Combination Computing of Support Vector Machine, Support Vector Regression and Molecular Docking for Potential Cytochrome P450 1A2 Inhibitors

Xi Chen, Lian-sheng Qiao, Yi-lian Cai, Yan-ling Zhang*, Gong-yu Li

Key Laboratory of TCM Foundation and New Drug Research School of Chinese Material Medica, Beijing University of Chinese Medicine, Beijing 100102, China

(Dated: Received on March 3, 2016; Accepted on August 10, 2016)

The computational approaches of support vector machine (SVM), support vector regression (SVR) and molecular docking were widely utilized for the computation of active compounds. In this work, to improve the accuracy and reliability of prediction, the strategy of combining the above three computational approaches was applied to predict potential cytochrome P450 1A2 (CYP1A2) inhibitors. The accuracy of the optimal SVM qualitative model was 99.432%, 97.727%, and 91.667% for training set, internal test set and external test set, respectively, showing this model had high discrimination ability. The $R^2$ and mean square error for the optimal SVR quantitative model were 0.763, 0.013 for training set, and 0.753, 0.056 for test set respectively, indicating that this SVR model has high predictive ability for the biological activities of compounds. According to the results of the SVM and SVR models, some types of descriptors were identified to be essential to bioactivity prediction of compounds, including the connectivity indices, constitutional descriptors and functional group counts. Moreover, molecular docking studies were used to reveal the binding poses and binding affinity of potential inhibitors interacting with CYP1A2. Wherein, the amino acids of THR124 and ASP320 could form key hydrogen bond interactions with active compounds. And the amino acids of ALA317 and GLY316 could form strong hydrophobic bond interactions with active compounds. The models obtained above were applied to discover potential CYP1A2 inhibitors from natural products, which could predict the CYPs-mediated drug-drug interactions and provide useful guidance and reference for rational drug combination therapy. A set of 20 potential CYP1A2 inhibitors were obtained. Part of the results was consistent with references, which further indicates the accuracy of these models and the reliability of this combinatorial computation strategy.

Key words: Support vector machine, Support vector regression, Molecular docking, CYP1A2 inhibitor

I. INTRODUCTION

In recent years, the computational approaches of support vector machine (SVM), support vector regression (SVR) and molecular docking are widely used in in vitro identification of candidate compounds [1]. For binary-class classifications, SVM constructs an optimal separating hyperplane between the positive and negative classes with the maximal margin method [2]. SVM is an efficient machine learning method, which has been extensively used in pattern recognition and classification study. The main advantage of SVM is suitable for the processing of small-sample learning problems, based on the structural risk minimization principle [3]. SVR provides robust models for finding quantitative formulation based on structural risk minimization inductive principle instead of empirical risk minimization principle [4]. Molecular docking was utilized to estimate binding affinity between ligands and protein to obtain the favorable orientation of ligand [5]. It has been proved that computational approaches are efficient and reliable for predicting the properties of compounds [6]. To improve the accuracy and reliability of prediction, SVM, SVR and molecular docking were combined to predict potential active compounds. The SVM model and SVR model of active compounds were constructed to predict the biological activities of compounds. And the molecular docking model was established to further refine the predicted results of SVM and SVR models. By comparing the three prediction models, the structural characteristics of active compounds were analyzed.

Cytochrome P450 (CYP450), a superfamily of membrane-bound heme proteins, plays an important role in the metabolism of drugs and xenobiotics. The CYP1A2 isotype is highly expressed in the liver and intestine and is involved in the metabolic processes of many drugs and environmental chemicals. Therefore, the development of potential CYP1A2 inhibitors is of great importance for the treatment of drug-drug interactions and the design of rational drug combination therapy.
role in drug metabolism [7]. As one of the key enzymes in CYP450, CYP1A2 is mainly expressed in hepatic tissues and accounts for approximately 15% of the total CYP450 content [8]. CYP1A2 metabolized about 5%–10% of the marketed drugs, some of which was related to drug-drug interactions (DDIs) [9, 10]. With CYP1A2 inhibition activity, the drug would restrict the metabolism of another interacting drug, and result in side effect [11]. Therefore, the prediction of CYP1A2 inhibitors is helpful for reducing the risk of side effect caused by DDIs [12].

In this work, SVM, SVR, and molecular docking were utilized to predict the potential CYP1A2 inhibitors from traditional Chinese medicine database (TCMD version 2009) to warn the CYP1A2-mediated DDIs. The SVM qualitative and SVR quantitative models discriminate the biological activity of potential CYP1A2 inhibitors. Then, molecular docking was employed to analyze the binding affinity by the appropriate binding conformations of CYP1A2-interacted compounds. The combinatorial computation strategy for discovering potential CYP1A2 inhibitors could give guidance for the discovery of the potential inhibitors of other CYPs.

II. MATERIALS AND METHODS

A. Data preparation

The data set of active compounds with CYP1A2 inhibition activity was derived from literatures [13, 14], while the data set of inactive compounds was collected from the drug bank database (http://www.drugbank.ca/). 133 active compounds and 111 inactive compounds of CYP1A2 were all utilized as data set for the SVM model construction. To establish the SVR quantitative model, 78 inhibitors with explicit active values among the 133 active compounds were selected as data set for the SVR model. Besides, 1481 molecular descriptors of all compounds were calculated by Dragon2.1 to describe molecular structural characteristics.

B. Support vector machine

1. Data set splitting of SVM

Selected by utilizing Kennard-Stone (KS) algorithm, 88 active compounds and 88 inactive compounds were chosen as SVM training set from the data set for the SVM [15]. A set of 33 active compounds and 11 inactive compounds were regarded as internal test set. A set of 12 active compounds and 12 inactive compounds were selected randomly as external test set to verify the reliability of established model. Then the optimal molecular descriptors subset was obtained by two effective algorithms BestFirst searching and CfsSubsetEval valuation from Weka3.7. BestFirst searches the whole space of descriptor subsets by greedy hill-climbing considering the addition or/deletion of all possible single descriptor. CfsSubsetEval evaluates the predictive ability of all descriptors individually and considers the redundancy degree between them. The combination of BestFirst (search method) and CfsSubsetEval (attribute evaluator) improves the efficiency of the variable selection technique.

2. Development of SVM models

The libsvm3.1 algorithm (http://www.csie.ntu.edu.tw/~cjlin/libsvm/) was used for SVM modeling. The radial basis function (RBF) was chosen as the kernel function of SVM model to make sure the minimization of spatial complexity change when the parameters were altered [16]. The two important parameters $(C, \gamma)$ of RBF kernels were used to find out the best compromise from the complex models [17]. Besides, two methods, including parallel grid search and 10-fold cross-validation, were used to compute optimal combination of molecular descriptors and identify appropriate $(C, \gamma)$ of the model. By combining three normalized data processing methods and four parameter optimization methods, 12 models were established to obtain the optimal model. The normalized data processing methods included non-normalized, $[0, 1]$ normalization and $[-1, 1]$ normalization, while the parameter optimization methods contained non-optimized, grid search (GS), genetic algorithm (GA) and Particle swarm optimization (PSO).

3. Validation of the SVM models

The internal test set and external test set were utilized to evaluate the performance of models using three evaluation indicators including accuracy (ACC), sensitivity (SE), and specificity (SP) [18]. The computing formulas are shown as Eq.(1)–Eq.(3):

$$\text{ACC} = \frac{\text{TN} + \text{TP}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}} \quad (1)$$
$$\text{SE} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (2)$$
$$\text{SP} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (3)$$

True positive (TP) represents the number of active inhibitors which are correctly predicted as active inhibitors. True negative (TN) represents the number of inactive inhibitors which are predicted as inactive inhibitors. False positive (FP) stands for the number of inactive inhibitors which are predicted as active inhibitors. False negative (FN) stands for the number of active inhibitors which are predicted as inactive inhibitors [19].
C. Support vector regression

The data set for SVR model was split into two parts based on the ratio of 4:1 by KS, including 62 compounds in training set and 16 compounds in test set. Three data processing methods and three parameter optimization methods were combined to optimize SVR models. Three data processing methods were composed of non-treatment, ScaleForSVR function in LibSVM and principal component analysis in SPSS, while three parameter optimization methods included GS, GA, and particle swarm optimization (PSO). Two evaluation indexes of correlation coefficient ($R^2$) and mean square error (MSE) were used to validate SVR models. $R^2$ closer to 1 represents the higher correlation between experimental active value and predicts active value of the compounds. A smaller MSE implies a more accurate prediction model.

D. Docking

1. Define binding site of CYP1A2

The crystal structure of human microsomal CYP1A2 with the inhibitor α-naphthoflavone (αNF) was chosen from the protein data bank (PDB entry: 2HI4) with the resolution of 1.95 Å. Then, the water molecules were deleted and hydrogen atoms were added using the Discovery Studio 4.0 (DS) to adjust the protein structure. The active binding pocket was defined according to the initial inhibitor αNF.

2. Molecular docking

LibDock is a rapid and semi-flexible docking algorithm for molecular docking of a large number of compounds [20]. Diverse conformations of all compounds were generated by the BEST mode, and the maximum conformations were set to 255 with the energy threshold of 20.0 kcal/mol. In order to evaluate the suitability of LibDock algorithms, the initial compound (αNF) was re-docked into the active binding pocket. The root-mean-square deviation (RMSD) value between the docking pose and initial conformation of αNF was calculated, which could indicate the reliability of the molecular docking model and the rationality of the docking parameter settings. In general, RMSD should be less than 2.00 Å. After the exploration of optimal docking model, the 133 active compounds in SVM data set were docked into the binding pocket to explore the binding affinity between CYP1A2 and inhibitors.

E. Combinatorial prediction of potential CYP1A2 inhibitors

The combination of the SVM model, SVR model and molecular docking model was utilized to predict the potential CYP1A2 inhibitors. The optimal SVM model was utilized to preliminarily distinguish the activities of the compounds from TCMD, which possesses 23033 natural compounds from 6735 natural products. The hit compounds by SVM model were further predicted by the optimal SVR model. The compounds with higher predicted activities than initial compound were reserved for further molecular docking analysis. Finally, the compounds, which had higher LibDock Scores than initial compound, were considered as the potential CYP1A2 inhibitors.

III. RESULTS AND ANALYSIS

A. Support vector machine

1. Molecular descriptors selection of SVM

The computed 1481 molecular descriptors could be divided into different classes, such as constitutional descriptors, functional groups, topological descriptors and so on. A set of 12 optimal molecular descriptors were selected by BestFirst and CfsSubsetEval algorithms, which were listed in Table I.

<table>
<thead>
<tr>
<th>Classification</th>
<th>Molecular descriptors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constitutional descriptors</td>
<td>Me, ARR, nDB</td>
</tr>
<tr>
<td>Connectivity indices</td>
<td>X3A</td>
</tr>
<tr>
<td>Geometrical descriptors</td>
<td>RCI</td>
</tr>
<tr>
<td>Getaway descriptors</td>
<td>HATS7v</td>
</tr>
<tr>
<td>Functional group counts</td>
<td>nCOOR, nNR2Ph, nOHPH,</td>
</tr>
<tr>
<td></td>
<td>nRORPh, nPhX, nCSSR</td>
</tr>
</tbody>
</table>

The value of accuracy, sensitivity, and specificity of model-6 were 97.727%, 100.000%, 83.330% in internal test set respectively, while the corresponding value in

DOI:10.1063/1674-0068/29/cjcp1603039 ©2016 Chinese Physical Society
external test was 91.667%, 96.970%, 100.000% respectively. These results were all above 80%, demonstrating that Model-6 has reliable ability to compute the properties of potential CYP1A2 inhibitors for further studies.

B. Support vector regression

1. Molecular descriptors selection of SVR

The 78 optimal molecular descriptors subset were selected by BestFirst and CfsSubsetEval algorithms to build SVR models. The molecular descriptors of CYP1A2 qualitative model are listed in Table II.

2. Construction and validation of SVR model

Nine SVR models were constructed based on the training set of SVR and validated by the test set (Table S2 in supplementary materials). Model-3 was selected as the optimal model with higher evaluation indicators, which was computed by non-treatment normalized data processing and GA parameter optimization. The optimum parameters of \((C, \gamma)\) for model-3 were \((1.360, 0.017)\). Two evaluation indicators of \(R^2\) and MSE, were 0.763, 0.013 in training set and 0.753, 0.056 in test set, respectively. The results in the trend of prediction are shown in Fig.2. The paired \(t\)-test showed that there were no significant difference between the predicted active value and the true active value of the compounds in training set and the test set \((P>0.05)\) (Table S3 and Table S4 in supplementary materials). The chart elucidated that the predicted active value of the compounds were consistent with the true active value, which demonstrated that Model-3 are reliable for computing the activity of potential CYP1A2 inhibitor.

The molecular descriptors for SVM and SVR modeling were analyzed and compared with literatures results. The selected molecular descriptors of SVM contained 5 types of descriptors, among which 3 types were
FIG. 2 The prediction results of the optimal SVR model. (a) Train set regression, (b) test set regression.

also found in SVR models. It indicated that these 3 types of descriptors were related to both compounds identification and bioactivity prediction. Constitutional descriptors are the most commonly-used descriptors, reflecting the molecular composition of compounds such as numbers of atoms, bonds, double bonds, aromatic ratio and aromatic bonds, which had also been used to construct the CYP1A2 models in Gaspar’s paper [21]. The functional group counts descriptors of the number of hydroxyl (\(N_{OH}\)), carboxyl (\(N_{COOH}\)), ester bond (\(N_{COOR}\)), aniline (\(N_{NR2Ph}\)), hydrogen bond donor groups (\(N_{CHDon}\)) gave the chemical information of compounds. Connectivity indices could measure the connectivity of all the atoms, individual atoms, molecular fragments, and molecular geometry, which had also been considered for the modeling of CYP1A2 inhibitors in other published models [22, 23]. These types of descriptors, which provide information of chemical properties and functional groups for ligands, are important for the identification of active and non-active compounds.

C. Docking analysis results

Calculated by LibDock algorithm between the redocking and initial poses of the αNF, the RMSD values was 0.397 Å (<2.000 Å), which indicated the docking model is reliable and valuable. The LibDock Score of the initial compound was 137.553, which was regarded as a reasonable threshold value for the evaluation of potential CYP1A2 inhibitors. According to docking pose analysis of 133 active compounds of SVM, the active compounds could form hydrogen bond interactions at CYP1A2 active site with the key amino acids GLY316, ALA317, THR124, ASP320 and hydrophobic bond interactions with the amino acids PHE226, ALA317, GLY316, which were consistent with results in Refs.[10, 24].

D. Analysis of potential CYP1A2 inhibitors

Firstly, a hit list of 7797 compounds was predicted as active compounds by optimal SVM qualitative model. Then, 163 compounds were retained by the optimal SVR quantitative model. Finally, 20 compounds with higher LibDock Scores than the initial compound were reserved by the computation of molecular docking model. Part of them has been verified by literatures, such as tanshinone IIA [25], quercetin [26], nuciferine [27], celastrol [28], and so on. Tanshinone IIA was used as an instance to analyze the prediction results. The key amino acids of the hydrogen bond interactions between tanshinone IIA and CYP1A2 include THR124 and ASP313, while the residues bound by hydrophobic interactions comprised of ASN312, PHE125, ALA317, GLY316, PHE260, PHE226, LEU497, LEU382, ILE386, and ASP313. The three-dimensional interactions between tanshinone IIA and the active pocket are shown in Fig.3. The interactions between potential CYP1A2 inhibitors and CYP1A2 enzyme are similar to the results of the active compounds interacting with CYP1A2 enzyme. It is indicated that those potential compounds have similar biological activity with CYP1A2 inhibitors in clinical trial.

Tanshinone IIA is an active ingredient extracted from the rhizome of the Danshen (Salvia miltiorrhiza Bunge). Studies have shown that some Danshen constituents may be associated with a number of clinically important DDIs leading to adverse side effects [29, 30]. Most of the tanshinones isolated from Danshen could competitively inhibit the metabolism of CYP1A2 substrates [31]. Tanshinone IIA might cause the DDIs by the inhibition of CYP1A2. Therefore, the DDIs should be considered for drugs metabolized by CYP1A2 during long-term treatment with tanshinone IIA.
IV. CONCLUSION

In recent years, the combinatorial computational strategy with higher efficiency and reliability was widely used for the prediction of potential compounds. CYP1A2 enzyme metabolizes a variety of clinical drugs. The inhibition of CYP1A2 would induce DDIs and undesirable side effect. The accurate prediction of CYP1A2 inhibitors plays predominant roles in reducing the risk of DDIs induced by CYP1A2. In this study, a strategy for discovering potential CYP1A2 inhibitors was carried out by integrating three computational methods, including SVM, SVR, and molecular docking. The optimal SVM qualitative model and SVR quantitative model were established to predict the biological activities of compounds from natural products, resulting in 163 compounds. The molecular docking was further employed to verify the computed results of SVM and SVR and described the binding affinity between CYP1A2 and inhibitors. The 20 compounds hit by all three models were regarded as potential CYP1A2 inhibitors, and tanshinone IIA may cause the DDIs by the inhibition of CYP1A2. Parts of the compounds have been verified by literatures which indicated the applicability and accuracy of the constructed models. The combinatorial computation strategy is efficient for predicting the potential CYP1A2 inhibitors from natural products.

Supplementary materials: Table S1 shows the results of the SVM qualitative model. Table S2 shows the results of the SVR quantitative model. Table S3 shows the predicted value and the experimental value of the compounds in the SVR training set. Table S4 shows the predicted value and the experimental value of the compounds in the SVR test set.

V. ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (No.81173522, No.81300094, and No.81573831) and Joint Construction Project of Beijing Municipal Commission of Education.