Journal of
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

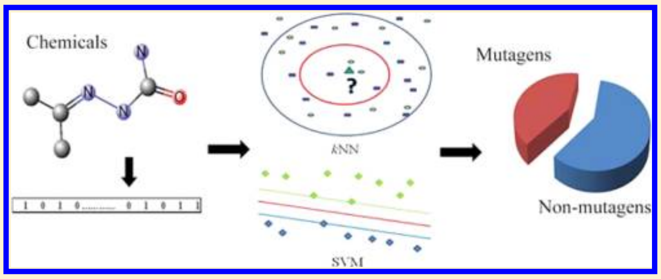
In silico Prediction of Chemical Ames Mutagenicity

Congying Xu,[†] Feixiong Cheng,[†] Lei Chen,[†] Zheng Du,[†] Weihua Li,[†] Guixia Liu,*[,†,‡] Philip W. Lee,[†] and Yun Tang*[,†]

[†]Shanghai Key Laboratory of New Drug Design, School of Pharmacy, East China University of Science and Technology, 130 Meilong Road, Shanghai 200237, China

[‡]State key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, 555 Zuchongzhi Road, Shanghai 201203, China

S *Supporting Information*



ABSTRACT: Mutagenicity is one of the most important end points of toxicity. Due to high cost and laboriousness in experimental tests, it is necessary to develop robust *in silico* methods to predict chemical mutagenicity. In this paper, a comprehensive database containing 7617 diverse compounds, including 4252 mutagens and 3365 nonmutagens, was constructed. On the basis of this data set, high predictive models were then built using five machine learning methods, namely support vector machine (SVM), C4.5 decision tree (C4.5 DT), artificial neural network (ANN), *k*-nearest neighbors (*k*NN), and naïve Bayes (NB), along with five fingerprints, namely CDK fingerprint (FP), Estate fingerprint (Estate), MACCS keys (MACCS), PubChem fingerprint (PubChem), and Substructure fingerprint (SubFP). Performances were measured by cross validation and an external test set containing 831 diverse chemicals. Information gain and substructure analysis were used to interpret the models. The accuracies of fivefold cross validation were from 0.808 to 0.841 for top five models. The range of accuracy for the external validation set was from 0.904 to 0.980, which outperformed that of Toxtree. Three models (PubChem-*k*NN, MACCS-*k*NN, and PubChem-SVM) showed high and reliable predictive accuracy for the mutagens and nonmutagens and, hence, could be used in prediction of chemical Ames mutagenicity.

■ INTRODUCTION

In the postgenomic era, in silico methods have played important roles in the drug discovery process. It is necessary for a drug to go to the market that the drug owns perfect ADMET (absorption, distribution, metabolism, excretion, toxicity) properties. Drug toxicology is one of the crucial research fields in the preclinical study. Toxicity is a leading cause of attrition at all stages of the drug development.[1] In silico prediction of compound toxicity which can reduce the expenses of the company and save a lot of time has attracted considerable attention. As we all know, the mutagenic effect has a close relationship with the carcinogenicity. Nowadays, the most widely used assay for testing the mutagenicity of compounds is the Ames experiment which was invented by a professor named Ames.[2] The Ames test is a short-term bacterial reverse mutation assay detecting a large number of compounds which can induce genetic damage and frameshift mutations.[3] The estimated interlaboratory reproducibility rate of Salmonella test data is only 85%.[4,5] This shows the intrinsic limitation of the in vitro test. Therefore, it is really necessary to develop a good model for mutagenicity prediction instead of in vitro tests.

Over the past decades, there have been mainly two kinds of models: one is the statistical model, and the other is the structural alerts based model. Many statistical learning algorithms have been used to build the models, such as support vector machine (SVM),[6−11] decision tree, random forest,[12] *k*-nearest neighbors (*k*NN),[6] artificial neural network (ANN),[13] and so on. These statistical models have an acceptable predictive capability, however, they are very complicated due to many steps such as descriptors calculation, descriptors selection, model building, and hard to explain. Some other researchers invent models based on special structure fragments which are responsible for the toxicity. Kazius et al. used 29 new toxicophore containing substructures to classify the mutagenicity of the investigated 4337 compounds with a total classification error of 18%.[14] Zheng et al. built a mutagenic probability prediction model by a novel molecular electrophilicity vector which was mainly described in terms of atomic electrophilicity combined with support vector machine learning algorithm. The predictive accuracy of this model is 90.13% for tenfold cross validation.[15] Ferrari et al. developed a cascade model with the max predictive accuracy of 82.1%, which integrated two different techniques together. That is to say, a statistical model was followed by further investigation for specific structural alerts in the "safe" subset of the prediction.[16]

Received: August 23, 2012
Published: October 2, 2012

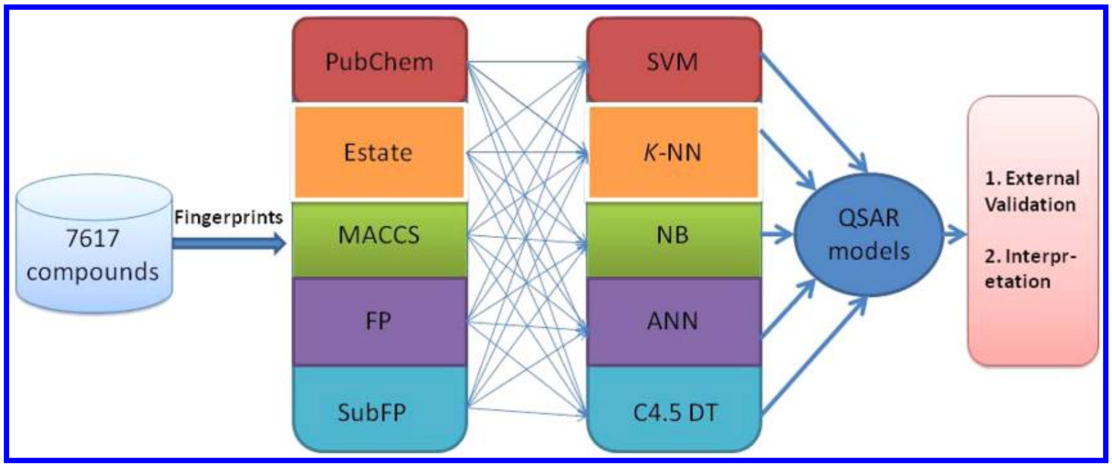


Figure 1. Workflow of our protocol: (ANN) artificial neural network; (SVM) support vector machine; (*k*NN) *k*-nearest neighbors; (C4.5 DT) C4.5 decision tree; (NB) naïve Bayes; (MACCS) MACCS keys; (SubFP) substructure fingerprints; (FP) CDK fingerprint; (PubChem) PubChem fingerprints; (Estate) Estate fingerprint.

In addition to the models of mutagenicity prediction published in the literatures, numerous commercial and open source software packages have also been available. Recently, Hillebrecht et al. compared the predictive power of four commonly used in silico tools: DEREK for windows (Dfw), Leadscope Model Applier (LSMA), MultiCASE (MC4PC), and Toxtree. As a general tendency, expert systems showed higher sensitivity and lower specificity compared with statistical-based tools, which displayed the opposite behavior. The combined use of an expert system with a statistical-based technique is recommended.[17]

Nevertheless, the predictive accuracy and robustness of the current models remain unsatisfactory. Due to the database limitation, the application domain is bounded, too. We observed a common phenomenon that many structural alerts based models have reasonably predictive accuracy of mutagenicity. This fact reflects a close relationship between the mutagenicity and 2D structure, which inspired us to use fingerprints as attributes for prediction of mutagenecity. Recently, we have used substructure-based methods in some end points successfully.[9,10] Molecular fingerprints are widely used in similarity searching,[18] virtual screening,[19] and classification.[20,21]

In this paper, we constructed a high-quality data set containing more than 8300 compounds with known mutagenicity property which is believed to be the largest public one up to now. On the basis of this data set, we proposed some binary models using different molecular fingerprints along with different statistical algorithms. In order to validate the effectiveness of our models, several predictions have been performed including cross validation and external validation. At last, information gain and substructure analysis were also performed to interpret models. A detailed protocol could be seen in Figure 1.

■ MATERIALS AND METHODS

Data Preparation. The training set for model building was collected from four papers.[7,19−21] The data set for external validation was extracted from the Web site of Lazar toxicity predictions.[22] The entire database was prepared as following. First, apart from the four false SMILES strings, duplicate molecules were removed from the five sources by using canonical SMILES. Second, molecules without clear E or Z configuration were removed. Third, inorganic compounds were omitted from the data set. The last step was to eliminate the tautomers and compounds with molecular weight less than 40 or more than 800 in the data set. When doing the data set curation, we followed one principle. For a given compound, if the experimental mutagenicity data varied in different sources, the compound was cleared out. For compounds without defined steric configuration or tautomers, if the experimental mutagenicity data was alike, then only one structure was kept, and the others were deleted.

Calculation of Molecular Fingerprints. Molecular fingerprints are widely used in similarity searching and classification. Five fingerprints were used in this work. They are CDK fingerprint (FP, 1024 bits), Estate fingerprint (Estate, 79 bits), MACCS keys (MACCS, 166 bits), PubChem fingerprint (PubChem, 881 bits), and Substructure fingerprint (SubFP, 307 bits). All the fingerprints were calculated by the PaDEL-Descriptor software.[23] Before the calculation, all the SMILES strings in the data set were processed by ChemAxon Standardizer[24] using the following options: add explicit hydrogens, aromatize, clean 2D, remove fragment.[25]

Machine Learning Methods. Five different methods, including SVM (support vector machine), C4.5 decision tree, ANN (artificial neural network), *k*NN (*k*-nearest neighbor), and NB (naïve Bayes), were used for model building. C4.5 decision tree, *k*NN, and NB were performed in Orange 2.0 (version 2.0b, freely available at http://www.ailab.si/orange/),[26] the SVM algorithm is provided by the open source LIBSVM [LIBSVM2.9 package].[27] ANN is performed in KNIME2.4.2 (freely available at http://www.knime.org).[28]

Artificial Neural Network (ANN). ANN is a flexible mathematical structure which is capable of identifying complex nonlinear relationship between input and output data sets.[29] In our work, the network consisted of three layers containing one input layer, one hidden layer, and one output layer in KNIME. KNIME has been implemented with the RProp algorithm for multilayer feedforward networks.[30] The number of neurons in hidden layer is ten.

k-Nearest Neighbor (kNN). For each test sample $Z = (x', y')$, the algorithm calculate the distance or similarity between each training sample (x, y) to determine the list of its nearest neighbor. Then it can be classified in accordance with the majority of the nearest neighbors. In order to reduce the impact of k value, a distance-weighted method is used. Itskowitz has confirmed that the model performance is determined by three factors: the variable selection, the value of k (the number of nearest neighbors), and the shape of the distance weighting function.[31] In our study, Euclidean distance and distance-weighted parameters have been chosen, and the k value is set for 3.

Support Vector Machine (SVM). SVM was first developed by Vapnik[32] as a general data modeling methodology, aiming at minimizing structure risk under the frame of VC theory. This algorithm constructs an optimal hyperplane separating two sets of positives and negatives. The performance of SVM for the classification depends on the combination of several parameters. In our study, a radial basis function (RBF) kernel function was chosen to seek the optimal pair of the penalty parameter C and different kernel parameter γ. A grid search script was used in the LIBSVM package.

Naïve Bayes (NB). The NB classifier method is a simple classification method based on the Bayes rule for the conditional probability. The prior probability can be straight forwardly estimated from the training set, while the marginal probability can be ignored, since it is the same to all of the classes.[33] Orange with the default setting was used to perform the NB classification.

C4.5 Decision Tree (C4.5 DT). C4.5 DT is a standard benchmark in machine learning. For this reason, it is incorporated in Orange. A decision tree takes as input an object or situation described by a set of properties and outputs a yes or no decision. An instance is classified by starting at the root node of the decision tree, testing the attribute specified by this node, and then moving down to the tree branch according to the value of the attribute.[34] The more detailed descriptions of C4.5 DT are available in the original literature.[35] All C4.5 DT parameters had their default values in Orange.

Performances Evaluation. All models were validated by 5-fold cross validation and a diverse external validation set. All the models were assessed by the counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The sensitivity (SE) which is the predictive accuracy of the mutagenic compounds, the specificity (SP) which means predictive accuracy of the nonmutagenic compounds, and the whole predictive accuracy (Q) which represents the total correct predictive accuracy of mutagens and nonmutagens were calculated with the following equations.[8]

$$Q = (TP + TN)/(TP + TN + FP + FN) \tag{1}$$

$$SP = TN/(TN + FP) \tag{2}$$

$$SE = TP/(TP + FN) \tag{3}$$

In addition, the receiver operating characteristic (ROC) curve where the true positive rate (or sensitivity) against the false positive rate (1-specificity) was plotted, and the area under the ROC (AUC) curve was also computed. The AUC is the probability of active compounds being ranked earlier than decoy compounds, and the value of AUC ranges from 0.5 (useless random classifiers) to 1 (perfect classifiers).[36]

Analysis of Privileged Substructures or Structural Alerts. The privileged substructure fragments and the structural alerts were analyzed using the information gain[37] and substructure fragment analysis.[38] The privileged structures known to have provided ligands for diverse receptors are capable to illustrate the biological mechanism.[39] If a substructure was more frequently presented in mutagenicity chemical class, this substructure was called a privileged substructure involved in chemical mutation. The frequency of a fragment in mutagens was defined as following:

$$\text{frequency of a fragment} = \frac{(N_{\text{fragment\_class}} \times N_{\text{total}})}{(N_{\text{fragment\_total}} \times N_{\text{class}})} \tag{4}$$

where $N_{\text{fragment\_class}}$ is the number of compounds containing the fragment in mutagens; N_{total} is the total number of compounds; $N_{\text{fragment\_total}}$ is the total number of compounds containing the fragment; and N_{class} is the number of mutagens.

The structure alerts are defined as molecular functional groups that are known to bring the toxicity. Their appearance in a chemical structure alerts the researchers to the potential toxicities of the test compounds.[40] Structural alerts (SAs) are the important predictive toxicity tools due to they are derived directly from mechanistic knowledge.[41] Here, we used information gain,[37] substructure fragment analysis,[38] and the MoSS module which searches for frequent molecular fragments in a set of molecules in KNIME[28] to accomplish this job. In MoSS module, the "minimum fragment size" value is an important parameter. Big values mean big fragments which are hard to find out, and small values mean small fragments which are too common to give useful information for mutagenicity. In our study, this value was set to 5. All the other parameters were default.

■ RESULTS

Data Set Analysis. In our study, in total 8348 compounds that were derived from 14 461 chemicals collected from five sources were used for the model building and validation. The training set contained 4252 mutagens and 3365 nonmutagens, while the external validation set included 614 mutagens and 117 nonmutagens (Table 1). To the best of our knowledge, this is the largest public Ames mutagenicity data set used in QSAR study. All data sets were given in Table S1 of the Supporting Information.

Table 1. Detailed Statistical Description of Chemicals Used in the Training Set and External Validation Set

data sets	mutagens	nonmutagens	total
training set	4252	3365	7617
external validation set	614	117	731
total	4866	3482	8348

Because of the importance of data quality to in silico prediction model, we have paid more attention to it. The five sources commonly contained some data sets such as CPDB[42] and CCRIS.[43] Therefore, the duplicated structures were removed when the original data sets were collected together, so that each compound was unique in the final data set used in this study. There are 3896 compounds occurred twice, among which only 20 ones have contradictory mutagenicity data. This small part of inconsistent compounds gives the good confidence of the data set.

The diversity of the database is very important to the predictive accuracy of the models. We have investigated the chemical space distribution by calculating the molecule weight (MW) and Ghose−Crippen LogKow (ALogP) of the training set and the external validation set. The distribution scatter diagram was presented in Figure 2. As shown in Figure 2, the

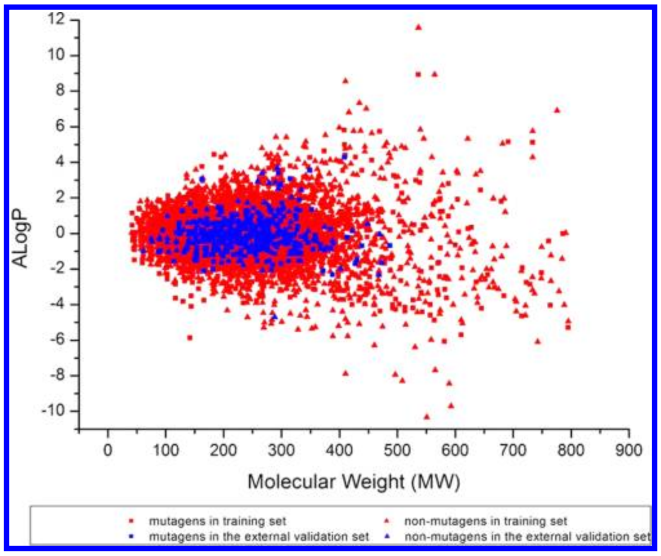


Figure 2. Diversity distribution of the training set (n = 7617 compounds), external validation set (n = 731 compounds). Chemical space defined by molecular weight (MW) as X-axis, and ALogP as Y-axis. In the picture, Red stands for the training set, blue stands for the external validation set, square represents the mutagens, and triangle represents the nonmutagens.

chemical space of the external validation set was within the scope of the training set. It was difficult to classify the mutagenic property of compounds according to the molecular weight (MW) and the ALogP due to their disordered relationship between the mutagens and nonmutagens.

Performances of Fivefold Cross Validation. The combinatorial predictive models were built using five different fingerprints along with five statistical algorithms. As a result, there were a total of 25 binary classification models generated by combination. According to the results of fivefold cross validation, we could draw two conclusions at least. The first one was that five machine learning methods differed greatly in prediction ability. The second one was that good models always used MACCS and PubChem as attributes. Especially, the publicly available version of MACCS consists of 166 structural feature bits,[44] while the length of PubChem is 881. MACCS molecular fingerprint is based on the well-defined structural fragments dictionary, which is full of structural information. Considering the amount of computation and prediction accuracy, MACCS was recommended for building the in silico Ames mutagenicity models. In total, the top five models were MACCS-*k*NN, PubChem-*k*NN, FP-SVM, MACCS-SVM, and PubChem-SVM. The detailed performances of these models were given in Table 2.

Performance of External Validation. The external validation set was used for testing the five best models. The best result was amazing at the highest accuracy of 98.0% using the PubChem fingerprint combined with *k*NN algorithm. Considering the values of SP and SE, two models (MACCS-SVM, FP-SVM) were not good enough due to their bad SE values. The detailed results can be seen in Table 3. We supposed that the high predictive accuracy of the external validation set was caused by the imbalance of the mutagens and nonmutagens in the data set, and wherein the ratio of mutagens and nonmutagens was 5.248. In order to verify this hypothesis, a balanced external validation has been constructed to test the prediction models as follows. Since original external validation set contained 614 mutagens and 117 nonmutagens, 117 nonmutagens have been kept in the balanced external validation set. The 614 mutagens have been grouped into 117 clusters according to the ECFP_6 fingerprint, each cluster-center molecule was chosen as a part of the balanced external validation set. In total, the balanced external validation set contained the 234 compounds (data set could be seen in Table S1 of the Supporting Information). The top three models (PubChem-*k*NN, MACCS-*k*NN, and PubChem-SVM) were used to test the balanced external validation set. The performance could be seen in Table S2 of the Supporting Information. As we can see, the lower accuracy is due to the

Table 2. Performances of Classification Models for the Fivefold Cross Validation Using Different Fingerprints and Modeling Methods[a]

model name	Q	SE	SP	AUC
PubChem-NB	0.658	0.664	0.652	0.691
Estate-NB	0.638	0.688	0.575	0.696
FP-NB	0.647	0.654	0.640	0.701
MACCS-NB	0.650	0.574	0.745	0.728
SubFP-NB	0.661	0.612	0.723	0.737
Estate-C4.5 DT	0.733	0.779	0.675	0.776
SubFP-C4.5 DT	0.737	0.786	0.676	0.782
Estate-ANN	0.722	0.768	0.662	0.787
FP-C4.5 DT	0.780	0.816	0.735	0.794
Estate-*k*NN	0.745	0.772	0.710	0.800
SubFP-*k*NN	0.743	0.787	0.687	0.804
SubFP-ANN	0.748	0.809	0.670	0.816
MACCS-C4.5 DT	0.804	0.839	0.759	0.820
PubChem-C4.5 DT	0.801	0.823	0.774	0.823
Estate-SVM	0.774	0.822	0.714	0.835
SubFP-SVM	0.776	0.826	0.711	0.839
FP-ANN	0.783	0.819	0.738	0.849
MACCS-ANN	0.789	0.836	0.729	0.850
FP-*k*NN	0.803	0.831	0.767	0.865
PubChem-ANN	0.802	0.843	0.750	0.868
MACCS-*k*NN	0.808	0.842	0.766	0.869
PubChem-*k*NN	0.821	0.852	0.782	0.879
FP-SVM	0.822	0.856	0.778	0.888
MACCS-SVM	0.841	0.865	0.811	0.901
PubChem-SVM	0.840	0.865	0.808	0.903

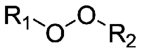
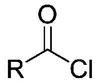
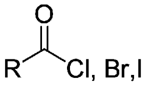
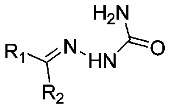
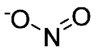
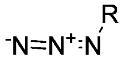
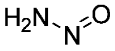
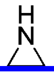
[a] ANN artificial neural network, SVM support vector machine, *k*NN *k*-nearest neighbors, C4.5 DT C4.5 decision tree, NB naïve Bayes, MACCS MACCS keys, SubFP substructure fingerprints, FP CDK fingerprint, PubChem PubChem fingerprints, Estate Estate fingerprint, SE sensitivity, SP specificity, Q overall predictive accuracy, AUC the area under receiver operating characteristic curve.

Table 3. Performance of Classification Models for the External Validation Set Using Different Fingerprints and Modeling Methods[a]

model name	Q	SE	SP	AUC
MACCS-*k*NN	0.959	0.985	0.821	0.956
PubChem-*k*NN	0.980	0.994	0.906	0.970
FP-SVM	0.904	0.982	0.496	0.903
MACCS-SVM	0.927	0.995	0.573	0.924
PubChem-SVM	0.952	0.995	0.726	0.949

[a] SVM support vector machine, *k*NN *k*-nearest neighbors, MACCS MACCS keys, SubFP substructure fingerprints, PubChem PubChem fingerprints, SE sensitivity, SP specificity, Q overall predictive accuracy, AUC the area under receiver operating characteristic curve.

Table 4. Eight Structural Alerts Using Information Gain (IG) Analysis and Frequency Value of Privileged Substructures

NO	SMARTS	Description	General structure	IG	F_m
1	[OX2H][OX2]	Hydroperoxide		0.001	1.535
2	[CX3;$([R0][#6]),$([H1R0])](=[OX1])[ClX1]	Acylchloride		0.002	1.605
3	[CX3;$([R0][#6]),$([H1R0])](=[OX1])[FX1,ClX1,BrX1,IX1]	Acylhalide		0.002	1.605
4[a]	[#7X2](=[#6])[#7X3][#6X3]([#7X3;!$([#7][#7])])=[OX1]	Semicarbazone		0.001	1.601
5	[NX2](=[OX1])[O;$([X2]),$([X1-])]	Nitrite		0.001	1.716
6	[NX1]~[NX2]~[NX2,NX1]	Azide		0.004	1.653
7	[NX2](=[OX1])N-*=O	Nitrosamide		0.004	1.684
8	[NX3H1r3]1[#6r3][#6r3]1	NH_aziridine		0.003	1.716

[a] The substructures were out of the Toxtree software.

balanced external validation set in some extent. At same time, the prediction results showed the stable robustness and precise prediction accuracy of the models.

Results of Structural Alerts. From the results of information gain analysis and frequency values of privileged substructures, we found 25 substructures responsible for the mutagenicity. They are alkyliodide, tertiary_arom_amine, secondary_mixed_amine, hydroperoxide, enolether, acylchloride, acylhalide, hydroxamic_acid, imidoacid, imidolactone, isothiourea, guanidine, semicarbazone, iminoarene, nitrite, nitroso, azide, nitrosamide, hydrazine, sulfonic_ester, epoxide, NH_aziridine, hetero_S, nitro, and hydroxylamine, respectively. Substructure names above are in the format of FP4 description (details can be seen in Table S3 of the Supporting Information). Among these 25 substructures, eight fragments can be considered as structural alerts according to their appearance in mutagens and nonmutagens. Seven structural alerts have been included in Toxtree (Table 4). The rest fragment is semicarbazone which was proved as a new structure alert. According to the result of MoSS in KNIME, we got seven fragments, in which one was incorrect from the view of the chemistry. The other six fragments and representative structures in the database were shown in Table 5.

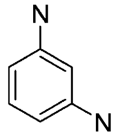
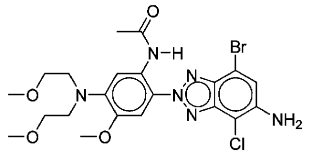
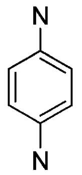
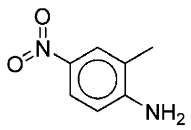
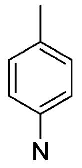
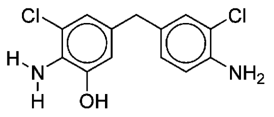
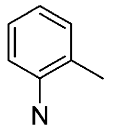
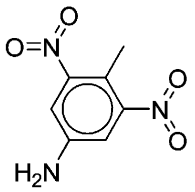
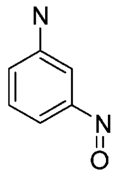
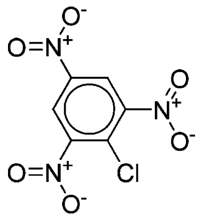
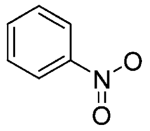
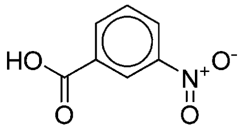
■ DISCUSSION

Comparison of Different Statistic Algorithms Used in Mutagenicity Prediction. Five algorithms (SVM, C4.5 DT, ANN, *k*NN, and NB) were used in this study. From the prediction performances, we could conclude that two algorithms, namely SVM and *k*NN, gave the best results. As is well-known, SVM has great power to fit the nonlinear relationship and seems to be something of a "gold standard" at the aspect of predictive accuracy.[45] Our emphasis is on the "*k*-nearest neighbor". We recommend that the outstanding performance of *k*NN in fivefold cross validation and external validation is because of its special algorithm and the unique end point of mutagenicity. If a compound had ability to binding covalently to DNA and caused DNA damage, it was usually positive in the standard mutagenicity assays. Compounds that are either electrophiles or can be activated to electorphilic reactive intermediates usually have some structure features.[46] Those features can be represented by the molecular fingerprints. A chemical can be classified in accordance with the majority of its nearest neighbors. The same category must have some structure similarity. For this reason, if the database was large enough and sufficiently diverse, models could own high prediction ability using the *k*NN algorithm in theory.

Comparison of Different Fingerprints Used in Model Building. From the values (Q, SE, SP, AUC) of the fivefold cross validation (Table 2), we could see that if using the same statistic algorithm, Estate fingerprint, and SubFP fingerprint did not perform well in trend. The length of the Estate fingerprint is only 79, which is too short to characterize molecules. Too much information loss led to the bad prediction. SubFP is a fingerprint method created from a set of SMARTS patterns defining functional groups. Each bit of SubFP corresponds to a particular chemical feature rather than to the general patterns. Therefore, features defined beforehand are often very restrictive to represent a large number of chemicals, which might be the reason why using SubFP as attribute for models was unreasonable.

Comparison with Toxtree Software. Toxtree was developed by Ideaconsult Ltd., which is a java-based, open-source, and freely used software for the public.[47] It is mainly based on structure alerts, famous for the toxicity prediction. It can also perform the "batch processing" of large amount of

Table 5. Results of MoSS Searched for Structural Alerts and Representative Structures

NO	Fragment	Representative structure	SMIlES
1			COCCN(CCOC)c1cc(NC(=O)C)c(cc1OC)n2nc3c(Br)cc(N)c(Cl)c3n2
2			Cc1cc(ccc1N)N(=O)=O
3			Nc1ccc(Cc2cc(O)c(N)c(Cl)c2)cc1Cl
4			Cc1c(cc(N)cc1N(=O)=O)N(=O)=O
5			[O-][N+](=O)c1cc(c(Cl)c(c1)[N+](=O)[O-])[N+](=O)[O-]
6			OC(=O)c1cccc(c1)[N+](=O)[O-]

compounds.[48] We have used the Toxtree to place our external validation set chemicals into mutagens and nonmutagens in order to compare with our models. The prediction outcomes of Toxtree were Q = 0.843, SE = 0.941, and SP = 0.675. The predictive accuracy value which is inferior to ours showed that our models are better.

Evaluation by Benchmark Data Set. Hansen and his co-workers had built a benchmark data set for in silico prediction of Ames mutagenicity. They also constructed four models using molecular descriptors on this new benchmark data set.[6] Here, we used this benchmark data set for validating our best models. We compared the results of our five best models (MACCS-*k*NN, PubChem-*k*NN, FP-SVM, PubChem-SVM, and MACCS-SVM) on the benchmark data set with the results from Hansen et al. under the same conditions. Comparison of the AUC values was shown in Table 6. From the table, we could conclude that using fingerprints as the attributes is comparable to molecular descriptor, according to the same methods: *k*NN and SVM. The highest value of AUC was 0.858 in our methods, which was almost equal to the results of Hansen et al. Especially, our *k*NN method combined with fingerprints has much better predictive ability than *k*NN combined with traditional molecular descriptors, in which the AUC value is 0.824 versus 0.790.

Table 6. Comparison of the the Area under Receiver Operating Characteristic Curve (AUC) Values between Our Five Best Models and the Work of Hansen et al.[6]

model name	AUC
FP-SVM	0.844
PubChem-SVM	0.858
MACCS-SVM	0.853
PubChem-*k*NN	0.824
M ACCS-*k*NN	0.824
SVM[a]	0.86 ± 0.01
GP[a]	0.84 ± 0.01
Random Forest[a]	0.73 ± 0.01
*k*NN[a]	0.79 ± 0.01

[a] The models of Hansen et al.[6]

Analysis of Substructural Alerts. Chemicals which may potentially interact with DNA nucleophilic centers are easily considered to be mutagenic, either directly or after metabolic activation. Acylhalides are potentially acylating agents toward DNA. In these substances, electron withdrawing effect of halogen atom increases the electrophilic character of the carbonyl carbon. Due to the large ring strain associated with the three-memberd ring, epoxides and aziridines may have the same mechanism by opening their rings. The two electrophilic carbons may react with nucleophilic centers of DNA, leading to alkylated products. Hydrazine derivatives can be activated by endogenous substances such as metal ions or enzymes such as cytochrome P450-dependent oxidases and flavin monooxygenases to form carbocations and carbon-centered radicals, resulting in active radical species causing DNA damage. Nitroso compounds require metabolic activation to form DNA adducts that are critical for their mutagenic and carcinogenic activity. The well-established major pathway is α-hydroxylation (adjacent to the *N*-nitroso group), catalyzed by cytochrome P450 enzymes. Azides also undergo metabolic oxidation by cytochrome P450 enzymes to generate monomethyltriazenes which are known alkylating agents and capable of methylating DNA.

Reactions of semicarbazone may be divided into two classes, of which one includes those reactions concerned with the Schiff base, such as hydrolysis reaction, the other includes reactions concerned with other linkages of the semicarbazide moiety. The products of hydrolysis are ketones or aldehydes and semicarbazide. Aldehydes or ketones contain potential carbocations which react as direct electrophiles to form adducts with DNA. Semicarbazide is known as one of the mutagens. As a result, considering semicarbazone as a structure alert is reasonable. However, because of the complexity of the mutagenicity, it is difficult to use a single structure alert to make a difference between mutagens and nonmutagens. In order to improve the confidence of our judgment, we can consider several structural alerts at the same time. If a compound contains more than one structure alert, it might be more possibly a mutagen (Table S4 in the Supporting Information).

Visual Analysis of Structural Alerts for Chemical Mutagenicity using KNIME. From the result of the MoSS, we can see that nitrobenzene derivatives and phenylamine derivatives are easily to be mutagens which react with DNA irreversibly. Aromatic nitroso compounds, aromatic *N*-oxide moiety, and nitroaromatic chemicals represent well-established classes of mutagens. The activation of compounds containing aromatic amine, nitro, and nitroso may produce the same intermediate aromatic hydroxylamine that can be further activated through enzymatic esterification to finally form electrophilic species (nitrenium ions). The nitrenium ions can bind covalently to bionucleophiles such as DNA.

■ CONCLUSION

In this study, we introduced binary Ames mutagenicity models with high predictive accuracy using fingerprints as attributes. The results of fivefold cross validation showed a great robustness of our models. Three models (PubChem-*k*NN, MACCS-*k*NN, and PubChem-SVM) which have high precision of prediction for both the mutagens and nonmutagens are full of practical value. All the tools used in this study for model building are free of charge and easily to be accessible. The results also demonstrate that fingerprints as attributes for classification models are powerful tools and it is possible to predict mutagenicity directly from 2D structures.

A comparison of prediction ability between Toxtree software and our models was also done. The results indicated that our models were better than Toxtree. In order to prove the effectiveness of our combinatorial methods, benchmark data set was employed to construct models. The predictive accuracy of the models based on the benchmark data set and our combinatorial methods was better than that of Hansen et al. when considering the same statistic algorithms. At the end of our study, we investigated the privileged substructure fragments and the structural alerts via information gain and MoSS module. Because of the limit of computational resource, we do not use some other excellent statistic algorithms like random forest in our study. According to the results of this paper, we suggest that the modeling methods used in this article can be promoted to other toxicity end points.

■ ASSOCIATED CONTENT

S Supporting Information

Additional details on materials and tables (Tables S1−S3). This material is available free of charge via the Internet at http://pubs.acs.org

■ AUTHOR INFORMATION

Corresponding Author

*Tel.: +86-21-64250811. Fax: +86-21-64253651. E-mail: gxliu@ecust.edu.cn (G.L.), ytang234@ecust.edu.cn (Y.T.).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We gratefully acknowledged the financial supports from the 863 Project (Grant 2012AA020308), the Fundamental Research Funds for the Central Universities (WY1113007), the State Key Laboratory of Drug Research (Grant SIMM1203KF-13), and the Shanghai Committee of Science and Technology (Grant 11DZ2260600).

■ REFERENCES

(1) Kramer, J. A.; Sagartz, J. E.; Morris, D. L. The application of discovery toxicology and pathology towards the design of safer pharmaceutical lead candidates. *Nat. Rev. Drug Discovery* **2007**, *6*, 636−649.

(2) Ames, B. N.; McCann, J.; Yamasaki, E. Methods for detecting carcinogens and mutagens with the Salmonella/mammalian-microsome mutagenicity test. *Mutat. Res.* **1975**, *31*, 347−364.

(3) Mortelmans, K.; Zeiger, E. The Ames Salmonella/microsome mutagenicity assay. *Mutat. Res.* **2000**, *445*, 29−60.

(4) Iurii Sushko, S. N.; Igor, V. T. Applicability domain for *in silico* models to achieve accuracy of experimental measurements. *J. Chemom.* **2010**, *24*, 202−208.

(5) Benigni, R.; Giuliani, A. Computer-assisted analysis of interlaboratory Ames test variability. *J. Toxicol. Environ. Health* **1988**, *25*, 135−148.

(6) Hansen, K.; Mika, S.; Schroeter, T.; Sutter, A.; ter Laak, A.; Steger-Hartmann, T.; Heinrich, N.; Muller, K. R. Benchmark data set for *in silico* prediction of Ames mutagenicity. *J. Chem. Inf. Model.* **2009**, *49*, 2077−2081.

(7) Cheng, F.; Yu, Y.; Shen, J.; Yang, L.; Li, W.; Liu, G.; Lee, P. W.; Tang, Y. Classification of Cytochrome P450 Inhibitors and non-Inhibitors using Combined Classifiers. *J. Chem. Inf. Model.* **2011**, *51*, 996−1011.

(8) Cheng, F.; Yu, Y.; Zhou, Y.; Shen, Z.; Xiao, W.; Liu, G.; Li, W.; Lee, P. W.; Tang, Y. Insights into molecular basis of cytochrome p450 inhibitory promiscuity of compounds. *J. Chem. Inf. Model.* **2011**, *51*, 2482−2495.

(9) Cheng, F.; Shen, J.; Yu, Y.; Li, W.; Liu, G.; Lee, P. W.; Tang, Y. *In silico* prediction of Tetrahymena pyriformis toxicity for diverse industrial chemicals with substructure pattern recognition and machine learning methods. *Chemosphere* **2011**, *82*, 1636−1643.

(10) Cheng, F.; Ikenaga, Y.; Zhou, Y.; Yu, Y.; Li, W.; Shen, J.; Du, Z.; Chen, L.; Xu, C.; Liu, G.; Lee, P. W.; Tang, Y. In silico assessment of chemical biodegradability. *J. Chem. Inf. Model.* **2012**, *52*, 655−669.

(11) Cheng, F.; Zhou, Y. D.; Li, J.; Li, W. H.; Liu, G.; Tang, Y. Prediction of Chemical-Protein Interactions: Multitarget-QSAR versus Computational Chemogenomic Methods. *Mol. BioSyst.* **2012**, *8*, 2373−2384.

(12) Guha, R. Flexible Web service infrastructure for the development and deployment of predictive models. *J. Chem. Inf. Model.* **2008**, *48*, 456−464.

(13) Votano, J. R.; Parham, M.; Hall, L. H.; Kier, L. B.; Oloff, S.; Tropsha, A.; Xie, Q.; Tong, W. Three new consensus QSAR models for the prediction of Ames genotoxicity. *Mutagenesis* **2004**, *19*, 365−377.

(14) Kazius, J.; McGuire, R.; Bursi, R. Derivation and validation of toxicophores for mutagenicity prediction. *J. Med. Chem.* **2005**, *48*, 312−320.

(15) Zheng, M.; Liu, Z.; Xue, C.; Zhu, W.; Chen, K.; Luo, X.; Jiang, H. Mutagenic probability estimation of chemical compounds by a novel molecular electrophilicity vector and support vector machine. *Bioinformatics* **2006**, *22*, 2099−2106.

(16) Ferrari, T.; Gini, G. An open source multistep model to predict mutagenicity from statistical analysis and relevant structural alerts. *Chem. Cent. J.* **2010**, *4*, S2.

(17) Hillebrecht, A.; Muster, W.; Brigo, A.; Kansy, M.; Weiser, T.; Singer, T. Comparative evaluation of *in silico* systems for ames test mutagenicity prediction: scope and limitations. *Chem. Res. Toxicol.* **2011**, *24*, 843−854.

(18) Willett, P. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

(19) Ewing, T.; Baber, J. C.; Feher, M. Novel 2D fingerprints for ligand-based virtual screening. *J. Chem. Inf. Model.* **2006**, *46*, 2423−2431.

(20) Wang, S.; Li, Y.; Wang, J.; Chen, L.; Zhang, L.; Yu, H.; Hou, T. ADMET evaluation in drug discovery. 12. Development of binary classification models for prediction of hERG potassium channel blockage. *Mol. Pharmaceutics* **2012**, *9*, 996−1010.

(21) Chen, L.; Li, Y.; Zhao, Q.; Peng, H.; Hou, T. ADME evaluation in drug discovery. 10. Predictions of P-glycoprotein inhibitors using recursive partitioning and naive Bayesian classification techniques. *Mol. Pharmaceutics* **2011**, *8*, 889−900.

(22) Lazar Toxcity Predictions. http://lazar.in-silico.de/models (accessed December 30, 2011).

(23) Yap, C. W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **2010**, *32*, 1466−1474.

(24) ChemAxon. http://www.chemaxon.com (accessed January 5, 2012).

(25) Fourches, D.; Muratov, E.; Tropsha, A. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J. Chem. Inf. Model.* **2010**, *50*, 1189−1204.

(26) Orange, version 2.0.b. http://www.ailab.si/orange/ (accessed November 20, 2011).

(27) Chang, C.; Lin, C.-J. LIBSVM, version 2.9. http://www.csie.ntu.edu.tw/~cjlin/libsvm (accessed November 28, 2011).

(28) KNIME, version 2.4.2. http://www.knime.org/ (accessed November 28, 2011).

(29) Basheer, I. A.; Hajmeer, M. Artificial neural networks: fundamentals, computing, design, and application. *J. Microbiol. Methods* **2000**, *43*, 3−31.

(30) Martin Riedmiller, H. B. A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm. *Proceedings of the IEEE International Conference on Neural Networks (ICNN)*; 1993, Vol. *16*, pp 586−591.

(31) Itskowitz, P.; Tropsha, A. kappa Nearest neighbors QSAR modeling as a variational problem: theory and applications. *J. Chem. Inf. Model.* **2005**, *45*, 777−785.

(32) Corinna Cortes, V. V. Support-Vector Networks. *Machine Learning* **1995**, *20*, 273−297.

(33) Sun, H. A naive bayes classifier for prediction of multidrug resistance reversal activity on the basis of atom typing. *J. Med. Chem.* **2005**, *48*, 4031−4039.

(34) Plewczynski, D.; Spieser, S. A.; Koch, U. Assessing different classification methods for virtual screening. *J. Chem. Inf. Model.* **2006**, *46*, 1098−1106.

(35) Quinlan, J. R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann Publishers; 1993.

(36) Perez-Garrido, A.; Helguera, A. M.; Borges, F.; Cordeiro, M. N.; Rivero, V.; Escudero, A. G. Two new parameters based on distances in a receiver operating characteristic chart for the selection of classification models. *J. Chem. Inf. Model.* **2011**, *51*, 2746−2759.

(37) Shen, J.; Cheng, F.; Xu, Y.; Li, W.; Tang, Y. Estimation of ADME properties with substructure pattern recognition. *J. Chem. Inf. Model.* **2010**, *50*, 1034−1041.

(38) Jensen, B. F.; Vind, C.; Padkjaer, S. B.; Brockhoff, P. B.; Refsgaard, H. H. In silico prediction of cytochrome P450 2D6 and 3A4 inhibition using Gaussian kernel weighted k-nearest neighbor and extended connectivity fingerprints, including structural fragment analysis of inhibitors versus noninhibitors. *J. Med. Chem.* **2007**, *50*, 501−511.

(39) Evans, B. E.; Rittle, K. E.; Bock, M. G.; DiPardo, R. M.; Freidinger, R. M.; Whitter, W. L.; Lundell, G. F.; Veber, D. F.; Anderson, P. S.; Chang, R. S.; et al. Methods for drug discovery: development of potent, selective, orally effective cholecystokinin antagonists. *J. Med. Chem.* **1988**, *31*, 2235−2246.

(40) Kruhlak, N. L.; Contrera, J. F.; Benz, R. D.; Matthews, E. J. Progress in QSAR toxicity screening of pharmaceutical impurities and other FDA regulated products. *Adv. Drug Delivery Rev.* **2007**, *59*, 43−55.

(41) Benigni, R.; Bossa, C. Structure alerts for carcinogenicity, and the Salmonella assay system: a novel insight through the chemical relational databases technology. *Mutat. Res.* **2008**, *659*, 248−261.

(42) CPDB. http://potency.berkeley.edu/cpdb.html (accessed November 14, 2011).

(43) CCRIS. http://toxnet.nlm.nih.gov./cgi-bin/sis/htmlgen?CCRIS (accessed November 18, 2011).

(44) Lounkine, E.; Hu, Y.; Batista, J.; Bajorath, J. Relevance of feature combinations for similarity searching using general or activity class-directed molecular fingerprints. *J. Chem. Inf. Model.* **2009**, *49*, 561−570.

(45) Chen, B.; Sheridan, R. P.; Hornak, V.; Voigt, J. H. Comparison of Random Forest and Pipeline Pilot Naive Bayes in Prospective QSAR Predictions. *J. Chem. Inf. Model.* **2012**, *52*, 792−803.

(46) Benigni, R. Structure-activity relationship studies of chemical mutagens and carcinogens: mechanistic investigations and prediction approaches. *Chem. Rev.* **2005**, *105*, 1767−1800.

(47) Jeliazkova, N. Toxtree, version 2.5.0; http://ihcp.jrc.ec.europa.eu/our_labs/computational_toxicology/qsar_tools/toxtree (accessed December 15, 2011).

(48) Pavan, M.; Worth, A. P. Publicly-accessible QSAR software tools developed by the Joint Research Centre. *SAR QSAR Environ. Res.* **2008**, *19*, 785−799.