Chemometrics and Intelligent Laboratory Systems 130 (2014) 109–114

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Combinations of fragment descriptors for improved prediction of CYP2C19 inhibitors

Li Chao [a,b], Hu Mei [a,b,*], Xianchao Pan [a], Wen Tan [a], Tengfei Liu [a], Li Yang [a,b]

[a] College of Bioengineering, Chongqing University, Chongqing 400044, China
[b] Key Laboratory of Biorheological Science and Technology, Ministry of Education, Chongqing 400044, China

ARTICLE INFO

Article history:
Received 13 June 2013
Received in revised form 19 October 2013
Accepted 22 October 2013
Available online 30 October 2013

Keywords:
CYP2C19
Inhibitor
Classification
Fragment
Support vector machine

ABSTRACT

The human cytochrome P450 (CYP450) superfamily plays an important role in drug–drug interactions, drug metabolism, and toxicity. Therefore, prediction of CYP450 inhibitors is extremely important in drug discovery and personal medicine. In this paper, characterized by fragment-based molecular hologram and MACCS descriptors, over 12,000 unique compounds with known CYP2C19 inhibitory activities were used to develop prediction models by partial least squares discriminant analysis (PLSDA) and support vector machine (SVM) methods. By combining two types of fragment-based descriptors, an optimal SVM model with an RBF kernel was obtained. The sensitivity, specificity, accuracy, and Matthews correlation coefficient (MCC) were 86.54%, 83.39%, 84.76%, and 0.6946 for a training set (n = 5,387), and 83.19%, 78.82%, 80.72%, and 0.6152 for a test set (n = 5,383), respectively. The optimal SVM model was further validated by an independent dataset (n = 1,470) with an overall accuracy of 82.38%. The results showed that these two types of fragment-based descriptors are, in some degree, complementary to each other, and can be combined to enhance model predictive power. In comparison with other 2-D or 3-D description methods, the combination of fragment descriptors seems extremely useful in constructing high-throughput screening models of CYP inhibitors in the process of drug discovery.

1. Introduction

It is a very common phenomenon that two or more drugs are co-administered for patients during disease treatment [1]. During drug treatment, the patients have an increased risk of exposure to potential adverse drug–drug interactions (DDIs) [2–4].

The human cytochrome P450 (CYP), a superfamily of heme-containing enzymes, plays an important role in mediating adverse DDIs. In the last several decades, several commercial drugs were withdrawn from the market due to CYP-mediated adverse DDIs, such as Seldane, Posicor, Hismanal, Propulsid, Lotronex, Baycol, and Seraone [1,5,6].

Polymorphic CYPs can catalyze the metabolism of a variety of endogenous and xenobiotic compounds. Although 57 isoforms are reported in the human genome, more than 90% of drugs are metabolized by 5 main CYP isoforms: CYP1A2, CYP2C9, CYP2C19, CYP2D6 and CYP3A4 [7]. The promiscuity with respect to substrates makes the CYP prone to being inhibited by a large amount of drugs, which may lead to clinically significant DDIs. Inhibition of CYP enzymes can lead to decreased elimination of compounds dependent on metabolism for systemic clearance and can lead to toxicity. Therefore, early prediction of CYP450 inhibitory activities of compounds may help to avoid the pursuit of drug candidates with these undesirable DDIs [8–11].

In the past several years, several in silico prediction models of CYP inhibitors have been reported. Based on extended connectivity fingerprints, Jensen et al. [12] reported classifiers for CYP2D6 and 3A4 inhibitors by using Gaussian kernel weighted k-nearest neighbor methods, and the sensitivities of a test set were only 59% and 65%, respectively. Based on common molecular descriptors, Dagliyan et al. [13] employed the mixed integer linear programming based hyperboxes method to construct classification models of CYP450 inhibitors. The mean accuracies for 8 CYP isoforms achieved 87.18% (based on the predicted binding free energy) and 88.09% (based on the predicted pIC_{50} values), respectively.

Recently, Novotarskyi et al. [9] applied associative neural networks, *k*-nearest neighbors, random tree, C4.5 tree, and support vector machine (SVM) to develop classifiers of CYP1A2 inhibitors based on constitutional, topological, physicochemical, 3-D descriptors, etc. The overall predictive accuracy of the best SVM model was 83% for a training set but only 68% for a test set. Based on Volsurf descriptors, Vasanthanathan et al. [14] applied SVM, random forest, *k*-nearest neighbors, and decision tree to predict CYP1A2 inhibitors, and the overall accuracy of the best SVM model was 75% for an internal test set and 67% for an external validation set. Cheng et al. [15] developed combined-classifier models for 5 major CYP isoforms (1A2, 2C9, 2C19, 2D6, and 3A4) based on a large data set with more than 24,700 unique compounds, and achieved

* Corresponding author at: Key Laboratory of Biorheological Science and Technology, Ministry of Education, Chongqing 400044, China. Tel.: +86 23 65112677.
E-mail address: meihu@cqu.edu.cn (H. Mei).

0169-7439/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.chemolab.2013.10.013

good results. For the combined-classifier models of CYP2C19, the overall predictive accuracies for a validation set ($n = 2691$) ranged from 79.0% to 81.0%. However, the combined-classifier method is relatively too complex and time consuming.

In general, the researches mentioned above still leave much to be desired. The first is the small sample size, which decreases the prediction capability of the model established. The second is high model complexity, which results in poor interpretability and being time consuming. The last is low prediction accuracy for external independent datasets.

In this paper, over 12,000 unique compounds with known CYP2C19 inhibitory activities were used to develop prediction models based on combined fragmental descriptors. CYP2C19 is one of the isoforms with marked genetic polymorphism. Many drugs are metabolized in vivo by CYP2C19 including antidepressants such as escitalopram and imipramine, anticonvulsants such as phenytoin, psychotropic drugs such as diazepam, and proton pump inhibitors such as omeprazole, lansoprazole, and rabeprazole [16,17]. By combining molecular hologram and MACCS description methods, an optimal SVM model with a radial basis function (RBF) kernel was obtained, of which the sensitivity, specificity, accuracy, and Matthews correlation coefficient (MCC) for training ($n = 5,387$) and test sets ($n = 5,383$) were 86.54%, 83.39%, 84.76%, and 0.6946 and 83.19%, 78.82%, 80.72%, and 0.6152, respectively. The predictive power of the SVM model was further tested by an independent test set ($n = 1,470$) with an overall accuracy of 82.38%.

To our knowledge, this is the first time that a simple and robust prediction model of CYP2C19 inhibitors has been established on such a large dataset. Moreover, the combination of fragment-based descriptors seems extremely useful in constructing high-throughput screening models of CYP inhibitors.

2. Principle and methods

2.1. Datasets

The 13,445 structures and inhibitor/non-inhibitor labels of CYP2C19 were collected from the PubChem database (AID: 1851). The inhibitory activities were measured by a standard protocol under the same experimental conditions against CYP2C19. Among the 13,445 compounds, 5913 were determined as inhibitors and 7532 as non-inhibitors. As shown in Fig. 1, the compounds were firstly pretreated and filtered by standard Lipinski's rule of five [18], and then divided into training ($n = 5,387$) and test sets ($n = 5,383$) randomly.

In addition, 9007 compounds collected from the PubChem database (AID: 899) were used to further verify the predictive power of resulting models. After the removal of duplicated compounds and pretreatment as illustrated in Fig. 1, a total of 1470 unique compounds were obtained. The statistical descriptions for training and test sets are presented in Table 1. The PubChem ID number, SMILES, and inhibitor/non-inhibitor labels of all 12,240 samples are shown in Table S1.

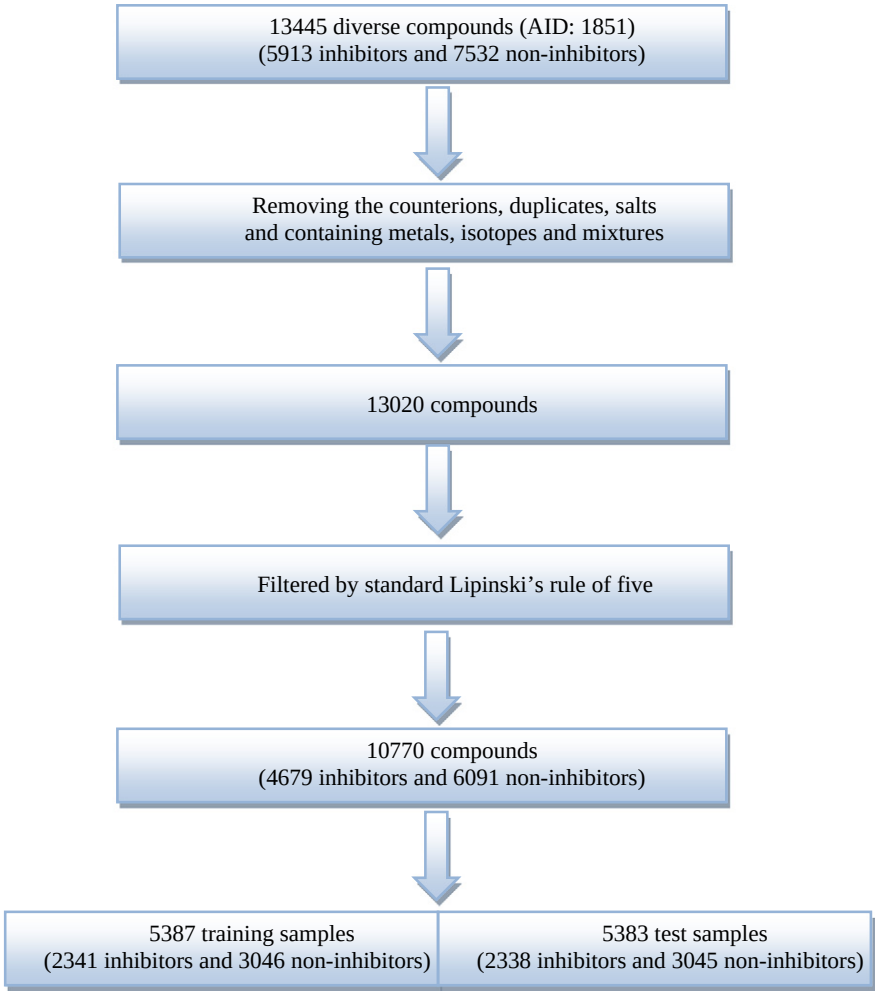

Fig. 1. The pretreatments of the dataset (AID: 1851).