ADME Evaluation in Drug Discovery. 8. The Prediction of Human Intestinal Absorption by a Support Vector Machine

Tingjun Hou*

Department of Chemistry and Biochemistry, Center for Theoretical Biological Physics, University of California at San Diego, La Jolla, California 92093

Junmei Wang

Encysive Pharmaceuticals, Inc., 7000 Fannin St., Houston, Texas 77030

Youyong Li

Materials and Process Simulation Center, California Institute of Technology, Pasadena, California 91125

Received June 12, 2007

Human intestinal absorption (HIA) is an important roadblock in the formulation of new drug substances. In silico models for predicting the percentage of HIA based on calculated molecular descriptors are highly needed for the rapid estimation of this property. Here, we have studied the performance of a support vector machine (SVM) to classify compounds with high or low fractional absorption (%FA > 30% or %FA ≤ 30%). The analyzed data set consists of 578 structural diverse druglike molecules, which have been divided into a 480-molecule training set and a 98-molecule test set. Ten SVM classification models have been generated to investigate the impact of different individual molecular properties on %FA. Among these studied important molecule descriptors, topological polar surface area (TPSA) and predicted apparent octanol−water distribution coefficient at pH 6.5 ($\log D_{6.5}$) show better classification performance than the others. To obtain the best SVM classifier, the influences of different kernel functions and different combinations of molecular descriptors were investigated using a rigorous training-validation procedure. The best SVM classifier can give satisfactory predictions for the training set (97.8% for the poor-absorption class and 94.5% for the good-absorption class). Moreover, 100% of the poor-absorption class and 97.8% of the good-absorption class in the external test set could be correctly classified. Finally, the influence of the size of the training set and the unbalanced nature of the data set have been studied. The analysis demonstrates that large data set is necessary for the stability of the classification models. Furthermore, the weights for the poor-absorption class and the good-absorption class should be properly balanced to generate unbiased classification models. Our work illustrates that SVMs used in combination with simple molecular descriptors can provide an extremely reliable assessment of intestinal absorption in an early in silico filtering process.

INTRODUCTION

High potency is not the exclusive factor for an efficacious drug. Other essential properties, especially absorption, distribution, metabolism, excretion and toxicity (ADMET), are also extremely important in the process of drug discovery. According to the analysis of the failed new chemical entities (NCEs), the leading causes of failures (∼50%−60%) are poor ADMET properties and adverse effects, which contribute significantly more than a "lack of efficacy" (∼30%).[1] The high failure rate of drug candidates at the late stage, which is due to the unfavorable ADMET properties, drives the shift of the drug discovery process from the "serial" diagram to the "parallel" diagram. Using the parallel diagram, it is expected that the efficacy, the selectivity, and the comprehensive ADMET properties can be assessed at the same stage. A recent analysis in 2000 shows that the attrition rate in the adverse pharmacokinetic and bioavailability aspects was significantly improved from ∼40% in 1991 to ∼10% in 2000. However, the attrition rate in the clinical safety issue and toxicology increased from ∼20% to ∼30%. Overall, the dropout of development candidates due to improper ADME/formulation, toxicology, and safety issues was decreased from ∼60% in 1991 to ∼45% in 2000. Recent developments in automation technology and experimental ADMET techniques also impel the applications of the "parallel" strategy, such as the Caco-2 permeability screening based on the three-day Caco-2 culture system,[2] high-throughput (HT) kinetic solubility assay,[3,4] and metabolic stability screening using microsomes or hepatocytes,[5] and liquid chromatography−mass spectroscopy (LCMS) and fluorogenic assays through cytochrome CYP inhibition for metabolism related to drug−drug interactions.[6,7] Undoubtedly, at the current stage, the HT technologies for ADMET assays can only be applied to limited properties. Another important issue that may affect the wide application of the "parallel" strategy is the poor predictivity of the HT assays caused by the intrinsic nature of the technology. For instance, HT kinetic solubility[3,4] or fluorescence-based CYP inhibition platforms[6,7] is significantly distinct from thermodynamic

* Corresponding author. E-mail: tingjunhou@hotmail.com.

10.1021/ci7002076 CCC: $37.00 © 2007 American Chemical Society
Published on Web 10/12/2007

solubility or LCMS-based CYP inhibition that have long been considered to be the industrial gold standards. The bottleneck of the current HT ADMET experimental techniques drives the need for development of in silico tools.[8−11] In silico approaches have great potentials to predict in vitro and in vivo properties and have advantages in that they save time and no experiments are required.

Although various routes are used for drug administration, oral dosing is the preferred one both in clinical and outpatient practices. For an orally administered drug, a high and stable bioavailability is indeed important for the successful development of a drug candidate. Oral bioavailability of a drug is related to many factors, such as dissolution in the gastrointestinal tract, intestinal membrane permeation, and intestinal/hepatic first-pass metabolism. The relationships between bioavailability and intestinal absorption show that the bioavailabilities of most compounds (64%) were primarily controlled by the intestinal absorption process.[12,13] Therefore, the prediction of intestinal absorption is the first step toward the prediction of human oral bioavailability. For an orally administrated drug across the intestinal epithelium, there are two important routes for permeation: passive diffusion and carrier-mediated influx via active transport mechanisms. It is assumed that, with only few exceptions, these orally administered drugs were transported across the intestinal epithelium predominantly by a passive transcellular process. Establishing a good in silico prediction model for intestinal absorption can greatly facilitate the progress of drug discovery programs. Several correlation and classification models have been developed to predict human intestinal absorption (HIA) by applying a variety of statistical and machine-leaning approaches.[10] A big problem for the prediction of HIA is that many published models were generated based on a small number of compounds (20−40), with only several exceptions;[14−16] thus, some of the models are only applicable to the limited ranges and are not statistically reliable. We have previously reported a large HIA database for 648 compounds, which is much larger than the other available databases.[13] The large database gives a solid basis for the development of reliable prediction models. Another important factor responsible for the quality of prediction models is the statistical methods to build the models. In response to increased accuracy demands, some machine-leaning methods, such as artificial neural networks (ANNs), genetic algorithms (GAs), and support vector machine (SVM),[10] are now being applied to the analysis of ADME data. In these past several years, SVM has attracted much attention and gained extensive applications, because of its remarkable generalization performance.[17−25] In many cases, SVMs have been observed to be consistently superior to other supervised learning methods.[23−25] Xue and co-workers have reported the application of SVM on the prediction of HIA.[18] In Xue's work, they classified a small data set of 198 molecules into chemical agents absorbable (HIA+) or nonabsorbable (HIA−) classes, using a %FA of 70% as criterion. Obviously, the classes they designed were unreasonable, according to the viewpoint of drug discovery, and the obtained prediction models cannot be applied as a practical filter. In the present investigation, SVM was used for the prediction of HIA based on the large and diverse data set using several simple molecular descriptors calculated from the molecular structure alone. The objective of this study was to establish a new and accurate classification SVM model and to confirm the possibility of predicting drug absorption based on molecular structures. The prediction results are extremely satisfactory in both training set and test set compounds, which proved that SVM was a powerful tool in the prediction of HIA.

METHODS AND MATERIALS

1. Data Set and Molecular Descriptors. The data set previously reported by us was used here.[13] In this data set, 648 chemical compounds were reported, among which 579 molecules were believed to be transported by passive diffusion. Here, 579 molecules were used for modeling passive transcellular absorption, without explicitly considering other factors. In our previous data set, gentamicin was indicated twice, and one duplicate was eliminated from the data set. Thus, in the current work, the entire data set used for model development includes 578 organic compounds. A %FA value of 30% was used as the criterion for dividing chemical agents into the poor-absorption (HIA−) and the good-absorption (HIA+) classes. The entire data set was divided into a training set of 480 molecules and a test set of 98 molecules, respectively. The HIA− and HIA+ classes have 78 and 500 compounds, respectively. Thus, the two classes are highly unbalanced. The experimental %FA values and the optimized three-dimensional (3-D) structures can be downloaded from the supporting website (http://modem.ucsd.edu/adme).

2. Molecular Descriptors. In the construction of classification models, 10 molecular descriptors were considered, including the topological polar surface area (TPSA), the octanol−water partitioning coefficient (log$P$), the apparent partition coefficient at pH 6.5 (log$D_{6.5}$), the number of violations of the four Rule-of-Five rules developed by Lipinski ($N_{\text{rule-of-5}}$),[3] the number of hydrogen bond donors (n_{HBD}), the number of hydrogen bond acceptors (n_{HBA}), the intrinsic solubility (log$S$), the number of rotatable bonds (n_{rot}), the molar volume (MV), and the molecular weight (MW). TPSA was calculated using the parameters originally proposed by Ertl et al.,[26] which was developed to calculate the polar surface area of a molecule based on its two-dimensional (2-D) molecular bonding information. The apparent partition coefficient, log$D$, was estimated based on the predicted log$P$ and pK_a calculated by ACDLABS 9.0.[27] ACDLABS can the predicted log$D$ values at pH values of 2, 5.5, 6.5, 7.4, and 10. Considering the best correlation between log$D$ and %FA was obtained at pH 6.5,[13] log$D$ at pH 6.5 was used here. The intrinsic solubility log S is the solubility for the neutral form of compounds. The molecular descriptors were calculated using ACDLABS 9.0.[27] It is well-known that molecules with positively charged N atoms always have very low %FA values. Therefore, here, a binary indicator, N+, was used to represent the existence of the positively charged N atom. If positively charged nitrogen was found in the molecule, N+ was defined to be 1; otherwise, N+ was defined to be 0. In the data set, 26 compounds were determined to have at least one positively charged N atom.

3. Support Vector Machine (SVM) Analysis. The SVM technique was applied to build the classification models of HIA. In the following text, we give a brief introduction of SVM. More-detailed descriptions can be found in articles

by Vapnik and Burges.[28−31] SVM originated as an implementation of Vapnik's Structural Risk Minimization (SRM) principle from statistical learning theory. A special property of SVMs is that they simultaneously minimize the empirical classification error and maximize the geometric margin; hence, they are also known as maximum margin classifiers. In linearly separable cases, SVM constructs two parallel hyperplanes on each side of the maximal separating hyperplane that maximizes the distance between the two parallel hyperplanes. An assumption is made that the larger the margin or distance between these parallel hyperplanes, the better the generalization error of the classifier will be. In our case, a vector corresponds to a chemical compound, and this vector is represented by $\mathbf{x}_i$, with molecular descriptors of this compounds as its components. This is done by finding a vector $\mathbf{w}$ and a parameter b that minimizes $||w||^2$ and satisfies the following conditions:

$$\mathbf{w}\cdot\mathbf{x}_i + b \geq +1 \quad \text{(for } y_i = +1\text{; class 1 (positive samples))} \tag{1a}$$

$$\mathbf{w}\cdot\mathbf{x}_i + b \leq -1 \quad \text{(for } y_i = -1\text{; class 2 (negative samples))} \tag{1b}$$

where y_i is the class index and $\mathbf{w}$ is a vector normal to the separating hyperplane. After the determination of $\mathbf{w}$ and b, a given vector $\mathbf{x}_i$ can be classified by

$$\text{sign}(\mathbf{w}\cdot\mathbf{x} + b) \tag{2}$$

The dual form of the SVM can be shown to be

$$\max \sum_{i=1}^{l} \alpha_i - \left(\frac{1}{2}\right)\sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i\alpha_j c_i c_j x_i^{\mathrm{T}} x_j \tag{3}$$

where the α terms constitute a dual representation for the weight vector in terms of the training set:

$$\mathbf{w} = \sum_{i=1}^{l} \alpha_i c_i x_i \tag{4}$$

The problem now is to minimize $^1/_2||w||^2$, subject to the conditions of eq 1. In practice, there often is no separting hyperplane that can split the data into either positive or negative samples. A soft margin method was then suggested by introducing a slack variable, ξ_i, which measures the degree of misclassification of the datum x_i:

$$c_i(\mathbf{w}\cdot\mathbf{x}_i - b) \geq 1 - \xi_i \qquad (\text{for } 1 \leq i \leq n) \tag{5}$$

The objective function then becomes a tradeoff between a large margin and a small error penalty. For a linear penalty function, the primal function $^1/_2||w||^2$ transforms to eq 6, where $C > 0$ is the penalty parameter of the error term:

$$\min_{w,b,\xi} \frac{1}{2} w^{\mathrm{T}} w + C \sum_{y_i=1}^{\xi_{i<}} \tag{6}$$

$$\text{s.t. } y_i(w^{\mathrm{T}} x_i + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0 \qquad (\text{for } i = 1, ..., l)$$

$$\min_{w,b,\xi} \frac{1}{2} w^{\mathrm{T}} w = C \sum_{y_i=1} \xi_i \tag{7}$$

$$\text{s.t. } y_i(w^{\mathrm{T}}\phi(x_i) + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0 \qquad (\text{for } i = 1, ..., l)$$

In nonlinearly separable cases, SVM maps the input variable X_i into a high-dimensional feature space using the function ϕ, and then eq 6 becomes eq 7. $K(x_i,x_j) = \phi(x_i)^{\mathrm{T}}\phi(x_j)$ is called the kernel function. There are four usually used kernel functions: linear, polynomial, the radial basis function (RBF), and sigmoid kernel. Generally, RBF is a reasonable first choice. After solving eq 7, SVM finds a linear separating hyperplane with the maximal margin in this higher dimensional space.

Here, the LIBSVM software developed by Chang and Lin was used for SVM analysis.[32] The quality of the SVM classification was measured by the quantity of true positives (*TP*), true negatives (*TN*), false positives (*FP*), false negatives (*FN*), sensitivity (which is the prediction accuracy for positive examples in this work and is denoted as *SE*), and specificity (which is the prediction accuracy for negative examples in this work and is denoted as *SP*). The prediction accuracy for the HIA+ and HIA− classes ($Q_{\mathrm{HIA+}}$ and $Q_{\mathrm{HIA-}}$) and the Matthews correlation coefficient (*C*) were also used to measure the prediction accuracies. The *C* values range from 0 to 1. A perfect prediction would lead to $C = 1$.

$$SE = \frac{TP}{TP + FN} \tag{8}$$

$$SP = \frac{TN}{TN + FP} \tag{9}$$

$$Q_{\mathrm{HIA+}} = \frac{TP}{TP + FP} \tag{10}$$

$$Q_{\mathrm{HIA-}} = \frac{TN}{TN + FN} \tag{11}$$

$$C = \frac{TP \times TN - FN \times FP}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}} \tag{12}$$

The SVM classifiers were extensively validated by two types of tests. First, the entire training set of 480 molecules was randomly divided into two subsets. The first subset consists of two-thirds of the samples (49 HIA− compounds and 271 HIA+ compounds) in the training set, and the second subset consists of the other samples in the training set (24 HIA− compounds and 136 HIA+ compounds). The second subset was used for the validation of the SVM classifier developed based on the first subset. The training−validation process was iteratively executed for 1000 times. This rigorous training−validation procedure was also applied for the selection of the best kernel function. Second, the SVM classifiers were further validated by the external test set of 98 molecules.

RESULTS AND DISCUSSIONS

1. SVM Classification Models Using Individual Molecular Descriptors. First, 10 classification models only

Table 1. Performance of the SVM Models Using Individual Molecular Descriptors

No.	descriptor	SE^{train}_{HIA+} [a]	SP^{train}_{HIA-} [a]	SE^{test}_{HIA+} [b]	SP^{test}_{HIA+} [b]	*SE*	*SP*	Q_{HIA+}	Q_{HIA-}	*C*
1	TPSA	93.2	82.0	93.1	81.4	93.1	81.7	97.8	57.9	0.645
2	$\log D_{6.5}$	93.8	73.8	93.6	72.2	93.7	73.0	96.8	57.3	0.601
3	N_{HBA}	91.7	76.7	91.6	74.0	91.7	75.4	97.0	55.1	0.568
4	N_{HBD}	92.1	70.5	91.8	69.6	92.0	70.1	96.4	50.2	0.537
5	$\log P$	93.5	59.0	93.5	57.0	93.5	58.0	95.1	50.8	0.486
6	$N_{rule-of-5}$	88.7	67.7	88.1	63.7	88.4	65.7	95.7	39.5	0.437
7	MW	94.6	41.8	94.5	38.7	94.5	40.3	93.2	46.0	0.370
8	N_{rot}	94.4	40.3	94.2	38.0	94.3	39.2	93.1	44.4	0.354
9	$\log S$	91.7	35.2	91.6	34.3	91.6	34.8	92.4	32.5	0.256
10	MV	86.7	27.7	86.1	21.6	86.4	24.7	90.9	17.3	0.095

[a] SE^{train}_{HIA+} and SP^{train}_{HIA-} are the average prediction accuracies for the HIA+ samples and the HIA− samples of the 1000 training groups. [b] SE^{test}_{HIA+} and SP^{test}_{HIA+} are the average prediction accuracies for the HIA+ samples and the HIA− samples of the 1000 validation groups.

using each individual descriptor were constructed. In SVM analysis, the RBF kernel function was used. Considering that the contribution of the positively charged nitrogen cannot be properly considered by these 10 molecular descriptors, the 26 compounds with positively charged N atoms were not included in the analysis. The 1000 times of training−validation procedure was applied to each SVM classifier by randomly splitting the 455 molecules in to a training group (24 HIA− and 203 HIA+) and a validation group (23 HIA− and 204 HIA+). The 10 SVM classifiers were ranked according to the average of 1000 *C* values (see Table 1). Overall, the averaged prediction accuracies for the training group are slightly better than those of the validation group.

Among the 10 studied molecular descriptors, TPSA shows the best classification performance. The SVM model using TPSA can correctly identify 93.1% of the HIA+ compounds and 81.4% of the HIA− compounds for the validated compounds. It was not a surprising result at all. Since van de Waterbeemd and Kansy correlated the polar surface area (PSA) of a series of CNS drugs to blood-brain partitioning first in 1992,[33] it has become the most popular parameter for the prediction of molecular transport properties.[13,34−37] The earlier work reported by us shows the correlation between TPSA and %FA ($r = -0.70$) is better than those between %FA and the other important molecular descriptors.[13] PSA or TPSA was assumed to be related with the hydrogen-bonding capability, and thus can account for the electrostatic interaction between drug molecules and the intestine. However, TPSA obviously performs better than the other two hydrogen-bonding descriptors (n_{HBD} and n_{HBA}). In regard to n_{HBD} and n_{HBA}, it seems that n_{HBA} is a little more important than n_{HBD} because the SVM classifier based on n_{HBA} is marginally better.

The performance of the apparent partition coefficient, $\log D_{6.5}$, is slightly worse than TPSA, but better than that of the other eight descriptors. Using this descriptor, the SVM model can correctly identify 93.6% of the HIA+ compounds and 72.7% of the HIA− compounds in the validation sets. The hydrophobic parameter, $\log D$, has long been known to be important for membrane permeation.[13,35,37] The octanol−water partitioning coefficient, $\log P$, was also usually used in the prediction of absorption or permeability, and it is even more popular than $\log D$, because $\log P$ can be precisely computed using atomic or fragment-addition approaches.[10] Compared with $\log P$, $\log D_{6.5}$ can give a better classification. We believe that, for partition processes in the body, the distribution coefficient $\log D$ for which there is an aqueous buffer at pH 6.5 provides a more meaningful description of lipophilicity, especially for ionizable compounds.

A interesting finding is that the $N_{rule-of-5}$ parameter, which is defined as the number of violations of the Rule of Five proposed by Lipinski,[3] does not show good performance for classification. Using $N_{rule-of-5}$, 11.9% of the HIA+ molecules and 36.3% of the HIA− molecules in the validation sets could not be correctly classified. Obviously, $N_{rule-of-5}$ performs worse than TPSA, $\log D_{6.5}$, N_{HBD}, N_{HBA}, or even $\log P$. This observation is somewhat strange, because the Rule of Five was widely applied to identify compounds with possible poor absorption and permeability, but our finding might reflect a fact that the impact of the Rule of Five usually may be overestimated. The classification analysis reported here is consistent with our previous observation.[13] In our recent work, the $N_{rule-of-5}$ classification was made by simply defining a cutoff value of 2 for $N_{rule-of-5}$ to be poorly absorbed or well absorbed. Compared with the performance of TPSA, the criterion of ≥ 2 is less reliable for identifying poorly absorbed molecules from the others.

Compared to the other six descriptors, the performances of MW, N_{rot}, $\log S$, and MV did not show good classification capabilities, indicated by the low Matthews correlation coefficients ($C < 0.4$). Among these four descriptors, two of them—MW and MV—are used to define the bulkiness properties of a molecule. Furthermore, the descriptor N_{rot}, which is used to define the number of rotatable bonds of a molecule, is also indirectly related to molecular bulkiness. The bulkiness property of a molecule is generally considered to be an important predictor of absorption, but the use of the bulkiness descriptor alone obviously cannot give an effective classification for HIA.

2. Classification SVM Models Using Multiple Molecular Descriptors. The aforementioned discussions demonstrate that HIA is primarily controlled by two molecular properties: TPSA and $\log D_{6.5}$. Some earlier studies show that a good prediction model for absorption usually is related to both of PSA (or TPSA) and $\log D$.[37−39] Other molecular properties certainly may also contribute to HIA. Therefore, we expect to develop a reliable SVM classification model using multiple molecular descriptors.

Recently, we have developed a RP classification model based on four molecular descriptors ($\log D_{6.5}$, TPSA, N_{HBD}, and MW) and one binary indicator (N+), using the same data set we used here. To give a direct comparison with the previous work, we first constructed the SVM classifiers, only considering these five descriptors. The 26 compounds with

Table 2. Performance of the SVM Classifiers, Using Different Kernel Functions and Different Molecular Descriptors

No.	descriptors	kernel	SE^{train}_{HIA+}	SP^{train}_{HIA-}	SE^{test}_{HIA+}	SP^{test}_{HIA+}	SE	SP	Q_{HIA+}	Q_{HIA-}	C
6	log$D_{6.5}$, TPSA, N+	linear	96.3	93.2	96.3	90.7	96.3	92.3	98.3	81.3	0.832
7	log$D_{6.5}$, TPSA, N+	polynomial	96.2	93.5	96.2	93.0	96.2	93.3	98.7	81.1	0.844
8	log$D_{6.5}$, TPSA, N+	RBF	96.6	92.3	96.5	91.0	96.6	91.9	98.4	82.0	0.838
9	log$D_{6.5}$, TPSA, N+	sigmoid	97.2	89.0	97.2	88.0	97.2	88.7	97.9	84.5	0.838
10	log$D_{6.5}$, TPSA, n_{HBD}, N+	linear	96.6	94.4	96.4	92.4	96.5	93.8	98.6	81.9	0.845
11	log$D_{6.5}$, TPSA, n_{HBD}, N+	polynomial	95.9	94.7	95.8	93.9	95.9	94.4	98.9	79.7	0.840
12	log$D_{6.5}$, TPSA, n_{HBD}, N+	RBF	96.9	94.0	96.8	93.2	96.8	93.8	98.8	83.7	0.862
13	log$D_{6.5}$, TPSA, n_{HBD}, N+	sigmoid	97.3	92.1	97.2	90.3	97.3	91.5	98.3	85.0	0.854
14	log$D_{6.5}$, TPSA, MW, N+	linear	96.3	95.4	96.2	92.2	96.3	94.3	98.6	81.0	0.838
15	log$D_{6.5}$, TPSA, MW, N+	polynomial	96.8	94.5	96.7	94.0	96.8	94.3	98.9	83.3	0.863
16	log$D_{6.5}$, TPSA, MW, N+	RBF	97.0	92.9	96.8	91.1	96.9	92.3	98.4	83.3	0.847
17	log$D_{6.5}$, TPSA, MW, N+	sigmoid	97.8	89.3	97.7	87.4	97.7	88.7	97.8	86.9	0.849
18	log$D_{6.5}$, TPSA, n_{HBD}, MW, N+	linear	97.1	94.9	96.9	91.9	97.0	93.9	98.5	84.1	0.857
19	log$D_{6.5}$, TPSA, n_{HBD}, MW, N+	polynomial	96.5	95.7	96.4	94.7	96.5	95.4	99.0	82.4	0.861
20	log$D_{6.5}$, TPSA, n_{HBD}, MW, N+	RBF	97.4	94.8	97.3	93.4	97.4	94.3	98.8	85.7	0.875
21	log$D_{6.5}$, TPSA, n_{HBD}, MW, N+	sigmoid	97.8	97.1	97.8	89.3	97.8	90.9	98.1	87.8	0.865
22	log$D_{6.5}$, TPSA, n_{HBD}, MW, $N_{rule-of-5}$, MV, N+	RBF	97.4	94.7	97.3	93.5	97.4	94.3	99.0	86.6	0.886

a positively charged N atom, which were not used in the aforementioned analysis, were also included in the following calculations. Considering the importance of TPSA and log$D_{6.5}$, they were used in all SVM classifiers.

For SVM, given the data set, the proper kernel function must be chosen to construct the best classifier.[32] This selection is very important, because the kernel function determines the sample distribution in the mapping space. Unfortunately, there are no successful theoretical techniques for determining the optimal kernel function and its parameters, so all of the four kernel functions supported by LIBSVM were tested in all cases. Each SVM classifier was developed and validated using the 1000 times of training−validation procedure introduced in the Methods and Materials section. The SVM classifiers, using different kernel functions and different molecular descriptors, are summarized in Table 2. Indicated by the average Matthews correlation coefficients, the linear kernel function performs worst in three cases. The performances of the other three kernel functions do not show a large difference. Another observation of our calculations is that the SVM classifier using three descriptors—log$D_{6.5}$, TPSA, and N+ (models 6−9 in Table 2)—is much better than those using one descriptor alone. By introducing two other descriptors (n_{HBD} and MW), the models will be improved obviously (models 10−21). Based only on the average Matthews correlation coefficients, model 20, which is based on the RBF kernel function, could be identified as the best model, which is based on five molecular descriptors. The overall correctness of classification of model 20 is very good, being 97.4% (the HIA+ class) and 94.8% (the HIA− class) for the training sets and 97.3% (the HIA+ class) and 93.4% (the HIA− class) for the validation sets.

We then tried to develop SVM classifiers by adding other molecular descriptors. We designed a systematic search to find the best combination of molecular descriptors. During the systematic search, all the descriptors used in model 20 were kept and then the other 5 molecular descriptors (N_{rot}, N_{HBA}, MV, $N_{rule-of-5}$, and log$S$) were added systematically. We then could generate 31 new models with 6−10 descriptors.

Figure 1 shows the distribution of the average Matthews correlation coefficients of the validation tests from the 1000 times of the training−validation process for these 31 new models. For most cases, the addition of new molecular descriptors could not improve the statistics of the SVM classifiers. For several cases, the qualities of the models could be improved slightly. According to the Matthews correlation coefficients, model 22 in Table 2 with seven molecular descriptors (N_{HBD}, log$D_{6.5}$, MW, MV, TPSA, $N_{rule-of-5}$, and N+) could be identified as the best.

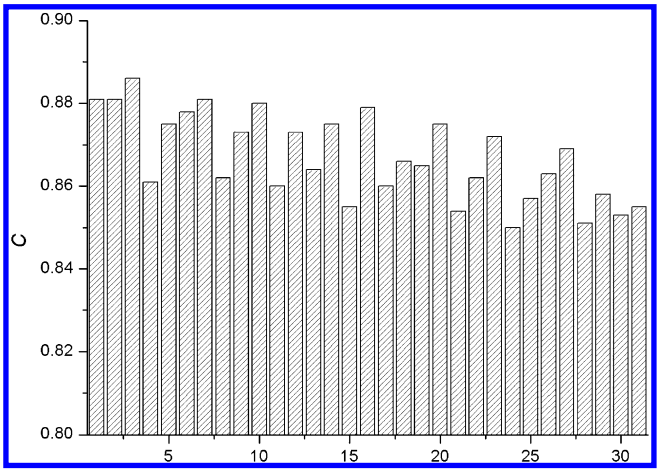

Figure 1. Distribution of the Matthews correlation coefficients (C) for the 31 support vector machine (SVM) classifiers generated by the systematic search.

One thing we need to mention is that, when we use the RBF kernel function, two parameters (C and γ) can usually affect the classification. It is not known beforehand which C and γ values are the best for one problem. To determine the best values of C and γ, a grid search based on 10-fold cross-validation was applied. $C = 32$ and $\gamma = 0.00781$ were suggested by the grid search. The final SVM classifier was then regenerated using all 480 molecules in the training set. The prediction of the %FA value seems to be really encouraging: prediction accuracies for the HIA+ class are 97.8% (398/407) and 94.5% (69/73), respectively. The SVM classifier was further tested by the predictions on the external test set of 98 molecules. The external test set includes 5 compounds in the HIA− class and 93 compounds in the HIA+ class. The performance of the SVM classifier on the test set is also very impressive. All 5 HIA− compounds were correctly classified, and only 2 HIA+ compounds were incorrectly classified.

We then examined the error sources of the final SVM classifier and found 15 compounds in the entire data set that

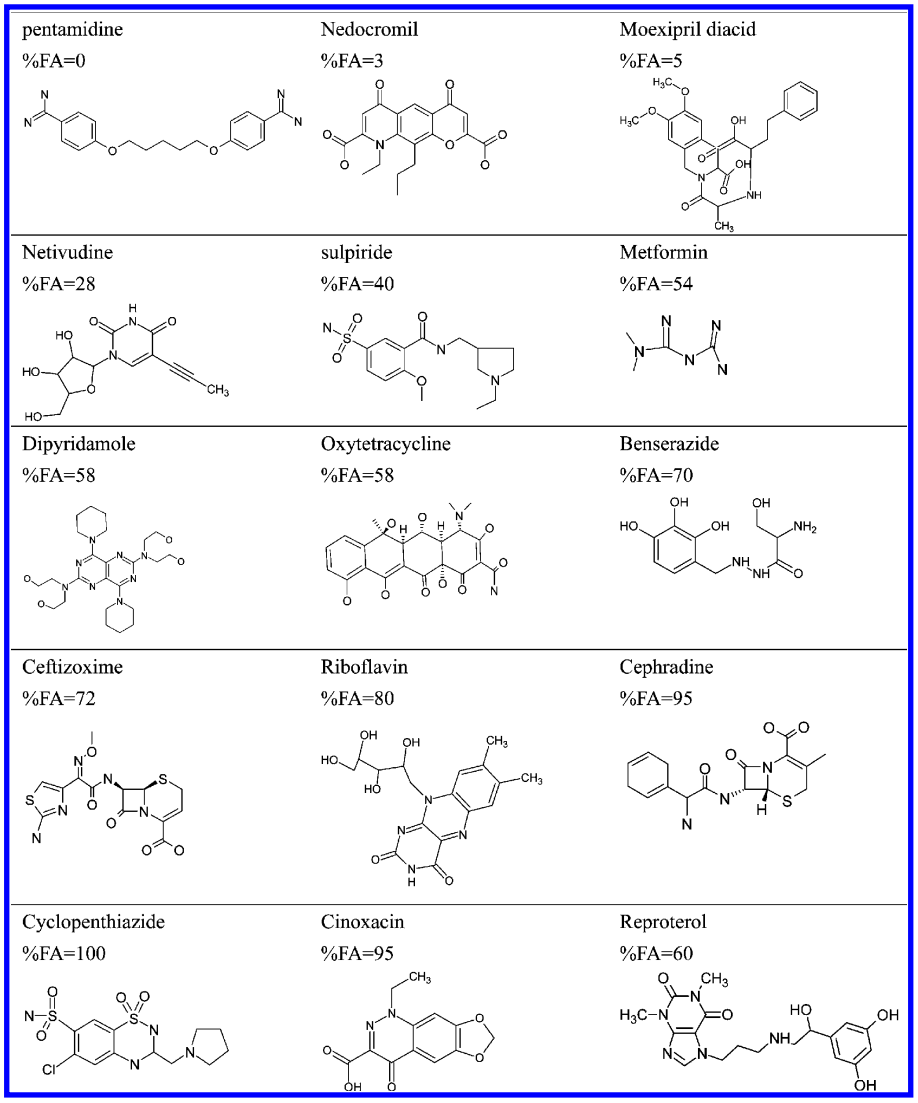

Figure 2. Structures of the 15 misclassified compounds by the best SVM classifier.

could not be correctly classified. The 2-D structures and the experimental %FA values for the misclassified compounds are shown in Figure 2. Analyzing the reasons of the misclassification of these compounds may give us valuable information for the further improvement of our models. We believe that three reasons may lead to the misclassification. The first possible explanation of misclassification may be the experimental error in the FA% values obtained from the literature. According to experiments, it is not rare that the experimental %FA values from different sources are inconsistent. Generally, the deviation observed for the experimental %FA could be as large as 20%.[13] Among these 15 misclassified compounds, the %FA values for netivudine, sulpiride, and metformin are 28%, 40%, and 54%, respectively. Therefore, it is understandable that these compounds are easily misclassified. The second possible explanation is that the apparent partition coefficient values for some compounds could not be correctly predicted theoretically. We reported a comparison between the experimental and predicted log*D* values.[37] The log*D* values for most compounds could be satisfactorily predicted, but some compounds still showed large prediction errors. The prediction errors of log*D* can directly lead to misclassification. For example, the predicted $\log D_{6.5}$ for cinoxacin is −3.8. However, according to the experiments,[40] log*D* at pH 7.4 for cinoxacin is −2.1. Considering the titration curve given by ACDLABS, the experimental log*D* at pH 6.5 for cinoxacin should be equal to or even larger than −2.1. If we used the experimental value of −2.1 in the prediction, this compound could be correctly classified into the HIA+ class. The third possible explanation is that some compounds shown in Figure 2 are not merely transported by passive diffusion, and some other routes may be involved. We expect that the readers of our paper can give more clues about the reasons for misclassification.

We have reported a recursive partitioning (RP) classification model, using the same data set that we used here. Therefore, it is fair to make a direct comparison of the performances of RP and SVM. The RP model is very predictive, indicated by the satisfactory predictions on the training (461/480) and test sets (95/98). The improvement of the RP model is not easy, because its prediction capability is quite good. However, compared to the RP model, the SVM classifier still performs better (correctly identifying 97.8% of the compounds in the HIA− class and 94.5% of the compounds in the HIA+ class). Moreover, the predictions for the test set using the SVM classifier is also better than those using the RP model. The total number of misclassified number was decreased from 22 of RP to 15 of SVM. Therefore, the SVM classifier gives more reliable predictions

than the RP model either based on the prediction for the training set or that for the test set.

3. Importance of the Large Data Set. The most important precondition for a good classification model is the reliable data set. Both the reliability of the data and the size of the data set are important. A large data set can guarantee the good statistics of the obtained models. The models derived from a small data set usually do not have good extendibility. Here, we can give a simple test of the influence of the size of the data set on the predictivity of the model. Niwa et al. applied a probabilistic neural network to generate a classification model for %HIA.[41] In his work, 67 molecules (11 HIA− and 56 HIA+) were used in the training set, 9 molecules (3 HIA− and 6 HIA+) were used as the test set for controlling the training, and 10 molecules (4 HIA− and 6 HIA+) were used as the prediction test, testing the actual classification capability. Here, we randomly selected a small training set (76 molecules) as the same number as that used in Niwa's work to construct the SVM classification model,[41] and the obtained model was applied to predict the absorption for the other molecules. The training−validation process was iteratively conducted 1000 times. The average accuracies for the HIA+ class and the HIA− class in the training set are 97.5% and 93.3%, respectively. The average accuracy for the HIA+ class in the test set is 97.3%, whereas that for the HIA− class in the test set is 86.8%. Using a small training set, the actual prediction clearly cannot be well-guaranteed, even if a very "reliable" model is constructed.

Another thing we want to highlight is that the data set we used is quite unbalanced. The degree of this type of unbalance is directly related to the criterion used to define the HIA+ and HIA− classes. Here, a %FA value of 30%, which was used by van de Waterbeemd and Kansy,[33] as the cutoff. Usually, a %FA value of 10% is also used in some cases.[34] In Xue et al.'s work, a cutoff value of 70% was used,[18] and in Niwa's work, a cutoff value of 50% was used.[41] Values of 50% or 70% certainly are not reasonable cutoffs for classification in practice. The cutoff value for distinguishing low absorption from high absorption is somewhat arbitrary, because the deviation observed for the experimental %FA was large. In our data set, only four compounds have %FA values of 20%−30%, so the predictions based on cutoffs of 20% or 30% do not have a large difference. Note that, because of the nature of the high unbalance of our data set, different penalty parameters should be applied to different classes. In LIBSVM, when we use different penalty parameters, eq 7 becomes

$$\min_{w,b,\xi} \frac{1}{2} w^{\mathrm{T}} w + C_+ \sum_{y_i=1} \xi_i + C_- \sum_{y_i=-1} \xi_i$$
$$\text{s.t. } y_i(w^{\mathrm{T}}\phi(x_i) + b) \geq 1 - \xi_i$$
$$\xi_i \geq 0 \qquad (\text{for } i = 1, ..., l) \tag{13}$$

where $C_+ = k_+C$ and $C- = k_-C$ are the penalty parameters of error terms for the (+) class and the (−) class, respectively; k_+ and k_- are the weight parameters for the (+) and (−) classes.

In the current work, considering the ratio of the compounds in the HIA+ class and those in the HIA− class, a weight parameter of 5.5 was used. That is to say, the penalty parameters for the HIA+ and HIA− classes are C and 5.5C, respectively. For an unbalanced database, the definition of different penalty parameters for different classes is quite important. The importance of the application of different penalty parameters was investigated by a simple calculation: instead of using different weight parameters, the same weight parameters were used for both classes. According to the calculations, the average accuracy for the HIA+ class of the training set is quite good (99.4%), but that for the HIA− class of the test set is not so good (73.1%). It is similar for the test set, as indicated by the high accuracy that is observed for the HIA+ class (99.3%) and the low accuracy that is observed for the HIA− class (72.3%). It is quite obvious that the SVM classifier is strongly biased to the HIA+ class when the same weight parameters were used for the different classes.

CONCLUSION

In this work, taking the advantage of the high-quality database, we developed a reliable support vector machine (SVM) model to discriminate the well-absorbed compounds and the poorly absorbed compounds. The comparisons of the classification models based on individual molecular descriptors show that two molecular descriptors (topological polar surface area (TPSA) and predicted apparent octanol−water distribution coefficient at pH 6.5 ($\log D_{6.5}$)) are essential for the prediction of intestinal absorption. The best SVM classifiers, based on seven molecular descriptors, demonstrate extremely good predictivity, as indicated by the high prediction accuracies for the training set and the test set. Our work gives solid evidence that the passive diffusion of intestinal absorption can be well-predicted by simple molecular descriptors. Moreover, SVM exhibited better performance than RP, indicating that the classification model based on SVM will have great applications on human intestinal absorption (HIA) predictions.

ACKNOWLEDGMENT

T.H. is supported by a CTBP postdoctoral scholarship.

Supporting Information Available: The molecular descriptors of 578-molecule data set are listed in Table S1 (PDF). This material also is available free of charge via the Internet at http://pubs.acs.org. The molecular 3-D structures and the experimental %FA values of the data set can be downloaded from the supporting website (http://modem.ucsd.edu/adme).

REFERENCES AND NOTES

(1) Kennedy, T. Managing the drug discovery/development interface. *Drug Discovery Today* **1997**, *2*, 436−444.
(2) Li, A. P. Screening for human ADME/Tox drug properties in drug discovery. *Drug Discovery Today* **2001**, *6*, 357−366.
(3) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.
(4) Bevan, C. D.; Lloyd, R. S. A high-throughput screening method for the determination of aqueous drug solubility using laser nephelometry in microtiter plates. *Anal. Chem.* **2000**, *72*, 1781−1787.
(5) Wienkers, L. C.; Heath, T. G. Predicting in vivo drug interactions from in vitro drug discovery data. *Nat. Rev. Drug Discovery* **2005**, *4*, 825−833.
(6) Crespi, C. L.; Miller, V. P.; Penman, B. W. Microtiter plate assays for inhibition of human, drug-metabolizing cytochromes P450. *Anal. Biochem.* **1997**, *248*, 188−190.

(7) 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*, 1005−1015.
(8) Beresford, A. P.; Segall, M.; Tarbit, M. H. In silico prediction of ADME properties: Are we making progress? *Curr. Opin. Drug Discovery* **2004**, *7*, 36−42.
(9) Hou, T. J.; Xu, X. J. Recent development and application of virtual screening in drug discovery: An overview. *Curr. Pharm. Des.* **2004**, *10*, 1011−1033.
(10) Hou, T. J.; Wang, J. M.; Zhang, W.; Wang, W.; Xu, X. Recent advances in computational prediction of drug absorption and permeability in drug discovery. *Curr. Med. Chem.* **2006**, *13*, 2653−2667.
(11) van de Waterbeemd, H.; Gifford, E. ADMET in silico modelling: Towards prediction paradise? *Nat. Rev. Drug Discovery* **2003**, *2*, 192−204.
(12) Hou, T. J.; Wang, J. M.; Zhang, W.; Xu, X. J. ADME evaluation in drug discovery. 6. Can oral bioavailability in humans be effectively predicted by simple molecular property-based rules? *J. Chem. Inf. Model.* **2007**, *47*, 460−463.
(13) Hou, T. J.; Wang, J. M.; Zhang, W.; Xu, X. J. ADME evaluation in drug discovery. 7. Prediction of oral absorption by correlation and classification. *J. Chem. Inf. Model.* **2007**, *47*, 208−218.
(14) Zhao, Y. H.; Le, J.; Abraham, M. H.; Hersey, A.; Eddershaw, P. J.; Luscombe, C. N.; Boutina, D.; Beck, G.; Sherborne, B.; Cooper, I.; Platts, J. A. Evaluation of human intestinal absorption data and subsequent derivation of a quantitative structure-activity relationship (QSAR) with the Abraham descriptors. *J. Pharm. Sci.* **2001**, *90*, 749−784.
(15) Wessel, M. D.; Jurs, P. C.; Tolan, J. W.; Muskal, S. M. Prediction of human intestinal absorption of drug compounds from molecular structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 726−735.
(16) Klopman, G.; Stefan, L. R.; Saiakhov, R. D. ADME evaluation. 2. A computer model for the prediction of intestinal absorption in humans. *Eur. J. Pharm. Sci.* **2002**, *17*, 253−263.
(17) Franke, L.; Byvatov, E.; Werz, O.; Steinhilber, D.; Schneider, P.; Schneider, G. Extraction and visualization of potential pharmacophore points using support vector machines: Application to ligand-based virtual screening for COX-2 inhibitors. *J. Med. Chem.* **2005**, *48*, 6997−7004.
(18) Xue, Y.; Li, Z. R.; Yap, C. W.; Sun, L. Z.; Chen, X.; Chen, Y. Z. Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1630−1638.
(19) 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*, 982−992.
(20) Jorissen, R. N.; Gilson, M. K. Virtual screening of molecular databases using a Support Vector Machine. *J. Chem. Inf. Model.* **2005**, *45*, 549−561.
(21) Byvatov, E.; Fechner, U.; Sadowski, J.; Schneider, G. Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1882−1889.
(22) Warmuth, M. K.; Liao, J.; Ratsch, G.; Mathieson, M.; Putta, S.; Lemmen, C. Active learning with support vector machines in the drug discovery process. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 667−673.
(23) Brown, M. P. S.; Grundy, W. N.; Lin, D.; Cristianini, N.; Sugnet, C. W.; Furey, T. S.; Ares, M.; Haussler, D. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Natl. Acad. Sci., U.S.A.* **2000**, *97*, 262−267.
(24) Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Comput. Chem.* **2001**, *26*, 5−14.
(25) Plewczynski, D.; Spieser, S. A. H.; Koch, U. Assessing different classification methods for virtual screening. *J. Chem. Inf. Model.* **2006**, *46*, 1098−1106.
(26) Ertl, P.; Rohde, B.; Selzer, P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.* **2000**, *43*, 3714−3717.
(27) ACDLABS v9.0, http://www.acdlabs.com, 2005.
(28) Vapnik, V.; Chapelle, O. Bounds on error expectation for support vector machines. *Neural Comput.* **2000**, *12*, 2013−2036.
(29) Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273−297.
(30) Scholkopf, B.; Sung, K. K.; Burges, C. J. C.; Girosi, F.; Niyogi, P.; Poggio, T.; Vapnik, V. Comparing support vector machines with Gaussian kernels to radial basis function classifiers. *IEEE Trans. Signal Process.* **1997**, *45*, 2758−2765.
(31) Burges, C. J. C. A tutorial on Support Vector Machines for pattern recognition. *Data Min. Knowl. Discovery* **1998**, *2*, 121−167.
(32) Chang, C. C.; Lin, C. J. LIBSVM: a library for support vector machine. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm, 2001.
(33) Vandewaterbeemd, H.; Kansy, M. Hydrogen-Bonding Capacity and Brain Penetration. *Chimia* **1992**, *46*, 299−303.
(34) Palm, K.; Stenberg, P.; Luthman, K.; Artursson, P. Polar molecular surface properties predict the intestinal absorption of drugs in humans. *Pharmaceut. Res.* **1997**, *14*, 568−571.
(35) Egan, W. J.; Merz, K. M.; Baldwin, J. J. Prediction of drug absorption using multivariate statistics. *J. Med. Chem.* **2000**, *43*, 3867−3877.
(36) Hou, T. J.; Xu, X. J. ADME evaluation in drug discovery. 3. Modeling blood-brain barrier partitioning using simple molecular descriptors. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2137−2152.
(37) Hou, T. J.; Zhang, W.; Xia, K.; Qiao, X. B.; Xu, X. J. ADME evaluation in drug discovery. 5. Correlation of Caco-2 permeation with simple molecular properties. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1585−1600.
(38) Linnankoski, J.; Makela, J. M.; Ranta, V. P.; Urtti, A.; Yliperttula, M. Computational prediction of oral drug absorption based on absorption rate constants in humans. *J. Med. Chem.* **2006**, *49*, 3674−3681.
(39) Winiwarter, S.; Bonham, N. M.; Ax, F.; Hallberg, A.; Lennernas, H.; Karlen, A. Correlation of human jejunal permeability (in vivo) of drugs with experimentally and theoretically derived parameters. A multivariate data analysis approach. *J. Med. Chem.* **1998**, *41*, 4939−4949.
(40) Austin, R. P.; Barton, P.; Cockroft, S. L.; Wenlock, M. C.; Riley, R. J. The influence of nonspecific microsomal binding on apparent intrinsic clearance, and its prediction from physicochemical properties. *Drug Metab. Dispos.* **2002**, *30*, 1497−1503.
(41) Niwa, T. Using general regression and probabilistic neural networks to predict human intestinal absorption with topological descriptors derived from two-dimensional chemical structures. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 113−119.

CI7002076