Estimation of ADME Properties with Substructure Pattern Recognition

Jie Shen, Feixiong Cheng, You Xu, Weihua Li,* and Yun Tang*

Department of Pharmaceutical Sciences, School of Pharmacy, East China University of Science and Technology, 130 Meilong Road, Shanghai 200237, China

Received March 16, 2010

Over the past decade, absorption, distribution, metabolism, and excretion (ADME) property evaluation has become one of the most important issues in the process of drug discovery and development. Since in vivo and in vitro evaluations are costly and laborious, in silico techniques had been widely used to estimate ADME properties of chemical compounds. Traditional prediction methods usually try to build a functional relationship between a set of molecular descriptors and a given ADME property. Although traditional methods have been successfully used in many cases, the accuracy and efficiency of molecular descriptors must be concerned. Herein, we report a new classification method based on substructure pattern recognition, in which each molecule is represented as a substructure pattern fingerprint based on a predefined substructure dictionary, and then a support vector machine (SVM) algorithm is applied to build the prediction model. Therefore, a direct connection between substructures and molecular properties is built. The most important substructure patterns can be identified via the information gain analysis, which could help to interpret the models from a medicinal chemistry perspective. Afterward, this method was verified with two data sets, one for blood-brain barrier (BBB) penetration and the other for human intestinal absorption (HIA). The results demonstrated that the overall predictive accuracies of the best HIA model for the training and test sets were 98.5 and 98.8%, and the overall predictive accuracies of the best BBB model for the training and test sets were 98.8 and 98.4%, which confirmed the reliability of our method. In the additional validations, the predictive accuracies were 94 and 69.5% for the HIA and the BBB models, respectively. Moreover, some of the representative key substructure patterns which significantly correlated with the HIA and BBB penetration properties were also presented.

1. INTRODUCTION

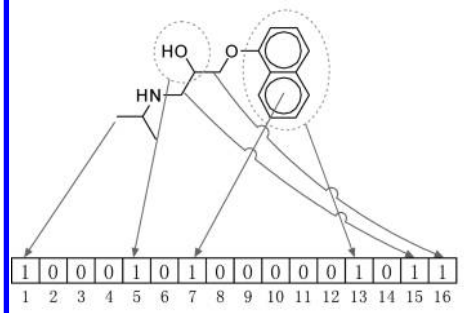
Nowadays, a consensus has been reached from both industries and academics that the importance of the absorption, distribution, metabolism, and excretion (ADME) properties for potential drug candidates is no less than their efficacy and specificity. Screening and optimizing ADME properties in the early stage of the drug development process are widely accepted.[1] In recent years, the attrition cost due to poor ADME properties in pharmaceutical industries has been greatly decreased.[2] However, experimental evaluation of ADME properties can still not meet the demands of lead discovery and optimization due to the time and cost effectiveness. Although numerous in vitro assay methods have been developed and improved over years,[3] the throughput capacity is nothing compared to the high-throughput activity assay and combinatorial synthesis, thus become a "bottleneck" in the process of drug discovery. Using in silico methods to evaluate the ADME properties has become a practicable alternative choice so far, which could break through the "bottleneck" in the high-throughput drug discovery process.[4−6]

To date, a large number of in silico ADME models have been developed, and increasing numbers of associated papers have been published.[1,5] Traditional methods tried to construct certain relationships between molecular descriptors and ADME properties using statistical methods or machine learning algorithms,[7] which include multiple linear regression (MLR),[8] partial least-squares (PLS),[9,10] genetic algorithms (GA),[11,12] artificial neural network (ANN),[13,14] decision tree (DT),[15] *k* nearest-neighbor (*k*NN),[16,17] and support vector machine (SVM).[16−19] Among them, SVM is a superior method which has been also widely applied in the drug discovery methods recently.[20] SVM is based on the theory of structural risk minimization (SRM), which covers both empirical risk and Vapnik−Cortes (VC) confidence. This means that reliable models could be built even in the condition of lack of samples.[21] Therefore, it is possible to obtain the models with better generalization ability through limited training samples using SVM. However, the interpretable ability is a drawback for SVM models in spite of their more accurate prediction ability than the models built via other methods. Such SVM models perform as a "black box", and only the prediction results can be provided.

Molecular description may be another potential problem in the model building. In traditional models, each molecule is represented as a vector with different types of molecular descriptors, which are usually numerical values that characterize properties of molecules. Such molecular descriptors perform as a "bridge" connecting the molecular structure and properties. The descriptors indeed provide important information of a molecule, and some of them can correlate well with the ADME properties, such as log *P* and polar surface area (PSA).[1,5] However, there are still some problems

* Corresponding authors. E-mail: whli@ecust.edu.cn (W. L.) and ytang234@ecust.edu.cn (Y. T.).

10.1021/ci100104j
Published on Web 05/18/2010



Substructure Fingerprint Dictionary

1: Primary_carbon	[CX4H3][#6]
2: Secondary_carbon	[CX4H2]([#6])[#6]
3: Tertiary_carbon	[CX4H1]([#6])([#6])[#6]
4: Quaternary_carbon	[CX4]([#6])([#6])([#6])[#6]
5: Secondary_alcohol	[OX2H][CX4H;!$(C([OX2H])[O,S,#7,#15])]
6: Ammonium	[N+;!$([N]~[!#6]);!$(N=*);!$([N]*~[#7,#8
7: Aromatic	a
8: Heterocyclic	[!#6;!R0]
9: Epoxide	[OX2r3]1[#6r3][#6r3]1
10: NH_aziridine	[NX3H1r3]1[#6r3][#6r3]1
11: Peptide_C_term	[NX3;$([N][CX3](=[OX1])[C][NX3,NX4+])][C
12: Spiro	[D4R;$(*(@*)(@*)(@*)@*)]
13: Annelated_rings	[R;$(*(@*)(@*)@*);!$([R2;$(*(@*)(@*)(@*
14: Bridged_rings	[R;$(*(@*)(@*)@*);!$([D4R;$(*(@*)(@*)(@*
15: C_ONS_bond	[#6]~[#7,#8,#16]
16: Rotatable_bond	[!$(*#*)&!D1]-!@[!$(*#*)&!D1]

Figure 1. Representation of a molecular substructure pattern fingerprint. (Propranolol, for example, with a dictionary of 16 substructure patterns.)

hampering the use of molecular descriptors in the model building. First, some descriptors cannot be obtained easily. The programs for descriptor calculation may not be accessed to everyone, and the calculation of some descriptors, for example, quantum chemical descriptors, requires a lot of computational resource and time. Second, some descriptors themselves are calculated based on empirical methods. The accuracy of these descriptors is, hence, questionable, which might lead to error accumulation from the source of the data. In addition, reasonable three-dimensional (3D) conformations are required for some 3D descriptors. However, it is usually a difficult task to determine the "best" 3D conformation of a molecule under certain conditions. Finally, descriptor selection is also a big challenge. There are thousands of descriptors available for model building. Although some efforts have been engaged in this problem,[12,22] the general instructive method of descriptor selection is still in controversy.

Since there are many adverse issues with the use of molecular descriptors, we put forward a new method to skip such a "bridge" and make a direct relationship between molecular structure and properties. In our method, a molecule is described using a fingerprint of structural keys, which is generated based on a predefined dictionary of substructure patterns. The fingerprint is represented as a Boolean array, in which each bit represents the presence (TRUE) or absence (FALSE) of a specific substructure pattern.[23] Thus each molecule corresponds to a definite fingerprint based on the predefined substructure pattern dictionary. Furthermore, the information gain (IG) method is introduced to determine the importance of each pattern. IG measures the information entropy of a classification system obtained for class prediction by knowing the presence or absence of a pattern in a molecule.[24] Finally, the prediction models are built with the SVM algorithm. Most importantly, the models built with our method are different from the "black box" model in previous studies. They are easy to interpret by recognizing the most important substructure patterns based on the IG analysis. This could help provide intuitive understanding for medicinal chemists.

To validate our method, two types of ADME properties, human intestinal absorption (HIA) and blood-brain barrier (BBB) penetration, are used as two case studies. The data set of HIA is from Hou's data set with 578 compounds, which were categorized into good intestinal absorption (HIA+) and poor intestinal absorption (HIA−).[18] The data set of BBB is from Adenot's data set with 1593 compounds, which were categorized into two classes of those that could (BBB+) and could not (BBB−) penetrate the BBB.[10] The substructure pattern fingerprint was generated based on two predefined substructure pattern dictionaries of FP4 and MACCS from Open Babel,[25] respectively. Furthermore, different IG thresholds for pattern selection and different kernel functions of SVM were also investigated exhaustively. From the results of the predictive accuracies, most of our models performed very well, and the key substructures which could affect the HIA and BBB penetration were identified. In addition, the generalization abilities of our models were evaluated by two external validation sets from other sources.

2. METHODS

2.1. Molecule Description. As mentioned above, each molecule is described using a fingerprint of structural keys, which is represented as a Boolean array. The predefined dictionary contains a SMARTS list of substructure patterns. There is a one-to-one correspondence between each SMARTS pattern and bit in the fingerprint. For a SMARTS pattern, if the specified substructure is present in the given molecule, the corresponding bit is set to "1"; conversely, it is set to "0" once the substructure is absent in the molecule. Figure 1 gives an example of the substructure pattern fingerprint of propranolol generated with a substructure fingerprint dictionary of 16 substructure patterns. In this study, two substructure dictionaries of FP4 and MACCS fingerprints were used. The dictionary of FP4 fingerprint contains 307 substructure patterns which can be divided into three categories. It was written in an attempt to represent the classification of organic compounds from the viewpoint of an organic chemist.[25] The MACCS fingerprint uses a dictionary of MDL keys,[26] which contains a set of 166 mostly common substructure features. These are referred to as the MDL public/MACCS keys. Both the definitions of FP4 and MACCS fingerprints are available in Open Babel.[25] (Table S1 and Table S2, Supporting Information).

2.2. Pattern Evaluation. Patterns in the dictionary are defined with a general purpose to cover representative substructures, as many as possible. However, not all of these patterns are necessary for modeling, especially within a small set of molecules. Therefore, a preliminary pattern filtering step is necessary and important. The IG of each pattern is

calculated to measure its effectiveness in a classification system, which is composed of two or more classes of molecules. The patterns with no or low IG values are discarded according to a predetermined threshold, and the remaining patterns compose a multidimensional vector representing each molecule. Meanwhile, important substructure patterns with major contributions to the classification system can also be identified, such as emerging chemical pattern,[27] which is only present with one class and absent with another class. IG is calculated based on the information entropy theory.[24] Suppose molecules in a classification system can be categorized into two categories based on the specific molecular property X, and therefore X has two possible values (1 = positive and −1 = negative). The possibility of molecules in the first category $P(X = 1)$ is p_1, and the possibility of the second category $P(X = -1)$ is p_0. The information entropy of X is

$$H(X) = -p_1\log_2 p_1 - p_0\log_2 p_0 \tag{1}$$

To investigate the contribution of a pattern T, we need to determine how much information entropy it brings into the system. Therefore, we can calculate the conditional entropy when pattern T is removed from the system using following equations:

$$H(X|T) = P(t)H(X|t) + P(\bar{t})H(X|\bar{t}) \tag{2}$$

where $P(t)$ is the possibility of pattern T being present in all molecules, and $P(\bar{t})$ is the possibility being of absent. The specific conditional entropy $H(X \mid t)$ is the entropy of X among only those molecules in which pattern T is present. Similarly, $H(X \mid \bar{t})$ is the entropy of X among only those molecules in which pattern T is absent. So they can be calculated easily using

$$H(X|t) = -p_1(t)\log_2 p_1(t) - p_0(t)\log_2 p_0(t) \tag{3}$$

$$H(X|\bar{t}) = -p_1(\bar{t})\log_2 p_1(\bar{t}) - p_0(\bar{t})\log_2 p_0(\bar{t}) \tag{4}$$

where $p_1(t)$ is the possibility of X in the first class among the molecules in which pattern T is present, and $p_0(t)$ is the possibility of X in the second class among only those molecules in which pattern T is present. While $\bar{t}$ indicates that the pattern T is absent. Obviously, the difference between the information entropy of the system X and the conditional entropy is the contribution of pattern T, called information gain (*IG*):

$$IG(T) = H(X) - H(X|T) \tag{5}$$

2.3. Model Training. The prediction models were trained using SVM. The formulation of SVM embodies the SRM principle, which has been shown to be superior to the traditional empirical risk minimization (ERM) principle, employed by conventional methods. SRM minimizes an upper bound on the expected risk, while ERM only minimizes the error on the training data. Such a difference features SVM with a greater ability of generalization, which is the goal of model building. Excellent descriptions of SVM and related theories can be found in the original publications of Vladimir N. Vapnik.[28,29]

In our method, each molecule is represented using a eigenvector $\mathbf{t}$, and the selected patterns $t_1, t_2, ..., t_n$ make up the components of $\mathbf{t}$. For SVM training, the category label y should be added. So the i^{th} molecule in the data set is defined as $\mathbf{M}_i = (\mathbf{t}_i, y_i)$, where $y_i = 1$ for the "positive" category, and $y_i = -1$ for the "negative" category. The purpose of SVM training is to find a hyperplane which could discriminate molecules from different categories. The formulation of the hyperplane is

$$f(\mathbf{t}) = \mathbf{w}^{\mathrm{T}}\cdot\mathbf{t} + b = 0 \tag{6}$$

where the vector $\mathbf{w}$ is known as the weight vector, and b is called the bias. The weight vector $\mathbf{w}$ can be expressed as a linear combination of the training vectors, for example, $\mathbf{w} = \sum_{i=1}^{n}\alpha_i y_i \mathbf{t}_i$. Then:

$$\begin{aligned} f(\mathbf{t}) &= \sum_{i=1}^{n}\alpha_i y_i \mathbf{t}_i^{\mathrm{T}}\cdot\mathbf{t} + b \\ &= \sum_{i=1}^{n}\alpha_i y_i \mathrm{K}(\mathbf{t}_i, \mathbf{t}) + b \end{aligned} \tag{7}$$

Another superiority of SVM is the application of kernel function. Kernel function is used to calculate the dot product of the two vectors in the high-dimensional space with the input of vectors from low-dimensional space, that is $\Phi(\mathbf{t}_i)\cdot\Phi(\mathbf{t}) = \mathrm{K}(\mathbf{t}_i,\mathbf{t})$. This increases greatly the computation efficiency and makes SVM able to deal with high-dimensional data. The commonly used kernel functions include polynomial kernel, Gaussian radial basis function kernel, and sigmoid kernel.

The determination of the hyperplane must consider the maximization of the geometric margin, which is expressed as $1/\|\mathbf{w}\|$. This leads to the following inequality constrained optimization problem:

$$\begin{aligned} &\underset{\mathbf{w},b}{\text{minimize}} \quad \frac{1}{2}\left\|\mathbf{w}\right\|^2 \\ &\text{subject to:} \quad y_i(\mathbf{w}^{\mathrm{T}}\cdot\mathbf{t}_i + b) \geq 1 \end{aligned} \tag{8}$$

In practice, the slack variable ξ and penalty coefficient C should be introduced to make a compromise between linear separability and maximal margin. So the optimization problem becomes

$$\begin{aligned} &\underset{\mathbf{w},b}{\text{minimize}} \quad \frac{1}{2}\left\|\mathbf{w}\right\|^2 + C\sum_{y_i=1}\xi_i + \frac{p_1}{p_0}C\sum_{y_i=-1}\xi_i \\ &\text{subject to:} \quad y_i(\mathbf{w}^{\mathrm{T}}\cdot\mathbf{t}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \end{aligned} \tag{9}$$

where the possibility of each category is introduced to deal with unbalanced data set. Using the method of Lagrange multipliers, the problem transforms to an equality constraint optimization problem, which is expressed in terms of variables α_i:

$$\begin{aligned} &\underset{\alpha_i}{\text{maximize}} \quad \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \mathrm{K}(\mathbf{t}_i, \mathbf{t}) \\ &\text{subject to:} \quad \sum_{y_i=1}\alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C; \\ &\qquad\qquad \sum_{y_i=-1}\alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq \frac{p_1}{p_0}C \end{aligned} \tag{10}$$

So, the decision rule of the SVM is

$$f(\mathbf{t}) = \mathrm{sgn}\left(\frac{1}{2}\sum_{i=1}^{n}\alpha_i K(\mathbf{t}_i, \mathbf{t}) + b\right) \quad (11)$$

The SVM module is provided by the LIBSVM package.[30] The parameters of kernel function and penalty are determined via grid search based on a five-fold cross validation.

2.4. Performance Measurement. A variety of approaches have been suggested to evaluate the quality of classification models.[31] Our models were assessed based on the counts of true positives (TP), false negatives (FN), true negatives (TN), false positives (FP), and the overall prediction accuracy (Q = (TP + TN)/(TP + FN + TN + FP)). Furthermore, the sensitivity (SE = TP/(TP + FN)), which is the prediction accuracy for positives, and the specificity (SP = TN/(TN + FP)), which is the prediction accuracy for negatives, were calculated to measure the prediction accuracies for the two classes. In addition, the Matthews correlation coefficient (C = (TP*TN − FN*FP)/[(TP + FN)(TP + FP)(TN + FN)(TN + FP)]$^{1/2}$) was calculated to measure the performances of models.

2.5. Data Sets. The HIA data set was collected from Hou's work.[11] This data set contained 578 compounds with FA% values. The threshold of 30% was used to divide molecules into HIA+ and HIA−. In order to make a comparison, these 578 (500 HIA+ and 78 HIA−) compounds were divided into a training set of 480 (407 HIA+ and 73 HIA−) compounds and a test set of 98 (93 HIA+ and 5 HIA−) compounds, in the same way as Hou's work.[18] Furthermore, to validate the generalization ability of the model built using our method, 634 oral drugs, which were not contained in the HIA data set, were collected from the DrugBank database[32] and composed of an external validation set. These drugs with oral dosage formulations were considered to be HIA+ compounds.

The BBB data set contained 1593 compounds from Adenot's data set, which have been categorized into BBB+ and BBB−.[10] The compounds in this data set were divided into a training set containing 1093 (832 BBB+ and 261 BBB−) compounds and a test set containing 500 (451 BBB+ and 49 BBB−) compounds, in the same way as Zhao's work.[15] In addition, an external validation set was collected from Li's data set,[17] in which 169 compounds were already included in the BBB data set, so the external validation set for the BBB models was composed of 246 compounds (155 BBB+ and 91 BBB−).

3. RESULTS AND DISCUSSION

3.1. Information Gain of the Patterns. The IG value of a pattern measures its contribution to the information entropy of a classification system. For the HIA data set with 578 compounds, the IG values were unevenly distributed ranging from 0 to 0.12 with both FP4 and MACCS fingerprints. For the BBB data set with 1593 compounds, the IG values unevenly ranged from 0 to 0.16 with the FP4 fingerprint and from 0 to 0.21 with the MACCS fingerprint (Figure 2 and Tables S1 and S2 in Supporting Information). According to the distributions of IG values, five selection thresholds were adopted respectively: 0, 0.001, 0.005, 0.01, and 0.02. These thresholds made the selection percentages of 52, 35, 23, 12, and 6% with the FP4 fingerprint for the HIA data set and 89, 62, 34, 24, and 14% with the MACCS fingerprint, respectively. For the BBB data set, the percentages of selection under different thresholds were 51, 34, 20, 14, and 12% with the FP4 fingerprint and 89, 70, 54, 46, and 33% with the MACCS fingerprint (Figure 2).

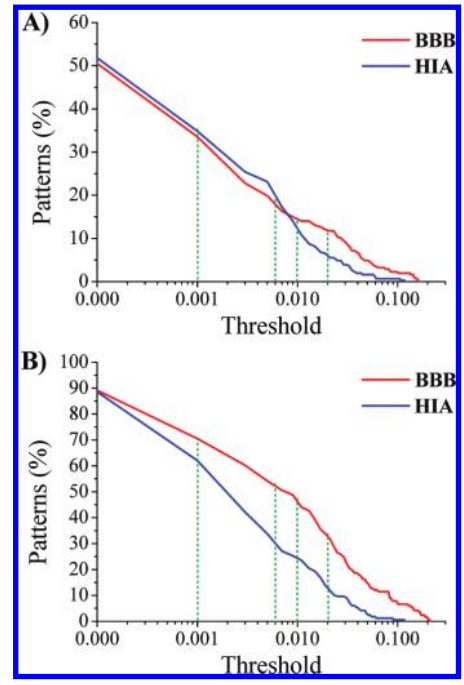


Figure 2. Percentage of selected patterns under different thresholds. (A) The molecule described using FP4 structural key. (B) The molecule described using MACCS structural key. Red line represented the BBB data set, and blue line represented the HIA data set. Threshold was plotted in logarithmic coordinates.

3.2. Model Training with Different Kernel Functions and Thresholds. The classification models of HIA were trained with a training set of 480 compounds. In which, 407 compounds with good intestinal absorption were set to be positive samples, and 73 compounds with poor intestinal absorption were set to be negative samples. The BBB models were trained with a training set of 1093 compounds, where 832 compounds which can penetrate the BBB were set to be positive samples and 261 compounds which can not penetrate the BBB were set to be the negative samples. Four different kernel functions, including linear, polynomial, radial basis function, and sigmoid, were used for the SVM training, respectively. The predictive accuracies of the models with different thresholds and kernels are listed in Table 1. It was encouraged that most of the models were fairly good with outstanding predictive accuracies for both training and test sets. The correctness of the best HIA models (models **A1** and **A2**) was 98.5% for the training set and 99% for the test set with FP4 fingerprint, where the thresholds of 0.001 were adopted. While using the MACCS fingerprint, the correctness achieved was 100% for the training set and 98% for the test set (models **A3** and **A4**). Among the BBB classification models, the correctness of the best models (model **B2**) was 97.4% and 98.2% for the training and test sets with the FP4 fingerprint, and the numbers also achieved were 99.6 and 98.2% with the MACCS fingerprint (model **B3**).

In order to investigate the number of patterns which impact the results, different thresholds were applied (Table 1). Although the differences among different thresholds were not very distinctive, it was necessary to pay attention on a few phenomena. In most cases, the correctness for the

Table 1. Predictive Accuracies of Models with Different Kernels and Thresholds

data set	pattern	threshold	predictive accuracy (%) of training and test sets under different kernels[a]							
			linear		polynomial		radial basis function		sigmoid	
HIA	FP4	0	97.7	95.9	97.7	96.9	99.0	96.9	99.4	93.9
		0.001	98.3	98.0	**A1**98.5	99.0	**A2**98.5	99.0	96.0	95.9
		0.005	96.7	95.9	98.3	99.0	97.9	98.0	96.9	95.9
		0.01	95.8	95.9	96.7	96.9	96.7	96.9	95.4	98.0
		0.02	93.8	96.9	94.6	95.9	93.8	96.9	93.8	96.9
	MACCS	0	98.5	98.0	**A3**100.0	98.0	**A4**100.0	98.0	98.1	98.0
		0.001	96.7	96.9	99.6	98.0	99.6	98.0	96.0	98.0
		0.005	97.1	99.0	95.2	98.0	96.3	96.9	95.2	98.0
		0.01	96.3	98.0	95.2	98.0	95.8	98.0	95.4	98.0
		0.02	92.1	95.9	92.5	95.9	92.1	95.9	92.1	95.9
BBB	FP4	0	95.0	97.6	98.9	96.6	99.5	96.6	95.8	97.4
		0.001	95.0	97.2	98.9	96.8	99.2	97.0	94.7	96.8
		0.005	94.6	97.0	**B1**96.9	98.2	**B2**97.4	98.2	94.8	97.2
		0.01	93.7	97.0	94.8	97.2	95.2	96.8	94.1	97.2
		0.02	94.4	96.8	93.5	97.4	94.1	97.6	94.1	97.0
	MACCS	0	96.7	97.6	100.0	98.0	100.0	98.0	96.3	97.6
		0.001	95.5	98.4	**B3**99.6	98.2	99.7	98.2	95.2	98.0
		0.005	95.3	97.4	97.3	98.0	**B4**98.8	98.4	95.1	97.6
		0.01	94.8	97.2	99.0	97.6	97.4	97.8	95.2	97.0
		0.02	95.0	97.0	97.3	97.6	96.6	97.6	94.9	97.4

[a] The predictive accuracy is presented through the overall concordance rate Q (%) for the training (first) and test (second) sets, respectively. **A1**−**A4** and **B1**−**B4** represent the models which were used for further investigations.

training set decreased with the increase of selection threshold. This phenomenon was especially noticeable in the case of radial basis function kernel. This was not unexpected because the threshold increased as the number of selected patterns decreased, while some information might be lost and caused the decrease of the correctness for the training set. However, such regularity did not appear in the test set. The model with the highest correctness for the training set might not have the highest correctness for the test set. This can be explained by "overfitting". For example, in the case of BBB modeling with the MACCS fingerprint, the threshold increased from 0 to 0.02, while the correctness of the models for the training set decreased from 100 to 96.6%. However, the correctness for the test set achieved the maximum at the threshold of 0.005, which made a selection of 54% patterns from the MACCS fingerprint. It suggested that the models would be overfitting if too many patterns were selected, and the models would miss information if the number of selected patterns was too small. Therefore, making a proper selection of patterns is necessary.

The data showed that the models with polynomial and radial basis function kernels performed better than those with linear and sigmoid kernels. Although the proper kernel function is important to a SVM model, there are no proper theories or methods to determine the optimal kernels besides cross-validation so far. Herein, we found that the polynomial and radial basis function kernels were proper, since models with the polynomial and the radial basis function kernels performed better than those with two other kernels in most cases (Table 1). Therefore, the models **A1**−**A4** and **B1**−**B4** were selected for further evaluation.

3.3. Further Investigations of Classification Models. Four HIA models (models **A1**−**A4**, marked in Table 1) and four BBB models (models **B1**−**B4**), which were generated with polynomial and radial basis function kernels, were selected for further investigation. Models **A1** and **A2** used the FP4 fingerprint dictionary, in which the pattern selection thresholds were 0.001. Models **A3** and **A4** used the MACCS fingerprint dictionary, and the pattern selection thresholds were 0. Models **B1** and **B2** used the FP4 fingerprint dictionary with a pattern selection threshold of 0.005. Model **B3** and **B4** used the MACCS fingerprint dictionary, and the pattern selection thresholds of 0.001 and 0.005 were adopted, respectively. Statistic details of these models and two published models are shown in Table 2. For the HIA models, the results suggested that models **A1** and **A2** performed better than those of models **A3** and **A4**, although models **A3** and **A4** made 100% accuracies for the training set. The pattern selection thresholds of models **A3** and **A4** were 0, so more patterns were selected. This might lead to "overfitting". Besides, the differences between fingerprint dictionaries might be another reason leading to such results. FP4 fingerprint covers a wide range of organic substructures which could provide more representative information for a molecule. Moreover, we can also find that the performances of our models (**A1** and **A2**) were better than that of Hou's model.[18]

In the case of BBB penetration, the performance of our models were also quite encouraging. The Matthews correlation coefficients of the best models (**B3** and **B4**) were 0.9924 and 0.8949 for the training and test sets, respectively, which were higher than those of Zhao's model.[15] However, the predictive accuracies of our models for the negative samples were not perfect. The specificity of the best model was 85.7%, which was slightly lower than the correctness of Zhao's model.

3.4. Validations of the Generalization Abilities. Generalization ability of a model is the most important of all, which decides the usefulness and the reliability of the models. In addition, it is also a validation of the modeling method. Therefore, external validation sets were also collected, and further evaluations of these models were performed. Here, the entire data set combining the training and test sets were used as a new training set for HIA and BBB, respectively. Two new HIA classification models were trained with the same parameters of models **A1** and **A2**, which were marked

Table 2. Overall Statistics of Models and Comparisons with Previous Published Models

data set	model	TP	TN	FP	FN	SE (%)	SP (%)	Q (%)	C
training set	**A1**	405	68	5	2	99.5	93.2	98.5	0.9428
	A2	405	68	5	2	99.5	93.2	98.5	0.9428
	A3	407	73	0	0	100.0	100.0	100.0	1.0000
	A4	407	73	0	0	100.0	100.0	100.0	1.0000
	Hou's work[a]	398	69	4	9	97.8	94.5	97.3	0.8986
test set	**A1**	92	5	0	1	98.9	100.0	99.0	0.9080
	A2	92	5	0	1	98.9	100.0	99.0	0.9080
	A3	91	5	0	2	97.8	100.0	98.0	0.8360
	A4	91	5	0	2	97.8	100.0	98.0	0.8360
	Hou's work[a]	91	5	0	2	97.8	100.0	98.0	0.8360
training set	**B1**	830	229	32	2	99.8	87.7	96.9	0.9137
	B2	830	235	26	2	99.8	90.0	97.4	0.9290
	B3	832	257	4	0	100.0	98.5	99.6	0.9899
	B4	832	258	3	0	100.0	98.9	99.7	0.9924
	Zhao's work[b]	*815*[c]	*246*	*15*	*17*	98.0	94.3	97.1	*0.9209*
test set	**B1**	450	41	8	1	99.8	83.7	98.2	0.8945
	B2	450	41	8	1	99.8	83.7	98.2	0.8945
	B3	449	42	7	2	99.6	85.7	98.2	0.8949
	B4	449	42	7	2	99.6	85.7	98.2	0.8949
	Zhao's work[b]	*443*	43	6	8	98.2	87.8	97.2	*0.8438*

[a] Ref 18. [b] Ref 15. [c] The numbers in italic type were calculated according to the predictive accuracies and the numbers of compounds in the original work.

Table 3. Performances of the Models on External Validation Data Sets

data set	model	TP	TN	FP	FN	SE (%)	SP (%)	Q (%)
HIA set	**A1′**	499	67	11	1	99.8	85.9	97.9
	A2′	499	70	8	1	99.8	89.7	98.4
valid. set	**A1′**	595	NA	NA	39	93.8	NA	93.8
	A2′	596	NA	NA	38	94.0	NA	94.0
BBB set	**B3′**	1282	301	9	1	99.9	97.1	99.4
	B4′	1282	304	6	1	99.9	98.1	99.6
valid. set	**B3′**	148	23	68	7	95.5	25.3	69.5
	B4′	149	19	72	6	96.1	20.9	68.3

[a] NA: no data were available.

as models **A1′** and **A2′**. Similarly, two new BBB classification models were also obtained with the same parameters of models **B3** and **B4**, which were marked as models **B3′** and **B4′**. The predictive accuracies of these models for the external validation sets are listed in Table 3. As can be seen, the HIA classification models **A1′** and **A2′** could identify 93.8 and 94% of compounds to be HIA+ from 634 oral drugs, respectively.

The overall predictive results of BBB models for the external validation set were not so good (Table 3). The predictive accuracies of models **B3′** and **B4′** were only 69.5 and 68.3%, respectively. Although the positive samples (BBB+ compounds) were well predicted, the models were not very predictive for the negative samples (BBB− compounds). First, poor specificity means that a certain number of the compounds with low log BB values are predicted as BBB+ compounds in our models. A possible reason is the uneven distribution of the positive and negative samples in the training set. Such uneven distribution is contributed by both the number and the chemical space of samples. There are 1283 and 310 positive and negative samples, respectively, in the training set, so the model may tend towards the positive samples. Furthermore, the BBB+ compounds in the training set cover diverse CNS active drugs, while only the drugs which obviously cannot penetrate the BBB, such as the amines, ammoniums IV, and some antibiotics, are included in the BBB− compounds.[15] The uneven distribution of the positive and negative samples in the chemical space may result from limited representative substructure patterns of the negative samples that can be identified by our model, which lead to the poor predictive accuracy of the BBB− compounds. Second, the compounds in the external validation set are taken from Li's work, in which each compound is defined as BBB+ and BBB− according to whether the experimentally determined BB ratio (the ratio of the steady-state concentrations of a drug in the brain and blood) was ≥ 0.1 or <0.1.[17] Such definition is different from that which used in the training set. In order to test the robustness of our method, the entire data set, containing 415 compounds (276 BBB+ and 139 BBB−), from Li's work were studied using our methods. With the same parameters of model **B3**, we obtained the correctness of 97.8% for the BBB+ compounds and 89.9% for the BBB− compounds. The Matthews correlation coefficient of the new model was 0.8911. Such a result not only demonstrated the reliability of our method but also indirectly proved that the log BB value of a compound could not be a sufficient criterion for CNS activity or inactivity.

3.5. Implications from Information Gain of Substructure Patterns. Besides the pattern selection, IG analysis can also provide us with more useful information about the two-class data set. A substructure pattern with a high IG value was considered to have a big contribution for the discrimination of which category it belongs to. Suppose that a substructure pattern was present with all of the "positive" molecules in a data set and absent with all of the "negative" molecules. The conditional entropy of this pattern was zero, according to eqs 2−4. Therefore, the IG value of this pattern achieved the maximum (equals to the information entropy of the system). This indicated that the two classes could be discriminated only via this substructure pattern.

Figure 3 showed some representative examples of substructure patterns with high IG values for the HIA and BBB data sets. For example, the pattern of tertiary aliphatic amine was present within 27.2% of the HIA+ compounds and only

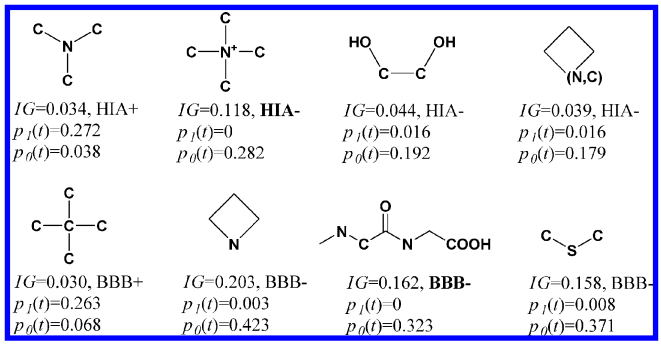


Figure 3. Representative substructure patterns with their possible classes identified using IG. Proportions of the compounds with pattern t in two different classes are $p_1(t)$ and $p_0(t)$, respectively.

3.8% of the compounds in the HIA− class. This implies that this substructure pattern could be one determinant feature of the molecules with good intestinal absorption. The pattern of ammonium was only present in the HIA− class, with a high IG value of 0.118. This is also an example of an emerging chemical pattern.[27] If a molecule contains the ammonium group, it should be classified into the HIA− class. This is in agreement with the common sense that molecules with positively charged nitrogen atoms always have very poor intestinal absorptions. Likewise, we also found other implications, for instance, the compounds with C-terminus of peptide (or just peptides) could not penetrate the BBB; and it might be also difficult for the compounds with azetidine analogs in penetrating the BBB (Figure 3).

3.6. Superiorities of Our Method. Most of the traditional models were built based on molecular descriptors, in which not only the proper descriptor calculators were required but also the correct 3D conformations were necessary in some cases. Thus, much more attention must be paid to the calculation and selection of molecular descriptors. In contrast, our models are built without using such molecular descriptors. Only the 2D structures or even 1D SMILES strings are needed for model building. The key factor which should be concerned in our methods is the molecular structures. Another advantage by our method is that no error accumulations exist in the data preparation process of the model building. In fact, the molecular descriptors are also derived from molecular structures, and they are highly interrelated. A lot of descriptors are calculated based on atom or substructure (or fragment) additive methods. Therefore, the information provided by the descriptors has been already included in the molecular structure or substructures. It is reasonable to predict the molecular properties directly from the molecular structure.

One of the big problems of the models built with SVM and other machine learning methods is the interpretation. The models perform as "black boxes", and therefore only the prediction results can be obtained. In our methods, IG analysis is adopted not only for pattern selection but also to help interpret the models from the chemistry perspective. Based on the IG value of each pattern, we can obtain the most important substructures which could influence the molecular property.

4. CONCLUSION

In this study, a substructure pattern recognition method is introduced to estimate absorption, distribution, metabolism, and excretion (ADME) properties of chemical compounds. In the method, molecules are described using substructure pattern fingerprints, which are generated based on a predefined pattern dictionary. And support vector machine (SVM) algorithm is applied to build the classification model. The substructure patterns can be recognized by the model, and the model can predict the ADME property of the molecules. It differs from conventional methods where the molecular descriptors are skipped, and a direct connection between the molecular structure and molecular properties is constructed via our method. These improvements make our method much simpler and more accurate compared with previous methods. In the case studies, two classification models of HIA and BBB penetration were generated using our method. Most of the models had high predictive accuracies and good generalization abilities. These results confirmed the reliability of our method. More importantly, the IG value of each substructure pattern can give us some useful implications from a medicinal chemistry perspective, which also helps make our models interpretable.

ACKNOWLEDGMENT

We thank Mr. Lian Duan for discussions with the technique details of this work. This work was supported by the Program for New Century Excellent Talents in University (grant no. NCET-08-0774), the 863 High-Tech Project (grant no. 2006AA020404), the 111 Project (grant no. B07023), the National S&T Major Project of China (grant no. 2009ZX09501-001), and the Shanghai Natural Scientific Foundation (grant no. 10ZR1407000).

Supporting Information Available: Structures and properties of the molecules used in this study; the IG values of each pattern in the FP4 pattern key (Table S1) and MACCS pattern key (Table S2). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Hou, T.; Wang, J. Structure-ADME relationship: still a long way to go? *Expert Opin. Drug Metab. Toxicol.* **2008**, *4*, 759–770.
(2) Kola, I.; Landis, J. Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discovery* **2004**, *3*, 711–715.
(3) Wang, J. L.; Skolnik, S. Recent Advances in Physicochemical and ADMET Profiling in Drug Discovery. *Chem. Biodiversity* **2009**, *6*, 1887–1899.
(4) van de Waterbeemd, H.; Gifford, E. ADMET in silico modelling: towards prediction paradise? *Nat. Rev. Drug Discovery* **2003**, *2*, 192–204.
(5) Tetko, I. V.; Bruneau, P.; Mewes, H. W.; Rohrer, D. C.; Poda, G. I. Can we estimate the accuracy of ADME-Tox predictions? *Drug Discovery Today* **2006**, *11*, 700–707.
(6) Huynh, L.; Masereeuw, R.; Friedberg, T.; Ingelman-Sundberg, M.; Manivet, P. In silico platform for xenobiotics ADME-T pharmacological properties modeling and prediction. Part I: Beyond the reduction of animal model use. *Drug Discovery Today* **2009**, *14*, 401–405.
(7) Sakiyama, Y. The use of machine learning and nonlinear statistical tools for ADME prediction. *Expert Opin. Drug Metab. Toxicol.* **2009**, *5*, 149–169.
(8) Young, R. C.; Mitchell, R. C.; Brown, T. H.; Ganellin, C. R.; Griffiths, R.; Jones, M.; Rana, K. K.; Saunders, D.; Smith, I. R.; Sore, N. E.; et al. Development of a new physicochemical model for brain penetration and its application to the design of centrally acting H2 receptor histamine antagonists. *J. Med. Chem.* **1988**, *31*, 656–671.
(9) Luco, J. M. Prediction of the brain-blood distribution of a large set of drugs from structurally derived descriptors using partial least-squares (PLS) modeling. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 396–404.
(10) Adenot, M.; Lahana, R. Blood-brain barrier permeation models: discriminating between potential CNS and non-CNS drugs including

P-glycoprotein substrates. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 239–248.
(11) Hou, T.; Wang, J.; Zhang, W.; Xu, X. ADME evaluation in drug discovery. 7. Prediction of oral absorption by correlation and classification. *J. Chem. Inf. Model.* **2007**, *47*, 208–218.
(12) Shen, J.; Du, Y.; Zhao, Y.; Liu, G.; Tang, Y. In Silico Prediction of Blood-Brain Partitioning Using a Chemometric Method Called Genetic Algorithm Based Variable Selection. *QSAR Comb. Sci.* **2008**, *27*, 704–717.
(13) Garg, P.; Verma, J. In silico prediction of blood brain barrier permeability: an Artificial Neural Network model. *J. Chem. Inf. Model.* **2006**, *46*, 289–297.
(14) Jung, E.; Kim, J.; Kim, M.; Jung, D. H.; Rhee, H.; Shin, J. M.; Choi, K.; Kang, S. K.; Kim, M. K.; Yun, C. H.; Choi, Y. J.; Choi, S. H. Artificial neural network models for prediction of intestinal permeability of oligopeptides. *BMC Bioinformatics* **2007**, *8*, 245.
(15) Zhao, Y. H.; Abraham, M. H.; Ibrahim, A.; Fish, P. V.; Cole, S.; Lewis, M. L.; de Groot, M. J.; Reynolds, D. P. Predicting penetration across the blood-brain barrier from simple descriptors and fragmentation schemes. *J. Chem. Inf. Model.* **2007**, *47*, 170–175.
(16) Zhang, L.; Zhu, H.; Oprea, T. I.; Golbraikh, A.; Tropsha, A. QSAR modeling of the blood-brain barrier permeability for diverse organic compounds. *Pharm. Res.* **2008**, *25*, 1902–1914.
(17) Li, H.; Yap, C. W.; Ung, C. Y.; Xue, Y.; Cao, Z. W.; Chen, Y. Z. Effect of Selection of Molecular Descriptors on the Prediction of Blood-Brain Barrier Penetrating and Nonpenetrating Agents by Statistical Learning Methods. *J. Chem. Inf. Model.* **2005**, *45*, 1376–1384.
(18) Hou, T.; Wang, J.; Li, Y. ADME evaluation in drug discovery. 8. The prediction of human intestinal absorption by a support vector machine. *J. Chem. Inf. Model.* **2007**, *47*, 2408–2415.
(19) Frölich, H.; Wegner, J. K.; Sieker, F.; Zell, A. Kernel Functions for Attributed Molecular Graphs - A New Similarity-Based Approach to ADME Prediction in Classification and Regression. *QSAR Comb. Sci.* **2006**, *25*, 317–326.
(20) Geppert, H.; Vogt, M.; Bajorath, J. Current Trends in Ligand-Based Virtual Screening: Molecular Representations, Data Mining Methods, New Application Areas, and Performance Evaluation. *J. Chem. Inf. Model.* **2010**, *50*, 205–216.
(21) Ivanciuc, O. Applications of Support Vector Machines in Chemistry. *Rev. Comput. Chem.* **2007**, *23*, 291–400.
(22) Soto, A.; Cecchini, R.; Vazquez, G.; Ponzoni, I. Multi-Objective Feature Selection in QSAR Using a Machine Learning Approach. *QSAR Comb. Sci.* **2009**, *28*, 1509–1523.
(23) Fingerprints - Screening and Similarity; Daylight Chemical Information Systems, Inc.: Aliso Viejo, CA; http://www.daylight.com/dayhtml/doc/theory/theory.finger.html. Accessed January 18, 2010.
(24) Sokolova, M.; Szpakowicz, S. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*; Olivas, E. S., Guerrero, J. D. M., Sober, M. M., Benedito, J. R. M., López, A. J. S., Eds.; IGI Global: New York, 2010; Vol. *II*, Chapter 15, pp 325−347.
(25) Open Babel; Free Software Foundation, Inc.: Boston, MA; http://openbabel.org/. Accessed January 18, 2010.
(26) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280.
(27) Auer, J.; Bajorath, J. Emerging Chemical Patterns: A New Methodology for Molecular Classification and Compound Selection. *J. Chem. Inf. Model.* **2006**, *46*, 2502–2514.
(28) Vapnik, V. N. *The Nature of Statistical Learning Theory*, 2nd ed.; Springer-Verlag: New York, 2000.
(29) Vapnik, V. N. *Statistical Learning Theory*; John Wiley and Sons, Inc.: New York, 1998.
(30) LIBSVM: a library for support vector machines; Department of Computer Science and Information Engineering, National Taiwan University: Taipei, Taiwan; http://www.csie.ntu.edu.tw/∼cjlin/libsvm. Accessed January 18, 2010.
(31) Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C. A.; Nielsen, H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **2000**, *16*, 412–424.
(32) Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* **2008**, *36*, D901–D906.

CI100104J