Prediction of Genotoxicity of Chemical Compounds by Statistical Learning Methods

H. Li,[†] C. Y. Ung,[†] C. W. Yap,[†] Y. Xue,[†,‡] Z. R. Li,[†,‡] Z. W. Cao,[§] and Y. Z. Chen*[,†]

Bioinformatics and Drug Design Group, Department of Computational Science, National University of Singapore, Blk SOC1, Level 7, 3 Science Drive 2, Singapore 117543, College of Chemistry, Sichuan University, Chengdu, 610064, People's Republic of China, and Shanghai Center for Bioinformation Technology, 100 Qinzhou Road, Shanghai, 200235, People's Republic of China

Received December 15, 2004

Various toxicological profiles, such as genotoxic potential, need to be studied in drug discovery processes and submitted to the drug regulatory authorities for drug safety evaluation. As part of the effort for developing low cost and efficient adverse drug reaction testing tools, several statistical learning methods have been used for developing genotoxicity prediction systems with an accuracy of up to 73.8% for genotoxic (GT+) and 92.8% for nongenotoxic (GT−) agents. These systems have been developed and tested by using less than 400 known GT+ and GT− agents, which is significantly less in number and diversity than the 860 GT+ and GT− agents known at present. There is a need to examine if a similar level of accuracy can be achieved for the more diverse set of molecules and to evaluate other statistical learning methods not yet applied to genotoxicity prediction. This work is intended for testing several statistical learning methods by using 860 GT+ and GT− agents, which include support vector machines (SVM), probabilistic neural network (PNN), *k*-nearest neighbor (*k*-NN), and C4.5 decision tree (DT). A feature selection method, recursive feature elimination, is used for selecting molecular descriptors relevant to genotoxicity study. The overall accuracies of SVM, *k*-NN, and PNN are comparable to and those of DT lower than the results from earlier studies, with SVM giving the highest accuracies of 77.8% for GT+ and 92.7% for GT− agents. Our study suggests that statistical learning methods, particularly SVM, *k*-NN, and PNN, are useful for facilitating the prediction of genotoxic potential of a diverse set of molecules.

Introduction

Adverse drug reactions (ADRs) are responsible for the failure of a substantial percentage of investigational drugs and the withdrawal of marketed drugs (*1*, *2*). Up to one-third of all drug failures are due to ADRs (*3*). A variety of toxicological tests and clinical safety evaluations need to be conducted and evaluated by the drug regulatory authorities for drug safety assessment. Because of the high cost of conducting toxicity tests and clinical trials, effort has been directed at developing low cost and efficient tools for predicting ADRs aimed at eliminating unsafe drug candidates in the early stages of drug development (*2*, *3*).

Genotoxicity is one of the ADRs closely evaluated in drug discovery and approval processes. The molecular mechanisms of genotoxicity include DNA intercalation by aromatic ring of a drug, DNA methylation, DNA adduct formation and strand break, and unscheduled DNA synthesis (*4*). Some genotoxic (GT+) compounds require metabolic activation, and their GT+ effects are mediated via N-dialkylation (*5*). These events subsequently result in chromosomal aberrations, micronuclei, sister chromatid exchanges, and cell death, which contribute to drug ADRs (*4*).

Tools for fast and efficient prediction of drug GT+ potential, particularly those based on computational methods, are being developed (*6*, *7*). For instance, expert systems that use structural alerts for predicting GT+ as well as other toxicological profiles are now commercially available. These include Deductive Estimation on Risk from Existing Knowledge (DEREK), Multiple Computer Automated Structure Evaluation (MCASE), and Toxicity Prediction by Komputer Assisted Technology (TOPKAT). Specific details about these computational databases can be found in the review by Greene (*8*). Quantitative structure−activity relationships (QSAR) have been developed for predicting the GT+ potential of several groups of related chemicals (*9*, *10*). However, the QSARs of a majority of chemical groups are yet to be determined, which hinders the practical application of this method.

Statistical learning methods have recently been explored as a new approach for genotoxicity prediction without the restriction on the features of structures or types of molecules (*11*−*13*). Instead of focusing on specific structural features or a particular group of related molecules, these methods classify molecules into GT+ and nongenotoxic (GT−) agents based on their general structural and physicochemical properties regardless of their structural and chemical types. Therefore, in principle, these methods are expected to be applicable to a diverse set of molecules. However, the performance of these methods can be practically limited by the quality of molecular descriptors, diversity of training and testing data, and the efficiency of statistical learning algorithm.

* To whom correspondence should be addressed. Tel: 65-6874-6877. Fax: 65-6774-6756. E-mail: yzchen@cz3.nus.edu.sg.
[†] National University of Singapore.
[‡] Sichuan University.
[§] Shanghai Center for Bioinformation Technology.

10.1021/tx049652h CCC: $30.25
Published on Web 06/04/2005

So far, three statistical learning methods, linear discriminate analysis (LDA), *k*-nearest neighbor classification (*k*-NN), and probabilistic neural networks (PNNs), have been used and achieved a prediction accuracy of up to 73.8% for GT+ and 92.8% for GT− agents, respectively (*11*−*13*). However, these methods have been developed and tested by using no more than 394 GT+ and GT− agents (*5*), which is significantly smaller in number and diversity than the 860 known GT+ and GT− agents found from our recent literature search. Therefore, there is a need to examine if a similar level of accuracy can be achieved for the more diverse set of molecules. It is also of interest to determine if the GT+ accuracy can be further improved by a training set composed of a more diverse set of GT+ agents. Moreover, other statistical learning methods such as support vector machines (SVM) (*14*, *15*) and decision tree (DT) (*16*) have shown promising potential, and it is useful to evaluate these methods.

This work is intended to evaluate several statistical learning methods by using 860 GT+ and GT− agents. These methods include SVM (*14*, *15*), PNN (*17*), *k*-NN (*18*), and DT (*16*). In particular, SVM is studied because of its good performance in a number of classification problems (*19*−*22*). SVM has been applied to the prediction of chromosome aberrations (*23*), torsade-causing potential of drugs (*24*), blood−brain barrier-penetrating agents (*19*, *20*), P-glycoprotein substrates (*25*), structure−activity relationship (SAR) of enzyme inhibition (*26*), and QSAR of antihistamines and antibacterials (*27*). Most of these studies have consistently demonstrated that SVM to various degrees gives better prediction accuracy than other supervised statistical learning methods (*19*−*22*).

A widely used feature selection method, recursive feature elimination (RFE), is used in this work for selecting the molecular descriptors relevant to the classification of GT+ and GT− agents. This method has recently gained popularity due to its effectiveness for discovering informative features or attributes in drug activity analysis (*28*, *29*), toxicological, pharmacokinetic, and pharmacodynamic properties (*25*, *30*). To adequately assess the prediction accuracy of the methods used in this work, two different evaluation methods are used. One is 5-fold cross-validation, which is a popular method for evaluating drug prediction systems (*21*, *31*), and the other is the use of an external independent validation set, which has been found to be equally useful for assessing drug prediction systems (*24*, *25*).

Materials and Methods

Selection of GT+ and GT− Agents. A total of 860 GT+ and GT− agents with known genotoxicity test results are selected from several sources including the 1999−2002 Physician's Desk Reference, National Toxicology Program, and a number of publications (*5*, *11*−*13*, *32*). Genotoxicity tests for generating these data include the pre-ICH four standard batteries (Ames test, in vitro cytogenetics, in vivo cytogenetics, and mouse lymphoma assay) and the salt-overly-sensitive (SOS) chromotest (which is a rapid alternative genotoxicity test based on the detection of the DNA damage through the SOS pathway) (*33*, *34*). Agents with genotoxicity test results are divided into GT+ and GT− groups according to whether these genotoxicity test results showed at least one positive finding. Under this definition, there are a total of 229 GT+ agents and 631 GT− agents, which are given in the Supporting Information. The three-dimensional (3D) structures of these compounds are generated by using the Concord software (*35*) and optimized by using the semiempirical AM1 method (*36*).

These compounds are further separated into training and testing sets by two different ways depending on the evaluation method used. For 5-fold cross-validation, these compounds are randomly divided into five subsets of approximately equal size. Four subsets are selected as the training set and the fifth as the testing set. This process is repeated five times such that every subset is selected as a testing set once. For evaluation by an independent validation set, these compounds are divided into training, testing, and independent validation sets based on their distribution in the chemical space. Chemical space is defined by the commonly used structural and chemical descriptors (*37*). Compounds of similar structural and chemical features are evenly assigned into separate sets. For those compounds without enough numbers of structurally and chemically similar counterparts, they are assigned, in order of priority, to the training and then the testing set, respectively. The training set is used for developing the prediction system, the testing set is used for optimizing the parameters of the system, and the independent validation set is used for assessing the accuracy of the system. The generated training, testing, and independent evaluation sets contain 577 (166 GT+, 411 GT−), 160 (36 GT+, 124 GT−), and 123 (27 GT+, 96 GT−) compounds, respectively.

Molecular Descriptors. Molecular descriptors have routinely been used in quantitative description of structural and physicochemical properties of molecules in a statistical study of drugs and small molecules (*37*−*41*). In this work, a set of 199 molecular descriptors were selected from the more than 1000 descriptors described in the literature after eliminating those descriptors that are obviously redundant or unrelated to the problem studied here. These descriptors, described in our earlier publications (*25*, *30*), include 143 topological, 31 quantum chemical, and 25 geometrical descriptors. They are computed from the 3D structure of each compound using our own designed molecular descriptor computing program. The remaining redundant and unrelated descriptors are further reduced by using the feature selection method (*28*, *29*, *42*).

Feature Selection Method. Feature selection methods have been introduced for the improvement of classification performance of statistical learning methods and for the selection of features relevant to the discrimination of two data sets (*21*, *28*, *29*, *42*−*44*). The RFE method has become a popular choice for feature selection in a variety of problems (*28*, *29*). Our study on torsade-causing potential of drugs (*30*), P-glycoprotein substrates (*25*), and human intestinal absorption (*30*) also showed that this method is useful for selecting features relevant to the study of toxicological and pharmacokinetic properties. Thus, the RFE method is used for feature selection in this work, and the details of the implementation of this method can be found in our earlier publications (*25*, *30*).

The feature selection procedure can be demonstrated by the following illustrative example of the development of a SVM classification system: This system is trained by using a Gaussian kernel function with an adjustable parameter σ. Sequential variation of σ is conducted against the whole training set to find a value that gives the best prediction accuracy. This prediction accuracy is evaluated by means of 5-fold cross-validation. In the first step, for a fixed σ, the SVM classifier is trained by using the complete set of features (molecular descriptors) described in the previous section. The second step involves the computation of the ranking criterion score $DJ(i)$ for each feature in the current set. All of the computed $DJ(i)$ scores are subsequently ranked in descending order. The third step involves the removal of the m features with smallest criterion scores. In the fourth step, the SVM classification system is retrained by using the remaining set of features, and the corresponding prediction accuracy is computed by means of 5-fold cross-validation. The first to fourth steps are then repeated for other values of σ. After the completion of these procedures, the set of features and parameter σ that gives the best prediction accuracy is selected.

The choice of the parameter m affects the performance of SVM as well as the speed of feature selection. Although it is desirable to remove one feature at a time ($m = 1$), this is often

difficult due to high CPU cost. It has been found that, in some cases, removal of several features at a time ($m > 1$) significantly improves computational efficiency without losing too much accuracy (*29*). Our studies on a subset of randomly selected GT+ and GT− agents in this work and compounds of different pharmacokinetic properties (*25*, *30*) suggested that the accuracy of a SVM system with $m = 5$ is only a few percentages smaller than that with $m = 1$, which is consistent with the findings from other studies (*28*, *44*). Thus, for computational efficiency, m =5 is used in this study.

Statistical Learning Methods. 1. SVM. The theory of SVM has been extensively described in the literature (*15*, *45*). Thus, only a brief description is given here. SVM is based on the structural risk minimization principle from statistical learning theory (*45*). In linearly separable cases, SVM constructs a hyperplane to separate two classes of molecules with a maximum margin. A molecule is represented by a vector $\boldsymbol{x}_i$, with the structural and physicochemical descriptors of this molecule as its components. Separation of the two classes of molecules is conducted by finding another vector $\mathbf{w}$ and a parameter b that minimizes $||\mathbf{w}||^2$ and satisfies the following conditions:

$$\mathbf{w} \cdot \boldsymbol{x}_i + b \geq +1, \text{ for } y_i = +1 \text{ class 1 (positive)} \quad (1)$$

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

where y_i is the class index, $\mathbf{w}$ is a vector normal to the hyperplane, $|b|/||\mathbf{w}||$ is the perpendicular distance from the hyperplane to the origin, and $||\mathbf{w}||^2$ is the Euclidean norm of $\mathbf{w}$. After the determination of $\mathbf{w}$ and b, a given vector $\boldsymbol{x}_i$ can be classified by:

$$\text{sign}[(\mathbf{w} \cdot \boldsymbol{x}) + b] \quad (3)$$

In nonlinearly separable cases, SVM maps the feature vectors into a higher dimensional feature space using a kernel function $K(\boldsymbol{x}_i,\boldsymbol{x}_j)$. An example of a kernel function is the Gaussian kernel, which has been extensively used in different studies with good results (*20*, *26*, *27*).

$$K(\boldsymbol{x}_i,\boldsymbol{x}_j) = e^{-||\boldsymbol{x}_j-\boldsymbol{x}_i||^2/2\sigma^2} \quad (4)$$

The linear SVM algorithm is then applied to the vectors in this hyperspace, and the following decision function is used to classify a vector:

$$f(\boldsymbol{x}) = \text{sign}[\sum_{i=1}^{l} \alpha_i^0 \, y_i \, K(\boldsymbol{x},\boldsymbol{x}_i) + b] \quad (5)$$

where the coefficients α_i^0 and b are determined by maximizing the following Langrangian expression:

$$\sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \, \alpha_j \, y_i \, y_j \, K(\boldsymbol{x}_i,\boldsymbol{x}_j) \quad (6)$$

under the following conditions:

$$\alpha_i \geq 0 \text{ and } \sum_{i=1}^{l} \alpha_i y_i = 0 \quad (7)$$

A positive or negative value from eq 5 indicates that the vector $\boldsymbol{x}$ belongs to the positive or negative class, respectively.

2. *k*-NN. k-NN measures the Euclidean distance between a to-be-classified vector $\boldsymbol{x}$ and each individual vector $\boldsymbol{x}_i$ in the training set (*18*, *46*). The Euclidean distances for the vector pairs are calculated using the following formula:

$$D = \sqrt{||\boldsymbol{x} - \boldsymbol{x}_i||^2} \quad (8)$$

A total of k number of vectors nearest to the vector $\boldsymbol{x}$ are used to determine its class, $f(\boldsymbol{x})$:

$$\hat{f}(\boldsymbol{x}) \leftarrow \text{argmax}_{v \in V} \sum_{i=1}^{k} \delta[v, f(\boldsymbol{x}_i)] \quad (9)$$

where $\delta(a,b) = 1$ if $a = b$ and $\delta(a,b) = 0$ if $a \neq b$, argmax is the maximum of the function, V is a finite set of vectors $\{v_1,...v_s\}$, and $\hat{f}(\boldsymbol{x})$ is an estimate of $f(\boldsymbol{x})$. Here, estimate refers to the class of the majority of the k-NNs.

3. PNN. PNN is a form of neural network designed for classification through the use of Bayes' optimal decision rule (*17*)

$$h_i \, c_i \, f_i(\boldsymbol{x}) > h_j \, c_j \, f_j(\boldsymbol{x}) \quad (10)$$

where h_i and h_j are the prior probabilities, c_i and c_j are the costs of misclassification, and $f_i(x)$ and $f_j(x)$ are the probability density function for classes i and j, respectively. An unknown vector $\boldsymbol{x}$ is classified into population i if the product of all of the three terms is greater for class i than for any other class j (not equal to i). In most applications, the prior probabilities and costs of misclassifications are treated as being equal. The probability density function for each class for a univariate case can be estimated by using the Parzen's nonparametric estimator (*47*)

$$g(\boldsymbol{x}) = \frac{1}{n\sigma} \sum_{i=1}^{n} W\left(\frac{\boldsymbol{x} - \boldsymbol{x}_i}{\sigma}\right) \quad (11)$$

where n is the sample size, σ is a scaling parameter, which defines the width of the bell curve that surrounds each sample point, $W(d)$ is a weight function, which has its largest value at $d = 0$, and $(\boldsymbol{x} - \boldsymbol{x}_i)$ is the distance between the unknown vector and a vector in the training set. The Parzen's nonparametric estimator was later expanded by Cacoullos (*48*) for the multivariate case.

$$g(x_1,..., x_p) = \frac{1}{n\sigma_1...\sigma_p} \sum_{i=1}^{n} W\left(\frac{x_1 - x_{1,i}}{\sigma_1},..., \frac{x_p - x_{p,i}}{\sigma_p}\right) \quad (12)$$

The Gaussian function is frequently used as the weight function because it is well-behaved, easily calculated, and satisfies the conditions required by Parzen's estimator. Thus, the probability density function for the multivariate case becomes

$$g(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} \exp\left[-\sum_{j=1}^{p}\left(\frac{x_j - x_{ij}}{\sigma_j}\right)^2\right] \quad (13)$$

The network architectures of PNN are determined by the number of compounds and descriptors in the training set. There are four layers in a PNN. The input layer provides input values to all neurons in the pattern layer and has as many neurons as the number of descriptors in the training set. The number of pattern neurons is determined by the total number of compounds in the training set. Each pattern neuron computes a distance measure between the input and the training case represented by that neuron and then subjects the distance measure to the Parzen's nonparameteric estimator. The summation layer has a neuron for each class, and the neurons sum all of the pattern neurons' output corresponding to members of that summation neuron's class to obtain the estimated probability density function for that class. The single neuron in the output layer then estimates the class of the unknown vector $\boldsymbol{x}$ by comparing all of the probability density function from the summation neurons and choosing the class with the highest probability density function.

4. C4.5 DT. C4.5 DT is a branch test-based classifier (*16*). A branch in a DT corresponds to a group of classes, and a leaf represents a specific class. A decision node specifies a test to be conducted on a single attribute value, with one branch and its subsequent classes as possible outcomes of the test. C4.5 DT uses recursive partitioning to examine every attribute of the data and rank them according to their ability to partition the

remaining data, thereby constructing a DT. A vector $\boldsymbol{x}$ is classified by starting at the root of the tree and moving through the tree until a leaf is encountered. At each nonleaf decision node, a test is conducted and the classification process proceeds to the branch selected by the test. Upon reaching the destination leaf, the class of the vector $\boldsymbol{x}$ is predicted to be that associated with the leaf.

The algorithm is a recursive greedy heuristic that selects descriptors for membership within the tree. Whether or not a descriptor is included within the tree is based on the value of its information gain. As a statistical property, information gain measures how well the descriptor separate training cases into subsets in which the class is homogeneous. Given that the descriptors in this study were all continuous variables, a threshold value had to be established within each descriptor so that it could partition the training cases into subsets. These threshold values for each descriptor were established by rank ordering the values within each descriptor from lowest to highest and repeatedly calculating the information gain using the arithmetical midpoint between all successive values within the rank order. The midpoint value with the highest information gain was selected as the threshold value for the descriptor. That descriptor with the highest information gain (information being the most useful for classification) was then selected for inclusion in the DT. The algorithm continued to build the tree in this manner until it accounted for all training cases. Ties between descriptors that were equal in terms of information gain were broken randomly (*49*).

Performance Evaluation. As in the case of all discriminative methods (*50*, *51*), the performance of statistical learning methods can be measured by the quantity of true positives (TP), true negatives (TN), false positives (FP), false negatives (FN), sensitivity, SE = TP/(TP + FN) × 100, which is the prediction accuracy for the GT+ compounds in this work, and specificity, SP = TN/(TN + FP) × 100, which is the prediction accuracy for the GT− compounds in this work. The overall prediction accuracy (Q) and Matthews correlation coefficient (C) (*52*) are also used to measure the prediction accuracies and can be given by:

$$Q = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{14}$$

$$C = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FN} \times \mathrm{FP}}{\sqrt{(\mathrm{TP} + \mathrm{FN})(\mathrm{TP} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})}} \tag{15}$$

Results and Discussion

Overall Prediction Accuracies. SVM, PNN, and k-NN are conducted by using our own software, and C4.5 DT was performed by using the code from Quinlan (*16*). Prediction results of SVM without RFE and SVM with RFE (SVM + RFE) by using 5-fold cross-validation are given in Table 1. The accuracies of SVM + RFE are 75.5% for GT+ agents and 90.6% for GT− agents, which are slightly better than the values of 69.4% for GT+ agents and 88.2% for GT− agents derived from SVM without RFE. The GT+ prediction accuracy is noticeably improved, which indicates the usefulness of RFE in selecting the proper set of features for the prediction of GT+ and GT− agents. The use of these RFE-selected descriptors also slightly improves the prediction accuracy of the other three statistical methods. The GT+ accuracies are improved from 70.4 to 74.1% for PNN and from 44.4 to 55.6% for DT, respectively, and that of k-NN remains roughly unchanged. The GT− accuracy of k-NN is improved from 82.2 to 86.5%, and those of PNN and DT are roughly unchanged. These results showed that feature selection by using RFE plays the important role in improving the prediction capability for the above methods in general. Similar prediction accuracies are also found from two additional 5-fold cross-validation studies conducted by using training−testing sets separately generated from different random number seed parameters.

Table 1. SVM and SVM + RFE Prediction Accuracy of the GT+ and GT− Agents by Using 5-Fold Cross-Validation[a]

		genotoxicity			nongenotoxicity				
method	cross-validation	TP	FN	SE (%)	TN	FP	SP (%)	Q (%)	C
SVM	1	32	17	65.3	109	11	90.8	83.4	0.59
	2	30	10	75.0	115	14	89.1	85.8	0.62
	3	32	13	71.1	119	21	85.0	81.6	0.53
	4	32	19	62.7	106	11	90.6	82.1	0.56
	5	32	12	72.7	107	18	85.6	82.2	0.56
	average			**69.4**			**88.2**	**83.0**	**0.57**
	SD			4.6			2.5	1.5	0.03
	SE			1.9			1.0	0.6	0.01
SVM + RFE	1	35	14	71.4	111	9	92.5	86.4	0.66
	2	32	8	80.0	118	11	91.5	88.8	0.69
	3	35	10	77.8	123	17	87.9	85.4	0.62
	4	35	16	68.6	109	8	93.2	85.7	0.65
	5	35	9	79.5	110	15	88.0	85.8	0.65
	average			**75.5**			**90.6**	**86.4**	**0.66**
	SD			4.6			2.3	1.2	0.02
	SE			1.9			0.9	0.5	0.01

[a] Predicted results are given in TP, FN, TN, FP, SE, which is the prediction accuracy for GT+, SP, which is the prediction accuracy for GT−, Q, and C. Statistical significance is indicated by SD (standard deviation) and SE (standard error), respectively.

Table 2. Comparison of the Prediction Accuracies of GT+ and GT− Agents Derived from Different Machine Learning Methods by Using the Independent Validation Set in This Work[a]

						GT+ accuracy	GT− accuracy	
method	parameter	TP	FN	TN	FP	SE (%)	SP (%)	Q (%)
C4.5 DT		15	12	72	24	55.6	75.0	70.7
PNN	$\sigma = 0.2$	20	7	77	19	74.1	80.2	78.9
k-NN	$k = 3$	19	8	83	13	70.4	86.5	82.9
SVM	$\sigma = 3$	21	6	89	7	77.8	92.7	89.4

[a] C4.5 DT, PNN, k-NN, SVM, and parameter σ is the width of Gaussian kernel function. Predicted results are given in TP, FN, TN, FP, SE, which is the prediction accuracy for GT+, SP, which is the prediction accuracy for GT−, and Q.

Table 2 gives the GT+ and GT− prediction accuracies derived from the four methods SVM, PNN, k-NN, and DT by using the independent validation set and the RFE-selected molecular descriptors. The GT+ accuracies are in the range of 55.6−77.8% and the GT− accuracies are in the range of 75.0−92.7%. Similar level of accuracies are obtained for SVM, PNN, and k-NN, with SVM giving the highest value of 77.8 and 92.7% for GT+ and GT− agents, respectively. DT appears to give substantially lower accuracies, which is concordant with other experimental comparison results (*53*, *54*). A possible reason for this lower accuracy is that DT uses information gain to find the optimum set of descriptors, which may not be the most effective approach for every problem. It has been pointed out that filter methods, such as information gain, may not be as efficient as wrapper methods, such as RFE, for determining the subset of descriptors relevant to a particular problem (*55*, *56*).

Relevance of Selected Features to Genotoxicty Study. Apart from the quality of data sets used, selection of descriptors relevant to genotoxicity study is important for optimizing the prediction system by reducing the noise in a statistical learning process. A total of 39 molecular

Table 3. Molecular Descriptors Selected from the RFE Method for SVM Classification of GT+ and GT− Agents

descriptors	description	class
Nrot	no. of rotatable bonds	simple molecular properties
ndonr	no. of H-bond donors	simple molecular properties
$^3\chi_C$	simple molecular connectivity χ indices for cluster	connectivity and shape
$^4\chi_{PC}$	simple molecular connectivity χ indices for path/cluster	connectivity and shape
$^3\chi^v_C$	valence molecular connectivity χ indices for cluster	connectivity and shape
$^4\chi^v_{PC}$	valence molecular connectivity χ indices for path/cluster	connectivity and shape
S(2)	atom type H estate sum for =NH	electrotopological state
S(4)	atom type H estate sum for $-NH_2$	electrotopological state
S(10)	atom type H estate sum for :CH: (sp^2, aromatic)	electrotopological state
S(13)	atom type H estate sum for CH_n (unsaturated)	electrotopological state
S(14)	atom type H estate sum for CH_n (aromatic)	electrotopological state
S(16)	atom type estate sum for $-CH_3$	electrotopological state
S(25)	atom type estate sum for =C<	electrotopological state
S(26)	atom type estate sum for :C:−	electrotopological state
S(27)	atom type estate sum for :C::	electrotopological state
S(30)	atom type estate sum for =NH	electrotopological state
S(34)	atom type estate sum for =N−	electrotopological state
S(35)	atom type estate sum for :N:	electrotopological state
S(41)	atom type estate sum for −O−	electrotopological state
Tradi	PetitJohn R2 index	electrotopological state
Tpeti	PetitJohn I2 index	electrotopological state
M	molecular dipole moment	quantum chemical properties
μ_{cp}	chemical potential	quantum chemical properties
χ_{en}	electronegativity index	quantum chemical properties
ω	electrophilicity index	quantum chemical properties
$Q_{H,max}$, $Q_{N,max}$, $Q_{O,max}$	most positive charge on H, N, O atoms	quantum chemical properties
$Q_{H,min}$	most negative charge on H atoms	quantum chemical properties
Rpc	relative positive charge	quantum chemical properties
Rnc	relative negative charge	quantum chemical properties
Rugty	molecular rugosity	geometrical properties
Gloty	molecular globularity	geometrical properties
Shpl	hydrophilic region	geometrical properties
Shpb	hydrophobic region	geometrical properties
Capty	capacity factor	geometrical properties
Hiwpl	hydrophilic integy moment	geometrical properties
Hiwpb	hydrophobic integy moment	geometrical properties
Hiwpa	amphiphilic moment	geometrical properties

descriptors are selected by the RFE method, as given in Table 3. Most of these are found to be relevant to the assessment of genotoxicity potential of molecules. For instance, an important characteristics of some GT+ agents is their ability to intercalate DNA (*12*). The selected electrotopological state descriptors S(10) and S(14) describe atom type H estate sum for :CH: sp^2 aromatic structures and atom type H estate sum for CH_n aromatic structures, respectively.

Many GT+ agents are known to structurally modify or form a covalent bond to DNA via chemical reactions. A substantial portion of the RFE selected descriptors are from the class of electrotopological state that describe characteristics of specific types of functional groups involved in DNA modification. There are also a substantial number of descriptors from the quantum chemical class that determine molecular dipole moment, chemical potential, electronegativity, electrophilicity, relative positive and negative charge, and the atomic charge on H, N, and O atoms in a molecule. These properties are important for describing features of chemical reactions involved in the modification of DNA.

The size, shape, and polar property of a molecule have also been found to play a role in genetic damages caused by GT+ agents (*12*). Eight of the selected descriptors are VolSurf descriptors (*39*). These are molecular rugosity, molecular globularity, capacity factor, hydrophilic and hydrophobic region, hydrophilic integy moment, hydrophobic moment, and amphiphilic moment. These descriptors primarily describe the size, shape, and polar property of a molecule. In general, VolSurf descriptors, which are one-dimensional descriptors extracted from the computed 3D molecular field maps, were developed specifically for pharmacokinetics and pharmacodynamics applications (*39*, *57*). It is thus not surprising that the VolSurf descriptors related to the molecular size, shape, and polar property are selected.

Molecular connectivity is another feature known to be important for discriminating between some GT+ compounds from their GT− analogues. For instance, 4-amino-3-nitro-2,5-dimethylaniline is a GT+ agent, while its analogue 4-amino-3-nitro-2,6-dimethylaniline is GT− (*58*). Four molecular connectivity descriptors, $^3\chi_C$, $^4\chi_{PC}$, $^3\chi_C{}^v$, and $^4\chi_{PC}{}^v$, are selected by RFE in this work. These descriptors are simple molecular connectivity χ indices for cluster, simple molecular connectivity χ indices for path/cluster, valence molecular connectivity χ indices for cluster, and valence molecular connectivity χ indices for path/cluster, respectively.

Performance Evaluation. To assess the performance of the statistical learning methods for genotoxicity prediction of the more diverse set of molecules, it is useful to examine whether the accuracy from these methods is at a similar level as those derived by the use of a significantly smaller set of molecules. It is noted that a direct comparison with results from previous studies is inappropriate because of the differences in the data set and molecular descriptors used. Although desirable, it is impossible to conduct a separate comparison using results directly from other studies without full information about the algorithms of molecular descriptors and classification methods used in each study. Nonetheless, a tentative comparison may provide some crude estimate regarding the approximate level of accuracy of the

Table 4. Overview of the Prediction Accuracies of GT+ and GT− Agents from This Work as with Those from Other Studies[a]

study (ref)	method	no. of compds	GT+ accuracy (%)	GT− accuracy (%)	Q overall accuracy (%)
Snyder RD (*5*)[b]	MCASE	394	48.1	95.1	89.6
	DEREK	394	51.9	75.1	73.6
	TOPKAT	394	43.4	88.1	81.7
Philip D. Mosier (*11*)	*k*-NN	140	66.7	92.9	85.0
Linnan He (*12*)	consensus model developed with *k*-NN, LDA, and PNN classifiers	227	73.8	84.3	81.2
Brian E. Mattioni (*13*)	*k*-NN	334	69.3	74.1	72.2
this work	C4.5	860	55.6	75.0	70.7
	PNN	860	74.1	80.2	78.9
	k-NN	860	70.4	86.5	82.9
	SVM	860	77.8	92.7	89.4

[a] Prediction accuracies of this work listed here are based on independent evaluation sets, which are similar to those based on 5-fold cross-validation. Because different groups used different sets of descriptors, the accuracies given in this table only reflect the relative efficiency of each method. [b] Best performance characteristics of the three programs were selected.

genotoxicity prediction systems studied in this work.

Table 4 gives the prediction results of the four statistical methods from this work along with those derived from previous studies. The GT+ accuracies of these four methods are comparable and in some cases slightly better than those of earlier studies derived from *k*-NN (*11*, *13*) and the consensus model developed with *k*-NN, LDA, and PNN (*12*). The GT− accuracies of these four methods are comparable to those of earlier studies (*5*, *11*−*13*). The results from all of these statistical learning methods are substantially better than those obtained by DEREK, TOPKAT, and MCASE programs (*5*). This is likely due to the capability of statistical learning methods for classification of a more diverse range of molecules than that of structural alert-based approaches.

Overall, our study suggests that statistical learning methods, particularly SVM, *k*-NN and PNN, are useful for genotoxicity assessment of a broad range of molecules. The prediction accuracy of these methods is at a similar level as those of earlier studies that were tested by using a much smaller number of molecules. Another advantage of these methods is that they do not require knowledge about the molecular mechanism or SAR of a particular drug property. Moreover, the classification speed of these methods is generally fast. For instance, the number of compounds, which can be classified per second by using SVM, *k*-NN, PNN, and DT methods, is approximately 4000, 3000, 2000, and 62000, respectively, on a P4 3.6 GHz machine. SVM typically uses a portion of the training set as support vectors for classification. In contrast, *k*-NN and PNN use the whole training set for classification. The number of support vectors of SVM is in the range of 45−75% of the training set. Thus, the classification speed of SVM is usually 25−55% faster than that of *k*-NN and PNN. On the other hand, the classification speed of SVM is slower than that of DT methods, which use a set of rules to reach a decision leaf.

There are six GT+ and seven GT− agents in the independent evaluation set that were misclassified by SVM, which are shown in Figures 1 and 2, respectively. The six misclassified GT+ compounds are mebendazole, clomiphene, lansoprazole, clarithromycin, imipramine, and ampicillin. From the study of Snyder et al. (*5*), ampicillin, imipramine, and lansoprazole were also misclassified by MCASE, DEREK, and TOPKAT. Clomiphene was misclassified by MCASE and TOPKAT but correctly classified by DEREK, which alerts the halogenated alkene structure (*5*). Mebendazole was misclassified by DEREK but predicted as equivocal genotoxicity by TOPKAT and by MCASE as GT+ with 57% probability (*5*). To the best of our knowledge, there is no computational study on clarithromycin, which has been found to be GT+ in the in vitro cytogenetics tests (*32*) but GT− in other assays such as bacterial mutation (Ames), mouse lymphoma assay (MLA), and in vivo cytogenetics.

DEREK is a knowledge-based expert system of qualitative estimation model (*8*). MCASE performs a quantitative prediction by generating each test molecule into 2−10 atom fragments by consideration of their physicochemical properties (*8*). TOPKAT uses electrotopological states as well as shape, symmetry, molecular weight, and logP as descriptors in a QSAR model for prediction (*8*). Although each of these methods is able to correctly predict one of the six GT+ compounds misclassified by our method, there are also GT+ compounds, such as naloxone and pentobarbital (*5*), correctly predicted by our method but misclassified by each of these methods. While all of the methods misclassified some of the GT+ compounds due to the general inadequacy for fully representing all of the properties of these molecules, each method appears to be more useful to specific types of compounds than other methods. For instance, clomiphene is correctly predicted by DEREK because of the use of knowledge-based alert for halogenated alkene structure, while it is misclassified by our method because of the lack of a descriptor to properly represent halogen atoms. Thus, the use of multiple methods may be useful to cover a more diverse set of compounds.

The seven misclassified GT− compounds are dansyltryptamine, ketotifin, 2-chloro-4-(4-methoxyphenyl)-3-phenylquinoline, ceftibuten, 5-chloro-1,3-dihydro-1,3,3-trimethylspiro, candesartan, and indinavir. Both candesartan and indinavir were correctly classified by MCASE, DEREK, and TOPKAT (*5*). Ketotifin was correctly classified by MCASE and DEREK but misclassified by TOPKAT (*5*). Ceftibuten was correctly classified by MCASE and TOPKAT but misclassed by DEREK (*5*). The first two compounds contain aromatic amines, the third contains an α,β-unsaturated ketone group, and the fourth is composed of an α,β-unsaturated amide group. These chemical groups can be easily distinguished from the structural alerts of genotoxicity (*59*) used in MCASE, DEREK, and TOPKAT, but they are not properly described by the commonly used molecular descriptors. This is perhaps the reason our method failed to correctly

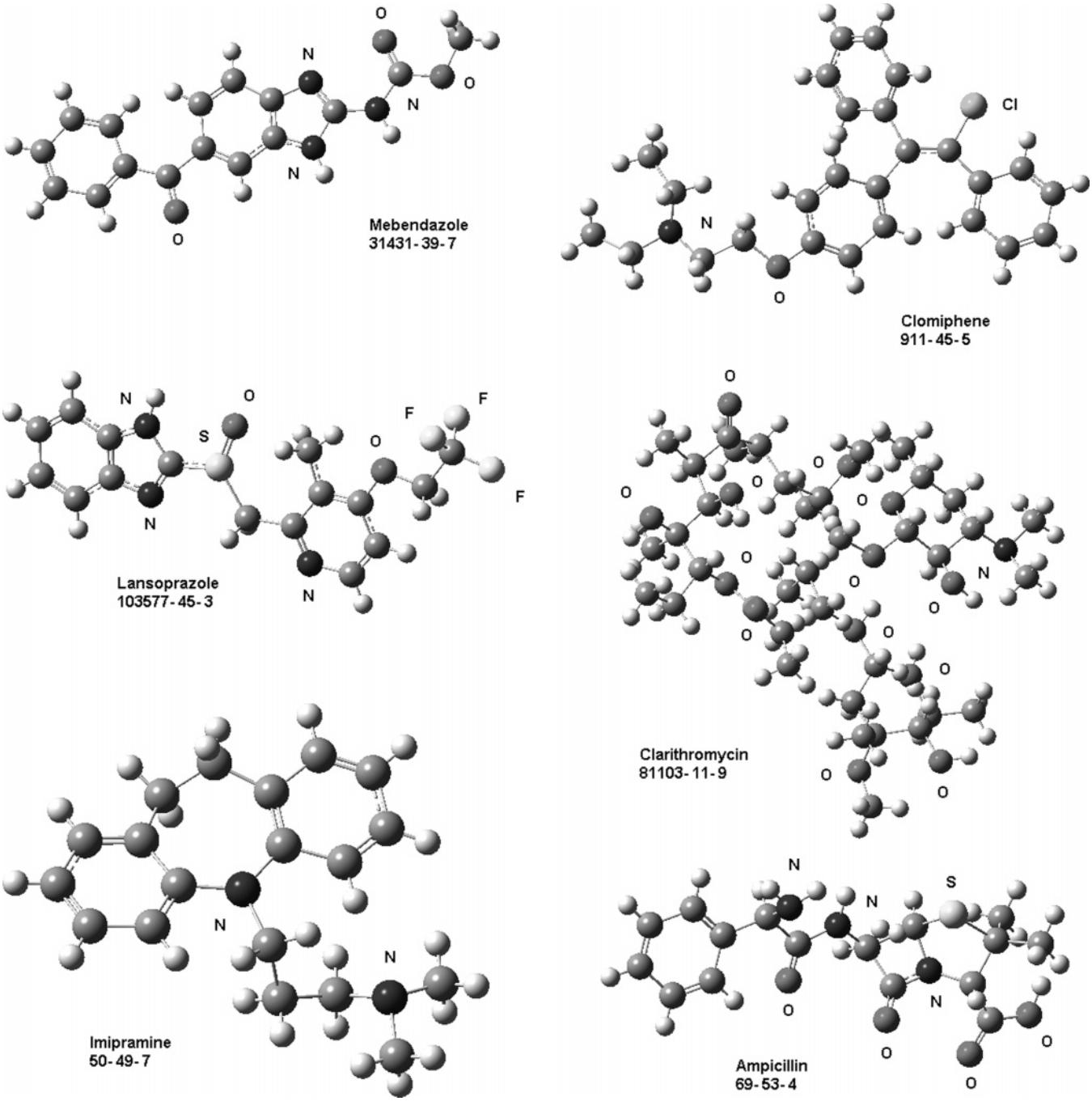

Figure 1. Six structures of misclassified GT+ agents in the independent validation set. Chemical names and relevant Chemical Abstracts Service (CAS) numbers of these compounds are shown in the figure. Heteroatoms (oxygen, nitrogen, fluorine, chlorine, and sulfur) are marked.

classify these four compounds. Dansyltryptamine, 2-chloro-4-(4-methoxyphenyl)-3-phenylquinoline, and 5-chloro-1,3-dihydro-1,3,3-trimethylspiro were correctly predicted by using LDA, *k*-NN, PNN, and their consensus model in an earlier study (*12*). These are polycyclic aromatic compounds that contain either halogen chlorine atom or aromatic amine and a N-dimethyl group. One possible reason for the correct prediction of these compounds in that study (*12*) is that it focused on polycyclic aromatic compounds only and thus was easier to select all of the relevant features without the concern of introducing noises for other types of chemical groups. In contrast, our study includes a diverse set of compounds, and our feature selection method can only pick up those descriptors that are both relevant to the polycyclic aromatic compounds and without significant noise to other types of compounds. It is also noted that there are polycyclic aromatic compounds, such as 9-aminophenanthrene and ethyl 5-hydroxy-2-methylindole-3-carb-oxylate, that were correctly predicted by our method and misclassified in the earlier study (*12*). This seems to suggest that the currently available descriptors may not be fully representative of the polycyclic aromatic compounds.

In general, the main reason for the SVM misclassification of these GT+ and GT− compounds is that none of the currently used descriptors adequately represents the compounds containing multirings with various heteroatoms such as nitrogen, oxygen, sulfur, fluorine, and chlorine. Currently used topological descriptors are capable of representing molecular shape, connectivity, and

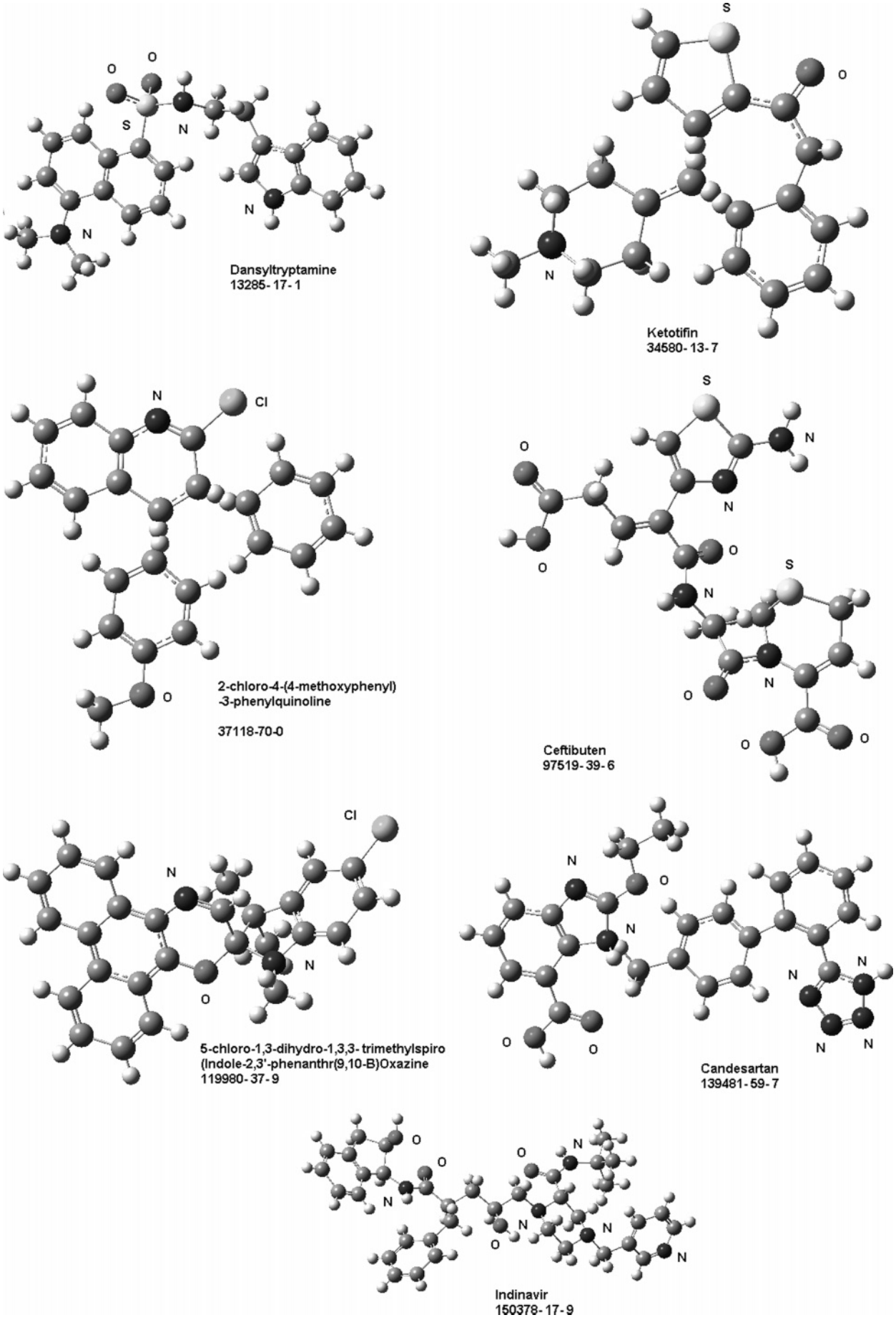

Figure 2. Seven structures of misclassified GT− agents in the independent validation set. Chemical names and relevant CAS numbers of these compounds are shown in the figure. Heteroatoms (oxygen, nitrogen, fluorine, chlorine, and sulfur) are marked.

some level of molecular flexibility (*23*, *60*−*62*). However, because of the limited coverage of the number of bond links in a heteroatom loop, these descriptors are not yet capable of describing the special features of a complex multiring structure that contains multiple heteroatoms. Another reason for the misclassification of some of these compounds is that none of the currently used descriptors can be used to fully represent molecules containing a long flexible chain. Therefore, there is a need to explore different combination of descriptors and to select more optimum set of descriptors by using more refined feature selection algorithms and parameters. However, indiscriminate use of many existing topological descriptors, which are overlapping and redundant to each others, may introduce noise as well as extending the coverage of some the aspects of these special features. Thus, it may be necessary to introduce more appropriate descriptors for representing these and other special features.

Conclusion

This study shows that statistical learning methods, particularly SVM, *k*-NN, and PNN, are useful for facilitating the prediction of GT+ potential of a diverse set of molecules without requiring the intrinsic mechanism knowledge of chemical compounds. The prediction accuracy of these methods may be further improved by introducing molecular descriptors that can better represent complex ring structures and flexible long chains and by selection of descriptors most relevant to genotoxicity prediction by means of more refined feature selection methods and parameters. Current efforts are directed at the improvement of the efficiency and speed of feature selection methods (*44*), which can further help to optimally select molecular descriptors and enable the development of more accurate and efficient computational tools for genotoxicity prediction. Moreover, recent works on the introduction of weighting function into SVM descriptors (*63*) may also be helpful in developing SVM into a practical tool for the prediction of toxicological properties of chemical agents.

Acknowledgment. This work was supported in part by grants from Shanghai Commission for Science and Technology (04DZ19850, 04QMX1450, 04DZ14005), and the "973" National Key Basic Research Program of China (2004CB720103, 2004CB715901).

Supporting Information Available: Table of all compound ID numbers, CAS numbers, names, and classes. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Johnson, D. E., and Wolfgang, G. H. (2000) Predicting human safety: Screening and computational approaches. *Drug Discovery Today 5*, 445−454.
(2) van de Waterbeemd, H., and Gifford, E. (2003) ADMET in silico modelling: towards prediction paradise? *Nat. Rev. Drug Discovery 2*, 192−204.
(3) Kennedy, T. (1997) Managing the drug discovery/development interface. *Drug Discovery Today 2*, 436−444.
(4) Bolzan, A. D., and Bianchi, M. S. (2002) Genotoxicity of streptozotocin. *Mutat. Res. 512*, 121−134.
(5) Snyder, R. D., Pearl, G. S., Mandakas, G., Choy, W. N., Goodsaid, F., and Rosenblum, I. Y. (2004) Assessment of the sensitivity of the computational programs DEREK, TOPKAT, and MCASE in the prediction of the genotoxicity of pharmaceutical molecules. *Environ. Mol. Mutagen. 43*, 143−158.
(6) Kramer, P. J. (1998) Genetic toxicology. *J. Pharm. Pharmacol. 50*, 395−405.
(7) Schwetz, B. A., and Casciano, D. A. (1998) Genetic toxicology: Impact on the next generation of toxicology. *Environ. Mol. Mutagen. 31*, 1−3.
(8) Greene, N. (2002) Computer systems for the prediction of toxicity: An update. *Adv. Drug Delivery Rev. 54*, 417−431.
(9) Cash, G. G. (2001) Prediction of the genotoxicity of aromatic and heteroaromatic amines using electrotopological state indices. *Mutat. Res. 491*, 31−37.
(10) Marchant, C. A. (1996) Prediction of rodent carcinogenicity using the DEREK system for 30 chemicals currently being tested by the National Toxicology Program. The DEREK Collaborative Group. *Environ. Health Perspect. 104*, 1065−1073.
(11) Mosier, P. D., Jurs, P. C., Custer, L. L., Durham, S. K., and Pearl, G. M. (2003) Predicting the genotoxicity of thiophene derivatives from molecular structure. *Chem. Res. Toxicol. 16*, 721−732.
(12) He, L., Jurs, P. C., Custer, L. L., Durham, S. K., and Pearl, G. M. (2003) Predicting the genotoxicity of polycyclic aromatic compounds from molecular structure with different classifiers. *Chem. Res. Toxicol. 16*, 1567−1580.
(13) Mattioni, B. E., Kauffman, G. W., Jurs, P. C., Custer, L. L., Durham, S. K., and Pearl, G. M. (2003) Predicting the genotoxicity of secondary and aromatic amines using data subsetting to generate a model ensemble. *J. Chem. Inf. Comput. Sci. 43*, 949−963.
(14) Cristianini, N., and Shawe-Taylor, J. (2000) *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press, New York.
(15) Burges, C. J. C. (1998) A tutorial on support vector machines for pattern recognition. *Data Min. Knowledge Discovery 2*, 127−167.
(16) Quinlan, J. R. (1993) *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA.
(17) Specht, D. F. (1990) Probabilistic neural networks. *Neural Networks 3*, 109−118.
(18) Johnson, R. A., and Wichern, D. W. (1982) *Applied Multivariate Statistical Analysis*, Prentice Hall, Englewood Cliffs, NJ.
(19) Doniger, S., Hofman, T., and Yeh, J. (2002) Predicting CNS permeability of drug molecules: Comparison of neural network and support vector machine algorithms. *J. Comput. Biol. 9*, 849−864.
(20) Trotter, M. W. B., Buxton, B. F., and Holden, S. B. (2001) Support vector machines in combinatorial chemistry. *Meas. Control 34*, 235−239.
(21) Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., and Haussler, D. (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics 16*, 906−914.
(22) Bock, J. R., and Gough, D. A. (2001) Predicting protein−protein interactions from primary structure. *Bioinformatics 17*, 455−460.
(23) Serra, J. R., Thompson, E. D., and Jurs, P. C. (2003) Development of binary classification of structural chromosome aberrations for a diverse set of organic compounds from molecular structure. *Chem. Res. Toxicol. 16*, 153−163.
(24) Yap, C. W., Cai, C. Z., Xue, Y., and Chen, Y. Z. (2004) Prediction of torsade-causing potential of drugs by support vector machine approach. *Toxicol. Sci. 79*, 170−177.
(25) Xue, Y., Yap, C. W., Sun, L. Z., Cao, Z. W., Wang, J. F., and Chen, Y. Z. (2004) Prediction of p-glycoprotein substrates by support vector machine approach. *J. Chem. Inf. Comput. Sci. 44*, 1497−1505.
(26) Burbidge, R., Trotter, M., Buxton, B., and Holden, S. (2001) Drug design by machine learning: Support vector machines for pharmaceutical data analysis. *Comput. Chem. 26*, 5−14.
(27) Czerminski, R., Yasri, A., and Hartsough, D. (2001) Use of support vector machine in pattern classification: Application to QSAR studies. *Quant. Struct.-Act. Relat. 20*, 227−240.
(28) Yu, H., Yang, J., Wang, W., and Han, J. (2003) Discovering compact and highly discriminative features or feature combinations of drug activities using support vector machines, *IEEE Computer Society Bioinformatics Conference (CSB'03)*, pp 220−228, Stanford, California.
(29) Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002) Gene selection for cancer classification using support vector machines. *Machine Learn. 46*, 389−422.
(30) Xue, Y., Li, Z. R., Yap, C. W., Sun, L. Z., Chen, X., and Chen, Y. Z. (2004) Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents. *J. Chem. Inf. Comput. Sci. 44*, 1630−1638.
(31) Trotter, M. W. B., and Holden, S. B. (2003) Support vector machines for ADME property classification. *QSAR Comb. Sci. 22*, 533−548.

(32) Snyder, R. D., and Green, J. W. (2001) A review of the genotoxicity of marketed pharmaceuticals. *Mutat. Res./Rev. Mutat. Res. 488*, 151−169.
(33) Quillardet, P. H. M. (1993) The SOS chromotest: A review. *Mutat. Res. 297*, 235−279.
(34) Vasilieva, S. (2002) SOS Chromotest methodology for fundamental genetic research. *Res. Microbiol. 153*, 435−440.
(35) Pearlman, R. S. *CONCORD User's Manual*, Tripos, St. Louis, MO.
(36) Dewar, M. J. S., Zoebisch, E. G., Healy, E. F., and Steward, J. P. P. (1985) AM1: A new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc. 107*, 3902−3909.
(37) Todeschini, R., and Consonni, V. (2000) *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim.
(38) Katritzky, A. R., and Gordeeva, E. V. (1993) Traditional topological indices vs electronic, geometrical and combined molecular descriptors in QSAR/QSPR research. *J. Chem. Inf. Comput. Sci. 33*, 835−857.
(39) Cruciani, G., Pastor, M., and Guba, W. (2000) VolSurf: A new tool for the pharmacokinetic optimization of lead compounds. *Eur. J. Pharm. Sci. 11*, S29−S39.
(40) Kier, L. B., and Hall, L. H. (1999) *Molecular Structure Description: The Electrotopological State*, Academic Press, San Diego.
(41) Karelson, M., Lobanov, V. S., and Katritzky, A. R. (1996) Quantum-chemical descriptors in QSAR/QSPR studies. *Chem. Rev. 96*, 1027−1043.
(42) Degroeve, S., De Baets, B., Van de Peer, Y., and Rouzé, P. (2002) Feature subset selection for splice site prediction. *Bioinformatics 18*, S75−S83.
(43) Bayada, D. M., Hamersma, H., and van Geerestein, V. J. (1999) Molecular diversity and representativity in chemical databases. *J. Chem. Inf. Comput. Sci. 39*, 1−10.
(44) Furlanello, C., Serafini, M., Merler, S., and Jurman, G. (2003) An accelerated procedure for recursive feature ranking on microarray data. *Neural Networks 16*, 641−648.
(45) Vapnik, V. N. (1995) *The Nature of Statistical Learning Theory*, Springer, New York.
(46) Fix, E., and Hodges, J. L. (1951) *Discriminatory Analysis: Nonparametric Discrimination: Consistency Properties*, USAF School of Aviation Medicine, Randolph Field, Texas.
(47) Parzen, E. (1962) On estimation of a probability density function and mode. *Ann. Math. Stat. 33*, 1065−1076.
(48) Cacoullos, T. (1966) Estimation of a multivariate density. *Ann. I. Stat. Math. 18*, 179−189.
(49) Carnahan, B., Meyer, G., and Kuntz, L.-A. (2003) Comparing statistical and machine learning classifiers: Alternatives for predictive modeling in human factors research. *Hum. Factors 45*, 408−423.
(50) Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A., and Nielsen, H. (2000) Assessing the accuracy of prediction algorithms for classification: An overview. *Bioinformatics 16*, 412−424.
(51) Roulston, J. E. (2002) Screening with tumor markers. *Mol. Pharmacol. 20*, 153−162.
(52) Matthews, B. W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta 405*, 442−451.
(53) Brown, M. P. S., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C., Furey, T. S., Manuel Ares, J., and Haussler, D. (2000) Knowledge-based analysis of microarray gene expression data using support vector machines. *Proc. Natl. Acad. Sci. U.S.A. 97*, 262−267.
(54) Huang, C., Davis, L. S., and Townshend, J. R. G. (2002) An assessment of support vector machines for land cover classification. *Int. J. Remote Sensing 23*, 725−749.
(55) Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T., and Vapnik, V. (2001) Feature selection for SVMs, *Advances in Neural Information Processing Systems*.
(56) Saeys, Y., Degroeve, S., Aeyels, D., Rouzé, P., and Van de Peer, Y. (2004) Feature selection for splice site prediction: A new method using EDA-based feature ranking. *BMC Bioinformatics 5*, 64.
(57) Crivori, P., Cruciani, G., Carrupt, P. A., and Testa, B. (2000) Predicting blood-brain barrier permeation from three-dimensional molecular structure. *J. Med. Chem. 43*, 2204−2216.
(58) Chung, K. T., Kirkovsky, L., Kirkovsky, A., and Purcell, W. P. (1997) Review of mutagenicity of monocyclic aromatic amines: Quantitative structure−activity relationships. *Mutat. Res. 387*, 1−16.
(59) Ashby, J. (1985) Fundamental structural alerts to potential carcinogenicity or noncarcinogenicity. *Environ. Mutagen. 7*, 919−921.
(60) Wegner, J. K., Fröhlich, H., and Zell, A. (2004) Feature selection for descriptor based classification models. 2. Human intestinal absorption (HIA). *J. Chem. Inf. Comput. Sci. 44*, 931−939.
(61) Luco, J. M. (1999) Prediction of the brain-blood distribution of a large set of drugs from structurally derived descriptors using partial least-squares (PLS) modeling. *J. Chem. Inf. Comput. Sci. 39*, 396−404.
(62) Basak, S. C., Gute, B. D., and Ghatak S. (1999) Prediction of complement-inhibitory activity of benzamidines using topological, and geometric parameters. *J. Chem. Inf. Comput. Sci. 39*, 255−260.
(63) Chapelle, O., Vapnik, V., Bousquet, O., and Mukherjee, S. (2002) Choosing multiple parameters for support vector machines. *Machine Learn. 46*, 131−159.

TX049652H