[Article] www.whxb.pku.edu.cn

基于机器学习方法的激素敏感脂肪酶抑制剂活性预测

吕 巍[1] 薛 英[1,2,*]

([1] 四川大学化学学院, 教育部绿色化学与技术重点实验室, 成都 610064;
[2] 四川大学生物治疗国家重点实验室, 成都 610041)

摘要: 脂肪组织中, 激素敏感脂肪酶(HSL)被认为是调节脂肪酸代谢的关键限速酶. HSL在糖尿病的发病过程中起重要作用. 抑制HSL活性有助于糖尿病的治疗, 因此探索新颖的HSL抑制剂变成当前研究的热门. 在激素敏感脂肪酶的作用机制和三维结构缺乏的情况下, 需要发展预测HSL抑制剂的方法. 本文采用几种机器学习方法(支持向量机((SVM)、*k*-最近相邻法(*k*-NN)和C4.5决策树(C4.5 DT))对已知的HSL抑制剂与非抑制剂建立分类预测模型. 252个结构多样性化合物(123个HSL抑制剂与129个HSL非抑制剂)被用于测试分类预测系统, 并用递归变量消除法选择与HSL抑制剂相关的性质描述符以提高预测精度. 本研究对独立验证集的总预测精度为75.0%–80.0%, HSL抑制剂的预测精度为85.7%–90.5%, 非HSL抑制剂的预测精度为63.2%–68.4%. 其中支持向量机方法给出最好的总预测精度80.0%. 本研究表明支持向量机等机器学习方法可以有效预测未知数据集中潜在的HSL抑制剂, 并有助于发现与其相关的分子描述符.

关键词: 激素敏感脂肪酶; 机器学习方法; 分子描述符; 递归变量消除法; 支持向量机
中图分类号: O641

Activity Prediction of Hormone Sensitive Lipase Inhibitors Based on Machine Learning Methods

LÜ Wei[1] XUE Ying[1,2,*]

([1]*Key Laboratory of Green Chemistry and Technology, Ministry of Education, College of Chemistry, Sichuan University, Chengdu 610064, P. R. China*; [2]*State Key Laboratory of Biotherapy, Sichuan University, Chengdu 610041, P. R. China*)

Abstract: Hormone-sensitive lipase (HSL) is known as the key rate-limiting enzyme responsible for regulating free fatty acids (FFAs) metabolism from adipose tissue. Recently, HSL has been found to be useful in the treatment of diabetes so the discovery of new HSL inhibitors (HSLIs) is of interest. Methods for the prediction of HSLIs are highly desired to facilitate the design of novel diabetes therapeutic agents because limited knowledge exists concerning the mechanism and three dimensional (D) structure of hormone-sensitive lipase. We have explored several machine learning methods (support vector machines (SVM), *k*-nearest neighbor (*k*-NN), and C4.5 decision tree (C4.5 DT)) to predict desirable HSLIs from a comprehensive set of known HSLIs and non-HSLIs. Our prediction system was tested using 252 compounds (123 HSLIs and 129 non-HSLIs) and these are significantly more diverse in chemical structure than those in other studies. The recursive feature elimination selection method was used to improve the prediction accuracy and to select the molecular descriptors responsible for distinguishing HSLIs and non-HSLIs. Prediction accuracies were 85.7%–90.5% for HSLIs, 63.2%–68.4% for non-HSLIs, and 75.0%–80.0% for all structures based on three kinds of machine learning methods using an independent validation set. SVM gave the best total accuracy of 80.0% for all the structures. This work suggests that machine learning methods such as SVM are useful to predict the potential HSLIs among unknown sets of compounds and to characterize the molecular descriptors associated with HSLIs.

Received: September 17, 2009; Revised: October 25, 2009; Published on Web: November 30, 2009.

[*] Corresponding author. Email: yxue@scu.edu.cn; Tel: +86-28-85418330.

The project was supported by the National Natural Science Foundation of China (20773089) and Scientific Research Foundation for the Returned Overseas Chinese Scholars, Ministry of Education, China (20071108-18-15).
国家自然科学基金(20773089)和教育部留学归国人员科研启动基金(20071108-18-15)资助项目

Key Words: Hormone-sensitive lipase; Machine learning method; Molecular descriptor; Recursive feature elimination; Support vector machine

Type 2 diabetes is a complex, multi-factorial, and chronic metabolic disease[1]. The prevenient researches presented that the important characteristics of the type 2 diabetes are higher levels of fatty acids in plasma and tissue[2]. It has been shown that the elevated level of plasma fatty acids (FAs) plays an important role in the pathogenesis of insulin resistance and type 2 diabetes[3,4].

In the period of low-energy in organism, lipids such as triglycerides are decomposed to release energy through their hydrolysis followed by oxidation, primarily β-oxidation. Accompanying this process, FAs and glycerol are also liberated. But in type 2 diabetic patients, adipose tissue would be decomposed during the period of high-energy, for example, nocturnal and postprandial periods[5], and the increase of FA levels in plasma is inspected obviously. So the energy can not be stored in the adipose tissue. Hormone-sensitive lipase (HSL) is a multifunctional tissue lipase which is called a component of the metabolic switch between glucose and FAs, as it is the rate-limiting enzyme in adipose tissue lipolysis and net FAs mobilization[6]. HSL can catalyze the fat metabolism to elevate levels of fatty acids. The activity of HSL is modulated *via* phosphorylation or dephosphorylation primarily controlled by many kinds of hormone, such as insulin, which can inactivate HSL[7]. Because of the insulin resistance in the type 2 diabetes patients, the active of HSL is advanced and the fat metabolism is accelerated, leading to elevated level of FAs. So inhibiting the activity of HSL will decrease in the release of FAs[8]. Such key role of HSL in the fat metabolism has led to the suggestion that HSL may be a potential therapeutic target for this disease[9].

Discovering the novel inhibitors of HSL (HSLIs) has been becoming a hot spot in the therapeutics of type 2 diabetes. During the last few years, a series of various classes of HSLIs were reported by different research groups, which contain 2*H*-isoxazol-5-ones[10], oxadiazolones[11], pyrrolopyrazinediones[12], carbazates[13], carbamoyltriazoles[14], and aryl boronic acids[15]. Few efforts have been directed at the development of computational methods for the prediction of HSLIs. Mutasem *et al.*[16] discovered some new potent HSLIs with quantitative structure-activity relationship (QSAR) method through screening the National Cancer Institute (NCI) list of compounds and their in-house built database of drugs and agrochemicals. Recently, structure- and mechanism-based drug design methods have been developed and applied to drug discovery projects[17]. However, the application of these methods for the HSLI prediction may be retarded by the following cases, that is, the lack of available HSL crystallographic structures for the new inhibitors to combine with, the complexity of the catalytic mechanism of the HSL in adipose tissue lipolysis, and millions of molecules in the compound libraries. This prompts us to explore the possibility of developing non-structure-based computational methods for predicting hormone-sensitive lipase inhibitors, which facilitates the identification of HSLIs in the early drug design phase without requiring the knowledge about their mechanisms, the intrinsic relationships between activities and molecular properties, and the structures of targeted proteins and other macromolecules and molecular assemblies.

It has been well shown that the machine learning (ML) methods are very useful tools for the classification of the pharmacodynamic, pharmacokinetic, and toxicological properties of drug agents[18]. High throughput screening (HTS) is a method of drug discovery or gene/protein function determination which can test or class tens of thousands of compounds against a particular through selecting for drug-like characteristics such as solubility, partition coefficient (lg*P*), molecular weight, and number of hydrogen bond donors/acceptors (Lipinski′s rule of 5)[19]. Recently, the ML methods have also been applied in HTS broadly. In this paper, we will use ML methods (the support vector machines (SVM), *k* nearest neighbor (*k*–NN), and C4.5 decision tree (C4.5 DT)) to study the classification prediction of HSLIs and non-HSLIs. It is of interest to improve the performance of SVM and to explore other machine learning methods for facilitating the classification prediction of HSLIs. The prediction precision of these ML methods relies on selecting appropriate subset of molecular descriptors suitable for distinguishing HSLIs and non-HSLIs. The recursive feature elimination (RFE) method[20], which has been extensively used in the feature selecting, was employed in this research for selecting the most relevant molecular descriptors. To assess the prediction accuracy of the models used in this work, two different evaluation methods were employed. One is five-fold cross validation and the other is evaluation by an independent validation set.

1 Materials and methods

1.1 Selection of HSLIs and non-HSLIs

A total of 260 HSLIs and non-HSLIs with known IC_{50} values was selected from a number of published papers[10,12–15,21] and their structures are supplied in Table S1 of Supporting Information (which is available free of charge *via* the internet at http://www.whxb.pku.edu.cn). Based on the tested experimental data in prevenient researches[12–15], when the IC_{50} value is lower than 500 $nmol\cdot L^{-1}$, the molecule will have better activity. And when the IC_{50} values are between 400 and 600 $nmol\cdot L^{-1}$, the activity of molecules is ambiguous. So we divided all molecules to three sets: one set includes 123 HSLIs (IC_{50}<400 $nmol\cdot L^{-1}$), the second set includes 129 non-HSLIs (IC_{50}>600 $nmol\cdot L^{-1}$). The last set includes 8 molecules (400 $nmol\cdot L^{-1} \leqslant IC_{50} \leqslant 600$ $nmol\cdot L^{-1}$) which are ambiguous between HSLIs and non-HSLIs. In these three sets, we only chose the first two sets as the tested sets.

The 2D structure of each of the compounds was generated by using ChemDraw[22] and was subsequently converted into 3D structure by using Corina[23] for calculating the quantum chemical properties. The 3D structure of each compound was manually inspected to ensure that the chirality of each chiral agent is properly generated. All the generated geometrics have been fully optimized without symmetry restrictions.

First all compounds were divided into training set, testing set, and independent validation set according to their distribution in the chemical space defined by their structural and chemical features. The ID of compound in every subset is supplied in Table S2 of Supporting Information. The training and testing sets were used to develop and optimize a statistical model, and the independent validation set is used for assessing the classification accuracy of the model. Then, all compounds in the training and testing sets were randomly divided into five subsets of approximately equal size. After training the SVM with a collection of four subsets, the performance of the SVM was tested against the fifth subset. This process was repeated five times, so that every subset was once used as the test data.

1.2 Molecular descriptors

Molecular descriptors are routinely used to quantitatively represent structural and physicochemical properties of molecules, which have been extensively applied in the structure-activity relationship(SAR)[24], QSAR[25], and other computational researches of pharmaceutical agents [26]. In this work, 198 molecular descriptors as described in the earlier studies[27] were used. These descriptors are given in Table S3 of Supporting Information, including 18 descriptors in the class of simple molecular properties, 27 descriptors in the class of molecular connectivity and shape, 97 descriptors in the class of electro-topological state, 31 descriptors in the class of quantum chemical properties, and 25 descriptors in the class of geometrical properties. They are computed from the 3D structure of each compound by using our own designed molecular descriptor computing program. When computing the quantum chemical descriptors and molecular surfaces, a semiempirical AM1 method widely used in QSAR and SLM models of compounds was used for preprocessing structural optimization. The irrelevant and redundant descriptors for HSLIs and non-HSLIs in the 198 molecular descriptors were eliminated by using feature selection method[20].

1.3 Feature selection method

In a dataset with a fixed number of samples, excessive descriptors may cause a prediction model to be over-fitted to lead to affect its performance. Therefore, feature selection methods have been employed to enhance the performance of ML methods by eliminating the molecular descriptors redundant and irrelevant to the discrimination of two datasets. The feature selection method, the recursive feature elimination (RFE), has been widely acknowledged because of its efficacy preformed in discovering informative feature molecular descriptors most relevant to prediction of antibacterial compounds[27], prediction of the human ether-a-go-go-related gene(hERG) potassium channel inhibitors[28], prediction of *tetrahymena pyriformis* toxicity chemicals[18]. Therefore, the RFE method combined with SVM was used in this work to determine a preferable set of descriptors relevant to the prediction of HSLIs and enhance the prediction accuracies of the models.

The computation procedure in this work can be outlined as follow: The Gaussian kernel was employed to train a SVM classification model with a series of variation of the parameter σ in the special region against the whole training dataset and the corresponding prediction accuracies were evaluated by 5-fold cross-validation. For a fixed parameter σ, in the first step, the SVM builds a model with the complete set of descriptors. The second step is ranking the contribution of the descriptors in the datasets based on a criterion score got from a scoring function. In the third step, the m lowest ranked descriptors are removed. Finally, the SVM classifier is retrained by using the remaining descriptors, and the corresponding prediction accuracy is computed by mean of five-fold cross validation. All the four steps are then repeated for other σ until all descriptors have been removed. After the completion of these procedures, the set of descriptors and parameter σ that gave the best prediction accuracy is selected.

The choice of the parameter m affects the performance of SVM as well as the speed of feature selection. To control the size of the selected descriptors, we only consider the number of descriptors smaller than one-fifth of the whole descriptors[29]. Our earlier studies[18,20,27] suggested that the performance of a SVM system with m=5 is only reduced by a few percentages smaller than that with m=1, which is consistent with the findings from the other study[30]. In this study, m=5 is used for the sake of computational efficiency.

1.4 Machine learning methods

There are a number of downloadable ML method software packages. For example, PHAKISO (http://www.phakiso.com/index.htm) and WEKA[31] (http://www.cs. waikato.ae.nz/~ml/weka) for a collection of ML method software, NeuNet (http://www.cormactech.com/neunet/index.html) for neural network, SVM-Light (http://svmlight.joachims.org) for SVM software are used in many work. We use our own programs to build SVM model for predicting the drug agents from HSLIs and non-HSLIs. And we compare the result of SVM model with the results of other ML methods.

1.4.1 SVM

The method of SVM has been extensively described in the articles[32]. Here we only briefly descript it. SVM is based on the structural risk minimization (SRM) principle from machine learning method. In linearly separable cases, SVM constructs a hyperplane which separates two different classes of molecules with a maximum margin. With regard to the nonlinearly problem, SVM projects feature vectors into a high-dimensional feature space and searches for a linear optimal separating hyperplane (decision boundary) in the new feature space. The transform can be done by using a kernel function that satisfied the MercerNs theorem.

1.4.2 *k*-NN

In *k*-NN, the Euclidean distance between an unclassified vector $\boldsymbol{x}$ and each individual vector $\boldsymbol{x}_i$ in the training set is measured[33]. A total of k number of vectors nearest to the unclassified vector $\boldsymbol{x}$ are used to determine the class of the unclassified vector. The class of the majority of the k-nearest neighbors is chosen as the predicted class of the unclassified vector $\boldsymbol{x}$.

1.4.3 C4.5 DT

C4.5 DT is a branch-test-based classifier[34]. A branch in a decision tree is in accordance with a group of classes and a leaf represents a specific class. A decision node specifies a test to be conducted on a single attribute value, with one branch and its subsequent classes as possible outcomes of the test. In C4.5 DT, recursive partitioning is used to examine every attribute of the data and rank them according to their abilities to partition the remaining data, thus constructing a decision tree.

1.5 Performance evaluation

As in the case of all discriminative methods[35], the performance of ML methods can be measured by the quantity of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). There are several accuracy functions for measuring prediction performance, which include sensitivity (SE), specificity (SP), the overall prediction accuracy (Q), and Matthews correlation coefficient (C) are given by Eq.(1−4), respectively.

$$Q=\frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}}\times100\% \quad (1)$$

$$C=\frac{\mathrm{TP\times TN-FN\times FP}}{\sqrt{\mathrm{(TP+FN)(TP+FP)(TN+FN)(TN+FP)}}}\times100\% \quad (2)$$

2 Results and discussion

2.1 Overall prediction accuracies and merit of the machine learning methods

SVM prediction of HSLIs is evaluated by the methods of both the use of 5-fold cross validation and independent validation set. Both of the methods appear to give consistent assessment about the prediction accuracy.

Firstly, through comparing the accuracies of SVM, which used 5-fold cross validation with and without the use of RFE of feature selection method, the feature selection method plays an important role in the performance of SVM for the prediction of HSLIs and non-HSLIs. The results are shown in Table 1. Through this method, we find 21 descriptors which are critical for SVM model. The 21 descriptors are shown in Table 2. The accuracies of SVM with RFE are 72.0% for HSLIs and 74.3% for non-HSLIs; the accuracies of SVM without RFE are 64.1% for HSLIs and 74.7% for non-HSLIs. And the overall prediction accuracy 73.3% obtained from SVM with RFE is substantially better than the value of 69.6% derived from SVM without RFE. It is obviously indicated that the method with RFE is substantially better than that derived from SVM without RFE, especially for HSLIs. The results show that the selection of appropriate molec-ular descriptors is important for the improvement of average prediction accuracy, but more important for implying which pharmacological features are more propitious to distinguish HSLIs and non-HSLIs.

Secondly, the accuracies of the databases are predicted through independent validation with all 198 descriptors and 21 descriptors selected by RFE. The results are shown in Table 3. These results suggest that the accuracies of independent validation set with 21 descriptors selected by RFE are obviously better than those with all descriptors. So these 21 descriptors are more important for our model. And it further proves the method with RFE is useful for our model.

Apart from the cross-validation method, the independent validation set also has frequently been used for testing the robustness of a model. In this work, an independent validation set with 21 HSLIs and 19 non-HSLIs is constructed from our existing datasets according to their distribution in the chemical space. Table 4 gives all prediction results of HSLIs and non-HSLIs derived from three ML methods k-NN, C4.5 DT and SVM by using the RFE selected descriptors and an independent validation set. For comparison, those results from SVM are also labeled in Table 4. By comparing the prediction accuracies from the three methods, we have obtained several results. For HSLIs, the accuracies of these methods are in the range of 85.7%−90.5% with

Table 1 Accuracies of HSLIs and non-HSLIs derived from SVM without and with the use of the feature selection method RFE by using five-fold cross validation

Method	Cross validation	HSLIs			non-HSLIs			Q(%)	C(%)
		TP	FN	SE(%)	TN	FP	SP(%)		
SVM	1	14	4	77.8	17	8	68.0	72.1	45.2
	2	10	5	66.7	12	6	66.7	66.7	33.2
	3	21	15	58.3	13	5	72.2	63.0	28.8
	4	14	3	82.4	17	7	70.8	75.6	52.4
	5	6	11	35.3	23	1	95.8	70.7	40.8
	average			64.1			74.7	69.6	40.1
	S.E.			8.4			5.4	2.2	4.2
SVM+RFE	1	15	3	83.3	14	11	56.0	67.4	39.7
	2	9	6	60.0	14	4	77.8	69.7	38.5
	3	25	11	69.4	12	6	66.7	68.5	34.4
	4	12	5	70.6	19	5	79.2	75.6	49.8
	5	13	4	76.5	22	2	91.7	85.4	69.7
	average			72.0			74.3	73.3	46.4
	S.E.			3.9			6.1	3.4	6.4

Statistical significance is indicated by S.E. (standard error). The number of HSLIs or non-HSLIs is TP+FN or TN+FP.

Table 2 Twenty-one molecular descriptors selected by the RFE feature selection method for the classification of HSLIs and non-HSLIs

Descriptor	Description	Class
ncof	count of F atoms	simple molecular property
ncocl	count of Cl atoms	simple molecular property
ncarb	count of C atoms	simple molecular property
nring	numbers of rings	simple molecular property
nrot	number of rotatable bonds	simple molecular property
${}^{3}\chi_{\mathrm{C}}$	simple molecular connectivity chi indices for cluster	molecular connectivity and shape
${}^{5}\chi_{\mathrm{CH}}$	simple molecular connectivity chi indices for cycles of 5 atom	molecular connectivity and shape
${}^{6}\chi_{\mathrm{CH}}$	simple molecular connectivity chi indices for cycles of 6 atom	molecular connectivity and shape
$5\chi_{\mathrm{CH}}^{\mathrm{V}}$	valence molecular connectivity chi indices for cycles of 5 atoms	molecular connectivity and shape
${}^{1}\kappa$	molecular shape kappa indices for one boned fragments	molecular connectivity and shape
S(15)	atom-type H Estate sum for AH_n (not C, N, O, S)	electrotopological state
S(18)	atom-type estate sum for $>CH_2$	electrotopological state
S(28)	atom-type estate sum for >C<	electrotopological state
S(48)	atom-type estate sum for $-PH_2$	electrotopological state
S(64)	atom-type estate sum for >Ge<	electrotopological state
T_{rmsd}	balaban RMSD index	electrotopological state
ε_{b}	hydrogen bond acceptor basicity (covalent HBAB)	quantum chemical property
η	absolute hardness	quantum chemical property
$Q_{\mathrm{H,\,SS}}$	sum of squares of charges on H atoms	quantum chemical property
dis1	length vectors (longest distance)	geometrical property
Hlb	hydrophilic-hydrophobic balance	geometrical property

SVM giving the best accuracy at 90.5%. For non-HSLIs, the accuracies are in the range of 63.2%–68.4% with SVM and *k*-NN giving the best accuracy at 68.4%. Lastly, for both HSLIs and non-HSLIs, the average accuracies are in the range of 75.0%–80.0% with SVM giving the best accuracy at 80.0%, *k*-NN giving the second best accuracy at 77.5% and C4.5 DT giving the worst accuracy at 75.0%. We could check whether a prediction system is over-fitting through a frequently used method which is to compare the prediction accuracies determined by using cross validation methods with those determined by using the independent validation set. Since descriptor selection is performed by using the cross validation method as the modeling testing sets, an over-fitted classification system is expected to have much higher prediction accuracy for the cross validation sets than that for the independent validation set. As shown in Tables 1 and 4, the prediction accuracies of the ML methods systems based on the ind ependent validation set and those based on the cross-validation method are similar. This work indicates the ML methods systems are unlikely to be overfitted.

Overall, our study suggests that ML methods, especially the SVM, are useful for facilitating the prediction of novel HSLIs from compounds with diverse structures. Another advantage of the SVM studied in this work is that they do not require the knowledge about the molecular mechanism or structure-activity relationship of a particular drug property.

2.2 Molecular descriptors associated with the diversity between HSLIs and non-HSLIs

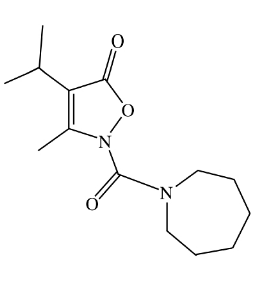
Selecting molecular descriptors which are most relevant to the prediction of HSLIs is important for optimizing the prediction models and for elucidating the molecular factors contributing to HSLIs. Commonly, QSAR models particularly design a group of specific descriptors to represent the studied HSLIs which have similar structural groups or structural alerts[16]. In this research, a total of 21 molecular descriptors are selected by RFE. These descriptors, given in Table 2, represent the structural and physicochemical properties associated with the diversity between HSLIs and non-HSLIs. Some of them are found to match or partially match those descriptors used in the published HSLIs QSAR models[16]. The pharmacophoric features, such as hydrogen bond acceptor, hydrogen bond donor, ring aromatic and hydrophobic, are primarily attributable to the HSL bioactivities. In our work, RFE method selects the descriptors, and ncof (count of F atoms), ncocl (count of Cl atoms) and ε_{b} (hydrogen bond acceptor basicity) are relative with hydrogen bond acceptor; nring (numbers of rings) and Hlb (hydrophilic-hydrophobic balance) are as the same as the results in these researches too. Mutasem *et al.* [16] constructed the QSAR model of HSLIs with several molecular descriptors, more of which are selected by RFE in our work.

Table 3 Comparison of the prediction accuracies of HSLIs and non-HSLIs obtained from SVM by using the independent validation set with all 198 descriptors and 21 descriptors selected by RFE

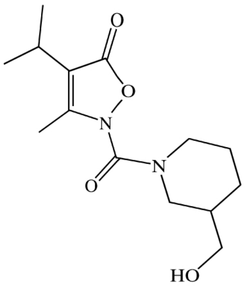
Set of descriptor	TP	FN	SE(%)	TN	FP	SP(%)	Q(%)
all descriptors	14	7	66.7	15	4	79.0	72.5
21 descriptors	19	2	90.5	13	6	68.4	80.0

Table 4 Comparison of the prediction accuracies of HSLIs and non-HSLIs derived from different machine learning methods by using the independent validation set

Method	Parameter	TP	FN	TN	FP	HSLIs SE(%)	non-HSLIs SP(%)	Q(%)
C4.5 DT	–	18	3	12	7	85.7	63.2	75.0
k-NN	k=14	18	3	13	6	85.7	68.4	77.5
SVM	σ=5	19	2	13	6	90.5	68.4	80.0



compound (**55**)



compound (**60**)

Fig.1 The Structures of the Misclassified HSLIs

For example, SssCH$_3$ (methyl) which are the electrotopological descriptors are selected to be the most correlative descriptors named S(18) (atom-type estate sum for >CH$_2$). Otherwise, shadow descriptors are geometric descriptors that characterize the shape of the molecules. In other research[16], shadow-*Y*length (the length of molecule in the *Y* dimension) is important and the dis1 (length vectors) is selected in our article according to shadow-*Y*length[16].

2.3 The misclassified compounds in the independent validation set

There are twelve molecules incorrectly classified by our SVM system with the independent validation set method. The prediction accuracy is 90.5% for HSLIs, 68.4% for non-HSLIs and 80.0% for all of them. And for HSLIs set, which is comprised of 21 molecules, there are 2 molecules which are predicted as non-HSLIs, on the other hand, for non-HSLIs set, which is comprised of 19 molecules, there are 6 molecules which are predicted as HSLIs. All of these misclassified molecules are shown in Fig.1 and Fig.2. From these two figures, we can see that the misclassified molecules are mainly with multiple rings and various hetero atoms such as nitrogen, oxygen, chlorine and fluorin. Examination of incorrectly predicted compounds suggests that using currently molecular descriptors may not be sufficient to properly discriminate the molecules with complex structures or chemical configurations. Therefore we should try to further explore different combination of descriptors and to select more optimal set of descriptors by using more refined feature selection algorithms. So it implies that further improvement and refinement of our molecular descriptors may be good for our prediction model.

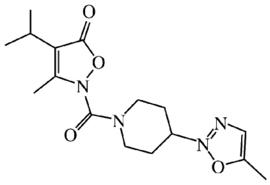
3 Conclusions

This study shows that machine learning methods, especially SVM, are useful for facilitating the prediction of HSLIs without the knowledge of mechanisms but only with the choice of specific molecular descriptors. But the current ML methods are limited in their ability to facilitate the study of the mechanism of predicted properties. Nevertheless, we believe in the near future, this weakness may be partially overcome by the development of regression-based ML methods. In addition, our study indicates that prediction accuracy of this model is affected by the molecular descriptors selected by RFE which can further help to optimally select molecular descriptors. To conclude, the availability of more extensive information about various HSLIs and associated mechanisms will facilitate the development of machine learning methods into practical tools for the prediction of different types of HSLIs in the early stage of drug development. Recent works on the introduction of weighting function into ML methods descriptors may also be applied to develop ML methods into a practical tool for the prediction of HSLIs and thus facilitate new drug development.

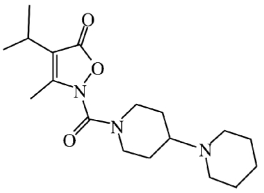
Supporting Information Available: The information of the investigated dataset is provided in Tables S1, S2 and S3. This information is available free of charge via the internet at http://www.whxb.pku.edu.cn.

References

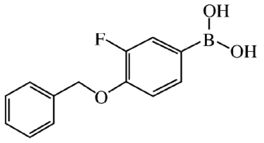
1 DeFronzo, R. A. *Diabetologia*, **1992, 35**: 389
2 Reaven, G. M.; Greenfield, M. S. *Diabetes*, **1981, 30**: 66
3 Unger, R. H. *Diabetes*, **1995, 44**: 863
4 Boden, G. *Diabetes*, **1997, 46**: 3



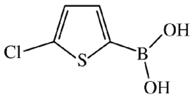
compound (**86**)



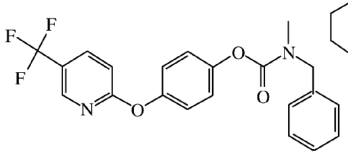
compound (**87**)



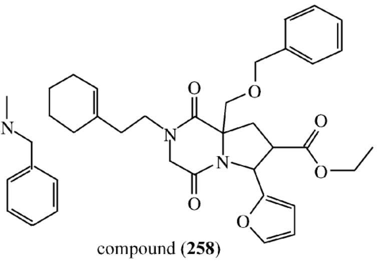
compound (**190**)



compound (**193**)



compound (**226**)



compound (**258**)

Fig.2 Structures of the misclassified non-HSLIs

5 Miles, J. M.; Wooldridge, D.; Grellner, W. J.; Windsor, S.; Isley, W. L.; Kelin, S.; Harris, W. S. *Diabetes*, **2003, 52**: 675

6 Kraemer, F. B.; Shen, W. J. *Nutr. Metab.*, **2006, 12**: 1

7 Kraemer, F. B.; Shen, W. J. *J. Lipid. Res.*, **2002, 43**: 1585

8 Langin, D.; Holm, C.; LaFontan, M. *Proc. Nutr. Soc.*, **1996, 55**: 93

9 Ye, J. *Endocr. Metab. Immune. Disord. Drug. Targets.*, **2007, 7**: 65

10 Lowe, D. B.; Magnuson, S.; Oi, N.; Campbell, A.; Cook, J.; Hong, Z.; Wang, M.; Rodriguez, M.; Achebe, F.; Kluender, H.; Wong, W. C.; Bullock, W. H.; Salhanick, A. I.; Witman, J. T.; Bowling, M. E.; Keiperb, C.; Clairmont, K. B. *Bioorg. Med. Chem. Lett.*, **2004, 14**: 3155

11 Shoenafinger, K.; Petry, S.; Mueller, G.; Barringhous, K. H. PCT Appl. WO12001/066531, 2001

12 Slee, D. H.; Bhat, A. S.; Nguyen, T. N.; Kish, M.; Lundeen, K.; Newman, M. J.; McConnell, S. J. *J. Med. Chem.*, **2003, 46**: 1120

13 de Jong, J. C.; Sorensen, L. G.; Tornqvistc, H.; Jacobsen, P. *Bioorg. Med. Chem. Lett.*, **2004, 14**: 1741

14 Ebdrup, S.; Sorensen, L. G.; Olsen, O. H.; Jacobsen, P. *J. Med. Chem.*, **2004, 47**: 400

15 Ebdrup, S.; Jacobsen, P.; Farrington, A. D.; Vedsø, P. *Bioorg. Med. Chem.*, **2005, 13**: 2305

16 Mutasem, O. T.; Lina, A. D.; Yasser, B.; Hiba, Z.; Suhair, S. *J. Med. Chem.*, **2008, 51**: 6478

17 Qiao, C.; Wilson, D. J.; Bennett, E. M.; Aldrich, C. C. *J. Am. Chem. Soc.*, **2007, 129**: 6350

18 Xue, Y.; Li, H.; Ung, C. Y.; Yap, C. W.; Chen, Y. Z. *Chem. Res. Toxicol.*, **2006, 19**: 1030

19 Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. *Adv. Drug Deliv. Rev.*, **1997, 23**: 3

20 Xue, Y.; Li, Z. R.; Yap, C. W.; Sun, L. Z.; Chen, X.; Chen, Y. Z. *J. Chem. Inf. Comput. Sci.*, **2004, 44**: 1630

21 Ebdrup, S.; Hanne, H. F. R.; Christian, F.; Poul, J. *J. Med. Chem.*, **2007, 50**: 5449

22 ChemDraw Version 9.0. Cambridge, USA: Cambridge Soft Corporation, 2004

23 Corina. Version 3.4. Erlangen, Germany: Molecular Networks GmbH Computerchemie, 2006

24 Yu, H.; Yang, J.; Wang, W.; Han, J. Discovering compact and highly discriminative features or feature combinations of drug activities using support vector machines. Proc. IEEE Comput. Soc. Bioinformatics Conf. (CSB), Washington, DC, USA: IEEE Computer Society, 2003: 220–228

25 Hu, J. Y.; Aizawa, T. *Water. Res.*, **2003, 37**: 1213

26 Byvatov, E.; Fechner, U.; Sadowski, J.; Schneider, G. *J. Chem. Inf. Comput. Sci.*, **2003, 43**: 1882

27 Yang, X. G.; Chen, D.; Wang, M.; Xue, Y.; Chen, Y. Z. *J. Comput. Chem.*, **2009, 30**: 1202

28 Wang, M.; Yang, X. G.; Xue, Y. *QSAR Comb. Sci.*, **2008, 27**: 1028

29 Leach, A. R.; Gillet, V. J. An Introduction to Chemoinformatics. New York: Springer, 2007

30 Furlanello, C.; Serafini, M.; Merler, S.; Jurman, G. *Neural Netw.*, **2003, 16**: 641

31 Garner, S. R. Weka. Version 3.4.12. Hamilton, New Zealand: University of Waikato, 2005

32 Vapnik, V. N. The nature of statistical learning theory. New York: Springer-Verlag, 1995

33 Johnson, R. A.; Wichern, D. W. Applied multivariate statistical analysis. New York: Prentice Hall, 1982

34 Quinlan, J. R. C4.5, programs for machine learning. San Mateo, CA: Morgan Kaufmann, 1992

35 Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C. A.; Nielsen, H. *Bioinformatics*, **2000, 16**: 412