochrome P-450 and implications for metabolic drug interactions. *Drug Metab. Dispos.* **1999**, *27* (3), 410–416.

(22) Nicolas, J. M.; Whomsley, R.; Collart, P.; Roba, J. In vitro inhibition of human liver drug metabolizing enzymes by second generation antihistamines. *Chem-Biol. Interact.* **1999**, *123* (1), 63–79.

(23) Chun, Y. J.; Ryu, S. Y.; Jeong, T. C.; Kim, M. Y. Mechanism-based inhibition of human cytochrome p450 1A1 by rhapontigenin. *Drug Metab. Dispos.* **2001**, *29* (4), 389–393.

(24) Walsky, R. L.; Gaman, E. A.; Obach, R. S. Examination of 209 drugs for inhibition of cytochrome p450 2C8. *J. Clin. Pharmacol.* **2005**, *45* (1), 68–78.

(25) Nakamura, T.; Kakinuma, H.; Umemiya, H.; Amada, H.; Miyata, N.; Taniguchi, K.; Bando, K.; Sato, M. Imidazole derivatives as new potent and selective 20-HETE synthase inhibitors. *Bioorg. Med. Chem. Lett.* **2004**, *14* (2), 333–336.

(26) Turpeinen, M.; Uusitalo, J.; Jalonen, J.; Pelkonen, A. Multiple P450 substrates in a single run: rapid and comprehensive in vitro interaction assay (vol 24, pg 123, 2005). *Eur. J. Pharm. Sci.* **2005**, *24* (4), 389–389.

(27) Obach, R. S.; Walsky, R. L.; Venkatakrishnan, K.; Gaman, E. A.; Houston, J. B.; Tremaine, L. M. The utility of in vitro cytochrome P450 inhibition data in the prediction of drug-drug interactions. *J. Pharmacol. Exp. Ther.* **2006**, *316* (1), 336–348.

(28) McKillop, D.; Back, D. J.; McCormick, A. D.; Evans, J. A.; Tjia, J. Preclinical and in vitro assessment of the potential of D0870, an antifungal agent, for producing clinical drug interactions. *Xenobiotica* **1999**, *29* (4), 395–408.

(29) Vickers, A. E. M.; Zollinger, M.; Dannecker, R.; Tynes, R.; Heitz, F.; Fischer, V. In vitro metabolism of tegaserod in human liver and intestine: Assessment of drug interactions. *Drug Metab. Dispos.* **2001**, *29* (10), 1269–1276.

(30) Taavitsainen, P.; Juvonen, R.; Pelkonen, O. In vitro inhibition of cytochrome P450 enzymes in human liver microsomes by a potent CYP2A6 inhibitor, trans-2-phenylcyclopropylamine (tranylcypromine), and its nonamine analog, cyclopropylbenzene. *Drug Metab. Dispos.* **2001**, *29* (3), 217–222.

(31) Cohen, L. H.; Remley, M. J.; Raunig, D.; Vaz, A. D. N. In vitro drug interactions of cytochrome P450: An evaluation of fluorogenic to conventional substrates. *Drug Metab. Dispos.* **2003**, *31* (8), 1005–1015.

(32) Moody, G. C.; Griffin, S. J.; Mather, A. N.; McGinnity, D. F.; Riley, R. J. Fully automated analysis of activities catalysed by the major human liver cytochrome P450(CYP) enzymes: assessment of human CYP inhibition potential. *Xenobiotica* **1999**, *29* (1), 53–75.

(33) Grimm, S. W.; Dyroff, M. C. Inhibition of human drug metabolizing cytochromes P450 by anastrozole, a potent and selective inhibitor of aromatase. *Drug Metab. Dispos.* **1997**, *25* (5), 598–602.

(34) Sai, Y.; Dai, R.; Yang, T. J.; Krausz, K. W.; Gonzalez, F. J.; Gelboin, H. V.; Shou, M. Assessment of specificity of eight chemical inhibitors using cDNA-expressed cytochromes P450. *Xenobiotica* **2000**, *30* (4), 327–343.

(35) Riley, R. J.; Parker, A. J.; Trigg, S.; Manners, C. N. Development of a generalized, quantitative physicochemical model of CYP3A4 inhibition for use in early drug discovery. *Pharm. Res.* **2001**, *18* (5), 652–655.

(36) Asano, T.; Kushida, H.; Sadakane, C.; Ishihara, K.; Wakui, Y.; Yanagisawa, T.; Kimura, M.; Kamei, H.; Yoshida, T. Metabolism of ipecac alkaloids cephaeline and emetine by human hepatic microsomal cytochrome P450s, and their inhibitory effects on P450 enzyme activities. *Biol. Pharm. Bull.* **2001**, *24* (6), 678–682.

(37) Obach, R. S. Inhibition of human cytochrome P450 enzymes by constituents of St. John's wort, an herbal preparation used in the treatment of depression. *J. Pharmacol. Exp. Ther.* **2000**, *294* (1), 88–95.

(38) Hutzler, J. M.; Walker, G. S.; Wienkers, L. C. Inhibition of cytochrome P450 2D6: Structure-activity studies using a series of quinidine and quinine analogues. *Chem. Res. Toxicol.* **2003**, *16* (4), 450–459.

(39) Walker, D. K.; Alabaster, C. T.; Congrave, G. S.; Hargreaves, M. B.; Hyland, R.; Jones, B. C.; Reed, L. J.; Smith, D. A. Significance of metabolism in the disposition and action of the antidysrhythmic drug, dofetilide - In vitro studies and correlation with in vivo data. *Drug Metab. Dispos.* **1996**, *24* (4), 447–455.

(40) Fonnepfister, R.; Meyer, U. A. Xenobiotic and Endobiotic Inhibitors of Cytochrome-P-450dbl Function, the Target of the Debrisoquine Sparteine Type Polymorphism. *Biochem. Pharmacol.* **1988**, *37* (20), 3829–3835.

(41) von Moltke, L. L.; Greenblatt, D. J.; Granda, B. W.; Giancarlo, G. M.; Duan, S. X.; Daily, J. P.; Harmatz, J. S.; Shader, R. I. Inhibition of human cytochrome P450 isoforms by nonnucleoside reverse transcriptase inhibitors. *J. Clin. Pharmacol.* **2001**, *41* (1), 85–91.

(42) Jones, B. C.; Hyland, R.; Ackland, M.; Tyman, C. A.; Smith, D. A. Interaction of terfenadine and its primary metabolites with cytochrome P450 2D6. *Drug Metab. Dispos.* **1998**, *26* (9), 875–882.

(43) Yu, J. L.; Paine, M. J. I.; Marechal, J. D.; Kemp, C. A.; Ward, C. J.; Brown, S.; Sutcliffe, M. J.; Roberts, G. C. K.; Rankin, E. M.; Wolf, C. R. In silico prediction of drug binding to CYP2D6: Identification of a new metabolite of metoclopramide. *Drug Metab. Dispos.* **2006**, *34* (8), 1386–1392.

(44) Kemp, C. A.; Flanagan, J. U.; van Eldik, A. J.; Marechal, J. D.; Wolf, C. R.; Roberts, G. C. K.; Paine, M. J. I.; Sutcliffe, M. J. Validation of model of cytochrome p450 2D6: An in silico tool for predicting metabolism and inhibition. *J. Med. Chem.* **2004**, *47* (22), 5340–5346.

(45) Venhorst, J.; Onderwater, R. C. A.; Meerman, J. H. N.; Commandeur, J. N. M.; Vermeulen, N. P. E. Influence of N-substitution of 7-methoxy-4-(aminomethyl)-coumarin on cytochrome P450 metabolism and selectivity. *Drug Metab. Dispos.* **2000**, *28* (12), 1524–1532.

(46) Shader, R. I.; Granda, B. W.; von Moltke, L. L.; Giancarlo, G. M.; Greenblatt, D. J. Inhibition of human cytochrome P450 isoforms in vitro by zafirlukast. *Biopharm. Drug Dispos.* **1999**, *20* (8), 385–388.

(47) Vaz, R. J.; Nayeem, A.; Santone, K.; Chandrasena, G.; Gavai, A. V. A 3D-QSAR model for CYP2D6 inhibition in the aryloxypropanolamine series. *Bioorg. Med. Chem. Lett.* **2005**, *15* (17), 3816–3820.

(48) Nnane, I. P.; Kato, K.; Liu, Y.; Long, B. J.; Lu, Q.; Wang, X.; Ling, Y. Z.; Brodie, A. Inhibition of androgen synthesis in human testicular and prostatic microsomes and in male rats by novel steroidal compounds. *Endocrinology* **1999**, *140* (6), 2891–2897.

(49) Nnane, I. P.; Njar, V. C. O.; Liu, Y.; Lu, Q.; Brodie, A. M. H. Effects of novel 17-azolyl compounds on androgen synthesis in vitro and in vivo. *J. Steroid Biochem.* **1999**, *71* (3−4), 145–152.

(50) Pelkonen, O.; Maenpaa, J.; Taavitsainen, P.; Rautio, A.; Raunio, H. Inhibition and induction of human cytochrome P450 (CYP) enzymes. *Xenobiotica* **1998**, *28* (12), 1203–1253.

(51) Handratta, V. D.; Vasaitis, T. S.; Njar, V. C. O.; Gediya, L. K.; Kataria, R.; Chopra, P.; Newman, D.; Farquhar, R.; Guo, Z. Y.; Qiu, Y.; Brodie, A. M. H. Novel C-17-heteroaryl steroidal CYP17 inhibitors/antiandrogens: Synthesis, in vitro biological activity, pharmacokinetics, and antitumor activity in the LAPC4 human prostate cancer xenograft model. *J. Med. Chem.* **2005**, *48* (8), 2972–2984.

(52) Hu, Q. Z.; Negri, M.; Jahn-Hoffmann, K.; Zhuang, Y.; Olgen, S.; Bartels, M.; Muller-Vieira, U.; Lauterbach, T.; Hartmann, R. W. Synthesis, biological evaluation, and molecular modeling studies of methylene imidazole substituted biaryls as inhibitors of human 17 alpha-hydroxylase-17,20-lyase (CYP17) - Part II: Core rigidification and influence of substituents at the methylene bridge. *Bioorg. Med. Chem.* **2008**, *16* (16), 7715–7727.

(53) Marechal, J. D.; Yu, J. L.; Brown, S.; Kapelioukh, I.; Rankin, E. M.; Wolf, C. R.; Roberts, G. C. K.; Paine, M. J. I.; Sutcliffe, M. J. In silico and in vitro screening for inhibition of cytochrome P450CYP3A4 by comedications commonly used by patients with cancer. *Drug Metab. Dispos.* **2006**, *34* (4), 534–538.

(54) *Marvin, version 4.1.7*; ChemAxon: Hungary, 2005.

(55) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **2005**, *26* (16), 1781–1802.

(56) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **1998**, *102* (18), 3586–3616.

(57) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19* (14), 1639–1662.

(58) Lutz, M. *Programming Python*; O'Reilly & Associates, Inc.: Sebastopol, CA, 1996.

(59) *The Open Babel Package, 2.0.1*. http://openbabel.sourceforge.net (accessed Jan 31, 2009).

(60) VCCLAB, Virtual Computational Chemistry Laboratory. http://www.vcclab.org (accessed Oct 1, 2008-Jan 31, 2009).

(61) Whitley, D. C.; Ford, M. G.; Livingstone, D. J. Unsupervised forward selection: A method for eliminating redundant variables. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (5), 1160–1168.

(62) Wold, S.; Sjostrom, M.; Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab.* **2001**, *58* (2), 109–130.

(63) *MINITAB Statistical Software, version 14*; Minitab: PA, 2003.

(64) Uney, F.; Turkay, M. A mixed-integer programming approach to multiclass data classification problem. *Eur. J. Oper. Res.* **2006**, *173* (3), 910–920.

(65) Weka: Waikato Environment for Knowledge Analysis; University of Waikato, New Zealand. http://www.cs.waikato.ac.nz/ml/weka/ (accessed Jan 6, 2009).

(66) EL-Manzalawy, Y. H. V. *WLSVM: Integrating LibSVM into Weka Environment*, 2005.

(67) Hall, M.; Frank, E. In *Combining Naive Bayes and Decision Tables*; Proceedings of the 21st Florida Artificial Intelligence Society Conference (FLAIRS), 2008.

(68) Poggio, T.; Girosi, F. In *Networks for approximation and learning*; Proc. IEEE, 1990; pp 1481−1497.

(69) Friedman, N.; Geiger, D.; Goldszmidt, M. Bayesian network classifiers. *Mach. Learn.* **1997**, *29* (2−3), 131–163.

(70) Friedman, J.; Hastie, T.; Tibshirani, R. Additive logistic regression: A statistical view of boosting. *Ann. Stat.* **2000**, *28* (2), 337–374.

(71) Dong, L. H.; Yuan, Y.; Cai, Y. D. Using bagging classifier to predict protein domain structural class. *J. Biomol. Struct. Dyn.* **2006**, *24* (3), 239–242.

(72) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; WILEY-VCH: Weinheim, 2000.

(73) Wiener, H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* **1947**, *69* (1), 1720) .

(74) Balaban, A. T.; Ciubotariu, D.; Medeleanu, M. Topological Indexes and Real Number Vertex Invariants Based on Graph Eigenvalues or Eigenvectors. *J. Chem. Inf. Comput. Sci.* **1991**, *31* (4), 517–523.

(75) Kier, L. B.; Hall, L. H.; Frazer, J. W. An Index of Electrotopological State for Atoms in Molecules. *J. Math. Chem.* **1991**, *7* (1−4), 229–241.

(76) Kier, L. B.; Hall, L. H. The meaning of molecular connectivity: A bimolecular accessibility model. *Croat. Chem. Acta* **2002**, *75* (2), 371–382.

CI900247T