

Prediction of drug-induced eosinophilia adverse effect by using SVM and naïve Bayesian approaches

Hui Zhang^{1,2} · Peng Yu¹ · Ming-Li Xiang² · Xi-Bo Li¹ · Wei-Bao Kong¹ · Jun-Yi Ma¹ · Jun-Long Wang¹ · Jin-Ping Zhang¹ · Ji Zhang^{1,3}

Received: 18 August 2014 / Accepted: 21 May 2015 / Published online: 5 June 2015
© International Federation for Medical and Biological Engineering 2015

Abstract Drug-induced eosinophilia is a potentially life-threatening adverse effect; clinical manifestations, eosinophilia–myalgia syndrome, mainly include severe skin eruption, fever, hematologic abnormalities, and organ system dysfunction. Using experimental methods to evaluate drug-induced eosinophilia is very complicated, time-consuming, and costly in the early stage of drug development. Thus, in this investigation, we established computational prediction models of drug-induced eosinophilia using SVM and naïve Bayesian approaches. For the SVM modeling, the overall prediction accuracy for the training set by means of fivefold cross-validation is 91.6 and for the external test set is 82.9 %. For the naïve Bayesian modeling, the overall prediction accuracy for the training set is 92.5 and for the external test set is 85.4 %. Moreover, some molecular descriptors and substructures considered as important for drug-induced eosinophilia were identified. Thus, we hope the prediction models of drug-induced eosinophilia built in this work should be applied to filter early-stage molecules

for potential eosinophilia adverse effect, and the selected molecular descriptors and substructures of toxic compounds should be taken into consideration in the design of new candidate drugs to help medicinal chemists rationally select the chemicals with the best prospects to be effective and safe.

Keywords Drug-induced eosinophilia · Support vector machine · Naïve Bayesian · Important features · Prediction

1 Introduction

Eosinophils are a type of leukocytes or white blood cells, which are part of the body's immune system components responsible for combating multicellular parasites and certain infections in vertebrates [5, 20]. The value of blood eosinophil above 600/cmm is defined as eosinophilia [23, 29], which was originally observed in patients treated with tryptophan-containing commercial products in the USA in 1898 [3, 15, 22]. Clinical manifestations, eosinophilia–myalgia syndrome (EMS), mainly include severe skin eruption, fever, hematologic abnormalities, and organ system dysfunction [1, 18, 24]. Presently, various factors have been found implicated as causes for eosinophilia, and exposure to drugs is considered as the most common causes, such as antipsychotic, antibacterial, antiviral, antithyroid, anticancer, and other medications [8, 11, 23]. Unfortunately, the drug-induced adverse effects are usually detected after the drug is introduced into the market or in phase III clinical trials. These experimental processes are very complicated, time-consuming, and costly [14, 34]. In particular, the drug-induced eosinophilia experimental evaluation processes would cause negative effect on human health, such as autoimmune

Electronic supplementary material The online version of this article (doi:10.1007/s11517-015-1321-8) contains supplementary material, which is available to authorized users.

✉ Hui Zhang
zhanghui123gansu@163.com

¹ College of Life Science, Northwest Normal University, Lanzhou 730070, Gansu, People's Republic of China

² State Key Laboratory of Biotherapy and Cancer Center, West China Hospital, West China Medical School, Sichuan University, Chengdu 610041, Sichuan, People's Republic of China

³ Bioactive Products Engineering Research Center for Gansu Distinctive Plants, Northwest Normal University, Lanzhou 730070, Gansu, People's Republic of China

diseases and end-organ failure, and even lead to mortality [8, 15]. Thus, using the cheaper, rapid, and accurate computational prediction methods as novel alternative techniques to evaluate the safety of candidate compounds prior to their synthesis would be a good choice. This may help medicinal chemists rationally select the chemicals with the best prospects to be effective and safe, and withdrawal of the suspected culprit chemicals in the early stage of drug development.

Presently, many academic institutions and pharmaceutical companies have realized the advantages of computational techniques and have been widely employed for the assessment of the pharmacokinetic properties and preclinical safety in the early stage of drug development [10, 13, 19, 21, 28]. Among these computational methods, the statistical and machine learning methods have been widely used in the prediction of adverse drug reactions (ADRs), and some of them have shown a good performance in the forecast of possible ADRs [6, 31, 32]. However, there were few reports of computational model of drug-induced eosinophilia. As far as we know, only González-Díaz et al. [8] constructed a computational prediction model of drug-induced eosinophilia using linear discriminant analysis (LDA) method. Thus, in this work, two statistical and machine learning methods, a modified method for support vector machine (SVM) [26] and the naïve Bayesian approach [2, 4], were considered to access drug-induced eosinophilia. For the modified SVM, the genetic algorithm (GA) is used for the feature selection [16], and conjugate gradient (CG) method is employed for the parameter optimization [12]. The naïve Bayesian classification model is a popular and mature machine learning method, which employs the versatile machine learning algorithms based on the Bayes' theorem and judges the plausibility of different candidate classes for a system [2, 4], and has been widely applied in the pharmaceutical industry [7, 17, 33].

The purpose of this investigation was to develop computer prediction models for drug-induced eosinophilia by using SVM and naïve Bayesian approaches, and identify some important molecular descriptors and substructures associated with compounds inducing eosinophilia. The generated prediction models will be validated by five-fold cross-validation and an external test set. We hope the established computational models should be employed for the prediction of drug-induced eosinophilia adverse effect in the early stage of drug development, and the molecular descriptors and substructures associated with drug-induced eosinophilia should be taken into consideration in the design of new candidate compounds to help medicinal chemists rationally select the chemicals with the best prospects to be effective and safe.

2 Materials and methods

2.1 Dataset collection

The biological activity and chemical structure of each of the compounds were extracted from the literature [8]. In this research, some compounds were deleted because of the Benzen was duplicate, the Nitroprusside is an inorganic compound, and the structures of Mustar vacilic and Nafaline were not found. Finally, the remaining 148 compounds were applied in this investigation. In order to compare with previous study, the same training set (107 agents) and test set (41 compounds) as those used in the literature [8] were applied. The structures of the training set (TrainingSet_107.sd) and test set (TestSet_41.sd) molecules are listed in the Supplementary Data.

2.2 Support vector machines (SVM)

The optimized SVM method, namely GA-CG-SVM, is a modified SVM modeling approach. Detailed description of the proposed GA-CG-SVM method can be found in our previous paper [30–32]. Here, we just make a short summary to the basic idea of SVM and GA-CG-SVM.

In SVM, each object is described by a vector x_i , and the class index is represented by the y_i . In linearly separable cases, two different classes of feature vectors can be correctly classified by

$$w \times x_i + b \geq +1, \quad \text{for } y_i = +1 \quad (1)$$

$$w \times x_i + b \leq -1, \quad \text{for } y_i = -1 \quad (2)$$

Here, w is a vector normal to the hyperplane, and b is a scalar quantity. The SVM attempts to find an optimal separating hyperplane with the maximum margin by solving the following optimization problem:

$$\text{Max}_{w,b} \frac{2}{\|w\|} \quad \text{Subject to } y_i(w \times x_i + b) - 1 \geq 0 \quad (3)$$

However, in the linearly non-separable cases, no hyperplane can be used to perfectly separate two sets of points. In this case, the nonnegative slack variables $\xi_i \geq 0$, $i = 1, \dots, m$. could be introduced. Such that

$$w \times x_i + b \geq +1 - \xi_i, \quad \text{for } y_i = +1 \quad (4)$$

$$w \times x_i + b \leq -1 + \xi_i, \quad \text{for } y_i = -1 \quad (5)$$

In order to find a hyperplane that provides the minimum number of training errors, the equation to be solved becomes:

$$\text{Max}_{w,b} \frac{2}{\|w\|} + C \sum_{i=1}^m \xi_i \quad \text{Subject to } y_i(w \times x_i + b) - 1 + \xi_i \geq 0 \quad (6)$$

Here, C is the penalty parameter, which should be predetermined by user.

The nonlinear (non-)separable cases could be easily transferred to linear cases through projecting the input variable into a new high-dimensional feature space by using a kernel function $K(x_i, x_j)$. Such as the radial basis function (RBF), which is the most widely used kernel function, it performed very well in most cases.

$$k(x_i, x_j) = \exp\left(-\gamma\|x_i - x_j\|^2\right) \quad (7)$$

The γ is a parameter which should be specified by user in advance.

An optimal C and γ can significantly improve the accuracy of SVM classification. Furthermore, the feature selection and parameter setting (C, γ) influence each other in SVM modeling. Thus, the combined scheme was used to handle the two problems: a genetic algorithm (GA)-based method is used for the feature selection, and a conjugate gradient (CG) method is used for the (C, γ) parameter optimizations.

2.3 Modeling details by GA-CG-SVM

All the structures of the prepared compounds were generated, and then, geometrical optimization of these compounds was calculated by using Accelrys Discovery Studio program package (Accelrys, San Diego, CA). The optimized 3D structure of each compound was manually inspected to ensure that each molecule was properly represented and is consistent with the one [8]. Molecular descriptors were calculated by using the online program PCLIENT [27].

The initial features were preprocessed whose purpose is to eliminate the redundancy and overlapping of the descriptors. Here, the following descriptors were removed: (1) descriptors with too many zero values, (2) descriptors with very small standard deviation values (<0.5 %), and (3) descriptors which are highly correlated with others (correlation coefficients >95 %). After the preprocessing, the descriptor values were scaled to a range of -1 to $+1$, which

is necessary since the different ranges of descriptor values will influence the quality of the SVM model generated.

2.3.1 Construction of the GA-CG-SVM model of drug-induced eosinophilia

A total of 107 compounds, including 71 toxic compounds and 36 non-toxic agents, were used as training set to train the SVM classification model of drug-induced eosinophilia. The following various molecule properties were initially calculated: 48 constitutional descriptors, 21 topological charge indices descriptors, 99 WHIM descriptors, 154 functional group counts descriptors, 119 topological descriptors, 150 RDF descriptors, 74 geometrical descriptors, and 31 molecular descriptors. These descriptors were firstly preprocessed for removing those redundant and unrelated properties. 186 molecular descriptors were selected and were subjected to being further reduced by using GA-CG method. Finally, eight molecular descriptors were selected (Table 1), and the optimized parameters (C, γ) are (4435.096191, 0.025620).

2.4 Naïve Bayesian model

The introduction of naive Bayes classification theory has been described in the literature [4, 26]. The naïve Bayesian classification approach is a popular and mature machine learning method, which could distinguish between compounds that are positives and those that are negatives with using molecular descriptors. In this investigation, the naïve Bayesian model was developed by using Discovery Studio (DS) version 3.1 (Accelrys Inc., San Diego, CA). The default physical property descriptors were used, including ALogP, molecular weight, number of H donors, number of H acceptors, number of rings, number of aromatic rings, and molecular fractional polar surface area. The cross-validation method of the training set was set to 5. The “Model Domain Fingerprint” was chosen as ECFP-6 [extended connectivity fingerprints, with a diameter of 6, were generated in Pipeline Pilot (SciTegic, Inc.)], because it could give the highest ROC curve. The ROC curve charts

Table 1 Molecular descriptors used in the SVM modeling for the prediction model of drug-induced eosinophilia adverse effect

Descriptor	Explanation
ZM2V	Second Zagreb index by valence vertex degrees
RDF015m	Radial distribution function—1.5/weighted by atomic masses
RDF035m	Radial distribution function—3.5/weighted by atomic masses
L3u	Third component size directional WHIM index/unweighted
E2u	Second component accessibility directional WHIM index/unweighted
E3s	Third component accessibility directional WHIM index/weighted by atomic Electropotential states
nHDon	Number of donor atoms for H bonds (with N and O)
BLTF96	Verhaar model of Fish baseline toxicity from MLOGP (mmol/l)

the true-positive rate (sensitivity) versus the false-positive rate (100 % specificity). Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. The other parameters were kept at their default values.

2.5 Statistical analysis

The predictive performances of statistical and machine learning models were assessed by overall prediction accuracy (Q); sensitivity (SE), the prediction accuracy for positive compounds; and specificity (SP), the prediction accuracy for negative compounds.

$$Q = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

$$SE = \frac{TP}{TP + FN} \quad (9)$$

$$SP = \frac{TN}{TN + FP} \quad (10)$$

where TP, true positives, is the number of positive instances which are correctly identified; TN, true negatives, is the number of negative instances which are correctly recognized; FP, false positives, is the number of the negative instances which are wrongly predicted as positives; FN, false negatives, is the number of positive instances which are wrongly predicted as negatives.

3 Results

3.1 SVM classification model of drug-induced eosinophilia

In this research, the fivefold cross-validation was employed for the training set to evaluate the stability and capacity of the established SVM model, and an external test set with 41 unique drugs was used to further assess the model's predictive power. For the training set, the overall prediction accuracy (Q , Table 2) is 91.6 %. Among these 71 toxic compounds, 67 agents were correctly predicted. The sensitivity (SE, Table 2) is 94.4 %. Of these 36 non-toxic agents, 31 compounds were correctly identified. The specificity (SP, Table 2) is 86.1 %. In order to evaluate whether the established SVM model could successfully recognize the external series as toxic agents and non-toxic agents, the external test set containing 41 compounds was applied. Table 3 shows the prediction results of the test set; of these 41 compounds, 34 were correctly classified. The overall prediction accuracy (Q , Table 3) for the test set is 82.9 %. For these 24 toxic compounds, 20 agents were correctly recognized. The sensitivity (SE, Table 2) is 83.3 %. For these 17 non-toxic compounds, 14 agents were correctly forecasted. The specificity (SP, Table 3) is 82.4 %. These results indicate the established SVM prediction model of drug-induced eosinophilia could successfully discriminate these agents as positives (toxic compounds) or negatives (non-toxic compounds).

Table 2 Fivefold cross-validation results of SVM and Bayesian models for the training set

Model name	Positives			Negatives			Q (%)
	TP	FN	SE (%)	TN	FP	SP (%)	
GA-CG-SVM	67	4	94.4	31	5	86.1	91.6
Bayesian model (descriptors + ECFP-6)	63	8	88.7	36	0	100	92.5
Bayesian model (simple descriptors)	60	11	84.5	31	5	86.1	85.0

TP True positive, *TN* true negative, *FP* false positive, *FN* false negative

SE (%): sensitivity, $SE = TP/(TP + FN)$; SP (%): specificity, $SP = TN/(TN + FP)$; Q (%): overall accuracy, $Q = (TP + TN)/(TP + TN + FP + FN)$

Table 3 Prediction results of the external test set

Model name	Positives			Negatives			Q (%)
	TP	FN	SE (%)	TN	FP	SP (%)	
GA-CG-SVM	20	4	83.3	14	3	82.4	82.9
Bayesian model (descriptors + ECFP-6)	18	6	75.0	17	0	100	85.4
Bayesian model (simple descriptors)	19	5	79.2	14	3	82.4	80.5

TP True positive, *TN* true negative, *FP* false positive, *FN* false negative

SE (%): sensitivity, $SE = TP/(TP + FN)$; SP (%): specificity, $SP = TN/(TN + FP)$; Q (%): overall accuracy, $Q = (TP + TN)/(TP + TN + FP + FN)$

3.2 The naïve Bayesian classification model of drug-induced eosinophilia

The Bayesian prediction model of drug-induced eosinophilia based on the same training set was successfully developed, in which the default physical property descriptors together with the extended connectivity fingerprint descriptor (ECFP-6) were applied (Bayesian model: descriptors + ECFP-6). The established naïve Bayesian prediction model was also evaluated by fivefold cross-validation method and an external test set. The best cutoff for this model is 0.014. The area under the ROC curve (AUC) is the ROC score, which is widely used as measure of a model discriminatory power. The maximum value for the ROC score of 1 indicates the model has a perfect prediction performance (100 % true-positive (TP) rate, and 0 % false-positive (FP) rate). The ROC score of 0.5 represents the model has no discriminative ability (i.e., 50 % true-positive (TP) rate and 50 % false-positive (FP) rate). In this work, the ROC score for the fivefold cross-validation in the training set is 0.858, which represents the established model has a good predictive power.

The fivefold cross-validation results of the training set for the model (Bayesian model: descriptors + ECFP-6) are given in Table 2. From Table 2, we can see that the prediction accuracy for these toxic compounds (SE, Table 2) is 88.7 %. For these non-toxic compounds, the specificity (SP, Table 2) is 100 %. The overall prediction accuracy (Q , Table 2) for the training set is 92.5 %. For the external test set, the ROC score is 0.973. The detail information of prediction results is shown in Table 3. As shown in Table 3, the total predictability (Q , Table 3) is 85.4 %. The model recognizes as toxic (SE, Table 3) 75.0 % of these compounds, that is, 18 chemicals out of 24. Moreover, the model correctly classifies 100 % of the non-toxic chemicals (SP, Table 3), that is, 17 agents out of 17. These results indicate the established naïve Bayesian prediction model (Bayesian model: descriptors + ECFP-6) of drug-induced eosinophilia could successfully recognize internal/external agents as positives or negatives. Furthermore, the other naïve Bayesian model based on the default physical property descriptors was established (Bayesian model: simple descriptors), in which the ECFP-6 fingerprint descriptor was removed. As shown in Tables 2 and 3, the prediction performance of the model (Bayesian model: simple descriptors) was significantly decreased, especially for the predictive capability for non-toxic compounds. The prediction accuracy for the training set and for the test set is 85.0 and 80.5 %, respectively. Figure 1 shows some fragments produced by the ECFP-6 descriptors. The Bayesian score is a measure of how different this is from the hit rate as a whole (the ratio that would be expected if the feature was occurring randomly across the toxic agents and non-toxic

agents), which represents the final contribution of a feature to the model prediction. The top 20 toxic/non-toxic fragments are listed in Fig. 1. The results suggested that combined with these fragments could significantly increase the overall accuracy of drug-induced eosinophilia prediction.

4 Discussion

In this investigation, the prediction models of drug-induced eosinophilia have been successfully developed by using the optimal SVM and naïve Bayesian approaches. For the SVM modeling, the overall prediction accuracy for the training set and for the test set is 91.6 and 82.9 %, respectively. For the naïve Bayesian modeling, the overall prediction accuracy for the training set and for the external test set is 92.5 and 85.4 %, respectively. All of these indicate the constructed SVM and naïve Bayesian models are suitable for predicting the drug-induced eosinophilia adverse effect and could be used as tools for screening compounds with eosinophilia adverse effect and reducing late-stage attrition rates in drug development process.

4.1 Molecular features important for drug-induced eosinophilia

The pathogenesis of drug-induced eosinophilia is very complex, and different mechanisms have been implicated in its development [9, 25]. Thus, investigation of important molecular descriptors of these compounds inducing eosinophilia is very necessary. Using simple natural descriptors depicting chemical–physical properties of chemical agents to establish the relationship between chemical agents and their bioactivities is an advantage of the statistical and machine learning methods, such as SVM and naïve Bayesian used here. In this research, the GA-CG method was used to select some important descriptors for drug-induced eosinophilia. Eight kinds of molecular descriptors, including 696 descriptors, were initially calculated. After those redundant and unrelated properties removed, 186 descriptors were obtained. Finally, eight important molecular descriptors were successfully selected from the 186 descriptors. Table 1 lists the selected descriptors and their definitions. From the results of this work, it can be seen that the GA-CG selected molecular descriptors are powerful to discriminate compounds causing and not causing eosinophilia. These selected descriptors can be roughly grouped into several categories: hydrogen-bonding descriptors (nHDon), molecular electronic property-related descriptors (E3s), molecular structural information-related descriptors (ZM2 V, L3u, E2u), lipophilicity-related descriptors (BLTF96), and molecular weight-related descriptors (RDF015 m, RDF035 m).

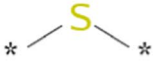
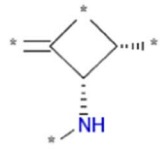
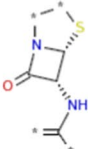

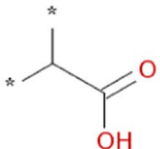
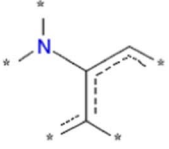
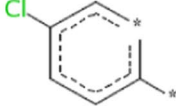
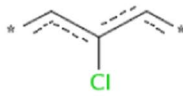

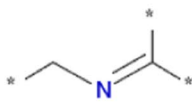
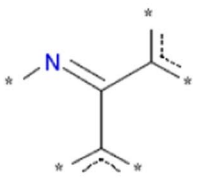
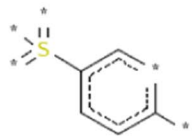
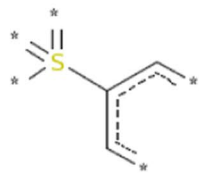
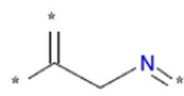
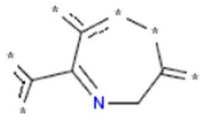
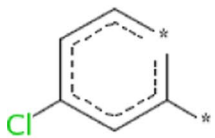
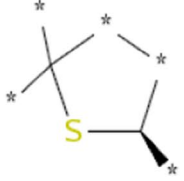
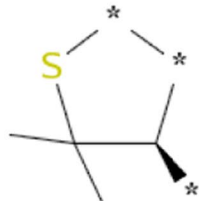
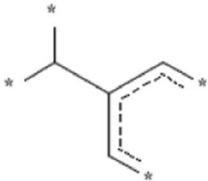
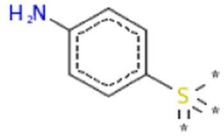
 <p>G1: 912478223 21 out of 21 toxic Bayesian Score: 0.240</p>	 <p>G2: -1567199489 16 out of 16 toxic Bayesian Score: 0.236</p>	 <p>G3: -655103622 14 out of 14 toxic Bayesian Score: 0.234</p>	 <p>G4: 2056644143 14 out of 14 toxic Bayesian Score: 0.234</p>	 <p>G5: -81842545 13 out of 13 toxic Bayesian Score: 0.233</p>
 <p>G6: -1236953626 13 out of 13 toxic Bayesian Score: 0.233</p>	 <p>G7: 1854732111 12 out of 12 toxic Bayesian Score: 0.231</p>	 <p>G8: -176494269 12 out of 12 toxic Bayesian Score: 0.231</p>	 <p>G9: 859433814 11 out of 11 toxic Bayesian Score: 0.229</p>	 <p>G10: 2090054846 11 out of 11 toxic Bayesian Score: 0.229</p>
 <p>G11: 1481235578 11 out of 11 toxic Bayesian Score: 0.229</p>	 <p>G12: -978131182 10 out of 10 toxic Bayesian Score: 0.227</p>	 <p>G13: -177264675 10 out of 10 toxic Bayesian Score: 0.227</p>	 <p>G14: -122376699 10 out of 10 toxic Bayesian Score: 0.227</p>	 <p>G15: 490215350 10 out of 10 toxic Bayesian Score: 0.227</p>
 <p>G16: 577592657 10 out of 10 toxic Bayesian Score: 0.227</p>	 <p>G17: -1806159325 10 out of 10 toxic Bayesian Score: 0.227</p>	 <p>G18: -193898895 9 out of 9 toxic Bayesian Score: 0.225</p>	 <p>G19: -176846085 8 out of 8 toxic Bayesian Score: 0.222</p>	 <p>G20: 431976397 8 out of 8 toxic Bayesian Score: 0.22</p>

Fig. 1 a ECFP-6 descriptors: some substructures that are important for drug-induced eosinophilia. *Each panel* shows the naming convention for each fragment, the numbers of molecules it is present in that are toxic agents, and the Bayesian score for the fragment. **b** ECFP-6

descriptors: some substructures that are absent from drug-induced eosinophilia compounds. *Each panel* shows the naming convention for each fragment, the numbers of molecules it is present in that are toxic compounds, and the Bayesian score for the fragment

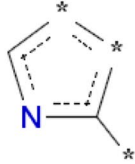
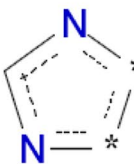
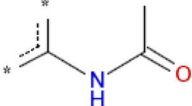
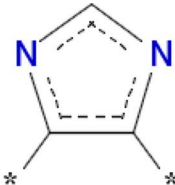
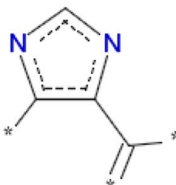
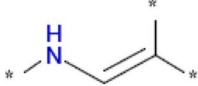
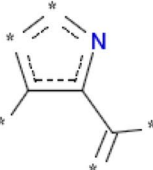
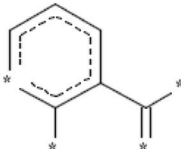
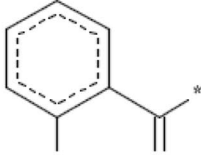
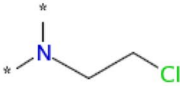



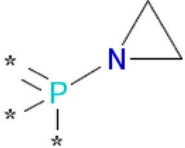
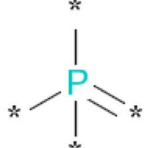
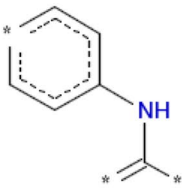
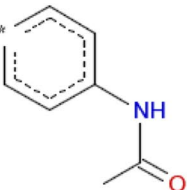
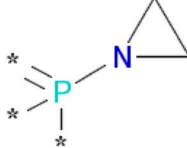
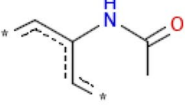
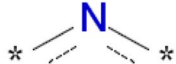
 <p>B1: -953984246 0 out of 4 toxic Bayesian Score: -1.412</p>	 <p>B2: -1021054081 0 out of 3 toxic Bayesian Score: -1.203</p>	 <p>B3: -1923054811 0 out of 3 toxic Bayesian Score: -1.203</p>	 <p>B4: 488354296 0 out of 3 toxic Bayesian Score: -1.203</p>	 <p>B5: -1464256689 0 out of 2 toxic Bayesian Score: -0.937</p>
 <p>B6: 470025226 0 out of 2 toxic Bayesian Score: -0.937</p>	 <p>B7: -782828288 0 out of 2 toxic Bayesian Score: -0.937</p>	 <p>B8: 1635415905 0 out of 2 toxic Bayesian Score: -0.937</p>	 <p>B9: 1993671714 0 out of 2 toxic Bayesian Score: -0.937</p>	 <p>B10: 692743133 0 out of 2 toxic Bayesian Score: -0.937</p>
 <p>B11: -1791034651 0 out of 2 toxic Bayesian Score: -0.937</p>	 <p>B12: -1790451017 0 out of 2 toxic Bayesian Score: -0.937</p>	 <p>B13: 103339584 0 out of 2 toxic Bayesian Score: -0.937</p>	 <p>B14: 1017218444 0 out of 2 toxic Bayesian Score: -0.937</p>	 <p>B15: -826638028 0 out of 2 toxic Bayesian Score: -0.937</p>
 <p>B16: 738938915 0 out of 2 toxic Bayesian Score: -0.937</p>	 <p>B17: 1776488 0 out of 2 toxic Bayesian Score: -0.937</p>	 <p>B18: -1104081458 0 out of 2 toxic Bayesian Score: -0.937</p>	 <p>B19: -742804400 0 out of 2 toxic Bayesian Score: -0.937</p>	 <p>B20: -152683720 1 out of 5 toxic Bayesian Score: -0.892</p>

Fig. 1 continued

4.2 Analysis of the toxic/non-toxic fragments produced by the ECFP-6 fingerprint descriptors

The molecular features considered as important for drug-induced eosinophilia have been identified by the GA-CG method. In order to better understand the structures of compounds inducing and not inducing eosinophilia, the ECFP-6 fingerprint descriptors were applied in the naïve Bayesian model to produce some substructures of toxic compounds and non-toxic compounds. Figure 1 shows some toxic fragments and non-toxic fragments generated by the ECFP-6 fingerprints. As shown in Fig. 1, some substructures that contribute to toxic compounds (Fig. 1a) and those that are not inducing eosinophilia (Fig. 1b) were identified. Figure 1a shows some substructures associated with toxic compounds, and a compound having any of these fragments was considered as a toxic agent, each panel represents the naming convention for each fragment, the numbers of molecules it is present in that are toxic agents, and the Bayesian score for the fragment. The Bayesian score takes account of the total number of occurrences of the feature, ensuring more weight is placed on features that occur more often and little weight on those for which there are very few occurrences. In further analysis of these fragments generated in toxic compounds and non-toxic compounds, we found that some fragments only appeared in compounds inducing eosinophilia, such as dimethylsulfane (G1), N-methylcyclobutanamine (G2), chlorobenzene (G7, G16), N-methylenemethanamine (G10, G11, G14, G15), and tetrahydrothiophene (G17, G18). Thus, these substructures of toxic compounds identified in this research might be associated with the drug-induced eosinophilia adverse effect and should be taken into consideration in the design of new candidate drugs to help medicinal chemists rationally select the chemicals with the best prospects to be effective and safe.

4.3 Comparison with previous prediction model of drug-induced eosinophilia

Presently, although a number of prediction models of the pharmacokinetic properties and toxicity have been developed and used in drug development, there were few reports of computational model for drug-induced eosinophilia. Only González-Díaz et al. [8] built a prediction model of drug-induced eosinophilia using linear discriminant analysis (LDA) method, which gave a good classification of 91.82 % for the training series and 88.1 % for the external validation series. In this study, the GA-CG-SVM gives 91.6 % for the training set and 82.9 % for the test set. The naïve Bayesian model could correctly classify 92.5 % of training set compounds and 85.4 % of test set agents. Prediction accuracies of the GA-CG-SVM model and naïve

Bayesian model established in this work are comparable to those of the LDA model built by González-Díaz et al. [8]. However, the GA-CG-SVM model could select some critical molecular descriptors for drug-induced eosinophilia, and the naïve Bayesian model could give some fragments that contribute to eosinophilia inductors and those that are not.

5 Conclusions

In this investigation, the prediction models of drug-induced eosinophilia adverse effect have been successfully developed by using SVM and naïve Bayesian approaches. A set of 107 compounds were used as the training set, and 41 agents were applied as the external test set. For the SVM modeling, the overall prediction accuracy for the training set by means of fivefold cross-validation is 91.6 and for the external test set is 82.9 %. For the naïve Bayesian modeling, the overall prediction accuracy for the training set and for the external test set is 92.5 and 85.4 %, respectively. Moreover, some molecular descriptors and substructures associated with the toxicity of eosinophilia compounds were identified. Thus, we hope the prediction models of drug-induced eosinophilia built in this work could be applied to filter early-stage molecules for this potential eosinophilia adverse effect. And the selected molecular descriptors and substructures of toxic compounds should be taken into consideration in the design of new candidate drugs and finally reduce attrition rate in later stages of drug development.

Acknowledgments This work was supported by the Project for Enhancing the Research Capability of Young Teachers in Northwest Normal University (NWNLU-LKQN-12-7).

Conflict of interest The authors declare that there are no conflicts of interest.

References

1. Allen JA, Varga J (2014) Encyclopedia of toxicology, 3rd edition from Philip Wexler. Elsevier, New York
2. Berger JO (1985) Statistical decision theory and Bayesian analysis. Springer, New York
3. Blackburn WD (1997) Eosinophilia myalgia syndrome. *Semin Arthritis Rheum* 26:788–793
4. Box GEP, Tiao GC (1973) Bayesian inference in statistical analysis. Addison-Wesley, Reading
5. Dent G, Loweth SC, Hasan AM, Leslie FM (2014) Synergic production of neutrophil chemotactic activity by colonic epithelial cells and eosinophils. *Immunobiology* 219:793–797
6. Ekins S (2014) Progress in computational toxicology. *J Pharm Toxicol Methods* 69:115–140

7. Ekins S, Williams AJ, Xu JJ (2010) A predictive ligand-based Bayesian model for human drug-induced liver injury. *Drug Metab Dispos* 38:2302–2308
8. González-Díaz H, Tenorio E, Castañedo N, Santana L, Uriarte E (2005) 3D QSAR Markov model for drug-induced eosinophilia—theoretical prediction and preliminary experimental assay of the antimicrobial drug G1. *Bioorg Med Chem* 13:1523–1530
9. Gotlib J (2005) Molecular classification and pathogenesis of eosinophilic disorders. *Acta Haematol* 114:7–25
10. Grime KH, Barton P, McGinnity DF (2013) Application of in silico, in vitro and preclinical pharmacokinetic data for the effective and efficient prediction of human pharmacokinetics. *Mol Pharmaceutics* 10:1191–1206
11. Hardman JG, Limbird LE, Gilman AG (1996) Goodman and Gilman's the pharmacological basis of therapeutics. McGraw-Hill, New York
12. Keerthi S, Sindhvani V, Chapelle O (2007) An efficient method for gradient-based adaptation of hyperparameters in SVM models. In: Schölkopf B, Platt J, Hofmann T (eds) *Advances in neural information processing systems ~20 (NIPS ~2006)*, Vancouver, Canada
13. Kimber I, Humphris C, Westmoreland C, Alepee N, Dal Negro G, Manou I (2011) Computational chemistry, systems biology and toxicology. *Harnessing the chemistry of life: revolutionizing toxicology. A commentary. J Appl Toxicol* 31:206–209
14. Li AP (2011) Drug discovery and development—present and future. In: Kapetanović I (ed) *Critical human hepatocyte-based in vitro assays for the evaluation of adverse drug effects*. InTech, USA
15. Lindgren CE, Walker LA, Bolton P (1991) L-tryptophan induced eosinophilia–myalgia syndrome. *J R Soc Health* 111:29–30
16. Lucasius CB, Kateman G (1993) Understanding and using genetic algorithms. Part 1. Concepts, properties and context. *Chemometr Intell Lab* 19:1–33
17. Magni P, Bellazzi R, Nauti A, Patrini C, Rindi G (2001) Compartmental model identification based on an empirical Bayesian approach: the case of thiamine kinetics in rats. *Med Biol Eng Comput* 39:700–706
18. Milaraa J, Martinez-Losac M, Sanzd C, Almuđervec P, Peiróc T, Serranoc A, Morcilloe EJ, Zaragozág C, Cortijoa J (2013) Bafetinib inhibits functional responses of human eosinophils in vitro. *Eur J Pharmacol* 715:172–180
19. Modi S, Hughes M, Garrow A, White A (2012) The value of in silico chemistry in the safety assessment of chemicals in the consumer goods and pharmaceutical industries. *Drug Discov Today* 17:135–142
20. Pereira MC, Oliveira DT, Kowalski LP (2011) The role of eosinophils and eosinophil cationic protein in oral cancer: a review. *Arch Oral Biol* 56:353–358
21. Selick HE, Beresford AP, Tarbit MH (2002) The emerging importance of predictive ADME simulation in drug discovery. *Drug Discov Today* 7:109–116
22. Sidransky H, Verney E, Cosgrove JW, Latham PS, Mayeno AN (1994) Studies with 1,1'-ethylidenebis(tryptophan), a contaminant associated with L-tryptophan implicated in the eosinophilia–myalgia syndrome. *Toxicol Appl Pharmacol* 126:108–113
23. Singh V, Gomez VV, Swamy SG, Vikas B (2009) Approach to a case of eosinophilia. *Ind J Aerospace Med* 53:58–64
24. Tefferi A (2005) Blood eosinophilia: a new paradigm in disease classification, diagnosis, and treatment. *Mayo Clinic Proc* 80:75–83
25. Valent P, Gleich GJ, Reiter A, Roufosse F, Weller PF, Hellmann A, Metzgeroth G, Leiferman KM, Arock M, Sotlar K, Butterfield JH, Cerny-Reiterer S, Mayerhofer M, Vandenberghe P, Haferlach T, Bochner BS, Gotlib J, Horny HP, Simon HU, Klion AD (2012) Pathogenesis and classification of eosinophil disorders: a review of recent developments in the field. *Expert Rev Hematol* 5:157–176
26. Vapnik V (1998) *Statistical learning theory*. Wiley, New York
27. VCCLAB (2005) Virtual computational chemistry laboratory. Available at : <http://www.vcclab.org>
28. Vedani A, Smiesko M (2009) In silico toxicology in drug discovery—concepts based on three-dimensional models. *Altern Lab Anim* 37:477–496
29. Weller PF (1991) The immunobiology of eosinophils. *N Engl J Med* 324:1110–1118
30. Yang SY, Huang Q, Li LL, Ma CY, Zhang H, Bai R, Teng QZ, Xiang ML, Wei YQ (2009) 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* 46:155–163
31. Zhang H, Chen QY, Xiang ML, Ma CY, Huang Q, Yang SY (2009) In silico prediction of mitochondrial toxicity by using GA-CG-SVM approach. *Toxicol In Vitro* 23:134–140
32. Zhang H, Li W, Xie Y, Wang WJ, Li LL, Yang SY (2011) Rapid and accurate assessment of seizure liability of drugs by using an optimal support vector machine method. *Toxicol In Vitro* 25:1848–1854
33. Zientek M, Stoner C, Ayscue R, Klug-McLeod J, Jiang Y, West M, Collins C, Ekins S (2010) Integrated in silico-in vitro strategy for addressing cytochrome P450 3A4 time-dependent inhibition. *Chem Res Toxicol* 23:664–676
34. Zurlo J, Rudacille D, Goldberg AM (1994) *Animals and alternatives in testing: history: science and ethics*. Mary Ann Liebert, New York