Prediction of hepatotoxicity of traditional Chinese medicine compounds by support vector machine approach

Ludi Jiang, Yusu He, Yanling Zhang*
School of Chinese Pharmacy
Beijing University of Chinese Medicine
Beijing, 100102, China

***Abstract*—In this study, based on literatures and web databases, 490 hepatotoxic compounds and 598 non-hepatotoxic compounds were selected as a data set for hepatotoxicity discriminative model generation. 1664 molecular descriptors, including physicochemical, charge distribution and geometrical descriptors, were calculated to characterize the molecular structure of liver toxic compounds. The combination of CfsSubsetEval valuation and BestFirst searching was used to choose molecular descriptors for model construction. With the help of support vector machine (SVM), a discriminative model with high accuracy was built. Meanwhile, the accuracy, sensitivity and specificity of this model were all above 80%. Besides, 23 traditional Chinese medicine compounds with hepatotoxicity were regarded as external validation, so as to further verify the model accuracy. Then, the present model was utilized to identify hepatotoxic compounds in Qingkailing injection. The results demonstrated that present study provides a reliable utility for the hepatotoxic compounds prediction in Chinese Medicinal Materials studies.**

Keywords—Support Vector Machine; hepatotoxicity; traditional Chinese medicine

I. INTRODUCTION

The liver, which has abundance of metabolizing enzymes, is the primary port of entry for ingested drugs [1]. Therefore, drugs can adversely affect the structure and functions of the liver. Drug-induced liver injury most frequently results in medication withdrawal from the market. Traditional Chinese medicine, consisted of complicated constituents, can exert therapeutic action, but cause adverse reactions at the same time. Because of hepatotoxicity, some traditional Chinese medicine, such as Zhuangguguanjie Pill, Baishi Pill and Zhixue Capsule, were notified by China's National Center for ADR Monitoring. Qingkailing injection, which is widely used in clinical, also resulted in hepatotoxicity [2, 3]. But the liver toxic compounds of this drug have not been found yet.

Computational toxicology, which studies the relationship between compound structure and toxicity, is widely applied in the toxicology field [4].For instance, quantitative structure-activity relationship (QSAR) [5], Bayesian [6], K-Nearest Neighbor (kNN) [7] and Support Vector Machine (SVM) [8] were used to predict drug toxicity. This type of methods can mitigate the time-consuming and high-cost problem, which caused by traditional toxicological experiments. In this study, a hepatotoxicity discriminative model was built by SVM. The purpose of our work is to extend the application sphere of this model and improve the prediction accuracy of hepatotoxic compounds screening, by using a training set containing diverse compounds. Moreover, this model was also applied to screen hepatotoxic compounds from Qingkailing injection.

This work was supported by the National Natural Science Foundation of China (No. 81173522). *Correspondence author: Y. Zhang (collean_zhang@163.com)

II. MATERIALS AND METHODS

A. Training and test set splitting

A universal set was constructed from 776 hepatotoxic compounds and 1892 non-hepatotoxic compounds, most of which have been reported by Nigel Greene and Chin Yee Liew [9, 10] (Table1).

By considering the diverse sources of the universal set, repetitive compounds between and within two groups should be removed; and compared with the number of hepatotoxic compounds, redundant non-hepatotoxic compounds were discarded. After that, a data set was obtained, which was comprised of 490 hepatotoxic compounds and 598 non-hepatotoxic compounds. In order to ensure the compounds of the training set had relatively good representative, training set and test set were respectively extracted from the data set, by using Kennard-Stone (KS) algorithm [11]. Thus, 872 compounds were chosen as training set, which contained 436 hepatotoxic compounds and 436 non-hepatotoxic compounds. Besides, 216 compounds were regarded as test set, which was comprised of 54 hepatotoxic compounds and 162 non-hepatotoxic compounds.

TABLE1. SOURCE DETAILS OF THE UNIVERSAL SET

Categories	**Source(literatures/database)**	**No.**	**To tal**
Hepatotoxic compounds (positive group)	①Compounds derived from "HH"(evidence of human hepatotoxicity)in literature [9]	181	77 6

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

	②Compounds derived from "training-positive" in literature [10]	654	
	③Search Drugbank (http://www.drugbank.ca/)with "hepatotoxicity", "liver toxicity"	59	
Non-hepatotoxic compounds (negative group)	①Compounds derived from "NH"(no evidence of hepatotoxicity in any species)in literature [9]	90	1892
	②Compounds derived from "training-negative" in literature [10]	433	
	③ Remove hepatotoxic compounds from "approved" list in Drugbank, and reserve the rest of compounds	1369	

B. Molecular descriptors

According to computational toxicology, molecular structure should be translated into molecular descriptors, which are used to describe different structural characteristics of molecule [12, 13]. In this study, 1664 molecular descriptors, which can be summarized as twenty different types of descriptors, such as Constitutional, Topological, Information Index, Connectivity Index, Topological Charge Index, and so on, were computed by E-Dragon [12, 14]. Because of the redundant features might be calculated in the molecular descriptors, data preprocessing should be applied: the molecular descriptor should be scrubbed, of which relative variance is less than 0.05 or equal values are more than 90% [15]. BestFirst and CfsSubsetEval [16], which are the two algorithms included in WEKA (Version 3.6.10) program package, were used for the selection of an optimal subset of preprocessed molecular descriptors for model construction. CfsSubsetEval evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. BestFirst searches the space of attribute subsets by greedy hillclimbing, which is augmented with a backtracking facility. Therefore, features which were highly correlated with the classification were chosen as the best subset.

C. Development of SVM

SVM (Support Vector Machines) is an important machine learning method, and is also a powerful tool for pattern recognition. In this study, Libsvm-Faruto-Ultimate (Version2.0), programmed by Faruto, was used to run the SVM algorithm [17]. While using SVM to solve factual classification problem, proper kernel function and relevant parameters should be selected. There are several kernel functions, such as Linear Kernel, Radial Basis Function Kernel (RBF Kernel), Polynomial Kernel and so on. In this paper, RBF Kernel was chosen as the kernel function of SVM. The RBF kernel can handle the case when the relation between class labels and attributes is nonlinear by mapping samples into a higher dimensional space nonlinearly. Besides, there are two parameters in RBF kernel, namely C and γ, which affect the precision of SVM classifier significantly. Thus, parallel grid search and 10-fold cross-validation were used to identify appropriate (C, γ), so as to make sure the classifier could predict test set accurately.

D. Validation of the SVM model

A test set was utilized to evaluate all the hepatotoxicity discriminative models. The evaluation indicators were presented as follow: accuracy (ACC), sensitivity (SE), and specificity (SP). Computational formulas of these indicators were shown in Equation (1-3).

$$Accuracy = (TN + TP) / (TP + FN + TN + FP) \quad (1)$$

$$Sensitivity = TP / (TP + FN) \quad (2)$$

$$Specificity = TN / (TN + FP) \quad (3)$$

TABLE 2. 23 HEPATOTOXIC COMPOUNDS OF TRADITIONAL CHINESE MEDICINE

Name	CAS	Source plant
Oxymatrine	16837-52-8	*Sophoraflavescens*
Senecionine	130-01-8	*Seneciojacobaea*
Senkirkine	2318-18-5	*Seneciojacobaea*
Riddelliine	23246-96-0	*Sassafras*
Lasiocarpine	303-34-4	*Heliotropiumlasiocarpum*
Camphor	76-22-2	*Sassafras*
Colchicine	64-86-8	*Colchicum autumnale*
Saikosaponin	20874-52-6	*Bupleurumchinense*
Cycasin	14901-08-7	*Cycascircinalis*
Diosbulbin B	20086-06-0	*Dioscoreabulbifera.*
Wilfordine	37239-51-3	*tripterygiumwilfordii*
Isatidine	15503-86-3	*Seneciobupleuroides*
Retroecine	480-85-3	*Seneciopseudoorientalis*
Retrorsine	480-54-6	*Senecio vulgaris*
Seneciphylline	480-81-9	*Seneciojacobaea*
Monocrotaline	315-22-0	*Crotalaria sessiiflora*
Heliotridine	520-63-8	*Heliotropiumeuropaeum*
Lycorine	476-28-8	*Lycoris radiate*
Dihydrolycorine	6271-21-2	*Lycorisradiata*
Cantharidin	56-25-7	*Mylabrisphalerata*
Punicalagin	65995-63-3	*Punicagranatum*
Toosendanin	58812-37-6	*Meliatoosendan*
Atractyloside	17754-44-8	*Atractylodesgummifera.*

From Equation (1-3), TP, TN, FN, and FP are the number of true positives, true negatives, false negatives, and false positives, respectively. Furthermore, 23 hepatotoxic compounds of traditional Chinese medicine (Table2) were collected from TOXNET (http://toxnet.nlm.nih.gov/) as external validation set to evaluate the reliability of the model while applying it in natural products research.

III. RESULTS AND ANALYSIS

A. Molecular descriptors selection

23 molecular descriptors were selected by using two feature selection algorithms in WEKA program package (Table 3). Then, a discriminative model was built based on these 23 molecular descriptors.

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

B. *Model construction and parameters selection*

The appropriate (C, γ) was identified by using parallel grid search and 10-fold cross-validation. The result is shown in Figure 1. The abscissa is the logarithm of C to the base 2, and the ordinate is the logarithm of γ to the base 2. Meanwhile, the best values were 0.32988, 0.10882, and 82.906% respectively.

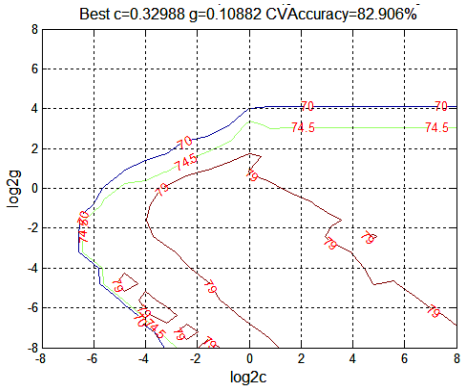

Fig.1. Contour chart of parameters optimization of hepatotoxicity discriminative model. The abscissa is the logarithm of C to the base 2, and the ordinate is the logarithm of γ to the base 2.

TABLE 3. NAMES OF MOLECULAR DESCRIPTORS

NO.	Abbreviation	Full name
1	AMW	average molecular weight
2	nAT	number of atoms
3	ZM1	first Zagreb index M1
4	MWC02	molecular walk count of order 02
5	X1Av	average valence connectivity index chi-1
6	IDM	mean information content on the distance magnitude
7	GATS1m	Geary autocorrelation-lag 1/weighted by atomic masses
8	ESpm01d	Spectral moment 01 from edge adj. matrix weighted by dipole moments
9	SEigp	Eigenvalue sum from polarizability weighted distance matrix
10	VRv1	Randic-type eigenvector-based index from van der Waals weighted distance matrix
11	RDF010u	Radial Distribution Function-1.0 / unweighted
12	Mor07u	3D-MoRSE-signal 07 / unweighted
13	Mor08u	3D-MoRSE-signal 08 / unweighted
14	Mor22u	3D-MoRSE-signal 22 / unweighted
15	HGM	geometric mean on the leverage magnitude
16	H1u	H autocorrelation of lag 1 / unweighted
17	HATSp	leverage-weighted total index / weighted by atomic polarizabilities
18	R5u	R autocorrelation of lag 5 / unweighted
19	RTv+	R maximal index / weighted by atomic van der Waals volumes
20	nHDon	number of donor atoms for H-bonds (N and O)
21	H-046	H attached to C0(sp3) no X attached to next C
22	Ui	unsaturation index
23	BLTF96	Verhaar model of Fish base-line toxicity from MLOGP

C. *Model validation*

The discriminative model was validated by test set. 7 positive compounds and 31 negative compounds were distinguished falsely. The sensitivity, specificity, and accuracy of this model were 87.04%, 80.86%, and 82.41% respectively. Besides, only 6 positive compounds of external validation set were distinguished falsely, with an accuracy of 73.91%. Corresponding compounds has shown in boldface in table 2.

The results demonstrated that this model provides a reliable utility for the hepatotoxic compounds prediction in traditional Chinese medicine study.

D. *The use of hepatotoxicity discriminative model in Qing kailing injection*

The hepatotoxicity discriminative model was applied to distinguish the 19 major compounds of Qingkailing injection [18] (Table 4). The result suggested that 5 of these compounds, namely Chenodeoxycholic acid, Desoxycholic acid, Geniposide, Hyodeoxycholic acid and Ursodeoxycholic acid may cause liver toxic, corresponding compounds has shown in boldface in table 4. Yue Ding [19] reported that 574 mg/kg Geniposide induced a delayed onset of hepatotoxicity in SD rats. Peizhen Song[20] reported that the lowest concentration of each bile acid in the mice feed is Chenodeoxycholic acid at 0.3% and Desoxycholic acid at 0.1%, which showed dose-response relationship between drug and hepatotoxicity.

TABLE 4. 19 MAJOR COMPOUNDS IN QINGKAILING INJECTION

Name	CAS	molecular weight	molecular formula
Isochlorogenic acid A	2450-53-5	516.46	$C_{25}H_{24}O_{12}$
Wogonoside	51059-44-0	460.40	$C_{22}H_{20}O_{11}$
Isochlorogenic acid C	32451-88-0	516.46	$C_{25}H_{24}O_{12}$
chlorogenic acid	327-97-9	354.30	$C_{16}H_{18}O_9$
cryptochlorogenic acid	905-99-7	354.31	$C_{16}H_{18}O_9$
neochlorogenic acid	906-33-2	354.31	$C_{16}H_{18}O_9$
Isochlorogenic acid B	14534-61-3	516.46	$C_{25}H_{24}O_{12}$
Scutellarin	27740-01-8	462.36	$C_{21}H_{18}O_{12}$
Chenodeoxycholic acid	474-25-9	392.57	$C_{24}H_{40}O_4$
Desoxycholic acid	83-44-3	392.57	$C_{24}H_{40}O_4$
Geniposide	24512-63-8	388.37	$C_{17}H_{24}O_{10}$
hyodeoxycholic acid	83-49-8	392.57	$C_{24}H_{40}O_4$
Cholic acid	81-25-4	408.58	$C_{24}H_{40}O_5$
Baicalin	21967-41-9	446.36	$C_{21}H_{18}O_{11}$
Ursodeoxycholic acid	128-13-2	392.57	$C_{24}H_{40}O_4$
Swertiamarin	17388-39-5	374.34	$C_{16}H_{22}$

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

			O_{10}
Shanzhisidemethyl ester	64421-28-9	406.38	$C_{17}H_{26}O_{11}$
Rutin	153-18-4	610.52	$C_{27}H_{30}O_{16}$
Luteoloside	5373-11-5	448.378	$C_{21}H_{20}O_{11}$

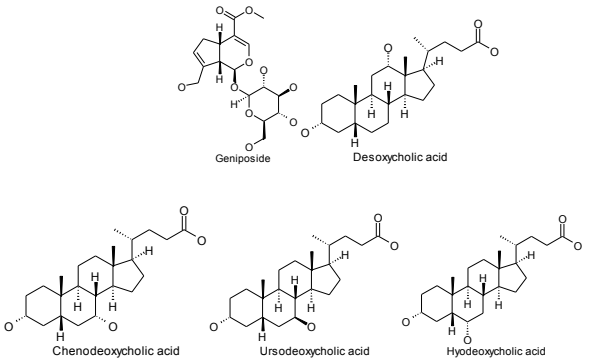

Fig.2. Structures of 5 compounds predicted by the hepatotoxicity discriminative model

Other 2 potential hepatotoxic compounds have not been verified yet. Although the false positive rate of the model might lead to such result, these 2 compounds share structural similarities with Chenodeoxycholic acid and Desoxycholic acid might be the major reason why the model predicated them as positive compounds. The structures are shown in Figure 2.

IV. CONCLUSION

Based on SVM, a hepatotoxicity discriminative model with high predicting accuracy was built and applied to distinguish hepatotoxic compounds in traditional Chinese medicine. Moreover, most of the experimental results were evidenced by literatures. This model can be widely used in prediction of hepatotoxicity of traditional Chinese medicine, because of the quickly and efficiently workability. However, present study is confined to qualitative research, and cannot distinguish toxic compounds based on different dose.

In summary, the discriminative model is regarded as a first step to screen hepatotoxicity compounds from traditional Chinese medicine or chemical drugs. The follow-up study can be a combination of discriminative model and toxicological test, which can help facilitate the study of safety in the clinical use of traditional Chinese medicines.

ACKNOWLEDGMENT

At the point of finishing this paper, the authors would like to express sincere thanks to the National Natural Science Foundation of China (No. 81173522) in Beijing University of Chinese Medicine.

REFERENCES

[1] Ignazio Grattagliano, Leonilde Bonfrate, Catia V Diogo, Helen H Wang, David QH Wang, Piero Portincasa. Biochemical mechanisms in drug-induced liver injury: certainties and doubts [J].World J Gastroenterol. 2009.

[2] Li Wang, Qiang Yuan, Gareth Marshall, Xiaohua Cui, Lan Cheng, Yuanyuan Li, Hongcai Shang, Boli Zhang, Youping Li. Adverse drug reactions and adverse events of 33 varieties of traditional Chinese medicine injections on National Essential medicines List (2004 edition) of China: an overview on published literatures[J].Journal of Evidence-Based Medicine.2010

[3] Baoqiu Li, Xin Dong, Guiqin Yang, Shihong Fang.Role of chlorogenic acid in the toxicity induced by Chinese herbal injections[J].Drug and Chemical Toxicology.2010.

[4] Yongliang Zhu, Zuguang Ye. Computational toxicology and its application in toxicity study of traditional Chinese medicine [J]. Chinese Journal of New Drugs 2011,20(24):2424-2429.

[5] Massarelli,Ilaria,Imbriani,Marcello;Coi,Alessio;Saraceno,Marilena;Carli ,Niccolò;Bianucci,Anna Maria.Development of QSAR models for predicting hepatocarcinogenic toxicity of chemicals.[J].European Journal of Medicinal Chemistry.2009.

[6] EkinsS,Williams AJ,Xu JJ. A predictive ligand-based bayesian model for human drug-induced liver injury [J]. DrugMetabolismandDisposition. 2010.38(12):2302-2308.

[7] Amie D. Rodgers, Hao Zhu, Dennis Fourches. Modeling liver-related adverse effects of drugs using kNN QSAR method. Chem Res Toxicol. 2011 April 19

[8] Cruz MonteagudoM, CordeiroMN, Borges F. Computational chemistry approach for the early detection of drug-induced idiosyncratic liver toxicity. Journal of computational chemistry 2008 Mar; 29(4):533-49.

[9] Nigel Greene, Lilia Fisk, Russell T. Naven;Reine R. Note, Mukesh L. Patel and Dennis J. Pelletier.Developing Structure Activity Relationships for the Prediction of Hepatotoxicity[J].Chemical Research in Toxicology.2010.

[10] Chin Yee Liew;Yen Ching Lim and Chun Wei Yap. Mixed learning algorithms and features ensemble in hepatotoxicity prediction [J].Journal of Computer-Aided Molecular Design.2011.

[11] Agus Saptoro;Moses O. Tadé;Hari Vuthaluru.A Modified Kennard-Stone Algorithm for Optimal Division of Data for Developing Artificial Neural Network Models[J].Chemical Product and Process Modeling.2012.

[12] Michael C, Hutter.Molecular Descriptors for Chemoinformatics [J].Chem Med Chem. 2010.

[13] Matthias Dehmer,Kurt Varmuza.Statistical modelling of molecular descriptors in QSAR/QSPR.[J].Reference and Research Book News.2012.

[14] MilanoChemometrics and QSAR Research Group. Dragon. http://michem.disat.unimib.it/chm

[15] Bin Huang. Prediction of blood-brain barrier penetrating drugs using supporting vector machine. Computers and Applied Chemistry,2009,26(2):188-189.

[16] I.H. Witten, Frank. Eibe. Data Mining: Practical Machine Learning Tools and Techniques, second ed. Morgan Kaufmann, San Francisco, 2005

[17] Yang Li. LIBSVM-faruto ultimate version: A toolbox with implements for support vector machines based on Libsvm [EB/OL]. http://www. ilovematlab.cn

[18] Jiayu Zhang, Qian Zhang, Hongxia Zhang. Rapid identification of 14 bile acids contained in Qingkailing injection by HPLC-ESI-MS /MS [J]. Journal of Chinese Materia Medica. 2013, 07: 990-994.

[19] Yue Ding, Tong Zhang, Jiansheng Tao, Liying Zhang, Jianrong Shi, Guang Ji.Potential hepatotoxicity of geniposide, the major iridoid glycoside in dried ripe fruits of Gardenia jasminoides (Zhi-zi).[J].Nat Prod Res.2013.

[20] Peizhen Song, Youcai Zhang, and Curtis D. Klaassen.Dose Response of Five Bile Acids on Serum and Liver Bile Acid Concentrations and Hepatotoxicty in Mice.[J].Toxicol Sci.2011.

978-1-4799-7294-4/14/$31.00 ©2014 IEEE