农药学学报 2010,12(4):477 -488
Chinese Journal of Pesticide Science http://www.nyxxb.com.cn

• Research report •

In silico prediction of terrestrial and aquatic toxicities for organic chemicals

CHENG Fei-xiong[1], SHEN Jie[1], LI Wei-hua[1], Philip W. LEE*[1,2], TANG Yun*[1]

(1. *Department of Pharmaceutical Sciences, School of Pharmacy, East China University of Science and Technology, 130 Meilong Road, Shanghai* 200237, *China*;
2. *Graduate School of Agriculture, Kyoto University, Kitashirakawa Oiwake-cho, Sakyo-ku, Kyoto* 606-8502, *Japan*)

Abstract: Qualitative classification and quantitative regression models for fathead minnow and honey bee toxicity prediction were developed using different chemoinformatics techniques such as substructure pattern recognition and different machine learning methods. Specifically, methods include support vector machine, C4.5 decision tree, *k*-nearest neighbors, random forest and naive bayes. Reliable predictive models were developed and all models were validated by the independent test set. The overall predictive accuracy of the classification models using support vector machine were 95.9% for the fathead minnow test set and 95.0% for the honey bee test set. The square of correlation coefficient of regression models were 0.878 for the fathead minnow test set and 0.663 for the honey bee test set using support vector machine regression algorithm. At last, some representative substructure patterns for characterizing fathead minnow and honey bee toxicity compounds, such as 1,2-diphenol, dialkylthioether, diarylether and phosphoric_ acid_ derivative were also identified via the information gain analysis. The approaches provide a useful strategy and robust tools in the screening of ecotoxicological risk or environmental hazard potential of chemicals.

Key words: fathead minnow toxicity; honey bee toxicity; quantitative structure-activity relationship (QSAR); substructure pattern recognition; information gain; support vector machine

有机化合物的陆地和水生环境毒性的计算机预测研究

程飞雄[1], 沈 杰[1], 李卫华[1], Philip W. LEE*[1,2], 唐 赟*[1]

(1. 华东理工大学 药学院 药物科学系,上海 200237;
2. 京都大学 农业研究院 日本 京都 606-8502)

摘 要: 采用子结构模式识别结合5种机器学习方法(包括支持向量机、C4.5决策树、*k*-最近邻法、随机森林法、和朴素贝叶斯法),分别构建了有机化合物对水生和陆地环境毒性评价的两个重要生物靶标——呆鲦鱼(Fathead minnow)和蜜蜂毒性的定性分类和定量回归预测模型。所有模型均通

Received: 2010-08-25; **Revised:** 2010-09-21.

Author introduction: CHENG Fei-xiong(1985-), Male, Anhui Province, doctor student, E-mail: fxcheng1985@gmail.com; * **Corresponding author:** TANG Yun(1968-), Male, Ph. D., Professor and Vice Dean of School of Pharmacy, East China University of Science and Technology. He has published more than 50 SCI scientific articles in computer-assisted drug design, chemoinformatics, computational biology etc. **Tel:** 021-64251052, **E-mail:** ytang234@ecust.edu.cn; Philip W. LEE, Male, American, Ph. D. Professor, Graduate School of Agriculture, Kyoto University, and also as Chair Professor, School of Pharmacy, East China University of Science and Technology. He has published more than 40 scientific articles in pesticides, drug metabolism and environmental chemistry. **E-mail:** philiplee2007@gmail.com.

Foundation items: Program for New Century Excellent Talents in University (Grant No. NCET-08-0774); the National S & T Major Project of China (Grant No. 2009ZX09501-001) and the 111 Project (Grant No. B07023).

过独立测试集验证。其中,利用支持向量机分类算法得到的分类模型对呆鲦鱼和蜜蜂毒性测试集的整体预测准确度分别达到95.9%和95.0%。采用支持向量机回归算法得到的回归模型,对呆鲦鱼和蜜蜂毒性测试集的预测相关系数的平方(R^2)分别达到0.878和0.663。最后,通过信息熵分析的方法,确定了一批能够代表性地表征呆鲦鱼和蜜蜂毒性的子结构模式,包括1,2-二酚、二烷基硫醚、二芳香醚和磷酸衍生物等。提出的方法为有毒化学品的生态风险评价提供了一种非常好的评价策略和可靠的工具。

关键词:呆鲦鱼毒性;蜜蜂毒性;定量结构-活性相关性(QSAR);子结构模式识别;信息熵;支持向量机

DOI:10.3969/j.issn.1008-7303.2010.04.18

中图分类号:X131;TQ450.1 文献标志码:A 文章编号:1008-7303(2010)04-0477-12

0 Introduction

Computational toxicology may help to significantly reduce the cost of experimental toxicity assessment and accelerate the environmental hazard assessment, which is highly recommended by US Environmental Protection Agency (EPA) and European Union (EU)[1]. EU approved a new regulation, called "Registration, Evaluation and Authorization of Chemicals (REACH)" on 1 June 2007, which advocates the use of non-animal testing methods and, in particular, quantitative structure-toxicity relationship (QSTR) and quantitative structure-activity relationship (QSAR) approaches in order to decrease the number and costs of animal testing (http://ec.europa.eu/environment/chemicals/reach/reach_intro.htm). Recently paradigm shift has been suggested in toxicology with a specific reference to computational methods as reliable support in toxicity assessment[2]. Investigations into the development and use of QSAR models to rapidly predict the ecotoxicity of xenobiotics, pesticides and industrial chemicals from their molecular structure and/or physicochemical properties have increased dramatically over the past decades in order to save time and money in the design of safer chemicals[3].

Fathead minnow (FHM) and honey bee (HB) are two commonly used test organisms for the assessment of environmental impact of toxicants. In the past decades, there were several QSAR models for fathead minnow toxicity (FHMT) and honey bee toxicity (HBT) prediction[4-6]. James *et al.* developed QSAR models to estimate the acute toxicity of 100 pesticides to *Apis mellifera* using multi-layer feed-forward neural network[4]. The root mean square residual values for the training set (89 chemicals) and external test set (11 chemicals) were 0.430 and 0.386, respectively. Vighi *et al.* proposed a QSAR model for estimating the acute toxicity of pesticides to *Apis mellifera*[5]. The usefulness of these models is limited because they were only designed for simulating the toxicity of organophosphorus pesticides. Tan *et al.* applied the support vector machines (SVM) and artificial neural networks (ANN) methods to predict the acute toxicity of 611 compounds to FHM based on molecular structure[6]. SVM model gave an averaged prediction accuracy of 95.5% for FHMT, 79.3% for non-FHMT and 91.0% for all samples; comparably, the ANN model results were 92.5%, 75.2% and 87.7%, respectively. Michielan *et al.* also presented a robust classification model for FHMT and the overall predictive precision was greater than 89.2% for test set[7]. The objective of this study is to generate a more reliable predictive tool for the FHMT and HBT prediction based on a vast diverse group of chemicals.

In this study, the diverse data set of 195 pesticides or pesticides-like molecules for HBT were collected from the US EPA ECOTOX database[8] and 554 compounds with training set and 295 compounds with test set for FHMT were collected from the work of Michielan *et al.*[7]. Models were developed using our recently developed substructure pattern recognition methods[9]. Different machine learning methods were also applied and evaluated, including support vector machine (SVM), C4.5 decision tree (C4.5 DT), *k*-nearest neighbors (*k*-NN), random forest (RF) and naive bayes (NB). The quantitative regression models for FHMT and HBT prediction were also developed based on support machine regression algorithm. Unlike

traditional methods, our methods made a direct connection between the chemical structure and the toxicity endpoints of compounds. Moreover, some representative substructure patterns for HBT and FHMT were identified based on the information gain (IG) analysis [9].

1 Materials and Methods

The entire workflow in this study was presented in Fig. 1.

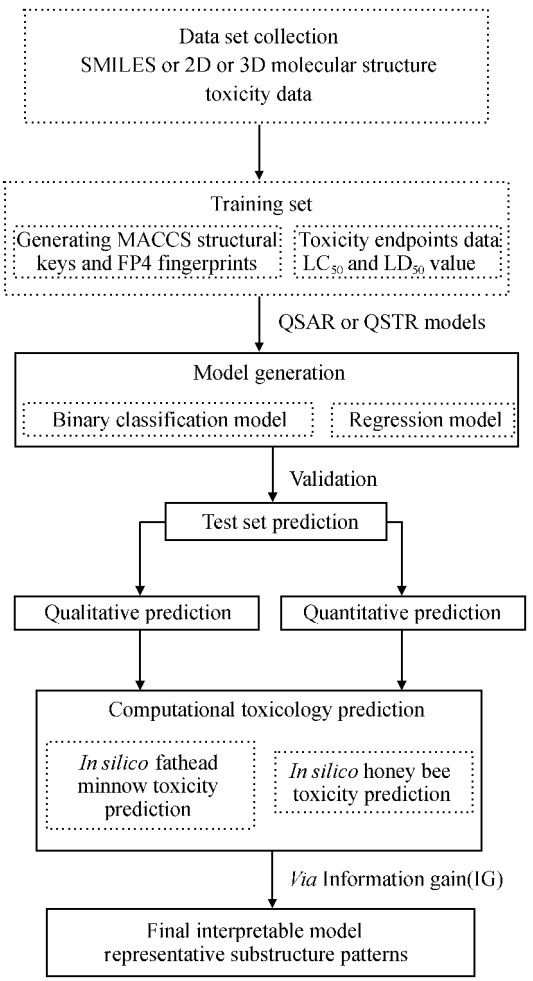


Fig. 1 The workflow of the qualitative classification models and quantitative regression models for fathead minnow and honey bee toxicity prediction

1.1 Data set collection and diversity analysis

1.1.1 *Fathead Minnow data set* 554 compounds served as the training set and other 295 compounds as the test set for FHMT were collected from EPA Fathead Minnow Acute Toxicity Database (EPAFHM) [7]. The training set and test set in our study were the same with Michielan *et al.* [7]. The FHMT endpoint of each compound was expressed as the concentration lethal to 50% of the organisms (LC_{50}) for FHM in 96-h flow-through exposure tests. The threshold value of LC_{50} = 0.5 mmol/L was chosen to divide the data set into high acute FHMT compounds and low acute FHMT ones [7]. Compounds with the value of LC_{50} < 0.5 mmol/L were assigned as high acute FHMT compounds, whereas others were assigned as low acute FHMT compounds. The statistical description of the entire compound data set was listed in Table 1. Compound name, SMILES and LC_{50} value can be found in the work of Michielan *et al.* [7].

Table 1 The statistical data for the entire fathead minnow and honey bee toxicity data set

Species	Training set		Test set		Total	
	P	N	P	N	P	N
fathead minnow	366	188	196	99	562	287
honey bee	76	79	23	17	99	96

Note: P, high acute honey bee toxicity or high acute fathead minnow toxicity compounds; N, low acute honey bee toxicity or low acute fathead minnow toxicity compounds.

1.1.2 *Honey Bee data set* 195 pesticides or pesticide-like molecules for HBT were collected from US EPA ECOTOX database [8]. The HBT endpoint to *Apis mellifera* was expressed as the dose lethal to 50% of the organisms (LD_{50}) in a 48-h exposure tests. The threshold value of LD_{50} = 100 μg/bee was chosen to designate high acute HBT compounds and low acute HBT compounds. Compounds with the value of LD_{50} < 100 μg/bee were assigned as high acute HBT compounds, while others were assigned as low acute HBT compounds. The next step, the entire data set was divided into 99 high acute HBT compounds and 96 low acute HBT compounds. 155 compounds (80% data for the entire data set) were randomly selected for the training set, which included 76 high acute HBT compounds and 79 low acute HBT compounds. And others 40 compounds (20% data for the entire data set) which included 23 high acute HBT compounds and 17 low acute HBT compounds were used as the test set. The training set and test set had good balance positive samples (high acute HBT compounds) and negative samples (low acute HBT compounds) based on this principle of allocation in

this study. The statistical description of the entire compound data set was presented in Table 1. CAS number and LD_{50} value of each compound were provided in US EPA ECOTOX database [8].

1.1.3 *Data set diversity analysis* The structural diversity of the data set was assessed by the calculation of the average Tanimoto similarity index (based on MDL PublicKeys). For this purpose, we used the educational version of Pipeline Pilot (version 6.0; SciTegic, San Diego, CA, 2005) to calculate molecular similarity matrix and derive the Tanimoto similarity index.

1.2 Substructure pattern recognition description

The recently developed substructure pattern recognition methods [9] to depict the entire data set were used. Each molecule is described as a bit string structural key. The predefined dictionary contains a SMARTS list of substructure patterns. There is a one-to-one correspondence between each SMARTS pattern and each bit in the pattern fingerprint. For a SMARTS pattern, if a specified substructure is present in the given molecule, the corresponding bit is set to "1"; conversely, it is set to "0" [9].

In this present study, MACCS structural keys and FP4 fingerprints were used. The MACCS structural keys use a dictionary of MDL PublicKeys [10], which contains a set of 166 most common substructure features and they are referred to as the MDL Public/MACCS keys. The dictionary of FP4 fingerprint contains 307 substructure patterns. The definitions of MACCS structure key and FP4 fingerprints are available in OpenBabel (http://openbabel.org/, Access Date: Jan. 18, 2010) [11].

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 were discarded according to a predetermined threshold and the remaining patterns composed of a multi-dimensional vector representing each molecule. Some representative substructure patterns of HBT and FHMT compounds were then identified based on IG analysis [9].

1.3 Machine learning methods

In this study, SVM, C4.5 DT, *k*-NN, RF and NB were selected for carrying out both HBT and FHMT classification and regression models. SVM was performed by LIBSVM 2.9 package [12] and LIBSVM 2.84 package [13]. C 4.5 DT, *k*-NN, RF and NB were performed in Orange 2.0 package (Version 2.0 b, freely available in the website <http://www.ailab.si/orange/>).

1.3.1 *Support vector machine (SVM)* Support vector machine (SVM), originally developed by Vapnik for pattern recognition, aims at minimizing the structural risk under the frame of VC theory [14]. Recently, it had been extended to the domain of regression problems [15]. In this study, support vector machine classification (SVMC) and support vector machine regression (SVMR) algorithms were selected for carrying out classification and regression modeling task, respectively. The classification models were built using SVM classification module provided by LIBSVM 2.9 package [12]. Regression models were built using the regression module provided by LIBSVM 2.84 package [12,16].

1.3.1.1 Support vector machine classification (SVMC) The classification problem can be restricted to consideration of the two-class problem without loss of generality. Details about the theory of SVM theory can be found in the literature [14]. Basically, in this study, each molecule was represented using a eigenvector t, and the selected patterns $t_1, t_2, \cdots, t_n$ make up the components of t. For SVM training, the category label y should be added. So the i^{th} molecule in the data set is defined as $M_i = (t_i, y_i)$, where $y_i = 1$ for the "positive" category and $y_i = -1$ for the "negative" category. SVM gives a decision function (classifier):

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

Where α_i is the coefficient to be learned and K is a kernel function. Parameter α_i is trained through maximizing the Lagrangian expression given below:

$$\underset{\alpha_i}{\text{maximize}} \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j y_i y_j K(t_i, t)$$

$$\text{subject to: } \sum_{y_i=1} y_i a_i = 0, \quad 0 \leqslant a_i \leqslant C \quad (2)$$

A superiority of SVM is that it can deal with high dimensional space with the input of vectors from low dimensional space by introducing kernel function. In this study, commonly-used kernel function of Gaussian radial basis function kernel was used. Radial

basis functions (RBF) kernel has paid significant attention, most commonly with a Gaussian of the form:

$$K(x,x') = \exp\left(-\frac{\| x - x' \|^2}{2\sigma^2}\right) \quad (3)$$

To obtain a SVMC models with optimal performance, the penalty parameter C and different kernels parameter γ were tuned based on the training set using the grid search strategy based on 5-fold cross-validation.

1.3.1.2 Support vector machine regression (SVMR)

SVM can also be applied to regression problems by the introduction of an alternative loss function [16]. The loss function must be modified to include a distance measure. Using a ε-insensitive loss function:

$$L_\varepsilon(y) = \begin{cases} 0 & \text{for} \quad |f(x) - y| < \varepsilon \\ |f(x) - y| - \varepsilon & \text{otherwise} \end{cases} \quad (4)$$

Similarly to classification problems, a non-linear model is usually required to adequately model data. In the same manner as the non-linear SVMC approach, a non-linear mapping can be used to map the data into a high dimensional feature space where linear regression is performed. The kernel approach is again employed to address the curse of dimensionality. The non-linear SVMR solution, using a ε-insensitive loss function, which is given by:

$$\max_{a,a^*} W(a,a^*) = \max_{a,a^*} \sum_{i=1}^{l} a_i^* (y_i - \varepsilon) - a_i (y_i + \varepsilon) - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} (a_i^* - a_i)(a_j^* - a_j) K(x_i, x_j) \quad (5)$$

with constraints,

$$0 \leqslant a_i, a_i^* \leqslant C, \quad i = 1, \cdots, l$$

$$\sum_{i=1}^{l} (a_i - a_i^*) = 0 \quad (6)$$

Solving equation 5 with constraints equation 6 determines the Lagrange multipliers, a_i, a_i^* and the regression function is given by,

$$f(x) = \sum_{SVs} (\bar{a}_i - \bar{a}_i^*) K(x_i, x) + \bar{b} \quad (7)$$

Where

$$< \bar{w}, x > = \sum_{i=1}^{l} (a_i - a_i^*) K(x_i, x_j)$$

$$\bar{b} = -\frac{1}{2} \sum_{i=1}^{l} (a_i - a_i^*) [K(x_i, x_r) + K(x_i, x_s)] \quad (8)$$

As with the SVMR the equality constraint may be dropped if the Kernel contains a bias term, b being accommodated within the Kernel function, and the regression function is given by:

$$f(x) = \sum_{i=1}^{l} (\bar{a}_i - \bar{a}_i^*) K(x_i, x) \quad (9)$$

A SVMR model contains three tuning parameters: Epsilon (ε) of the loss function, C of the constraints. These parameters were also selected based on the training set using the grid search strategy by 5-fold cross-validation. The negative logarithm of LC_{50} for FHMT (pLC_{50}) and the negative logarithm of LD_{50} for HBT (pLD_{50}) were used as the dependent variable to develop regression models.

1.3.2 *Random forest* (*RF*) RF is a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest [17]. RF models consist of an ensemble of decision trees, each obtained by splitting object collections until terminal nodes contain only objects of the same class. In this study, models were trained using a number of binary strings from computed MACCS structural key and FP4 fingerprint, with the objective of modeling whether a given compound is correctly fit to the high acute HBT ($y_i = 1$) or low acute HBT ($y_i = -1$) and high acute FHMT ($y_i = 1$) or low acute FHMT ($y_i = -1$).

1.3.3 *k-nearest neighbors* (*k-NN*) *k*-NN is a method for classifying objects based on closest training examples in the feature space. In this study, the nearness is measured by hamming distance matrix and the standard protocol of 3-NN is implemented simply as follows: 1) to calculate distances between an unknown object (y) and all the objects in the training set; 2) to select 3 objects from the training set most similar to object y, according to the calculated distances; and 3) to classify object y with the group to which the majority of the 3 objects belongs.

1.3.4 *C* 4.5 *decision tree* (*C* 4.5 *DT*) The program C 4.5 DT is a successor of the basic Iterative Dichotomiser 3 (ID3) decision tree learning algorithm developed by Ross Quinlan [18]. C 4.5 defines the possible decision tree by means of a hill-climbing search based on the statistical property measure called information gain. The elements of the tree generated by ID3 and C 4.5 are either leafs or decision nodes.

The leaf shows a class, and the decision node specifies the test to be implemented on an attribute value, with one branch and sub-tree for each possible result of the test. The detail descriptors of C 4.5 can be found in original literature [18].

1.3.5 *Naïve bayes* (*NB*) Bayesian classification is a statistical method that allows the user to categorize instances in a data set based on the equal and independent contributions of their attributes [19]. A NB classifier is generated using a training set to provide the prior evidence that an instance belongs to a certain class. An example of this would be a training set of *B* molecules where *A* of the molecules are known to be high acute toxicity and the remainder are known to be low acute toxicity against a given organisms. These molecules can be used to train the classifier such that it is able to distinguish the high acute toxicity molecules from the low acute toxicity molecules. The prior probability of a molecule being toxicity, P [*A*], is given by equation 10. In this study, the NB classifier can be generated using the MACCS structural keys and FP4 fingerprint described above.

$$P[A] = \frac{A}{B} \tag{10}$$

1.4 Performance of models

All models were validated by the independent test set. The classification models for high and low acute HBTs, high and low acute FHMTs were evaluated based on the counts of true positives (*TP*), true negatives (*TN*), false positives (*FP*), false negatives (*FN*). *TP* represents the number of high acute HBT and FHMT compounds predicted correctly. *TN* represents the number of low acute HBT and FHMT compounds predicted correctly. *FP* represents the number of low acute HBT and FHMT compounds predicted wrongly. And *FN* represents the number of high acute HBT and FHMT compounds predicted wrongly. Furthermore, the sensitivity [$SE = TP/(TP + FN)$], which is the prediction accuracy for high acute HBT and FHMT compounds, and the specificity [$SP = TN/(TN + FP)$], which is the prediction accuracy for low acute HBT and FHMT compounds, were calculated. The overall accuracy (*Q*), F-measure (*F*) and Matthews correlation coefficient (*C*) were also calculated by the equation 11, 12 and 13.

$$Q = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

$$F = \frac{2TP}{2TP + FP + FN} \tag{12}$$

$$C = \frac{TP \times TN - FN \times FP}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}} \tag{13}$$

The overall performance of regression models was evaluated by measuring the square of correlation coefficient (R^2), root mean square error (*RMSE*) calculated from the following equations:

$$R^2 = 1 - \frac{\sum (y_i - y_j)^2}{\sum (y_i - y_m)^2} \tag{14}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n_s} (y_i - y_j)^2}{n_s}} \tag{15}$$

where, y_i, y_j and y_m represent the experimental value, predicted value and the mean of dependent variable, respectively. n_s is the number of molecules in data set of regression equation.

In addition, a receiver operating characteristic (ROC) curve was also employed to graphically present the model behavior in a visual way. A ROC curve had been proved to be a valuable way to evaluate the quality of a binary classifier. At last, a plot of trade-off between the true positive rate (sensitivity, *y*-axis) and false positive rate (1-specificity, *x*-axis) was shown.

2 Results and Discussion

2.1 Data set diversity analysis

If compounds have the smaller Tanimoto similarity index, they have good diversity. The average Tanimoto similarity indexes of our data set were 0.123 for the FHMT training set and 0.149 for the FHMT test set; The average Tanimoto similarity index were 0.239 for the HBT training set and 0.314 for HBT test set. The results showed that the entire data set of the FHMT and HBT had good chemical diversity.

2.2 Performance of binary classification models

2.2.1 *Binary classification models for fathead minnow toxicity* In this study, the FHMT classification models were built using the training set composed of 544 compounds with five different

machine learning methods, including SVM, *k*-NN, C 4.5 DT, RF and NB. All models were validated by a test set containing 295 compounds. The detail statistical results of *SE*(%), *SP*(%), *Q*(%), *F*, *C* and AUC values for test set were listed in Table 2, where the *Q* value was 95.9%, 99.3%, 91.2%, 93.6% and 75.9% with MACCS structural keys using SVM, *k*-NN, C4.5 DT, RF and NB algorithms, respectively. Comparing the five different machine learning methods, the performance of *k*-NN and SVM was better than others. As shown in Table 2 and the ROC curve in Fig. 2, the model performance using MACCS structural keys were better than FP4 fingerprint. It could be explained that MACCS structural keys which

Table 2 Performance of classification models for fathead minnow toxicity test set using five different machine learning methods

Models	Methods	*TP*	*TN*	*FP*	*FN*	*SE*/%	*SP*/%	*Q*/%	*F*	*C*	AUC
FHMT-MACCS	SVM	192	91	8	4	98.0	91.9	95.9	0.970	0.908	0.998
	k-NN	195	98	1	1	99.5	99.0	99.3	0.995	0.985	0.999
	RF	186	90	9	10	94.9	90.9	93.6	0.951	0.856	0.988
	C4.5	185	84	15	11	94.4	84.9	91.2	0.934	0.801	0.931
	NB	154	70	29	42	78.6	70.7	75.9	0.813	0.479	0.819
FHMT-FP4	SVM	187	78	21	9	95.4	78.8	89.8	0.926	0.768	0.956
	k-NN	183	95	4	13	93.4	96.0	94.2	0.956	0.876	0.984
	RF	180	75	24	16	91.8	75.8	86.4	0.900	0.691	0.955
	C4.5	180	72	27	16	91.8	72.7	85.4	0.893	0.666	0.853
	NB	166	76	23	30	84.7	76.8	82.0	0.862	0.605	0.871
Michielan *et al.*'s work		183	81	18	14	92.9	81.8	89.2	0.92	0.755	

Note: FHMT-MACCS represents the FHMT classification models built by MACCS structural keys; FHMT-FP4 represents the FHMT classification models built by FP4 fingerprints. SVM(support vector machine), C4.5 DT(C4.5 decision tree), *k*-NN(*k*-nearest neighbor), RF(random forest), *NB*(naive bayes); *TP*(true positives), *TN*(true negatives), *FP*(false positives), *FN*(false negatives), *SE*(sensitivity), *SP*(specificity), *Q*(overall predictive accuracy), *F*(F-measure), *C*(Matthews correlation coefficient) and AUC(the area under receiver operating characteristic curve).

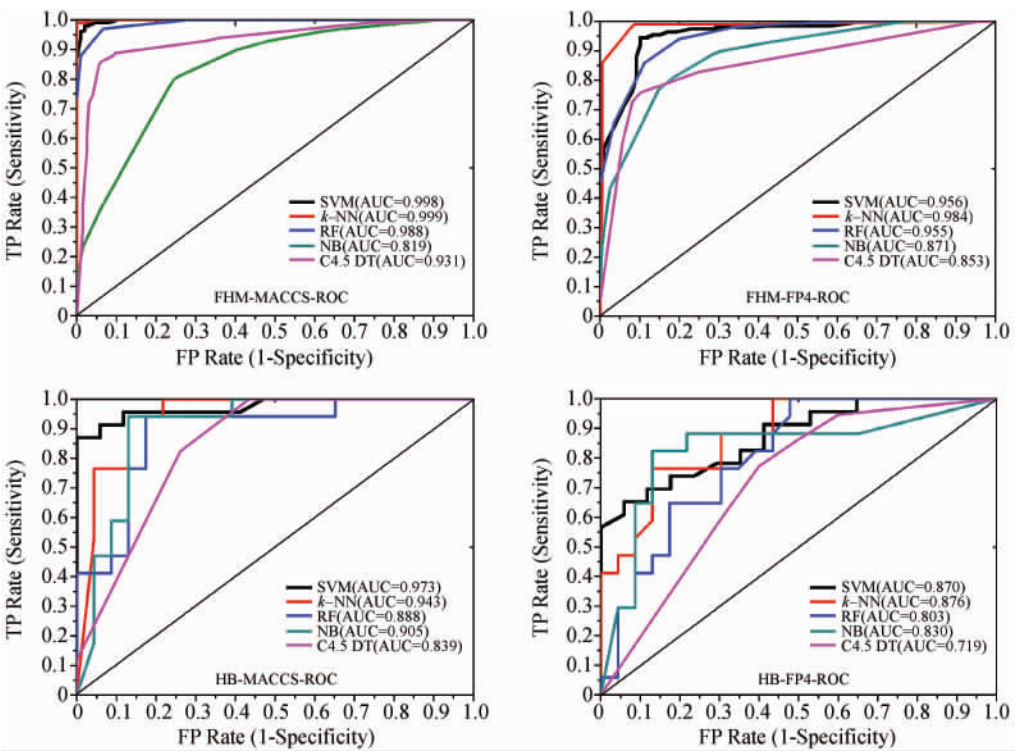


Fig. 2 Representation of receiver operating characteristics(ROC) curves with five different machine learning methods

SVM(support vector machine), C 4.5 DT(C 4.5 decision tree), *k*-NN(*k*-nearest neighbor), RF(random forest), NB(naive bayes) and AUC(the area under receiver operating characteristics curves)

were the mostly common substructure features may be better than the FP4 fingerprint which was written in an attempt to represent the classification of organic compounds from the viewpoint of an organic chemist.

2.2.2 *Binary classification models for honey bee toxicity* The HBT classification models were developed using the training set containing 155 compounds and validated by a test set with 40 compounds. The detail statistical results of HBT classification models for test set using five different machine learning methods were listed in Table 3. The performance of HBT classification models was some

Table 3 Performance of classification models for honey bee toxicity test set using five different machine learning methods

Models	Methods	*TP*	*TN*	*FP*	*FN*	*SE* /%	*SP* /%	*Q* /%	*F*	*C*	AUC
HBT-MACCS	SVM	21	17	0	2	91.3	100.0	95.0	0.955	0.904	0.973
	k-NN	19	16	1	4	82.6	94.1	87.5	0.884	0.759	0.943
	RF	18	16	1	5	78.3	94.1	85.0	0.857	0.717	0.888
	C4.5	14	15	2	9	60.9	88.2	72.5	0.718	0.496	0.839
	NB	20	16	1	3	87.0	94.1	90.0	0.909	0.803	0.905
HBT-FP4	SVM	16	14	3	7	69.6	82.4	75.0	0.762	0.514	0.870
	k-NN	16	13	4	7	69.6	76.5	72.5	0.744	0.455	0.876
	RF	15	14	3	8	65.2	82.4	72.5	0.732	0.473	0.803
	C4.5	14	13	4	9	60.9	76.5	67.5	0.683	0.371	0.719
	NB	19	14	3	4	82.6	82.4	82.5	0.844	0.646	0.830

Note: HBT-MACCS represents the HBT classification models built by MACCS structural keys; HBT-FP4 represents the HBT classification models built by FP4 fingerprints. SVM (support vector machine), C4.5 DT (C4.5 decision tree), *k*-NN (*k*-nearest neighbor), RF (random forest), NB (naïve bayes); *TP* (true positives), *TN* (true negatives), *FP* (false positives), *FN* (false negatives), *SE* (sensitivity), *SP* (specificity), *Q* (overall predictive accuracy), *F* (F-measure), *C* (Matthews correlation coefficient) and AUC (the area under receiver operating characteristic curve).

little different to FHMT classification models. The SVM was performed better than other algorithms in HBT classification models, but the *k*-NN was performed better than others in FHMT classification models study. The sensitivity and specificity of HBT classification models using SVM and MACCS structure keys were 91.3% and 100.0%, respectively. It showed that the performance of SVM method was obvious better than other four kinds of machine learning methods when developing HBT classification models. The advantage of SVM is not only to obtain good statistical performance, but also can be applied when some experimental data were lost. SVM method typically used a portion of training set as support vectors for classification. If the lost experimental data are the non-support vectors, it can not affect the performance of models.

2.3 Performance of regression models

2.3.1 *Regression models for fathead minnow toxicity* The pLC_{50} value of FHMT were used as the dependent variable and MACCS structural keys and FP4 fingerprint for each compound were used as the independent variables to develop FHMT regression model. The estimated square of correlation coefficient (R^2) and *RMSE* for FHMT regression model were listed in Table 4. The R^2 and *RMSE* using MACCS structural keys for fathead minnow toxicity test set were 0.878 and 0.258, respectively. Comparing the performance of the regression models, MACCS structural keys were also better than FP4 fingerprints. Fig. 3 showed the plot of the predictive R^2 for the fathead minnow toxicity test set using MACCS structural keys and FP4 fingerprints, respectively. As shown in Fig. 3, we currently investigated the chemical and toxicological reasoning behind the four outliers, such as 5,5-dimethyl-1,3-cyclohexanedione, malononitrile, 2,6-diphenylpyridine and 2,3-methylene bis(3,4,6-trichlorophenol).

Table 4 Performance of regression models for fathead minnow and honey bee data set using support vector machine regression algorithm

Data set	Species	MACCS		FP4	
		R^2	*RMSE*	R^2	*RMSE*
Training set	FHM	0.881	0.213	0.647	0.601
	HB	0.833	0.323	0.854	0.290
Test set	FHM	0.878	0.258	0.653	0.804
	HB	0.663	1.11	0.422	1.95

Note: FHM: fathead minnow; HB: honey bee; R^2: the square of correlation coefficient; *RMSE*: root mean square error.

2.3.2 *Regression models for honey bee toxicity* The pLD_{50} value of HBT were used the dependent

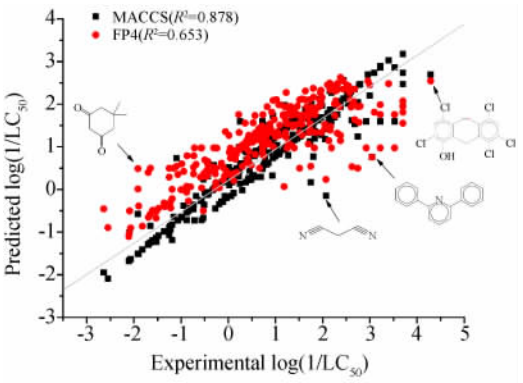


Fig. 3 The plot showed the square of correlation coefficient(R^2) of support vector machine regression models for fathead minnow toxicity test set using MACCS structural keys and FP4 fingerprints, respectively

variable and MACCS structural keys and FP4 fingerprints for each compound were used as the independent variables respectively to develop HBT regression model. The R^2 and *RMSE* for HBT predictive regression models were listed in Table 4. The R^2 and *RMSE* using MACCS structural keys for honey bee toxicity test set were 0.663 and 1.11, respectively. Comparing the performance of the HBT regression models, MACCS structural keys were obvious better than FP4 fingerprint which was in agreement with the results of FHMT prediction regression model. Fig. 4 showed the plot of R^2 for honey bee toxicity test set using MACCS structural keys and FP4 fingerprints, respectively. As shown in Fig. 4, we also further investigated the chemical and toxicological reasoning behind the three outliers, such

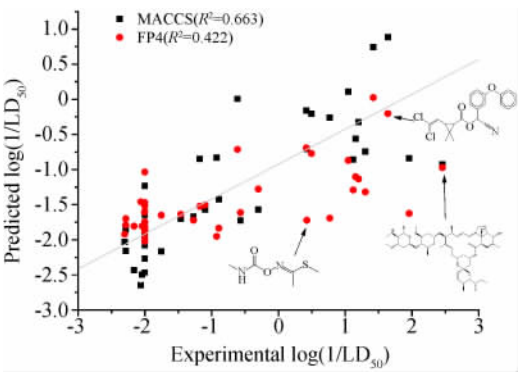


Fig. 4 The plot showed the square of correlation coefficient(R^2) of support vector machine regression models for honey bee toxicity test set using MACCS structural keys and FP4 fingerprints, respectively

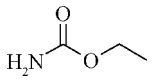
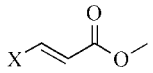
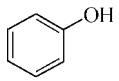
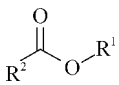
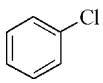
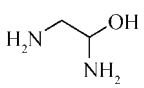
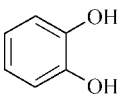
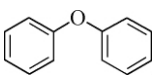
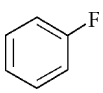
as mythomyl CAS 16752-77-5, emamectin benzoate CAS 155569-91-8 and *beta*-cypermethrin CAS 52315-07-8.

2.4 Identifying key substructure patterns

Some representative substructure patterns for FHMT and HBT compounds were identified by our previous developed substructure pattern recognition method [9]. The representative substructure patterns, the frequency of patterns and IG value were listed in Tables 5 and 6. As listed in Table 5, the patterns of urethane, vinylogous_halide, phenol, carboxylic_ester, aldehyde and arylchloride were present more frequently in high acute FHMT compounds than in low acute FHMT compounds. However, the patterns of primary_amide and 1,2-aminoalcohol were present more frequently in low acute FHMT compounds than in high acute FHMT compounds class. The patterns of 1,2-diphenol, dialkylthioether, diarylether and arylfluoride were only present in high acute FHMT compounds class. As listed in Table 6, the patterns of trifluoromethyl, amide, urea and carboxylic_acid were present more frequently in low acute HBT compounds class than high acute HBT compounds class. The pattern of nitrile, dialkylthioether, chloroalkene and sulfenic_derivative were present more frequently in high acute HBT compounds class than low acute HBT compounds class. Furthermore, phosphoric_acid_derivative was only present in high acute HBT compounds class (evidenced by the organophosphate insecticides) and vinylogous_amide was only present in low acute HBT compounds class. If one pattern was only present in toxicity class, this pattern was called structural alert. That is, if a compound has a pattern of 1,2-diphenol, dialkylthioether, diarylether and arylfluoride, it has a higher potential to exhibit toxicity for FHM. If a compound has the patterns of phosphoric_acid_derivative, it has a higher potential to exhibit toxicity for HB.

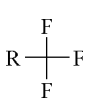
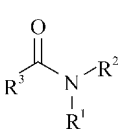
The interpretation of QSAR models is an important issue. In this study, the diverse FHMT and HBT data covered a wide range of toxicity mechanism, which ranged from narcosis Ⅰ, narcosis Ⅱ, or narcosis Ⅲ, electrophile/proelectrophile reactivity, and CNS seizure mechanisms (including AChE inhibition) [20-21]. These complex toxicity mechanisms can be explained by representative

Table 5 Some representative substructure patterns with their possible classes for fathead minnow toxicity(FHMT) were identified by Information Gain analysis

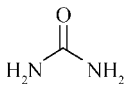
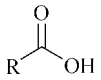
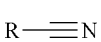
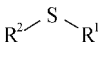
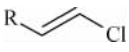
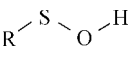
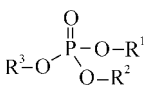
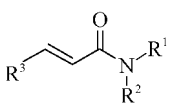
Substructure	SMARTS of pattern	Descriptions	N_P	N_N	$P(t)$	$N(t)$	IG
	[#7X3] [#6](=[OX1]) [#8X2] [#6]	Urethan	10	1	0.018	0.003	0.003
	[#6X3](=[OX1]) [#6X3]=,: [#6X3] [FX1,ClX1,BrX1,IX1]	Vinylogous_halide	15	2	0.027	0.007	0.003
	[OX2H] [c]	Phenol	111	25	0.198	0.087	0.016
	[CX3;$([R0][#6]),$([H1R0])](=[OX1]) [OX2] [#6;!$(C=[O,N,S])]	Carboxylic_ester	68	9	0.121	0.031	0.018
	[$([CX3H][#6]),$([CX3H2])]=[OX1]	Aldehyde	72	6	0.128	0.021	0.028
	[Cl] [c]	Arylchloride	89	10	0.158	0.035	0.029
	[CX3;$([R0][#6]),$([H1R0])](=[OX1]) [NX3H2]	Primary_amide	1	13	0.002	0.045	0.019
	[OX2H] [CX4;!$(C([OX2H])[O,S,#7,#15,F,Cl,Br,I])] [CX4;!$(C([N])[O,S,#7,#15])] [NX3;!$(NC=[O,S,N])]	1,2-Aminoalcohol	1	16	0.002	0.056	0.024
	[OX2H] [c] [c] [OX2H]	1,2-Diphenol	6	0	0.011	0	0.004
	[SX2]([CX4;!$(C([OX2])[O,S,#7,#15,F,Cl,Br,I])]) [CX4;!$(C([OX2])[O,S,#7,#15])]	Dialkylthioether	15	0	0.027	0	0.011
	[c] [OX2] [c]	Diarylether	18	0	0.032	0	0.013
	[F] [c]	Arylfluoride	19	0	0.034	0	0.014

Note: N_P are the number of high acute FHMT compounds in entire data set with pattern t, N_N are the number of low acute FHMT compounds in entire data set with pattern t, $P(t)$ and $N(t)$ are the proportion of the compounds with pattern t in high acute and low acute FHMT compounds, respectively.

Table 6 Some representative substructure patterns with their possible classes for honey bee toxicity(HBT) were identified by Information Gain analysis

Substructure	SMARTS of pattern	Descriptions	N_P	N_N	$P(t)$	$N(t)$	IG
	[FX1] [CX4;!$([H0][Cl,Br,I]);!$([F][C]([F])([F])[F])]([FX1])([FX1])	Trifluoromethyl	7	13	0.071	0.135	0.008
	[CX3;$([R0][#6]),$([H1R0])](=[OX1]) [#7X3;$([H2]),([H1][#6;!$(C=[O,N,S])]),([#7]([#6;!$(C=[O,N,S])])[#6;!$(C=[O,N,S])])]	Amide	4	10	0.040	0.104	0.011

续表(Continued)

Substructure	SMARTS of pattern	Descriptions	N_P	N_N	$P(t)$	$N(t)$	IG
	[#7X3;!$([#7][#6])][#6X3](=[OX1])[#7X3;!$([#7][#6])]	Urea	1	9	0.010	0.094	0.029
	[CX3;$([R0][#6]),$([H1R0])](=[OX1])[OX2][#6;!$(C=[O,N,S])]	Carboxylic_acid	4	8	0.040	0.188	0.042
	[NX1]#[CX2]	Nitrile	8	2	0.081	0.021	0.014
	[SX2]([CX4;!$(C([OX2])[O,S,#7,#15,F,Cl,Br,I])])[CX4;!$(C([OX2])[O,S,#7,#15])]	Dialkylthioether	7	1	0.071	0.010	0.019
	[ClX1][CX3]=[CX3]	Chloroalkene	12	1	0.121	0.040	0.042
	[SX2;$([H1]),$([H0][#6])][#6]	Sulfenic_derivative	17	1	0.172	0.010	0.067
	[PX4D4](=[#6])([#6])([#6])[#6]	Phosphoric_acid_derivative	28	0	0.283	0	0.157
	[#6X3](=[OX1])[#6X3]=,:[#6X3][#7X3;$([H2]),$([H1][#6;!$C=[O,N,S])],$([#7]([#6;!$(C=[O,N,S])])[#6;!$(C=[O,N,S])])]	Vinylogous_amide	0	8	0	0.083	0.043

Note: N_P are the number of high acute HBT compounds in entire data set with pattern t, N_N are the number of low acute HBT compounds entire data set with pattern t, $P(t)$ and $N(t)$ are the proportion of the compounds with pattern t in high acute and low acute HBT compounds, respectively.

substructure patterns identified by IG analysis. For example, the patterns of aldehyde, arylchloride, phenol, 1,2-diphenol and dialkylthioether have more potential toxicity to FHM and HB, because these pattern can covalent bind with biological macromolecules or can react with nucleophilic groups ($-NH_2$, $-OH$, $-SH$) in biological macromolecules such as DNA and proteins and result in narcosis or electrophile/proelectrophile reactivity toxicity[20,22]. As listed in Table 6, the pattern of phosphoric_acid_derivative was on represent in high acute BHT compounds class, which was in agreement with findings of Christine *et al.* that phosphoric_acid_derivative easily take place oxidative phosphorylation uncoupling with organisms[21]. These meaningful substructures can potentially provide scaffolds and be interpreted by chemists to gain understanding and guide modification information to reduce FHMT and HBT. Thus, our models had higher information content than historical descriptors employing exhaustive structural features.

2.5 Comparison with previous reported models

A direct comparison of our results with previous studies is inappropriate, because the data set, used data description methods were different between the various models. Nevertheless, a simple comparison of the model statistics could provide some basic information about the accuracy of the various FHMT and HBT prediction methodologies. As listed in Table 2, the same training set and test set in the work of Michielan *et al.* were used in this work. Comparing the results of our models with Michielan *et al.*[7], the performance of our models was better than Michielan' work. The overall predictive accuracy of 99.3% and the *SP* of 99.0% using *k*-NN and MACCS structural keys were significantly higher than 89.2% and 81.8% in the work of Michielan *et al.*, respectively. Tan *et al.* reported a SVM model for FHMT using 611 compounds, which gave an average *SE* 95.5%, *SP* 79.3% and *Q* 91.0%[6]. The *SE*, *SP* and *Q* value of our SVM model using the MACCS structural keys for

FHM test set were 98.0%, 91.9% and 95.9% respectively, which was obvious better than models performance of Tan *et al.*.

3 Conclusion

In this study, the robust classification and regression models for FHMT and HBT prediction were developed using different machine learning methods and substructure pattern recognition method. All models were validated by independent test set and the performances of our methods were better than literature reports. Five different machine learning methods including SVM, C4.5 DT, *k*-NN, RF and NB were evaluated here. The performances of FHMT and HBT classification models showed that SVM and *k*-NN algorithms were the superior algorithms than others. The average predictive accuracy of the FHMT classification models to test set with MACCS structural keys was 95.9% and 99.3% for SVM and *k*-NN algorithms, respectively. The average predictive accuracy and AUC of ROC curve for HBT test using SVM with MACCS structural keys was 95.0% and 0.973, respectively. The square of correlation coefficient of regression models were 0.878 for FHMT test set and 0.663 for HBT test set using MACCS structural keys and support machine regression algorithm. Moreover, some representative substructure patterns for FHMT and HBT compounds were identified, which can be applied to guide modification information for chemical detoxification. This study provided a useful strategy and robust tool for evaluating toxicological properties of industrial chemicals and pesticides in the environmental hazard assessment.

谨以此文敬贺李正名院士八十华诞!

Reference:

[1] HENGSTLER J G, FOTH H, KAHL R, *et al*. The reach concept and its impact on toxicological sciences [J]. *Toxicology*, 2006, 220(2-3): 232-239.

[2] COLLINS F S, GRAY G M, BUCHER J R. Toxicology Transforming environmental health protection [J]. *Science*, 2008, 319(5865): 906-907.

[3] DEVILLERS J. A decade of research in environmental QSAR [J]. *SAR QSAR Environ Res*, 2003, 14(1): 1-6.

[4] JAMES D, MINH H P D, AXEL D, *et al*. Modeling the acute toxicity of pesticides to *Apis mellifera* [J]. *B Insectol*, 2003, 56(1): 103-109.

[5] MARCO V, MARCELLA M G, DAVIDE C. QSARs for toxicity of organophosphorous pesticides to Daphnia and honeybees [J]. *Sci Total Environ*, 1991, 109-110: 605-622.

[6] TAN N X, LI P, RAO H B, *et al*. Prediction of the acute toxicity of chemical compounds to the fathead minnow by machine learning approaches [J]. *Chemometr Intell Lab Syst*, 2010, 100(12): 66-73.

[7] MICHIELAN L, PIREDDU L, FLORIS M, *et al*. Support vector machine(SVM) as alternative tool to assign acute aquatic toxicity warning labels to chemicals [J]. *Mol Inf*, 2010, 29: 51-64.

[8] Environmental Protection Agency. Ecotox database: http://cfpub.epa.gov/ecotox/ [S/OL]. United States, (2010-05)

[9] SHEN J, CHENG F X, XU Y, *et al*. Estimation of ADME properties with substructure pattern recognition [J]. *J Chem Inf Model*, 50(6): 1034-1041.

[10] DURANT J L, LELAND B A, HENRY D R, *et al*. Reoptimization of MDL keys for use in drug discovery [J]. *J Chem Inf Comput Sci*, 2002, 42(6): 1273-1280.

[11] Open Babel. http://openbabel.org/ [S/OL]. (2010-01-18).

[12] CHANG C C, LIN C J. LIBSVM: a library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm [S/OL]. (2010-01-18).

[13] LI H D, LIANG Y Z, XU Q S. Support vector machines and its applications in chemistry [J]. *Chemom Intell Lab Syst*, 2009, 95: 188-198.

[14] CORINNA C, VLADIMIR V. Support-vector networks [J]. *Machine Learning*, 1995, 20: 273-297.

[15] VAPNIK V, GOLOWICH S, SMOLA A A. Support vector method for function approximation, regression estimation, and signal processing [M] // MOZER M, JORDAN M, PETSCHE T. Advances in Neural Information Processing Systems 9. Cambridge, MA, 1997: 281-287.

[16] SMOLA A J. Regression estimation with support vector learning machines [D]. Munchen, Technische University Munchen, 1996.

[17] BREIMAN L. Random Forests [J]. *Machine Learning*, 2001, 45(1): 5-32.

[18] QUINLAN J R. C4.5: Programs for Machine Learning [M]. Morgan Kaufmann Publishers, 1993.

[19] WATSON P. Naive Bayes classification using 2D pharmacophore feature triplet vectors [J]. *J Chem Inf Model*, 2008, 48(1): 166-178.

[20] PAPA E, VILLA F, GRAMATICA P. Statistically validated QSARs, based on theoretical descriptors, for modeling aquatic toxicity of organic chemicals in *Pimephales promelas* (fathead minnow) [J]. *J Chem Inf Model*, 2005, 45(5): 1256-1266.

[21] CHRISTINE L R, STEVEN P B, STEVEN J B, *et al*. Predicting modes of toxic action from chemical structure: Acute toxicity in the fathead minnow (*Pimephales promelas*) [J]. *Environ Toxicol Chem*, 1997, 16(5): 948-967.

[22] LIPNICK R L. Outliers: their origin and use in the classification of molecular mechanisms of toxicity [J]. *Sci Total Environ*, 1991, 109-110: 131-153.

(Ed. JIN S H)