Send Orders for Reprints to reprints@benthamscience.ae

Current Drug Metabolism, **2018,** *19,* **000-000**

REVIEW ARTICLE

Survey of Machine Learning Techniques for Prediction of the Isoform Specificity of Cytochrome P450 Substrates

Yi Xiong[1], Yanhua Qiao[2], Daisuke Kihara[3,4], Hui-Yuan Zhang[1], Xiaolei Zhu[2,*] and Dong-Qing Wei[1,*]

[1]*State Key Laboratory of Microbial Metabolism, and School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China;* [2]*School of Life Sciences, Anhui University, Hefei, Anhui 230601, China;* [3]*Department of Biological Science, Purdue University, West Lafayette, IN 47907, USA;* [4]*Department of Computer Science, Purdue University, West Lafayette, IN 47907, USA*

ARTICLE HISTORY

Received: January 19, 2018
Revised: August 05, 2018
Accepted: August 06, 2018

DOI:
10.2174/1389200219666181019094526

Abstract: ***Background:*** Determination or prediction of the absorption, distribution, metabolism, and excretion (ADME) properties of drug candidates and drug-induced toxicity plays crucial roles in drug discovery and development. Metabolism is one of the most complicated pharmacokinetic properties to be understood and predicted. However, experimental determination of the substrate binding, selectivity, sites and rates of metabolism is time- and recourse- consuming. In the phase I metabolism of foreign compounds (i.e., most of drugs), cytochrome P450 enzymes play a key role. To help develop drugs with proper ADME properties, computational models are highly desired to predict the ADME properties of drug candidates, particularly for drugs binding to cytochrome P450.

Objective: This narrative review aims to briefly summarize machine learning techniques used in the prediction of the cytochrome P450 isoform specificity of drug candidates.

Results and Conclusion: Both single-label and multi-label classification methods have demonstrated good performance on modelling and prediction of the isoform specificity of substrates based on their quantitative descriptors.

Keywords: Cytochrome P450, drug metabolism, isoform specificity, machine learning, single-label classification, multi-label classification.

1. INTRODUCTION

During the process of drug discovery stage, it is widely recognized that either determining or predicting the inappropriate absorption, distribution, metabolism, and excretion (ADME) properties of drug candidates and drug-induced toxicity can help prevent the failure of pharmaceutical compounds in clinical trials [1]. ADME properties describe the disposition of a pharmaceutical compound within an organism. All the four processes influence what the body does to a drug and the kinetics of drug exposure to the tissues, and hence influence the performance and pharmacological activity of the compound as a drug. Among the ADME properties, drug metabolism is a key determinant of metabolic stability, drug-drug interactions, and drug toxicity [2-4]. Metabolism refers to a process whereby the body converts a drug that has been absorbed by the body from its original form to a new form (called a metabolite). Most of the drugs that enter the body in the liver are metabolized by the cytochrome P450 enzymes.

The Cytochrome P450 (CYP450) is a ubiquitous heme-containing and biotransformation enzymes responsible for the metabolism of a wide variety of drugs, xenobiotics and endogenous compounds [5-7]. In phase I metabolism, CYP450 isoforms chemically modify a large variety of substrates mainly by oxidation reactions in order to make their products more water-soluble and easier to be excreted from the body. The Human Genome Project identified 57 different active genes, which encode cytochrome P450 isoforms. They share the same fold, and can be categorized into 18 families and 43 subfamilies based on the similarity of their primary sequences [8, 9]. The first three families (CYP1-3) are generally responsible for metabolizing exogenous substances such as drugs, whereas the rest of CYP families are usually involved in the metabolism of endogenous substances [10]. Several main isoforms of particular importance for drug metabolism are CYP450 1A2, 2C19, 2C8, 2C9, 2D6, 2E1, and 3A4, which cover the majority of all possible metabolism routes. Generally, those different CYP isoforms are responsible for metabolizing chemically or structurally different drugs. For example, due to its spacious binding site, CYP 3A4 is capable of metabolizing high-volume and lipophilic xenobiotics, indeed, at least 422 drugs. Moreover, in many cases, the same drug can potentially be metabolized by multiple CYP450 isoforms. In a recent compilation of interaction data between CYP isoforms and substrates by us (unpublished data), it was estimated that a single substrate can be potentially metabolized by about two different isoforms (1646 isoform-substrate interactions among the 776 chemically different substrates). Isoform specificity of cytochromes is manifested in the following ways: (i) substrate selection, i.e., specific substrate metabolized by specific isoform; (ii) Regioselectivity, i.e., multiple sites of a substrate may be oxidized by more than one isoenzyme; (iii) Rate of conversion to products, i.e., while a specific substrate may be oxidized by two different isoforms with the same site of metabolism on the substrate, the Km value associated with catalytic process or reaction differs. In this review, we focus on the solutions to address the first problem. Experimental determination of the isoform specificity of substrates is both time- and resource-consuming. This motivates the development of high throughput computational methods to classify a compound according to the isoform by which it is metabolized. Based on various molecular properties, prediction of the metabolic profile of drugs is a matter of wide interest.

In recent years, many different in silico approaches have been developed to predict the CYP isoform specificity of drug molecules. Generally, two major categories of computational approaches

*Address correspondence to these authors at the Room 4-321, Life Science Building, Shanghai Jiao Tong University, 800 Dongchuan Road, 200240, Minhang District, Shanghai, China; Tel:/Fax: +86 21-34204573; Emails: xlzhu_mdl@hotmail.com, dqwei@sjtu.edu.cn

are well studied and applied to the prediction of potential CYP450 enzyme isoforms involved in the metabolism of a small molecule. The first group of approaches is protein structure-based methods, based on the available three-dimensional structures of macromolecules to directly evaluate the interaction details between CYP450 enzymes and drug molecules [11, 12]. However, application of these approaches is limited to cases that high resolution tertiary structures of substrate-bound-form of CYP450 enzymes are available. The second group of approaches is ligand-based methods, which consider the structural similarity of ligands to known substrates. The most commonly used ligand-based approach is quantitative structure-activity relationship (SAR or QSAR) model, which aims to establish a correlation between descriptors encoded in the information of molecular structures of ligands and biological activities (i.e., whether or not the compound is the substrate of a specific CYP450 enzyme isoform) for a given compound with structurally and biologically characterized [13].

Various SAR or QSAR models had been published to classify substrates and nonsubstrates for a particular type of CYP450 enzyme isoform using a variety of algorithms, including multiple linear regression, partial linear least squares, neural networks, support vector machines (SVM), and other machine learning techniques [14-21]. However, these traditional QSAR models based on single-label classification algorithms can consider only a single type of CYP450 enzyme isoform at a time, and deal exclusively with nonoverlapping classes. With the increasing availability of high-quality data of other CYP450 enzyme isoforms, it is preferable to classify a compound into the maximum possible types of CYP450 enzyme isoform substrates. The SuperCYP database by Preissner *et al.* [9] provides interaction data on various types of CYP450 enzyme isoforms and their associated substrates which can be metabolized by more than one isoform. Multi-label classification is used for assigning a target object to multiple classes, and suitable for prediction of the metabolism profile of CYP450 substrates because some substrates are known to interact with multiple CYP450 isoforms.

In this article, we discuss key aspects of various approaches that are available to predict CYP450 enzyme isoform specificity of substrate molecules. Firstly, we introduce the dataset of substrates for the main types of CYP450 enzyme isoforms. Then, we summarize the feature selection and representation techniques used to describe the training and testing data sets as the input of the classification algorithms. Finally, we show several state-of-the-art classification strategies and algorithms. Fig. (**1**) shows an overflow of the classification framework to predict the CYPs isoform specificity of substrates.

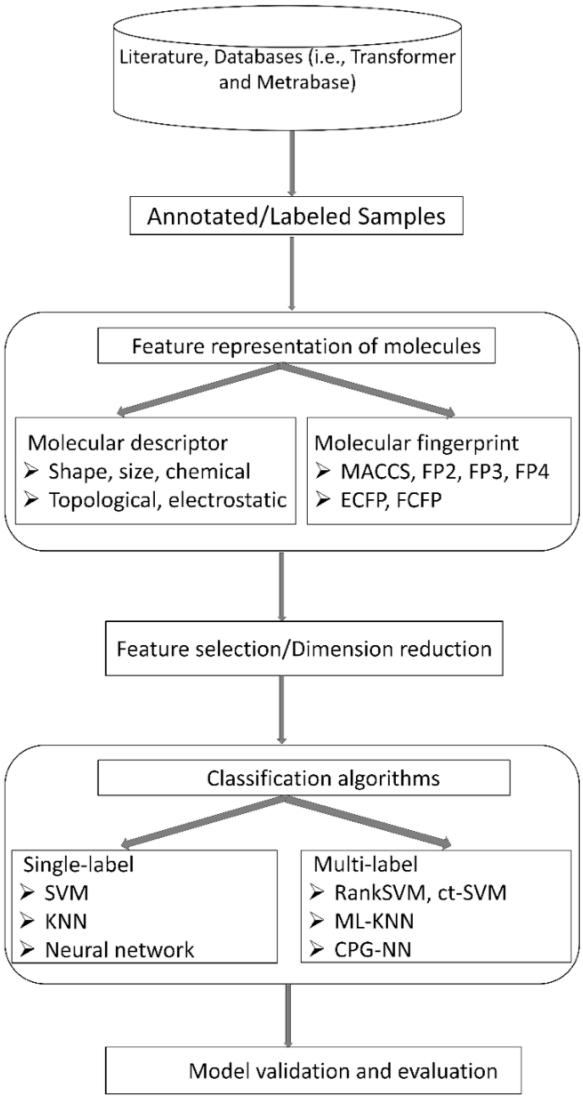

Fig. (1). An overflow of the classification framework to predict the CYPs isoform specificity of substrates.

2. DATABASES AND THE BENCHMARK DATA SETS

For any classification model, high-quality data sets play a critical role in training, which can lead to better prediction performance. In bioinformatics applications, most studies constructed their benchmark data sets based on public database and/or published literature [22-35]. During the last decade, several databases were developed to focus on CYPs and drug metabolism, such as SuperCYP [9], Transformer [36], and Metrabase [37]. SuperCYP was originally developed to contain information about drugs and cytochrome-drug interactions of 57 human CYPs. Each drug was attributed to those CYPs that are involved in drug metabolism as a substrate, inhibitor, or inducer. In recent years, SuperCYP was updated and enhanced to the new database named Transformer (metabolism of xenobiotics database) to include many other important enzymes in the metabolism of xenobiotics, such as phase II enzymes or transporters [36]. As of Oct 2016, the Transformer database contains integrated information on the three phases of biotransformation of about 2,800 drugs, and 350 food and herbal ingredients, involved in 4,009 phase I reactions. Metrabase v1.0 [37] is an integrated cheminformatics and bioinformatics resource, which includes the interaction data on 20 transporters from ATP-binding cassette (ABC) and solute carrier families and a limited set of 13 CYPs, involving 3,438 chemical compounds (small molecule substrates and modulators). The Metrabase database aims to provide comprehensive structural, physicochemical and biological data that can be used to infer the relationships between these transporters or CYP450 enzymes and their ligands. A summary of CYP-drug interaction related databases is shown in Table **1**.

3. FEATURE REPRESENTATIONS OF SUBSTRATES

Any classification model requires a mathematical representation of molecules through computation of structural descriptors or fingerprints. For small molecules, varieties of molecular descriptors (physicochemical, topological, etc.) and fingerprints are calculated to represent their properties. Table **2** is a list of freely available tools to calculate various molecular descriptors of substrates. Four different types of fingerprints can be generated by Open Babel [38]. They are FP2, FP3, FP4, and MACCS, calculated from a one-dimensional (1D) representation of a compound, which is the Simplified Molecular Input Line Entry Specification (SMILES) [39]. These fingerprints are binary strings, which encode the presence or absence of substructural fragments. For example, the MACCS fingerprint consists of 166 structural keys based on

Table 1. Summary of CYP-drug interaction related databases.

Database	Release Date	Description	Web Site
SuperCYP	2010	About 1,170 drugs, 2,785 cytochrome-drug interactions and about 1,200 alleles	http://bioinformatics.charite.de/supercyp
Transformer	October 2016	Integrated information on the three phases of biotransformation (modification, conjugation and excretion) of 3000 drugs and more than 350 relevant food ingredients and herbs, which are catalyzed by 400 proteins	http://bioinformatics.charite.de/transformer
Metrabase	2015 v1.0	11,649 interaction records involving nearly 3,500 small molecule substrates and modulators of 20 transport proteins and 13 cytochrome P450 enzymes	http://www-metrabase.ch.cam.ac.uk

Table 2. A list of freely available software to calculate the molecular descriptors of substrates.

Software	Description	Web Site
Open Babel	Calculates four types of fingerprints, FP2, FP3, FP4, and MACCS	http://openbabel.org/
RDKit	Open source toolkit for cheminformatics, including descriptor generation for machine learning	http://www.rdkit.org/
Dragon	Calculates 5,270 molecular descriptors, covering most of the various theoretical approaches	http://www.talete.mi.it/
Chemistry Development Kit (CDK)	A collection of modular Java libraries for processing chemical information, such as the calculation of various molecular descriptors	https://cdk.github.io/

SMARTS patterns covering most of the important chemical fragments. Given any two substrates, their fingerprint similarity was defined by Tanimoto coefficient or Jaccard Coefficient using the bits set in the two fingerprints [40], which is defined as:

$$TC = \frac{c}{a+b-c} \quad (1)$$

where a and b are the number of bits set in their fragment bit-strings, with *c* of these bits being set in both of the fingerprints.

Moreover, the family of Morgan fingerprints (known as circular fingerprints) and their functional invariants are built by applying the Morgan algorithm to a set of user-supplied atom invariants. They are generated by the RDKit, a python toolkit (http://www.rdkit.org/). They are the circular fingerprint equivalents of the well-known extended connectivity fingerprint (ECFP) and functional-class fingerprint (FCFP), respectively, which are designed for a wide variety of applications such as molecular characterization, similarity searching in drug discovery, and structure-activity modeling. The number after "ECFP" or "FCFP" (such as ECFP4 or FCFP4) corresponds to the parameter of the diameter of the atom environments used for calculation of the fingerprint, whereas in the case of Morgan fingerprints the number denotes the size of a radius parameter. Therefore, Mr2 are roughly equivalent to ECFP4 or FCFP4.

The TOpological MOlecular COMputer Design (TOMOCOMD) approach can be used to construct 2D and 3D graph-theoretical descriptors using discrete mathematics and linear algebra theory to chemical structures. In this approach, a molecule is represented as a pseudograph [41, 42]. In other studies [43-47], 2D and 3D Zernike descriptors can be effectively used to represent molecular surface properties of compounds and for rapid compound comparison.

4. FEATURE SELECTION

Once compounds are represented with a set of features, it is important to select the most informative features to remove redundancy in the representation and make algorithms efficient. As suggested in previous studies [48-57], effective feature selection or dimensionality reduction is necessary for reducing the computational time and complexity of the classification models, in order to provide more insights into the data abundance. Feature selection is a challenging problem because it aims to find the best or optimal feature combination in the vast space of possible combinations from a large number of candidate descriptors. In a realistic application, researchers often design heuristic algorithms to search for an approximate solution to get the balance between accuracy and efficiency. Based on the manner of selection, feature selection strategies are categorized into three main groups, which are a filter, wrapper, and embedding. Genetic algorithm (GA) is a powerful feature selection method, which is a heuristic search and stochastic algorithm inspired by the natural process of evolution. It has been widely used in variable selection for drug metabolism related models [14, 17, 18, 21, 58].

5. SINGLE-LABEL CLASSIFICATION METHODS

Single-label classification models are based on the assumption that each compound or substrate is metabolized by a single predominant CYP450 isoform. Most of these single-label based models used the supervised classification algorithms, such as support vector machine (SVM), decision tree, K-Nearest Neighbor (KNN), and neural network (NN), to build the quantitative relationship between the features (descriptors of substrates) and the class labels (isoform specificity) [14-19].

SVM algorithms were most popularly used to develop classification models in the isoform specificity and other bioinformatics applications [59-64]. SVM maps input data nonlinearly into a high dimensional feature space and separate them by a hyperplane into two classes. Intuitively, a good separation is achieved by selecting a hyperplane that separates the two classes with the maximal distance from any one of the given samples. The larger the margin is, the lower the generalization error of the classifier on unknown samples is. The use of a kernel function allows the algorithm to fit the maximum-margin hyperplane in a transformed feature space. Recently, deep neural network architectures such as convolutional and long short-term memory networks gain increasing popularity as

machine learning tools, which achieve great success especially in image recognition and speech recognition. In the bioinformatics community, deep learning will be also useful for classifying substrates to CYP450 isoform(s).

Moreover, unsupervised machine learning algorithms can also be successfully applied to the prediction of substrate selectivity for the major CYP450 isoforms. Korolev and Balakin *et al.* [65] developed a computational method for prediction of isozyme-specific groups of substrate molecules based on Kohonen Self-organizing Map (SOM). The advantage of using the unsupervised Kohonen learning strategy is that it does not depend on the construction of a negative training set, which is hard to be correctly defined for a set of non-substrates for a specific enzyme.

6. MULTI-LABEL CLASSIFICATION METHODS

Multi-label classification of biological data remains to be a challenging problem, and is widely used in various applications [66-70]. Although certain major CYP isoforms mediate the metabolism of a large number of substrates, they can exhibit overlapping substrate specificities with other

CYP isoforms. In comparison to the single-label classification strategy, the multi-label classification strategy has a unique property that a substrate is simultaneously assigned to at least two classes of isoforms. Michielan *et al.* [20] used three different multi-label classification algorithms, which include counterpropagation neural network (CPG-NN), cross-training with SVM (ct-SVM), and multi-label k-nearest-neighbor (ML-KNN), to classify the compounds simultaneously in multiple classes of isoforms. In addition to ML-KNN, Wei *et al.* [21] used two other multi-label algorithms, including back propagation (BP) neutral network and RankSVM. ML-KNN is derived from the popular K-nearest neighbor algorithm. For each instance in the testing set, its KNNs in the training set are firstly identified. Then, according to statistical information gained from the label sets of these neighboring instances, i.e. the number of neighboring instances belonging to each possible class, maximum a posteriori principle is utilized to determine the label set for the testing instance [71]. RankSVM algorithm is extended from standard SVM, and is a learning retrieval function that employs pairwise ranking algorithms to adaptively sort results based on its relevance to a specific query. The choice of a suitable kernel function is crucial to the performance of RankSVM. The linear RankSVM is more efficient, whereas the kernel RankSVM takes longer running time but yields higher accuracy on the same dataset [72].

A summary of single-label and multi-label classification methods for prediction of the CYP450 isoform specificity of substrates is listed in Table **3**.

Table 3. Summary of machine-learning-based models for prediction of the isoform specificity of substrates.

Ref	Isoform	Descriptor	Feature Selection	Strategy	Algorithm	Performance
[14]	CYP3A4,2D6, 2C9	A total of 1497 1D, 2D, and 3D molecular descriptors, calculated by DRAGON Web version 3.0	Genetic algorithm	Single label	PM-CSVM, PM-CSVM	MCC value of 0.849, 0.852,0.851 for the substrates of CYP3A4, 2D6, 2C9
[65]	38 CYPs (enzyme-specific groups)	Sixty molecular descriptors calculated by using Cerius and ChemoSoft	Principal component analysis	Single label	Kohonen SOM	An accuracy of 76.7%
[15]	CYP3A4, 2D6, 2C9	Topological Autocorrelation, 3D Autocorrelation, Global Molecular, Shape/Size-Related Descriptors, and Substructure Counts	BestFirst or ExhaustiveSearch	Single label	Multinomal logistic regression, decision tree, SVM	The accuracy remains at 78.5-82.4% on the external validation set
[20]	CYP1A2, 2C19, 2C8, 2C9, 2D6, 2E1, 3A4	Molecular shape, functional-group-count descriptors, 2D, 3D spatial autocorrelation descriptors, autocorrelation molecular electrostatic potential descriptors	BestFirst automatic variable selection	Single label, multi label	ct-SVM, ML-KNN, CPG-NN	The MCC remains at 0.44-0.70 by multi-label classification
[16]	CYP2C9, 2D6, 3A4	Pharmacophore maps and chemical features	Manual	Single label	Decision tree	An accuracy of 76.67%~82%
[17]	CYP3A4, 2C9	Classical molecular descriptors and binary fingerprints	Genetic algorithm	Single label	CART, KNN, N-Nearest Neighbor	An average of sensitivity and specificity remains 75%~78% on the test sets
[21]	CYP1A2, 2A6, 2B6, 2C9, 2C19, 2D6, 2E1, 3A4	A total of 193 molecular descriptors were calculated including topological, geometrical, electrostatic and other physicochemical descriptors by the package SAMM	Genetic algorithm	Single label, multilabel	Decision tree, neural network, ML-KNN, Rank-SVM	~90% classification accuracy on single label system, ~80% prediction accuracy on multi-label system
[18]	CYP3A4, 2D6, 1A2, 2C9, 2C19	1D, 2D and 3D descriptors belonging to different categories such as mass, surface area, volume, moment of inertia, dipole, molar refractivity, lipoles, connectivity, electrostatic parameters, Chemistry Development Kit (CDK) molecular descriptors	Genetic algorithm	Single label	SVM	An average accuracy of 86.02% using fivefold cross-validation, 70.55% on an independent dataset
[19]	CYP1A2, 2C9, 2C19, 2D6, 2E1, 3A4	Constitutional descriptors, topological and electrotopological descriptors, and descriptors relating to hydrophobicity, electronic properties, hydrogen bonding, and molecular ionization	Manual	Single label	Decision tree model by multiobjective recursive partitioning analysis	An average accuracy of 88% using cross-validation, 84.3% on an independent dataset

7. MODEL VALIDATION AND EVALUATION

In supervised machine learning framework, it is important to split a benchmark dataset into two parts, one for training a model and the other for testing the model. To evaluate the performance of classification models, the validation methods are mainly consisting of *k*-fold cross-validation, leave-one-out cross-validation, and independent tests. In *k*-fold cross-validation, the sample set is randomly partitioned into *k* subsets with equal size. Of the *k* subsets, one subset is selected as the validation data for testing the model, and the remaining *k*-1 subsets are used as training data. The cross-validation process is then repeated *k* times (the folds), with each of the *k* subsets used exactly once as the validation data. The results from *k* folds are finally averaged to generate a single estimation metric. Leave-one-out cross-validation (LOOCV) involves using a single sample from the sample set as the validation data, and the remaining samples as the training data. This is repeated such that each sample in the sample set is used once as the validation data. This is the same as a *k*-fold cross-validation with *k* being equal to the number of samples in the original sample set. Leave-one-out cross-validation is computationally expensive when the number of samples in the training set is too large.

In order to assess the classification performance, various threshold-dependent metrics can be utilized. They are accuracy (*ACC*), sensitivity (*SN*, also called recall), specificity (*SP*), precision (*PR*), Matthew's correlation coefficient (*MCC*) and F-measure (F_1). The set of metrics have been used by a series of studies [73-77]. The receiver operating characteristic (ROC) curve is a plot of the sensitivity versus (1-specificity) for a binary classifier at varying thresholds. The area under the curve (*AUC*) can be used as a threshold-independent measure of classification performance. It is a nontrivial task to assess the quality of prediction for heavily unbalanced data sets. On the unbalanced data sets, the accuracy and AUC of ROC curve can present overly optimistic assessments of the performance of an algorithm. Instead, the precision-recall curve is a plot of the recall versus precision for a binary classifier at varying thresholds.

CONCLUSION

Determination or prediction of the properties of drug metabolism plays key roles in the early stage of drug discovery and development. In this review, we focused on the ligand-based methods by using machine learning techniques for prediction of CYP450 isoform specificity of substrates. We systematically and briefly summarized the four essential components, which consist of the construction of golden standard datasets, feature selection and representation, classification algorithms, and model validation and evaluation, to constitute a complete model for classification and prediction of CYP450 enzyme-substrate selectivity.

LIST OF ABBREVIATIONS

1D	=	One-dimensional
ACC	=	Accuracy
ADME	=	Absorption, distribution, metabolism, and excretion
AUC	=	Area under the curve
BP	=	Back propagation
CART	=	Classification and regression tree
CDK	=	Chemistry Development Kit
CPG-NN	=	Counterpropagation NN
ct-SVM	=	Cross-training with SVM
CYP450	=	Cytochrome P450
ECFP	=	Extended connectivity fingerprint
F1	=	F-measure
FCFP	=	Functional-class fingerprint
GA	=	Genetic algorithm
KNN	=	K-nearest neighbor
LOOCV	=	Leave-one-out cross-validation
MCC	=	Matthew's correlation coefficient
ML	=	Multi-label
NN	=	Neural network
PM-CSVM	=	Positive majority consensus SVM
PP-CSVM	=	Positive probability consensus SVM
PR	=	Precision
QSAR	=	Quantitative SAR
ROC	=	Receiver operating characteristic
SAR	=	Structure-activity relationship
SMILES	=	Simplified Molecular Input Line Entry Specification
SN	=	Sensitivity
SOM	=	Self-organizing map
SP	=	Specificity
SVM	=	Support vector machine
TOMOCOMD	=	TOpological MOlecular COMputer Design

CONSENT FOR PUBLICATION

Not applicable.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

This work was supported by the funding from National Natural Science Foundation of China for Young Scholars (Grant No. 31601074 and 21403002), National Key Research Program (Contract No. 2016YFA0501703), Shanghai Key Laboratory of Intelligent Information Processing (Contract No. IIPL-2016-005), and Shanghai Jiao Tong University School of Medicine (Contract No. YG2015QN34).

REFERENCES

[1] Cheng F, Li W, Liu G, Tang Y. In silico ADMET prediction: recent advances, current challenges and future trends. Curr Top Med Chem 2013;13(11):1273-89.

[2] Tyzack JD, Mussa HY, Williamson MJ, Kirchmair J, Glen RC. Cytochrome P450 site of metabolism prediction from 2D topological fingerprints using GPU accelerated probabilistic classifiers. J Cheminform 2014;6:29.

[3] Nielsen LM, Linnet K, Olsen L, Rydberg P. Prediction of cytochrome p450 mediated metabolism of designer drugs. Curr Top Med Chem 2014;14(11):1365-73.

[4] Zaretzki J, Bergeron C, Huang TW, Rydberg P, Swamidass SJ, Breneman CM. RS-WebPredictor: a server for predicting CYP-mediated sites of metabolism on drug-like molecules. Bioinformatics 2013;29(4):497-8.

[5] Lewis DF. Human cytochromes P450 associated with the phase 1 metabolism of drugs and other xenobiotics: a compilation of substrates and inhibitors of the CYP1, CYP2 and CYP3 families. Curr Med Chem 2003;10(19):1955-72.

[6] Zheng M, Luo X, Shen Q, Wang Y, Du Y, Zhu W, *et al.* Site of metabolism prediction for six biotransformations mediated by cytochromes P450. Bioinformatics 2009;25(10):1251-8.

[7] Li L, Xiong Y, Zhang ZY, Guo Q, Xu Q, Liow HH, *et al.* Improved feature-based prediction of SNPs in human cytochrome P450 enzymes. Interdisciplinary sciences, computational life sciences 2015;7(1):65-77.

[8] Ingelman-Sundberg M. The human genome project and novel aspects of cytochrome P450 research. Toxicol Appl Pharmacol 2005;207(2 Suppl):52-6.

[9] Preissner S, Kroll K, Dunkel M, Senger C, Goldsobel G, Kuzman D, *et al.* SuperCYP: a comprehensive database on Cytochrome P450 enzymes including a tool for analysis of CYP-drug interactions. Nucleic Acids Res 2010;38(Database issue):D237-43.

[10] Sim SC, Ingelman-Sundberg M. The Human Cytochrome P450 (CYP) Allele Nomenclature website: a peer-reviewed database of CYP variants and their associated effects. Hum Genomics 2010;4(4):278-81.

[11] Lewis DF, Ito Y. Human CYPs involved in drug metabolism: structures, substrates and binding affinities. Expert Opin Drug Metab Toxicol 2010;6(6):661-74.
[12] Kesharwani SS, Nandekar PP, Pragyan P, Rathod V, Sangamwar AT. Characterization of differences in substrate specificity among CYP1A1, CYP1A2 and CYP1B1: an integrated approach employing molecular docking and molecular dynamics simulations. J Mol Recognit 2016;29(8):370-90.
[13] Shaikh N, Sharma M, Garg P. Selective Fusion of Heterogeneous Classifiers for Predicting Substrates of Membrane Transporters. J Chem Inf Model 2017;57(3):594-607.
[14] Yap CW, Chen YZ. Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines. J Chem Inf Model 2005;45(4):982-92.
[15] Terfloth L, Bienfait B, Gasteiger J. Ligand-based models for the isoform specificity of cytochrome P450 3A4, 2D6, and 2C9 substrates. J Chem Inf Model 2007;47(4):1688-701.
[16] Ramesh M, Bharatam PV. CYP isoform specificity toward drug metabolism: analysis using common feature hypothesis. Journal of molecular modeling 2012;18(2):709-20.
[17] Nembri S, Grisoni F, Consonni V, Todeschini R. In Silico Prediction of Cytochrome P450-Drug Interaction: QSARs for CYP3A4 and CYP2C9. International journal of molecular sciences 2016;17(6).
[18] Mishra NK, Agarwal S, Raghava GP. Prediction of cytochrome P450 isoform responsible for metabolizing a drug molecule. BMC Pharmacol 2010;10:8.
[19] Yamashita F, Hara H, Ito T, Hashida M. Novel hierarchical classification and visualization method for multiobjective optimization of drug properties: application to structure-activity relationship analysis of cytochrome P450 metabolism. J Chem Inf Model 2008;48(2):364-9.
[20] Michielan L, Terfloth L, Gasteiger J, Moro S. Comparison of multilabel and single-label classification applied to the prediction of the isoform specificity of cytochrome p450 substrates. J Chem Inf Model 2009;49(11):2588-605.
[21] Zhang T, Dai H, Liu LA, Lewis DFV, Wei DQ. Classification Models for Predicting Cytochrome P450 Enzyme-Substrate Selectivity. Molecular informatics 2012;31(1):53-62.
[22] Zhang W, Qu QL, Zhang YQ, Wang W. The linear neighborhood propagation method for predicting long non-coding RNA - protein interactions. Neurocomputing 2018;273:526-34.
[23] Zhu X, Xiong Y, Kihara D. Large-scale binding ligand prediction by improved patch-based method Patch-Surfer2.0. Bioinformatics 2015;31(5):707-13.
[24] Zhang W, Xiong Y, Zhao M, Zou H, Ye X, Liu J. Prediction of conformational B-cell epitopes from 3D structures by random forests with a distance-based feature. BMC Bioinformatics 2011;12:341.
[25] Chen W, Yang H, Feng P, Ding H, Lin H. iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. Bioinformatics 2017;33(22):3518-23.
[26] Chen W, Feng P, Yang H, Ding H, Lin H, Chou KC. iRNA-AI: identifying the adenosine to inosine editing sites in RNA sequences. Oncotarget 2017;8(3):4208-17.
[27] Zhang W, Chen Y, Liu F, Luo F, Tian G, Li X. Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data. BMC Bioinformatics 2017;18(1):18.
[28] Rudik A, Dmitriev A, Lagunin A, Filimonov D, Poroikov V. SOMP: web server for in silico prediction of sites of metabolism for drug-like compounds. Bioinformatics 2015;31(12):2046-8.
[29] Zhang W, Liu X, Chen Y, Wu W, Wang W, Li X. Feature-derived graph regularized matrix factorization for predicting drug side effects. Neurocomputing 2018;287:154-62.
[30] Peng S, You R, Wang H, Zhai C, Mamitsuka H, Zhu S. DeepMeSH: deep semantic representation for improving large-scale MeSH indexing. Bioinformatics 2016;32(12):i70-i9.
[31] Liu K, Peng S, Wu J, Zhai C, Mamitsuka H, Zhu S. MeSHLabeler: improving the accuracy of large-scale MeSH indexing by integrating diverse evidence. Bioinformatics 2015;31(12):i339-47.
[32] Wei YQ, Bi DX, Wei DQ, Ou HY. Prediction of Type II Toxin-Antitoxin Loci in Klebsiella pneumoniae Genome Sequences. Interdisciplinary sciences, computational life sciences 2016;8(2):143-9.
[33] Zhang W, Yue X, Lin W, Wu W, Liu R, Huang F, *et al*. Predicting drug-disease associations by using similarity constrained matrix factorization. BMC Bioinformatics 2018;19(1):233.
[34] Zhang W, Yue X, Liu F, Chen Y, Tu S, Zhang X. A unified frame of predicting side effects of drugs by using linear neighborhood similarity. BMC Syst Biol 2017;11(Suppl 6):101.
[35] Zhang W, Zou H, Luo L, Liu Q, Wu W, Xiao W. Predicting potential side effects of drugs by recommender methods and ensemble learning. Neurocomputing 2016;173:979-87.
[36] Hoffmann MF, Preissner SC, Nickel J, Dunkel M, Preissner R, Preissner S. The Transformer database: biotransformation of xenobiotics. Nucleic Acids Res 2014;42(Database issue):D1113-7.
[37] Mak L, Marcus D, Howlett A, Yarova G, Duchateau G, Klaffke W, *et al*. Metrabase: a cheminformatics and bioinformatics database for small molecule transporter data analysis and (Q)SAR modeling. J Cheminform 2015;7:31.
[38] O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An open chemical toolbox. J Cheminform 2011;3:33.
[39] Hatherley R, Brown DK, Musyoka TM, Penkler DL, Faya N, Lobb KA, *et al*. SANCDB: a South African natural compound database. J Cheminform 2015;7:29.
[40] Keum J, Yoo S, Lee D, Nam H. Prediction of compound-target interactions of natural products using large-scale drug and protein information. BMC Bioinformatics 2016;17 Suppl 6:219.
[41] Speck-Planche A, Cordeiro MN. Review of current chemoinformatic tools for modeling important aspects of CYPs-mediated drug metabolism. Integrating metabolism data with other biological profiles to enhance drug discovery. Current drug metabolism 2014;15(4):429-40.
[42] Marrero-Ponce Y. Linear indices of the "molecular pseudograph's atom adjacency matrix": definition, significance-interpretation, and application to QSAR analysis of flavone derivatives as HIV-1 integrase inhibitors. J Chem Inf Comput Sci 2004;44(6):2010-26.
[43] Shin WH, Zhu X, Bures MG, Kihara D. Three-dimensional compound comparison methods and their application in drug discovery. Molecules 2015;20(7):12841-62.
[44] Hu B, Zhu X, Monroe L, Bures MG, Kihara D. PL-PatchSurfer: a novel molecular local surface-based method for exploring protein-ligand interactions. International journal of molecular sciences 2014;15(9):15122-45.
[45] Venkatraman V, Chakravarthy PR, Kihara D. Application of 3D Zernike descriptors to shape-based ligand similarity searching. J Cheminform 2009;1:19.
[46] Zhu X, Shin WH, Kim H, Kihara D. Combined Approach of Patch-Surfer and PL-PatchSurfer for Protein-Ligand Binding Prediction in CSAR 2013 and 2014. J Chem Inf Model 2016;56(6):1088-99.
[47] Shin WH, Bures MG, Kihara D. PatchSurfers: Two methods for local molecular property-based binding ligand prediction. Methods 2016;93:41-50.
[48] Xu Q, Xiong Y, Dai H, Kumari KM, Xu Q, Ou HY, *et al*. PDC-SGB: Prediction of effective drug combinations using a stochastic gradient boosting algorithm. J Theor Biol 2017;417:1-7.
[49] Xiong Y, Liu J, Zhang W, Zeng T. Prediction of heme binding residues from protein sequences with integrative sequence profiles. Proteome science 2012;10 Suppl 1:S20.
[50] Yao Y, Zhang T, Xiong Y, Li L, Huo J, Wei DQ. Mutation probability of cytochrome P450 based on a genetic algorithm and support vector machine. Biotechnol J 2011;6(11):1367-76.
[51] Xiong Y, Xia J, Zhang W, Liu J. Exploiting a reduced set of weighted average features to improve prediction of DNA-binding residues from 3D structures. PLoS One 2011;6(12):e28440.
[52] Niu Y, Zhang W. Quantitative prediction of drug side effects based on drug-related features. Interdisciplinary sciences, computational life sciences 2017;9(3):434-44.
[53] Feng P, Chen W, Lin H. Identifying Antioxidant Proteins by Using Optimal Dipeptide Compositions. Interdisciplinary sciences, computational life sciences 2016;8(2):186-91.
[54] Zou Q, Wan S, Ju Y, Tang J, Zeng X. Pretata: predicting TATA binding proteins with novel features and dimensionality reduction strategy. Bmc Systems Biology 2016;10(4):114.
[55] Zou Q, Zeng J, Cao L, Ji R. A novel features ranking metric with application to scalable visual and bioinformatics data classification. Neurocomputing 2016;173:346-54.

[56] Yu L, Sun X, Tian SW, Shi XY, Yan YL. Drug and Nondrug Classification Based on Deep Learning with Various Feature Selection Strategies. Current Bioinformatics 2018;13(3):253-9.
[57] Qiao Y, Xiong Y, Gao H, Zhu X, Chen P. Protein-protein interface hot spots prediction based on a hybrid feature selection strategy. BMC Bioinformatics 2018;19(1):14.
[58] Dai H, Xu Q, Xiong Y, Liu WL, Wei DQ. Improved Prediction of Michaelis Constants in CYP450-Mediated Reactions by Resilient Back Propagation Algorithm. Current drug metabolism 2016;17(7):673-80.
[59] Li D, Ju Y, Zou Q. Protein Folds Prediction with Hierarchical Structured SVM. Current Proteomics 2016;13(2):79-85.
[60] Soyemi J, Isewon I, Oyelade J, Adebiyi E. Inter-Species/Host-Parasite Protein Interaction Predictions Reviewed. Current Bioinformatics 2018;13(4):396-406.
[61] Xia JF, Zhao XM, Song J, Huang DS. APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility. BMC Bioinformatics 2010;11(1):174.
[62] Xiong Y, Liu J, Wei DQ. An accurate feature-based method for identifying DNA-binding residues on protein surfaces. Proteins 2011;79(2):509-17.
[63] Sun Y, Xiong Y, Xu Q, Wei D. A hadoop-based method to predict potential effective drug combination. Biomed Res Int 2014;2014:196858.
[64] Wang W, Liu J, Xiong Y, Zhu L, Zhou X. Analysis and classification of DNA-binding sites in single-stranded and double-stranded DNA-binding proteins using protein information. IET Syst Biol 2014;8(4):176-83.
[65] Korolev D, Balakin KV, Nikolsky Y, Kirillov E, Ivanenkov YA, Savchuk NP, *et al*. Modeling of human cytochrome p450-mediated drug metabolism using unsupervised machine learning approach. J Med Chem 2003;46(17):3631-43.
[66] Zou Q, Chen W, Huang Y, Liu X, Jiang Y. Identifying Multi-Functional Enzyme by Hierarchical Multi-Label Classifier. Journal Of Computational And Theoretical Nanoscience 2013;10(4):1038-43.
[67] Zhang W, Zhu X, Fu Y, Tsuji J, Weng Z. Predicting human splicing branchpoints by combining sequence-derived features and multi-label learning methods. BMC Bioinformatics 2017;18(Suppl 13):464.
[68] You R, Zhang Z, Xiong Y, Sun F, Mamitsuka H, Zhu S. GOLabeler: improving sequence-based large-scale protein function prediction by learning to rank. Bioinformatics 2018;34(14):2465-73.
[69] Zhang W, Liu F, Luo L, Zhang J. Predicting drug side effects by multi-label learning and ensemble learning. BMC Bioinformatics 2015;16:365.
[70] Yuan Q, Gao J, Wu D, Zhang S, Mamitsuka H, Zhu S. DrugE-Rank: improving drug-target interaction prediction of new candidate drugs or targets by ensemble learning to rank. Bioinformatics 2016;32(12):i18-i27.
[71] Zhang ML, Zhou ZH. ML-KNN: A lazy learning approach to multi-label leaming. Pattern Recognition 2007;40(7):2038-48.
[72] Lee CP, Lin CJ. Large-scale linear rankSVM. Neural Comput 2014;26(4):781-817.
[73] Chen W, Feng PM, Lin H, Chou KC. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Res 2013;41(6):e68.
[74] Chen W, Feng P, Lin H. Prediction of ketoacyl synthase family using reduced amino acid alphabets. J Ind Microbiol Biotechnol 2012;39(4):579-84.
[75] Bai LY, Dai H, Xu Q, Junaid M, Peng SL, Zhu X, *et al*. Prediction of Effective Drug Combinations by an Improved Naive Bayesian Algorithm. International journal of molecular sciences 2018;19(2).
[76] Chou KC. Some remarks on protein attribute prediction and pseudo amino acid composition. J Theor Biol 2011;273(1):236-47.
[77] Feng P, Zhang J, Tang H, Chen W, Lin H. Predicting the Organelle Location of Noncoding RNAs Using Pseudo Nucleotide Compositions. Interdisciplinary sciences, computational life sciences 2017;9(4):540-4.