Computers in Biology and Medicine 51 (2014) 122–127

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

A prediction model of drug-induced ototoxicity developed by an optimal support vector machine (SVM) method

Shu Zhou [a], Guo-Bo Li [a], Lu-Yi Huang [a], Huan-Zhang Xie [b], Ying-Lan Zhao [a], Yu-Zong Chen [a], Lin-Li Li [b,*], Sheng-Yong Yang [a,**]

[a] State Key Laboratory of Biotherapy and Cancer Center, West China Hospital, West China Medical School, Sichuan University, Sichuan 610041, PR China
[b] West China School of Pharmacy, Sichuan University, Sichuan 610041, PR China

ARTICLE INFO

Article history:
Received 3 December 2013
Accepted 9 May 2014

Keywords:
Drug-induced ototoxicity
Support vector machine
Naïve Bayesian
Recursive partitioning
Classification

ABSTRACT

Drug-induced ototoxicity, as a toxic side effect, is an important issue needed to be considered in drug discovery. Nevertheless, current experimental methods used to evaluate drug-induced ototoxicity are often time-consuming and expensive, indicating that they are not suitable for a large-scale evaluation of drug-induced ototoxicity in the early stage of drug discovery. We thus, in this investigation, established an effective computational prediction model of drug-induced ototoxicity using an optimal support vector machine (SVM) method, GA-CG-SVM. Three GA-CG-SVM models were developed based on three training sets containing agents bearing different risk levels of drug-induced ototoxicity. For comparison, models based on naïve Bayesian (NB) and recursive partitioning (RP) methods were also used on the same training sets. Among all the prediction models, the GA-CG-SVM model II showed the best performance, which offered prediction accuracies of 85.33% and 83.05% for two independent test sets, respectively. Overall, the good performance of the GA-CG-SVM model II indicates that it could be used for the prediction of drug-induced ototoxicity in the early stage of drug discovery.

1. Introduction

Clinical studies have shown that some drugs, such as amikacin, gentamicin and neomycin, could lead damage to the ear, specifically the cochlea or auditory nerve and sometimes the vestibular system [1–3]. Symptoms of the drug-induced ototoxicity vary considerably from drug to drug and person to person, ranging from mild imbalance to total incapacitation, and tinnitus to total hearing loss. The drug-induced ototoxicity may be either reversible and temporary, or irreversible and permanent. The bad thing is that currently there is no effective treatment to reverse the effects of ototoxicity if permanent damage happens to the ear [1,2]. Therefore, drug-induced ototoxicity, like other drug toxicities, is also an important issue needed to be considered in drug discovery.

Currently, several methods based on chemical biology have been used to evaluate drug-induced ototoxicity [4]. However, these methods require a number of experiments, most of which are time-consuming and expensive, indicating that they are not suitable for a large-scale evaluation of drug-induced ototoxicity in the early stage of drug discovery. Lately emerging zebrafish-based methods seemed faster and cheaper [5,6]. However, due to the relatively large species difference between zebrafish and humans, these methods are prone to misjudge the drug-induced ototoxicity [4,7]. Therefore, it is highly needed to develop more efficient and fast methods for a large-scale evaluation of drug-induced ototoxicity in drug discovery.

Computational methods have been thought as a faster and cheaper strategy and have been successfully used in the prediction of various pharmacokinetic and toxic properties of drugs. For example, Wang et al. developed prediction models of human ether-a-go-go related gene (hERG) potassium channel blockage using the naïve Bayesian (NB) and recursive partitioning (RP) methods [8]. Burton et al. used the RP method to develop prediction models of Cytochromes P450 2D6 and 1A2 inhibitors [9]. Ma et al. developed a prediction model of drug oral bioavailability by the support vector machine (SVM) method [10]. More computational prediction models of pharmacokinetic and toxic properties could be found in the literature [11–14]. However, as far as we know, there is no report of computational model for the prediction of drug-induced ototoxicity. Therefore, we shall, in this investigation, develop a computational classification prediction model of drug-induced ototoxicity using the SVM method [15,16]. Here we chose the SVM method because SVM has been demonstrated to be one of the best statistical learning methods and has

* Corresponding author.
** Corresponding author. Tel.: +86 28 85164063; fax: +86 28 85164060.
E-mail addresses: ysylilinli@sina.com (L.-L. Li), yangsy@scu.edu.cn (S.-Y. Yang).

http://dx.doi.org/10.1016/j.compbiomed.2014.05.005
0010-4825/

shown better performances in the development of prediction models of pharmacokinetic and toxic properties than many other methods [17–21]. Specifically, we used in this study a modified version of SVM, namely GA-CG-SVM [22], in which the genetic algorithm (GA) [23]is used for the feature selection and the conjugate gradient (CG) [24] method for the parameter optimization. A main advantage of GA-CG-SVM is that it can concurrently optimize the features and SVM parameters. GA-CG-SVM has been demonstrated to outperform the common SVM method [10,17]. For comparison, NB and RP methods will also be used to build prediction models of drug-induced ototoxicity; the two methods were chosen since they are also very good and popular methods for classification modeling.

2. Materials and methods

2.1. GA-CG-SVM

SVM is a supervised machine learning method, which has been widely used for classification and regression analysis. It has shown a good performance in solving a number of biological classification problems [25–27]. Nevertheless, some issues that may have important influence on the quality of established models are often not or insufficiently considered in SVM modeling, such as feature selection and parameter optimization. Our group recently developed an optimal SVM method, termed GA-CG-SVM, in which feature selection and parameter optimization are efficiently handled in SVM modeling. Detailed algorithms for GA-CG-SVM have already been described in a previous paper [22]. Here we just make a brief summary for the basic idea of GA-CG-SVM.

Supposing a given training set $\{(x_1,y_1),(x_2,y_2),...(x_i,y_i)\}$ where x_i represents a vector of n real numbers (features or descriptors) and y_i is the class that vector x_i belongs to. The purpose of SVM classification is to find an optimal hyperplane to separate these classes with the maximum margin, which needs to solve the following optimization problem:

$$\underset{w,b}{\text{Max}}\frac{2}{||w||} \text{ subject to } y_i(wx_i+b)-1\geq 0 \quad (1)$$

where $2/||w||$ is the margin.

For linearly separable cases, it is easy to find an optimal separating hyperplane by a classifying determination function. For linearly non-separable cases, there is no hyperplane that can be used to perfectly separate two sets of points (See Supplementary Figure S1). Therefore, non-negative slack variables $\xi_i\geq 0,\ i=1,....,m$ were introduced. The equation to be solved becomes:

$$\underset{w,b}{\text{Max}}\frac{2}{||w||}+C\sum_{i=1}^{m}\xi_i \text{ subject to } y_i(wx_i+b)-1+\xi_i\geq 0 \quad (2)$$

where C is a user predetermined penalty parameter.

For nonlinear (non-) separable cases, the basic idea is to project the input data set x_i into a high-dimensional feature space via a nonlinear manner using a kernel function [28]. Until now, many kernel functions have been suggested for this purpose. Among them, the radial basis function (RBF) is widely used and performed very well in most cases [29]. Thus, the RBF kernel function was also selected in this study.

$$k(x_i,x_j)=exp(-\gamma||x_i-x_j||^2) \quad (3)$$

where the parameter γ denotes the width of Gaussian kernel.

As mentioned previously, the selection of penalty parameter C and kernel parameter γ in SVM modeling has significant influence on the predictive accuracy of an SVM model. Thus, we used the conjugate gradient method to optimize the two parameters. Additionally, the selection of features is also of importance to the prediction ability of an SVM model [30]. The GA is a very popular optimization algorithm, which is based on the Darwinian evolutionary idea of natural selection and genetics in biological systems. GA has been widely used to solve a range of diverse problems such as data mining and optimization [31,32]. Here, we applied GA for the feature selection in SVM modeling. Finally, it is worth mentioning that an integrated scheme for the simultaneous treatment of both feature selection (by GA) and parameter optimization (by CG) was adopted. This is important since it has been shown that the feature selection and parameter setting influence each other in SVM modeling [33].

2.2. The naïve Bayesian (NB) classifier

A naive Bayesian classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions [34]. In NB, each object is described by an n-dimensional vector $F=(f_1,f_2,....,f_n)$, where $(f_1,f_2,....,f_n)$ represents n features. Objects belonging to the first class are each labeled a value of $CL_1=1$, while those in the second class are assigned $CL_2=-1$. Based on the Bayes' theorem, we got

$$p(CL_i|F)=\frac{p(F|CL_i)p(CL_i)}{p(F)} \quad (4)$$

where $p(CL_i|F)$ denotes the posterior probability, and $p(CL_i)$ represents the prior probability. $p(F|CL_i)$ and $p(F)$ are conditional probability and marginal probability, respectively. $p(F)$ is constant for all classes. $p(F|CL_i)$ and $p(F)$ can be learned from a training set.

In this study, NB was used for the development of prediction models of drug-induced ototoxicity. CL_1 and CL_2 represent ototoxic drug class and non-ototoxic drug class, respectively. The naive Bayesian classifier was constructed using the Discovery Studio 3.1 software package.

2.3. The recursive partitioning (RP) classifier

RP is a statistical method for multivariable analysis, which is also a popular classification method [35,36]. RP produces a decision tree that strives to correctly classify members of the population based on several dichotomous dependent variables. At each node of the decision tree, the data are split into two subsets based on a particular descriptor and corresponding splitting value, which are decided by an automated statistical analysis of the entire data set. The splitting process continues until no more significant nodes are obtained or a threshold value for stopping is reached. Let X stand for a set of independent properties (molecular properties) and Y represent a dependent property (ototoxicity class). With RP method, created decision trees can recursively partition data according to the relationship between X and Y values. In this study, 5-fold cross validation was used, splits were scored using Gini index and the minimum number of samples at each node was set to 10 to avoid excessive partitioning. Furthermore, the maximum tree depth was changed from 2 to 10 systematically in order to find a better RP model. Our recursive partitioning classifiers were built using the Discovery Studio 3.1 software package.

2.4. Data set

A total of 572 small molecule compounds (positive) that have been reported to bear ototoxicity were collected from reference [37], which contains the most updated collections of ototoxic drugs. 347 drug molecules (negative) for which there is no report to bear ototoxicity were collected from DrugBank [38]. We used the "Generate Training and Test Data" module in Discovery Studio

to randomly select 20% compounds (121 positives and 63 negatives) to form an independent test set (called TS1) in advance. The remaining positives and negatives were taken to construct three training sets, namely dataset I, II and III; these compounds could occur in different sets. The biggest difference of these datasets is the risk or strength of ototoxicity of positive compounds. According to Bauman, the ototoxic drugs were divided into five classes (class 1, 2, 3, 4 and 5) based on their risk or strength of ototoxicity [37]. Compounds in class 1 have the lowest risk; compounds in class 5 have the highest risk; compounds in class 2–4 have an increasing risk. Dataset I includes positive compounds of all of the risk classes (class 1, 2, 3, 4 and 5). Dataset II contains positive compounds of risk class 2, 3, 4 and 5. Dataset III includes positive compounds of higher risk classes (class 3, 4 and 5). The finally formed dataset I consists of 451 positives and 284 negatives (see Supplementary Table S1). Dataset II contains 252 positives and 284 negatives (see Supplementary Table S2). Dataset III includes 64 positives and 284 negatives (see Supplementary Table S3). Test set TS1 (121 positives and 63 negatives) is presented in Supplementary Table S4.

To further evaluate the established models, we constructed another independent test set, called TS2. TS2 contains 19 positives and 40 negatives (see Supplementary Table S5). Positives in TS2 were collected from Hazardous Substances Data Bank (HSDB) of TOXNET toxicology data network [39]. Negatives in TS2 were again collected from DrugBank [37]. There is no overlap between TS2 and any one of training set I, II, III and TS1.

2.5. Molecular descriptors

Molecular descriptors used in this study were calculated by Discovery Studio 3.1 software package. Initially, a total of 237 molecular descriptors for each molecule were calculated; these descriptors cover a variety of molecular properties, including topological descriptors, element counts, AlogP, surface area and volume, and so on. After that, a preprocessing procedure was conducted to these calculated descriptors. The preprocessing procedure includes two steps. In the first step, some "bad" descriptors were removed. For example, descriptors with too many zero values, bearing a very small standard deviation value with others ($<0.5\%$), or having a high correlation coefficient with others ($>95\%$), were deleted from the descriptor list. In the second step, all the descriptor values for the remaining descriptors were scaled to a range of -1 to 1. These preprocessed descriptors were further optimized by GA (for GA-CG-SVM) or a Monte Carlo (MC) method (for NB and RP), a feature selection method similar to that used in reference [40]. Finally, these selected descriptors were used for the development of prediction models of drug-induced ototoxicity.

2.6. Data analysis and model validation

To assess the performance of the established prediction models, the following quantities were calculated: true positives (TP, true ototoxic drugs), true negatives (TN, true non-ototoxic drugs), false positives (FP, false ototoxic drugs) and false negatives (FN, false non-ototoxic drugs). Sensitivity SE=TP/(TP+FN) and specificity SP=TN/(TN+FP) are the prediction accuracy for ototoxic drugs and non-ototoxic drugs, respectively. The overall accuracy (Q) was calculated by the equation: $Q=(TP+TN)/(TP+TN+FP+FN)$. The Matthew's correlation (MCC) was calculated by the following equation:

$$MCC = (TP \times TN - FN \times FP)/\sqrt{(TP+FN)(TP+FP)(TN+FN)(TN+FP)}$$

3. Results

3.1. SVM prediction models of ototoxicity by using GA-CG-SVM

Three prediction models, namely GA-CG-SVM model I, II and III, were developed using GA-CG-SVM based on dataset I, II and III, respectively. The compositions of these three datasets are shown in Supplementary Table S6. Numbers of descriptors, which were optimized by GA, are 27, 32 and 23 for GA-CG-SVM model I, II and III, respectively. Table 1 shows the 5-fold cross validation results for the established models. Among the three models, the GA-CG-SVM model III, which was developed based on dataset III, had the highest overall prediction accuracy (91.38%). Nevertheless, the prediction accuracy for positives was just 59.38%. Though GA-CG-SVM model I, which was developed based on dataset I, gave the highest prediction accuracy for positives (84.70%), it offered the lowest overall prediction accuracy (82.31%). Comparatively speaking, the GA-CG-SVM model II, which was developed based on dataset II, is relatively superior; it gave an overall prediction accuracy of 86.75% and a prediction accuracy of 82.94% for positives.

3.2. Validation of the established GA-CG-SVM models by an independent test set

A good SVM model is not only able to classify training set correctly but also capable of categorizing external agents that are outside of the training set. Thus, an independent validation set, TS1, was further used to assess the predictability of these SVM models just built. The predicted results for TS1 are also shown in Table 1. For GA-CG-SVM model I, the calculated SE and SP were 63.64% and 88.89%, respectively, and the overall accuracy was 72.28%. For GA-CG-SVM model II, the calculated SE, SP and the overall accuracy were 81.82%, 92.06% and 85.33%, respectively. The calculated SE, SP and the overall accuracy for GA-CG-SVM model III were 40.50%, 100% and 60.87%, respectively. Clearly, among the three models, the GA-CG-SVM model II has a superior performance in terms of the overall accuracy and the accuracy for positives (SE).

3.3. Comparison with models built by the naïve Bayesian and recursive partitioning methods

For comparison, we also developed prediction models of drug-induced ototoxicity using naïve Bayesian and recursive partitioning

Table 1
Predicted results for the training sets and test set TS1 of the three GA-CG-SVM models.

Model	Training set								Test set							
	TP	TN	FP	FN	SE	SP	Q	MCC	TP	TN	FP	FN	SE	SP	Q	MCC
GA-CG-SVM model I	382	223	61	69	0.8470	0.7852	0.8231	0.6291	77	56	7	44	0.6364	0.8889	0.7228	0.5004
GA-CG-SVM model II	209	256	28	43	0.8294	0.9014	0.8675	0.7344	99	58	5	22	0.8182	0.9206	0.8533	0.7072
GA-CG-SVM model III	38	280	4	26	0.5938	0.9859	0.9138	0.6894	49	63	0	72	0.4050	1.0000	0.6087	0.4347

Table 2
Comparison of the predicted results for test set TS1 of GA-CG-SVM, NB and RP models.

Model	Method	TP	TN	FP	FN	SE	SP	Q
GA-CG-SVM model I	GA-CG-SVM	77	56	7	44	0.6364	0.8889	0.7228
GA-CG-SVM model II	GA-CG-SVM	99	58	5	22	0.8182	0.9206	0.8533
GA-CG-SVM model III	GA-CG-SVM	49	63	0	72	0.4050	1.0000	0.6087
RP model I (RP-9)	Recursive partitioning	92	49	14	29	0.7603	0.7778	0.7663
RP model II (RP-6)	Recursive partitioning	101	46	17	20	0.8347	0.7302	0.7989
RP model III (RP-4)	Recursive partitioning	98	36	27	23	0.8099	0.5714	0.7283
NB model I	Naïve Bayesian	87	48	15	34	0.7190	0.7619	0.7337
NB model II	Naïve Bayesian	102	38	25	19	0.8430	0.6032	0.7609
NB model III	Naïve Bayesian	90	41	22	31	0.7438	0.6508	0.7120

Table 3
Predicted results for the test set TS2 of GA-CG-SVM model II, RP and NB models.

Model	Method	TP	TN	FP	FN	SE	SP	Q
GA-CG-SVM model II	GA-CG-SVM	15	34	6	4	0.7895	0.8500	0.8305
RP model I (RP-6)	Recursive partitioning	14	30	10	5	0.7368	0.7500	0.7458
RP model II (RP-9)	Recursive partitioning	15	31	9	4	0.7895	0.7750	0.7797
RP model III (RP-6)	Recursive partitioning	16	29	11	3	0.8421	0.7250	0.7627
NB model I	Naïve Bayesian	13	31	9	6	0.6842	0.7750	0.7458
NB model II	Naïve Bayesian	14	29	11	5	0.7368	0.7250	0.7288
NB model III	Naïve Bayesian	14	30	10	5	0.7368	0.7500	0.7458

methods. The same datasets and initial descriptors as those used in the development of GA-CG-SVM models were used. The initial descriptors were preprocessed in advance, followed by an optimization by the MC method. The predicted results for the RP models based on the three datasets with different tree depths are shown in Supplementary Table S7, S8 and S9, respectively. For dataset I, the RP-9 model (tree depth: 10) had the highest overall accuracy (Supplementary Table S7 and Table 2, 76.63%). For dataset II, the highest overall accuracy (Supplementary Table S8 and Table 2, 79.89%) corresponds to the RP-6 model (tree depth: 7). For dataset III, the RP-4 model (tree depth: 5) offered the highest overall prediction accuracy (Supplementary Table S9 and Table 2, 72.83%). For the naïve Bayesian models, the overall prediction accuracy was 73.37% for NB model I (developed based on dataset I), 76.09% for NB model II (developed based on dataset II) and 71.20% for NB model III (developed based on dataset III). For comparison, all the predicted results of models developed by GA-CG-SVM, RP and NB are summarized in Table 2. From Table 2, we can see that the overall prediction accuracy of GA-CG-SVM model II was still the highest one among those of all the models established here.

3.4. Further validation of GA-CG-SVM mode II by a new independent test set.

To further evaluate the GA-CG-SVM model II, we re-constructed a new independent test set, TS2, which contains 19 positives and 40 negatives. Different from TS1, the positive compounds in TS2 were collected from HSDB of TOXNET toxicology data network [37]. Table 3 shows the predicted results of GA-CG-SVM model II to TS2. The SE and SP are 78.95% and 85.00%, respectively. The overall prediction accuracy is 83.05%. Again, for comparison, the prediction results of three RP models (RP model I, II, and III, details see Supplementary Tables S10, S11, and S12) and three NB models (NB model I, II, and III, details see Supplementary Table S13) are also summarized in Table 3. Clearly, the GA-CG-SVM model II still outperforms all the RP and NB models established here in terms of the prediction accuracy.

4. Discussion

Drug-induced ototoxicity is a severe adverse effect of drugs. However, the molecular mechanisms of drug-induced ototoxicity are extremely complicated and far from being established. Thus it is difficult to predict the ototoxic potential of new drugs using traditional statistical methods or structure-toxicity methods. Dealing with those systems with complex mechanisms is the strong point of SVM approach. The SVM method has shown promising capability for solving a number of biological classification problems [25–27]. However, some problems still exist in SVM modeling, i.e., feature selection and parameters optimization. GA-CG-SVM is a modified SVM method that can handle simultaneously the feature selection and parameters optimization [22]. Previous studies have also shown that the GA-CG-SVM method could give higher prediction accuracy than traditional methods [10,17,25]. The good performance of SVM model established here once again demonstrates the capability of GA-CG-SVM method in dealing with complicated systems. In the established GA-CG-SVM model II, 32 different molecular descriptors (see Table 4) were selected by the GA algorithm, which covers diverse molecular properties including molecular structural information, lipophilicity, hydrogen bonding feature, molecular electronic properties, molecular aromatic functions and molecular polar surface area. The involvement of so many molecular descriptors reflects at least to some extent that the drug-induced ototoxicity is affected by many complicated factors.

In the SVM modeling, we elaborately constructed the training sets. The ototoxic drugs were firstly assigned into three training datasets (I, II and III) according to their risk or strength. Of these datasets, dataset III contains a relatively stronger condition, which means that positive compounds in dataset III are definitely able to induce ototoxicity and have a relatively higher potency. On the contrary, dataset I contains a weaker condition, which means that positive compounds in dataset I also include those that are just suspected to be able to induce ototoxicity or have a lower potency, in addition to compounds that are definitely able to induce ototoxicity and have a relatively higher potency. Dataset II contains a medium condition. Based on the constructed datasets, three

Table 4
Molecular descriptors optimized by the GA-CG-SVM method in the establishment of the GA-CG-SVM model II together with their explanations.

Descriptor	Explanation
C_Count	Number of carbon atoms
Cl_Count	Number of chlorine atoms
H_Count	Number of hydrogen atoms
N_Count	Number of nitrogen atoms
O_Count	Number of oxygen atoms
S_Count	Number of sulfur atoms
ALogP	Log of the octanol–water partition coefficient using Ghose and Crippen's method
Apol	Sum of atomic polarizabilities
FormalCharge	Formal charge of an atom
LogD	The octanol–water partition coefficient calculated taking into account the ionization states of the molecule
HBA_Count	The number of hydrogen bond accepting groups in the molecule
HBD_Count	Number of hydrogen bond donating groups
Num_PositiveAtoms	Atoms with a positive charge
Num_NegativeAtoms	Atoms with a negative charge
Num_BridgeHeadAtoms	A bridgehead atom connects a bridge to a ring
Num_AromaticBonds	Bonds in aromatic ring systems
Num_BridgeBonds	Bonds in bridgehead ring systems
Num_RingAssemblies	Number of ring assemblies
Num_RingS6	Number of rings of size 5
Num_RingS7	Number of rings of size 6
Num_StereoBonds	Number of stereo bonds
Num_AliphaticDoubleBonds	Number of aliphatic double bonds
Num_TerminalRotomers	Number of terminal rotomers
Num_H_Donors	Number of hydrogen bond donors
Molecular_PolarSASA	Calculates the polar solvent accessible surface area for each molecule using a 2D approximation
Molecular_Fractional Polar SASA	The ratio of the polar solvent accessible surface area divided by the total solvent accessible surface area
E_DIST_equ	Raph-theoretical info content descriptors
IAC_Mean	Number of nitrogen and oxygen atoms
IC	Graph-theoretical info content descriptors
CHI_V_3_C	Connectivity Indices
Kappa_2	Kappa shape indices
Kappa_3	Kappa shape indices

prediction models of drug-induced ototoxicity were then established. The GA-CG-SVM model II, which was developed based on dataset II, was superior compared with other models. These results reflect the fact that the training set is critical to the quality of generated SVM models, although we cannot conclude that a training set containing a medium condition must be better than that containing a stronger or weaker condition.

Finally, GA-CG-SVM displayed a better performance in prediction of drug-induced ototoxicity than NB and RP. In prediction of some other pharmacokinetic and toxic properties, this method also showed excellent performance. Even so, we still cannot guarantee that GA-CG-SVM must perform well and outperform other methods in various application fields. In fact, many factors may have influence on the quality of models established by a specific modeling method. These factors include the size, diversity, and representativeness of training sets, the number of descriptors, and the link between descriptor values and properties of samples. Among all the mentioned factors, what is worth mentioning is the size of training set. For majority of the modeling methods, a large training set benefits to the generation of models with a high quality. Nevertheless, in SVM, the size of training set is not decisively important on the model quality if samples constituting the supporting vectors have already been included in the training set [41].

In summary, we have established an effective prediction model of drug-induced ototoxicity using GA-CG-SVM. The GA-CG-SVM model II, which is the best GA-CG-SVM model among the three GA-CG-SVM models developed based on different training sets, gave an overall prediction accuracy of 85.33% and 83.05% for the independent test set TS1 and TS2, respectively. A comparison analysis showed that the GA-CG-SVM model II outperformed the RP and NB models, which were developed by the recursive partitioning and naïve Bayesian methods, respectively. All of these results indicate that the GA-CG-SVM model II is a good prediction model of drug-induced ototoxicity. It is expected that the GA-CG-SVM model established here can be used as an effective screening tool for identifying the ototoxic drugs and non-ototoxic drugs in the early stage of drug discovery.

Conflict of interest statement

None declared.

Acknowledgments

This work was supported by the 863 Hi-Tech Programs (2012AA020301, 2012AA0203), National Natural Science Funds for Distinguished Young Scholar (81325021) and National Science and Technology Major Project (2012ZX09501001–003).

Appendix A. Supplementary information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.compbiomed.2014.05.005.

References

[1] V.P. Garcia, F.A. Martinez, E.B. Agusti, L.A. Mencia, V.P. Asenjo, Drug-induced otoxicity: current status, Acta Otolaryngol. 121 (2001) 569–572.
[2] M. Bisht, S.S. Bist, Ototoxicity: The Hidden Menace, Indian J, Otolaryngol. Head Neck Surg. 63 (2011).
[3] L.P. Rybak, Drug ototoxicity, Ann. Rev. Pharmacol. Toxicol. 26 (1986) 79.
[4] J.G. Yorgason, W. Luxford, F. Kalinec, in vitro and in vivo models of drug ototoxicity: studying the mechanisms of a clinical problem, Expert Opin. Drug Metab. Toxicol. 7 (2011) 1521–1534.
[5] L.L. Chiu, L.L. Cunningham, D.W. Raible, E.W. Rubel, H.C. Ou, Using the zebrafish lateral line to screen for ototoxicity, J. Assoc. Res. Otolaryngol. 9 (2008) 178–190.
[6] C. Ton, C. Parng, The use of zebrafish for assessing ototoxic and otoprotective agents, Hearing Res. 208 (2005) 79–88.
[7] H.C. Ou, F. Santos, D.W. Raible, J.A. Simon, E.W. Rubel, Drug screening for hearing loss: using the zebrafish lateral line to screen for drugs that prevent and cause hearing loss, Drug Discov. Today 15 (2010) 265–271.
[8] S. Wang, Y. Li, J. Wang, L. Chen, L. Zhang, H. Yu, T. Hou, ADMET evaluation in drug discovery. 12. Development of binary classification models for prediction of hERG potassium channel blockage, Mol. Pharm. 9 (2012) 996–1010.
[9] J. Burton, I. Ijjaali, O. Barberan, F. Petitet, D.P. Vercauteren, A. Michel, Recursive partitioning for the prediction of cytochromes P450 2D6 and 1A2 inhibition: importance of the quality of the dataset, J. Med. Chem. 49 (2006) 6231–6240.
[10] C.Y. Ma, S.Y. Yang, H. Zhang, M.L. Xiang, Q. Huang, Y.Q. Wei, Prediction models of human plasma protein binding rate and oral bioavailability derived by using GA-CG-SVM method, J. Pharmaceut. Biomed. 47 (2008) 677–682.
[11] Y. Li, Y. Wang, J. Ding, Y. Wang, Y. Chang, S. Zhang, In silico prediction of androgenic and nonandrogenic compounds using random forest, QSAR Comb. Sci. 28 (2009) 396–405.
[12] P. Garg, J. Verma, In silico prediction of blood brain barrier permeability: an artificial neural network model, J. Chem. Inf. Model. 46 (2006) 289–297.
[13] S. Vilar, L. Santana, E. Uriarte, Probabilistic neural network model for the in silico evaluation of anti-HIV activity and mechanism of action, J. Med. Chem. 49 (2006) 1118–1124.
[14] E. Deconinck, M.H. Zhang, D. Coomans, Y. Vander Heyden, Classification tree models for the prediction of blood–brain barrier passage of drugs, J. Chem. Inf. Model. 46 (2006) 1410–1419.
[15] B.E. Boser, I.M. Guyon, V.N. Vapnik, A Training Algorithm for Optimal Margin Classifiers, ACM, New York, 1992.
[16] V. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 2000.
[17] L. Zhong, C.Y. Ma, H. Zhang, L.J. Yang, H.L. Wan, Q.Q. Xie, L.L. Li, S.Y. Yang, A prediction model of substrates and non-substrates of breast cancer resistance

protein (BCRP) developed by GA-CG-SVM method, Comput. Biol. Med. 41 (2011) 1006–1013.
[18] Y. Xue, C.W. Yap, L.Z. Sun, Z.W. Cao, J. Wang, Y.Z. Chen, Prediction of P-glycoprotein substrates by a support vector machine approach, J. Chem. Inf. Model. 44 (2004) 1497–1505.
[19] S. Doniger, T. Hofmann, J. Yeh, Predicting CNS permeability of drug molecules: comparison of neural network and support vector machine algorithms, J. Comput. Biol. 9 (2002) 849–864.
[20] H. Li, C.W. Yap, C.Y. Ung, Y. Xue, Z.W. Cao, Y.Z. Chen, Effect of selection of molecular descriptors on the prediction of blood–brain barrier penetrating and nonpenetrating agents by statistical learning methods, J. Chem. Inf. Model. 45 (2005) 1376–1384.
[21] C. Xu, F. Cheng, L. Chen, Z. Du, W. Li, G. Liu, P.W. Lee, Y. Tang, In Silico prediction of chemical ames mutagenicity, J. Chem. Inf. Model. 52 (2012) 2840–2847.
[22] S.Y. Yang, Q. Huang, L.L. Li, C.Y. Ma, H. Zhang, R. Bai, Q.Z. Teng, M.L. Xiang, Y. Q. Wei, An integrated scheme for feature selection and parameter setting in the support vector machine modeling and its application to the prediction of pharmacokinetic properties of drugs, Artif. Intell. Med. 46 (2009) 155–163.
[23] C.B. Lucasius, G. Kateman, Understanding and using genetic algorithms Part 1. Concepts, properties and context, Chemometr. Intell. Lab. 19 (1993) 1–33.
[24] J.R. Shewchuk, An Introduction to the Conjugate Gradient Method Without the Agonizing Pain, Carnegie Mellon University, Pittsburgh, PA, 1994.
[25] X. Hu, A. Yan, In Silico models to discriminate compounds inducing and noninducing toxic myopathy, Mol. Inform. 31 (2012) 27–39.
[26] J.M. Kriegl, T. Arnhold, B. Beck, T. Fox, A support vector machine approach to classify human cytochrome P450 3A4 inhibitors, J. Comput. Aidmol. Des. 19 (2005) 189–201.
[27] H. Zhang, M.L. Xiang, C.Y. Ma, Q. Huang, W. Li, Y. Xie, Y.Q. Wei, S.Y. Yang, Three-class classification models of logS and logP derived by using GA-CG-SVM approach, Mol. Divers. 13 (2009) 261–268.
[28] M.W. Trotter, S.B. Holden, Support vector machines for ADME property classification, QSAR. Comb. Sci. 22 (2003) 533–548.
[29] C.H. Wu, G.H. Tzeng, Y.J. Goo, W.C. Fang, A real-valued genetic algorithm to optimize the parameters of support vector machine for predicting bankruptcy, Expert Syst. Appl. 32 (2007) 397–408.
[30] H. Frohlich, O. Chapelle, B. Scholkopf, Feature Selection for Support Vector Machines by Means of Genetic Algorithm, 15th IEEE, 2003.
[31] Q. Guo, W. Wu, F. Questier, D. Massart, C. Boucon, S. De Jong, Sequential projection pursuit using genetic algorithms for data mining of analytical data, Anal. Chem. 72 (2000) 2846–2855.
[32] S.D. Pickett, D.V. Green, D.L. Hunt, D.A. Pardoe, I. Hughes, Automated lead optimization of MMP-12 inhibitors using a genetic algorithm, ACS Med. Chem. Lett. 2 (2010) 28–33pp 2 (2010) 28–33.
[33] C.L. Huang, C.J. Wang, A GA-based feature selection and parameters optimization for support vector machines, Expert Syst. Appl. 31 (2006) 231–240.
[34] K.M. Leung, Naive Bayesian Classifier, Polytechnic University Department of Computer Science/Finance and Risk Engineering, 2007.
[35] X. Xia, E. Maliski, J. Cheetham, L. Poppe, Solubility prediction by recursive partitioning, Pharm. Res. 20 (2003) 1634–1640.
[36] E.F. Cook, L. Goldman, Empiric comparison of multivariate analytic techniques: advantages and disadvantages of recursive partitioning analysis, J. Chronic. Dis. 37 (1984) 721–731.
[37] N.G. Bauman, Ototoxic Drugs Exposed: The Shocking Truth About Prescription Drugs, Medications, Chemicals and Herbals That Can (and Do) Damage Our Ears, 3rd Ed., Integrity First Publications, Pennsylvania (PA), USA, 2010.
[38] D.S. Wishart, C. Knox, A.C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, M. Hassanali, DrugBank: a knowledgebase for drugs, drug actions and drug targets, Nucleic Acids Res. 36 (2008) D901–D906.
[39] TOXNET, Toxicology Data Network, ⟨http://toxnet.nlm.nih.gov/⟩ (accessed March 2014).
[40] M. Dramiński, A. Rada-Iglesias, S. Enroth, C. Wadelius, J. Koronacki, J. Komorowski, Monte Carlo feature selection for supervised classification, Bioinformatics 24 (2008) 110–117.
[41] R. Koggalage, S. Halgamuge, Reducing the number of training samples for fast support vector machine classification, Neural Inf. Process.-Lett. Rev. 2 (2004) 57–65.