Computational Prediction of the Isoform Specificity of Cytochrome P450 Substrates by an Improved Bayesian Method

Hao Dai

Center for Excellence in Molecular Cell Science, Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, Shanghai, China

Yu-Xi Zheng

State Key Laboratory of Microbial Metabolism, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China

Xiao-Qi Shan

State Key Laboratory of Microbial Metabolism, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China

Yan-Yi Chu

State Key Laboratory of Microbial Metabolism, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China

Wei Wang

School of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai, China

Yi Xiong (✉ xiongyi@sjtu.edu.cn)

State Key Laboratory of Microbial Metabolism, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China <https://orcid.org/0000-0003-2910-6725>

Dong-Qing Wei (✉ dqwei@sjtu.edu.cn)

State Key Laboratory of Microbial Metabolism, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China

Research Article

Keywords: Cytochrome P450, Substrate-specificity, Naïve Bayes, Improved Bayesian Method

Posted Date: May 23rd, 2019

DOI: <https://doi.org/10.21203/rs.2.9738/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Cytochrome P450 (CYP) is the most important drug-metabolizing enzyme in human beings. Each CYP isoform is able to metabolize a large number of compounds, and if patients take more than one drugs during the treatment, it is possible that some drugs would be metabolized by the same CYP isoform, leading to potential drug-drug interactions and side effects. Therefore, it is necessary to investigate the isoform specificity of CYP substrates. In this study, we constructed a data set consisting of 10 major CYP isoforms associated with 776 substrates, and used machine learning methods to construct the predictive models based on the features of structural and physicochemical properties of substrates. We also proposed a new method called Improved Bayesian method, which is suitable for small data sets and is able to construct more stable and accurate predictive models compared with other traditional machine learning models. Based on this method, the predictive performance of our method got the accuracy of 86% for the independent test, which was significantly better to the existing models. We believe that our proposed method will facilitate the understanding of drug metabolisms and help the large-scale analysis of drug-drug interactions.

Introduction

Cytochrome P450 (CYP450, CYP) is the most important drug-metabolizing enzymes in human beings [1], which involves in the metabolism of more than 90% clinical drugs [2, 3]. CYP450 is a superfamily of monooxygenases, which contains heme and can increase the water solubility and polarity of drugs to make them easy to excrete by oxidation reaction. 57 CYP isoforms have been discovered in human beings[4], and based on the similarity of amino acids, CYP450 can be divided to different families and subfamilies, in which xenobiotics are mainly metabolized by the families of CYP1, CYP2 and CYP3 [5]. The number of CYP isoforms is limited but the xenobiotics are various, so each CYP isoform is able to metabolize large number of compounds, which is quite different from the high specificity of most enzymes. In many cases, people will take more than one drugs during the treatment. Then it is possible that some drugs are metabolized by the same CYP isoform and lead to drug-drug interactions [6, 7], in which the safety and efficacy of a drug can be altered by the co-administration of another drug, and this may result in serious side effects [6] and even death of the patients [8, 9]. Therefore, it is quite important to know which CYP isoform will metabolize a drug candidate in the drug discovery stage [10].

In the past decades, both experimental and computational approaches have been developed to identify the isoform specificity of CYP450 substrate. Experimental assay is accurate but the process such as the bioluminescent and fluorescent assays [11] are quite resource-consuming. A variety of computational models including structure-based and feature-based methods have also been developed, in which the structure-based methods such as molecular dynamics simulation [12-14], quantum mechanical methods [15, 16] and pharmacophore modeling [17, 18] are based on the crystal structures of CYPs. However, some CYP structures are still unavailable, and because of the heme in the active site, the force field is difficult to construct, which makes these methods quite complex and time-consuming. The feature-based methods employ various physiochemical and structural features of substrates to predict which CYP

isoforms will participate in the metabolism reactions based on machine learning technologies, which have been widely used in bioinformatics [19-21] and drug discovery [22-24]. These methods only employ the structures of substrates and are easy to develop and use, which are suitable for the analysis of drug metabolism [25] and drug-drug interactions [26] in large scale.

A large number of predictive models have been developed to predict the isoform specificity of CYP450 substrate using various machine learning techniques such as logistic regression, decision tree, naïve Bayes, k-nearest neighbors (KNN), support vector machine (SVM) and artificial neural network (ANN) [27-32]. Block and Henry [27] constructed independent classifiers using the descriptors including molecular connectivity, E-state indices, Spartan descriptors and structural keys, but these classifier cannot get the isoform specificity straight and the accuracy is just about 70-80%. Michielan et al. [28] developed several single-label and multi-label classification models with the accuracy of 75-95%, but the results seemed to come from the large number of negative samples because the sensitivity and precision was usually low and the more positive samples were used, the worse predictive performance was gotten. Zhang et al. [32] used decision tree, BP neural network and ML-kNN to construct classification models, and if a substrate was metabolized by only one CYP isoform, the predictive performance is more than 90%, but if a substrate could be metabolized by more CYP isoforms, the performance would be down obviously. And above of all, these studies used small number of samples to construct the models, which limited the application fields and cause difficulties when the approaches were scaled to larger data set.

In this paper, we developed a new method called Improved Bayesian method, which not only used the samples whose biological activities are known, but also the samples whose activities are unknown, and two-steps training was used in the model construction. This method may be quite suitable for small data set, and is able to construct more stable and accurate predictive models to small data set compared with other machine learning models. Based on the Improved Bayesian method, we used the metabolism information of 776 compounds and up to 10 CYP isoforms (CYP 1A1, 1A2, 2A6, 2B6, 2C8, 2C9, 2C19, 2D6, 2E1, 3A4) from Meta-CYP database to construct the predictive models of isoform specificity of CYP450 substrates. Compared with naïve Bayesian method, k-nearest neighbors and support vector machine, we found the predictive performance of our method was obviously better than others and got the accuracy of 86% for the independent test, which was a significant improvement to the existing models.

Results

Feature selection

In this paper, a total of 176 descriptors representing various structural and physicochemical properties of substrates were subject to the feature selection. As we constructed 10 separate models for different CYP isoforms, the selected descriptors were different for each model. As shown in Figure 1, CYP2A6 and CYP2E1 prefer to metabolize small molecules and CYP3A4 prefers to metabolize large molecules, but it has little correlation between molecular weight and whether a molecule can be metabolized by CYP2D6,

so in the model of CYP2A6, 2E1 and 3A4, the descriptor molecular weight was selected, but in the model of CYP2D6, this descriptor was not selected.

Some descriptors were selected by most models. As shown in Table 1, the descriptors about molecular size and shape seem to appear frequently, such as *chi0*, *VDistMa*, *weinerPol*, *Kier1*, *zagreb* and *Std_dim3*, which implies molecular size and shape may have significant influence over the isoform specificity of CYP substrates. Other features such as hydrogen bond or molecular surface area seem also important, but the descriptor *logP* (Log of the octanol/water partition coefficient) did not be selected by any model, which is quite different from the studies on the other biological activities of CYP [18, 33-38]. Maybe the mechanism of isoform specificity of CYP substrates is somewhat different from other biological activities.

Model comparison

Four methods including Improved Bayesian method, naïve Bayesian method, k-nearest neighbors (kNN) and support vector machine (SVM) were used to construct our classification models based on the validation set. The selected descriptors were used as input and whether a molecule can be metabolize by the CYP isoform was output (represented by 1 or 0). Leave-one-out cross-validation was used to evaluate the predictive performance, and the result is represented by AUC of ROC curve, which is shown in Table 2, Figure 2 and Figure 3.

We can see Improved Bayesian method is superior to other methods in all models, while naïve Bayesian method seems to make the worst predictive performance. The results support that our method is a significant improvement to naïve Bayesian method and it is even better than some other machine learning methods such as SVM and kNN. The predictive performance also seems different among the models of different CYP isoforms. The accuracy of CYP2C8, 2C9 and 2C19 is lower than other models while CYP1A1, 2D6 and 2E1 make the best performance. This is possibly due to the different selected descriptors in each model.

Independent test

Besides the leave-one-out cross validation test, we also conducted an independent test on the models constructed by the improved Bayesian method, which simulates a true prediction since the model trained on one dataset is tested on an unseen dataset. The 600 molecules in the validation set were used to train the models and the other 176 molecules were used to test the models. As shown in Table 3, the predictive performance is similar to those of the leave-one-out cross validation test and the accuracy is up to 86% on average, suggesting that our method performed well on the independent test set.

Discussion

As the major drug-metabolizing enzymes in human beings, cytochrome P450s have been considered as the important factor of drug-drug interactions [6]. In many cases, people will take multiple drugs during the treatment, and if two drugs are metabolized by the same CYP isoform, mutual inhibition may occur and the blood concentration will abnormally increase when the patient still follow the scheduled treatment doses, which may lead to serious side effect [6]. In China, tradition Chinese medicines (TCM) are widely used, and sometimes, doctors will prescribe TCM combined with other medicines to improve the curative effect [39], but the problem of drug-drug interactions is usually neglected. The absence of rapid analysis tools is obviously a significant reason, and the computational methods with high accuracy are certainly a good solution.

In this work, we constructed novel predictive models to predict the isoform specificity of CYP substrates based on Improved Bayesian method. This method not only used the samples whose biological activities are known, but also used the samples whose activities are unknown (background samples). Compared with naïve Bayesian method, the relationship of features is taken into account and we no longer need to assume the independence of conditional probability. Compared with most machine learning methods, as the relationship of features is determined by the background samples, only small amount of model variables will be changed during the training, which may lead to more stable and reliable predictive performance and less probability of overfitting. In biostatistics and bioinformatics, we often face the dataset with small sample size but large number of features, which is usually hard to analyze, but with the help of background samples, we may construct much better models. Of course, there is still some limitation in our method. We have to assume that the correlation of all features is independent of the biological activity. If this assumption is not tenable, our method is inappropriate.

The common CYP isoform of two drugs is certainly a cause of drug-drug interactions, but there are still many other causes [40]. For example, some drugs can inhibit or induce the activity of CYP isoform straight, and these drugs are called inhibitors or inducers. Some drugs can interact with the transcription factors and influence the expression of CYP isoforms. All these phenomena increase the difficulty in accurate prediction of drug-drug interactions caused by CYPs, which need further investigation in future.

Methods

Improved Bayesian method

In this paper, we will introduce a new method to construct our classification models, which is called Improved Bayesian method. In bioinformatics, due to the limitation of experimental technology, we often face the dataset with small sample size. For example, if we had 1000 drug candidates, maybe only 100 candidates were known the corresponding CYP isoforms that metabolize them, and the activities of the other 900 candidates were unknown. In most machine learning methods, we will only use the 100 compounds as samples to construct the model, but in our method, we will use the other 900 samples as

well. In other words, we will not only use the samples whose biological activities are known (activity samples), but also the samples whose biological activities are unknown (background samples).

As shown in Figure 4, the model will be trained by two steps. First, the model is trained by the background samples, and we can get the distribution of each feature and correlations of all features. Next, after training with the activity samples, we can get the relationship between the target and all features. Thus, the activity samples will just change a small amount of variables in the model, which may lead to more stable and reliable predictive performance and less probability of over-fitting. We do not have to construct a simple model for a small data set, and a more accurate model may be built.

We will introduce the procedures of our Improved Bayesian method in detail.

(a) Make all features follow normal distribution similarly, and calculate the correlation coefficient matrix $\mathbf{R}$ based on the normalized features

Denote $F(x)$ is the cumulative distribution function of feature X . $\Phi(x)$ and $\Phi^{-1}(x)$ are the distribution function and inverse distribution function of standard normal distribution respectively. Let

$$\hat{X} = \Phi^{-1}(F(x)) \quad (1)$$

Then

$$P(\hat{X} \leq x) = P(\Phi^{-1}(F(X)) \leq x) = P(X \leq F^{-1}(\Phi(x))) = F(F^{-1}(\Phi(x))) = \Phi(x) \quad (2)$$

Thus, $\hat{X}$ follows standard normal distribution. Note that $F(x)$ is calculated based on the background samples.

Next, we need to calculate the correlation coefficient matrix $\mathbf{R}$ based on the normalized features. Of course, $\mathbf{R}$ is also calculated based on the background samples. Take note that the correlation coefficient may be meaningless if the variables do not follow normal distribution, so the normalization procedure is essential.

(b) Estimate the joint distribution function of all features

Here, we need two basic assumptions in our Improved Bayesian method. First, the correlation coefficient of two features is not dependent on the biological activity. For example, if a molecule has more atoms, its molecular weight is usually larger, and the positive correlation between the atomic number and molecular weight is irrelevant to the CYP specificity of these molecules. Second, if random variables $X_1, X_2, \dots, X_n$

follow normal distribution, n dimensional random variable $(X_1, X_2, \dots, X_n)$ follows n dimensional normal distribution approximately. Obviously, the two assumptions are reasonable and easy to satisfy.

Based on the assumptions, the joint distribution function of $(X_1, X_2, \dots, X_n)$ is

$$\begin{aligned}
 & F(x_1, x_2, \dots, x_n) \\
 &= P(X_1 \leq x_1, \dots, X_n \leq x_n) = P\left(F_1^{-1}\left(\Phi\left(\hat{X}_1\right)\right) \leq x_1, \dots, F_n^{-1}\left(\Phi\left(\hat{X}_n\right)\right) \leq x_n\right) \\
 &= P\left(\hat{X}_1 \leq \Phi^{-1}\left(F_1(x_1)\right), \dots, \hat{X}_n \leq \Phi^{-1}\left(F_n(x_n)\right)\right) \approx \Phi_n\left(\Phi^{-1}\left(F_1(x_1)\right), \dots, \Phi^{-1}\left(F_n(x_n)\right)\right)
 \end{aligned} \tag{3}$$

(c) Construct the classification models based on the activity samples.

Denote $P(A)$ is the probability of event A . Each sample is described by n features $X_1, X_2, \dots, X_n$, and belongs to one of the two classes C_0 or C_1 . Then, for a sample $x = (x_1, x_2, \dots, x_n)$, we can calculate the probability that it belongs to C_k ($k = 0$ or $k = 1$) based on Bayesian formula:

$$P\left(C_k \mid X_1 = x_1, \dots, X_n = x_n\right) = \frac{P\left(C_k\right) P\left(X_1 = x_1, \dots, X_n = x_n \mid C_k\right)}{P\left(X_1 = x_1, \dots, X_n = x_n\right)} \tag{4}$$

In naïve Bayesian classification, we should assume the independence of conditional probability, and then

$$P\left(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n \mid C_k\right) = \prod_{i=1}^n P\left(X_i = x_i \mid C_k\right) \tag{5}$$

But in our method, based on equation (1), (3), (4), we can calculate the probability straight.

(6)

$$\begin{aligned}
P(C_k | X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) &= \frac{P(C_k) f(x_1, x_2, \dots, x_n | C_k)}{f(x_1, x_2, \dots, x_n)} \\
&= P(C_k) \frac{\frac{\partial^n}{\partial x_1 \dots \partial x_n} F(x_1, x_2, \dots, x_n | C_k)}{\frac{\partial^n}{\partial x_1 \dots \partial x_n} F(x_1, x_2, \dots, x_n)} \\
&\approx P(C_k) \frac{\frac{\partial^n}{\partial x_1 \dots \partial x_n} \Phi_n(\Phi^{-1}(F_1(x_1)), \dots, \Phi^{-1}(F_n(x_n)) | C_k)}{\frac{\partial^n}{\partial x_1 \dots \partial x_n} \Phi_n(\Phi^{-1}(F_1(x_1)), \dots, \Phi^{-1}(F_n(x_n)))} \\
&= P(C_k) \frac{\varphi_n(\Phi^{-1}(F_1(x_1)), \dots, \Phi^{-1}(F_n(x_n)) | C_k) \cdot \prod_{i=1}^n \frac{f_i(x_i)}{\varphi(\Phi^{-1}(F_i(x_i)))}}{\varphi_n(\Phi^{-1}(F_1(x_1)), \dots, \Phi^{-1}(F_n(x_n))) \cdot \prod_{i=1}^n \frac{f_i(x_i)}{\varphi(\Phi^{-1}(F_i(x_i)))}} \\
&= P(C_k) \frac{\varphi_n(\hat{x}_1, \hat{x}_2, \dots, \hat{x}_n | C_k)}{\varphi_n(\hat{x}_1, \hat{x}_2, \dots, \hat{x}_n)}
\end{aligned}$$

where $F_i(x)$ and $f_i(x)$ are the distribution function and density function of X_i respectively. $\Phi(x)$ and $\varphi(x)$ are the distribution function and density function of standard normal distribution respectively. $\Phi_n(x)$ and $\varphi_n(x)$ are the joint distribution function and joint density function of n dimensional normal distribution respectively.

Take note that $\varphi_n(x)$ can be expressed as

(7)

$$\varphi_n(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} \|\boldsymbol{\Sigma}\| \|\mathbf{R}\|^{\frac{1}{2}}} \exp \left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}) \boldsymbol{\Sigma}^{-1} \mathbf{R}^{-1} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})^T \right]$$

where $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_n)$ is standard deviation matrix, $\mu = (\mu_1, \dots, \mu_n)$ is mean vector, R is correlation coefficient matrix and $x = (x_1, x_2, \dots, x_n)$.

As R is real symmetric matrix, it must exist orthogonal matrix $P = (\alpha_1, \alpha_2, \dots, \alpha_n)$, which makes

$$P^T R P = \Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n) \quad (8)$$

and

$$R^{-1} = P \Lambda^{-1} P^T \quad (9)$$

where λ_i and α_i are the eigenvalues and corresponding eigenvectors of R

Let

$$\beta = (\beta_1, \beta_2, \dots, \beta_n) = (\hat{x} - \mu) \Sigma^{-1} P = \begin{pmatrix} \frac{\hat{x}_1 - \mu_1}{\sigma_1} & \frac{\hat{x}_2 - \mu_2}{\sigma_2} & \dots & \frac{\hat{x}_n - \mu_n}{\sigma_n} \end{pmatrix} P \quad (10)$$

$$\gamma = (\gamma_1, \gamma_2, \dots, \gamma_n) = (\hat{x} - \mu_k) \Sigma_k^{-1} P = \begin{pmatrix} \frac{\hat{x}_1 - \mu_{k1}}{\sigma_{k1}} & \frac{\hat{x}_2 - \mu_{k2}}{\sigma_{k2}} & \dots & \frac{\hat{x}_n - \mu_{kn}}{\sigma_{kn}} \end{pmatrix} P \quad (11)$$

where μ_i and σ_i are the mean and the standard deviation of feature i in all samples respectively. μ_{ki} and σ_{ki} are the mean and the standard deviation of feature i in the samples belonging to class C_k .

Plug equation (7)-(11) into equation (6), and take logarithm, we can get

$$\begin{aligned} H_k &= \ln P(C_k | X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) \\ &= \frac{1}{2} \beta \Lambda^{-1} \beta^T - \frac{1}{2} \gamma \Lambda_k^{-1} \gamma^T + \ln P(C_k) \frac{|\Sigma|}{|\Sigma_k|} = \sum_{i=1}^n \left(\frac{\beta_i^2 - \gamma_i^2}{2\lambda_i} + \ln \frac{\sigma_i}{\sigma_{ki}} \right) + \ln P(C_k) \end{aligned} \quad (12)$$

Then, the larger H_k is, the more probability that $x = (x_1, x_2, \dots, x_n)$ belongs to C_k . We can compare H_0 and H_1 to determine whether x belongs to C_0 or C_1 , or we can plot receiver operating characteristic curve (ROC curve) and select the best threshold to determine the classification.

Dataset

We collected 776 compounds from the Meta-CYP database, and each compound can be metabolized by at least one of the ten CYP isoforms: CYP1A1, 1A2, 2A6, 2B6, 2C8, 2C9, 2C19, 2D6, 2E1 and 3A4. The number of compounds metabolized by each CYP isoform and their distribution are shown in Table 4. We can see the total number is up to 1646, which means each compound will be metabolized by about two CYP isoforms on average. The set of 776 compounds were randomly divided into two data sets. The first set was composed of 600 samples, taken as the validation set for feature selection and model construction. The rest 176 samples constituted the independent test set for the model test. Meta-CYP database and the analysis tools based on our proposed methods are currently under construction by our laboratory, in which the information used in this paper was mainly from SuperCYP (<http://bioinformatics.charite.de/supercyp/index.php>) [41], CYP450 Engineering Database (<https://cyped.biocatnet.de>) [42, 43] and PubChem Compound (<https://www.ncbi.nlm.nih.gov/pccompound>) [44]. The detailed information about this data set is available in Supplementary Material. In this paper, background samples were gotten from DrugBank database (<http://www.drugbank.ca>) [45, 46], including the approved, experimental and investigational drugs, and 7114 compounds in all.

It should be noted that inorganic compounds and polypeptides are not including in our dataset, because the features describing these compounds are quite different from other small molecules. So the models constructed in this study are not suitable for the two types of compounds.

Molecular descriptors calculation

We employed the molecular descriptors representing various structural and physicochemical properties of the compounds including atomic composition, molecular size, shape, weight, hydrogen bond, hydrophobicity, acidity or basicity, charge distribution and energy of chemical bonds to construct our models. To accurately calculate these molecular descriptors, the chemical structures of all compounds were first drawn by SAMM, an in-house molecular mechanics and drug design package. Next, all structures were further energy minimized by the MMFF94 force field[47]. At last, the molecular descriptors were calculated based on the energy-minimized structures by SAMM. After removing the descriptors that were highly correlated or had the same value in almost all compounds, 176 molecular descriptors including 134 two-dimensional and 42 three-dimensional descriptors were gotten and would be subject to further feature selection.

Feature selection

Feature selection is performed prior to the construction of the predictive models in order to eliminate the redundant features and identify the most relevant features [48-50]. Minimum Redundancy Maximum Relevance (MRMR)[51] was applied to select the best features in this paper, which is a powerful method for getting useful information from complex systems through maximizing the mutual information between the selected features and the classification target, and in the meanwhile, minimizing the correlation of the features themselves[52]. In this work, we selected the top 20 features with the highest scores in the model construction.

Model comparison and evaluation

Besides Improved Bayesian method, we also used naïve Bayesian method, k-nearest neighbors and support vector machine[53] to construct our classification models based on the validation set for model comparison. In naïve Bayesian method, the Laplace smoothing factor was 1. The parameters of k-nearest neighbors was $k = 5$, Euclidean distance and distance weighting function $= 1/d$. The kernel function in support vector machine was radial basis function and other parameters were optimized by LIBSVM[54].

For the constructed models, we would use AUC (area under the curve) of receiver operating characteristic curve (ROC) for model evaluation. Besides that, we would calculate specificity, sensitivity, accuracy as well.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \text{ Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \text{ Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

where TP, FP, TN, FN means true positive, false positive, true negative and false negative respectively.

Conclusions

In this study, we employed machine learning methods to predict the CYP isoform specificity of 776 compounds with their corresponding CYP enzymes, including CYP 1A1, 1A2, 2A6, 2B6, 2C8, 2C9, 2C19, 2D6, 2E1 and 3A4. The improved Bayesian method was proposed and applied in the model construction based on the features of structural and physicochemical properties of substrates. In comparison with the naïve Bayesian method, k-nearest neighbors and support vector machine, it was found that the predictive models built by our method achieved the best accuracy, which is up to 86% accuracy on average on the independent test. Using such a model, we can efficiently and reliably estimate the metabolic paths of a given compound involved in CYP-mediated metabolism of these ten CYPs. We believed that our proposed method will be potentially useful in large-scale *in silico* drug screening and drug-drug interaction analysis.

Acknowledgements

This work was supported by the fundings from National Key Research and Development Program of China (Contract No. 2016YFA0501703), and National Natural Science Foundation of China (Grant No. 31601074, 61872094, and 61832019).

References

1. Guengerich FP: **Cytochrome P450s and other enzymes in drug metabolism and toxicity.** *AAPS J* 2006, **8**(1):E101-111.
2. Tomaszewski P, Kubiak-Tomaszewska G, Pachecka J: **Cytochrome P450 polymorphism—molecular, metabolic, and pharmacogenetic aspects. II. Participation of CYP isoenzymes in the metabolism of endogenous substances and drugs.** *Acta Pol Pharm* 2008, **65**(3):307-318.

3. Evans WE, Relling MV: **Pharmacogenomics: translating functional genomics into rational therapeutics.** *Science* 1999, **286**(5439):487-491.
4. Lewis DFV: **57 varieties: the human cytochromes P450.** *Pharmacogenomics* 2004, **5**(3):305-318.
5. Guengerich FP: **Cytochromes P450, drugs, and diseases.** *Mol Interv* 2003, **3**(4):194-204.
6. Ito K, Iwatsubo T, Kanamitsu S, Ueda K, Suzuki H, Sugiyama Y: **Prediction of pharmacokinetic alterations caused by drug-drug interactions: metabolic interaction in the liver.** *Pharmacol Rev* 1998, **50**(3):387-411.
7. Chen L, Chu C, Zhang YH, Zheng MY, Zhu LC, Kong XY, Huang T: **Identification of Drug-Drug Interactions Using Chemical Interactions.** *Current Bioinformatics* 2017, **12**(6):526-534.
8. Marroum PJ, Uppoor RS, Parmelee T, Ajayi F, Burnett A, Yuan R, Svadjian R, Lesko LJ, Balian JD: **In vivo drug-drug interaction studies-A survey of all new molecular entities approved from 1987 to 1997.** *Clin Pharmacol Ther* 2000, **68**(3):280-285.
9. Yuan R, Parmelee T, Balian JD, Uppoor RS, Ajayi F, Burnett A, Lesko LJ, Marroum P: **In vitro metabolic interaction studies: Experience of the Food and Drug Administration&ast.** *Clin Pharmacol Ther* 1999, **66**(1):9-15.
10. Lounkine E, Keiser MJ, Whitebread S, Mikhailov D, Hamon J, Jenkins JL, Lavan P, Weber E, Doak AK, Cote S *et al*: **Large-scale prediction and testing of drug activity on side-effect targets.** *Nature* 2012, **486**(7403):361-U391.
11. Veith H, Southall N, Huang R, James T, Fayne D, Artemenko N, Shen M, Inglese J, Austin CP, Lloyd DG *et al*: **Comprehensive characterization of cytochrome P450 isozyme selectivity across chemical libraries.** *Nat Biotechnol* 2009, **27**(11):1050-1055.
12. Park H, Lee S, Suh J: **Structural and dynamical basis of broad substrate specificity, catalytic mechanism, and inhibition of cytochrome P450 3A4.** *J Am Chem Soc* 2005, **127**(39):13634-13642.
13. Seifert A, Tatzel S, Schmid RD, Pleiss J: **Multiple molecular dynamics simulations of human p450 monooxygenase CYP2C9: the molecular basis of substrate binding and regioselectivity toward warfarin.** *Proteins* 2006, **64**(1):147-155.
14. Zhang T, Liu LA, Lewis DFV, Wei DQ: **Long-Range Effects of a Peripheral Mutation on the Enzymatic Activity of Cytochrome P450 1A2.** *J Chem Inf Model* 2011, **51**(6):1336-1346.
15. Friesner RA, Guallar V: **Ab initio quantum chemical and mixed quantum mechanics/molecular mechanics (QM/MM) methods for studying enzymatic catalysis.** *Annu Rev Phys Chem* 2005, **56**:389-427.
16. Bathelt CM, Zurek J, Mulholland AJ, Harvey JN: **Electronic structure of compound I in human isoforms of cytochrome P450 from QM/MM modeling.** *J Am Chem Soc* 2005, **127**(37):12900-12908.
17. de Groot MJ, Ekins S: **Pharmacophore modeling of cytochromes P450.** *Adv Drug Deliv Rev* 2002, **54**(3):367-383.
18. Ekins S, de Groot MJ, Jones JP: **Pharmacophore and three-dimensional quantitative structure activity relationship methods for modeling cytochrome P450 active sites.** *Drug Metab Dispos* 2001,

29(7):936-944.

19. Liu B, Long R, Chou KC: **iDHS-EL: identifying DNase I hypersensitive sites by fusing three different modes of pseudo nucleotide composition into an ensemble learning framework.** *Bioinformatics* 2016, **32**(16):2411-2418.
20. Liu B, Wang S, Wang X: **DNA binding protein identification by combining pseudo amino acid composition and profile-based protein representation.** *Sci Rep* 2015, **5**:15479.
21. Zou Q, Xing P, Wei L, Liu B: **Gene2vec: Gene Subsequence Embedding for Prediction of Mammalian N6-Methyladenosine Sites from mRNA.** *RNA* 2019, **25**(2):205-218.
22. Zhu X, Xiong Y, Kihara D: **Large-scale binding ligand prediction by improved patch-based method Patch-Surfer2.0.** *Bioinformatics* 2015, **31**(5):707-713.
23. Shin WH, Zhu X, Bures MG, Kihara D: **Three-dimensional compound comparison methods and their application in drug discovery.** *Molecules* 2015, **20**(7):12841-12862.
24. Liu H, Luo LB, Cheng ZZ, Sun JJ, Guan JH, Zheng J, Zhou SG: **Group-sparse Modeling Drug-kinase Networks for Predicting Combinatorial Drug Sensitivity in Cancer Cells.** *Current Bioinformatics* 2018, **13**(5):437-443.
25. Crivori P, Poggesi I: **Computational approaches for predicting CYP-related metabolism properties in the screening of new drugs.** *Eur J Med Chem* 2006, **41**(7):795-808.
26. 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.
27. Block JH, Henry DR: **Evaluation of descriptors and classification schemes to predict cytochrome substrates in terms of chemical information.** *J Comput Aided Mol Des* 2008, **22**(6-7):385-392.
28. 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):2588-2605.
29. 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):1688-1701.
30. 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):364-369.
31. 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):982-992.
32. Zhang T, Dai H, Liu LA, Lewis DFW, Wei DQ: **Classification Models for Predicting Cytochrome P450 Enzyme-Substrate Selectivity.** *Mol Inf* 2012, **31**(1):53–62.
33. Klon AE: **Machine learning algorithms for the prediction of hERG and CYP450 binding in drug development.** *Expert Opin Drug Metab Toxicol* 2010, **6**(7):821-833.
34. Lewis DFW, Eddershaw PJ, Dickins M, Tarbit MH, Goldfarb PS: **Structural determinants of cytochrome P450 substrate specificity, binding affinity and catalytic rate.** *Chem Biol Interact* 1998, **115**(3):175-

35. Mishra NK: **Computational modeling of P450s for toxicity prediction.** *Expert Opin Drug Metab Toxicol* 2011, **7**(10):1211-1231.
36. Novotarskyi S, Sushko I, Korner R, Pandey AK, Tetko IV: **A comparison of different QSAR approaches to modeling CYP450 1A2 inhibition.** *J Chem Inf Model* 2011, **51**(6):1271-1280.
37. Smith DA, Ackland MJ, Jones BC: **Properties of cytochrome P450 isoenzymes and their substrates part 2: properties of cytochrome P450 substrates.** *Drug Discov Today* 1997, **2**(11):479-486.
38. Szklarz GD, Paulsen MD: **Molecular modeling of cytochrome P450 1A1: enzyme-substrate interactions and substrate binding affinities.** *J Biomol Struct Dyn* 2002, **20**(2):155-162.
39. Chen KC, Lu R, Iqbal U, Hsu KC, Chen BL, Nguyen PA, Yang HC, Huang CW, Li YC, Jian WS *et al.*: **Interactions between traditional Chinese medicine and western drugs in Taiwan: A population-based study.** *Comput Methods Programs Biomed* 2015, **122**(3):462-470.
40. Pelkonen O, Turpeinen M, Hakkola J, Honkakoski P, Hukkanen J, Raunio H: **Inhibition and induction of human cytochrome P450 enzymes: current status.** *Archives of Toxicology* 2008, **82**(10):667-715.
41. Preissner S, Kroll K, Dunkel M, Senger C, Goldsobel G, Kuzman D, Guenther S, Winnenburg R, Schroeder M, Preissner R: **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-D243.
42. Fischer M, Knoll M, Sirim D, Wagner F, Funke S, Pleiss J: **The Cytochrome P450 Engineering Database: a navigation and prediction tool for the cytochrome P450 protein family.** *Bioinformatics* 2007, **23**(15):2015-2017.
43. Sirim D, Wagner F, Lisitsa A, Pleiss J: **The cytochrome P450 engineering database: Integration of biochemical properties.** *BMC biochemistry* 2009, **10**(27).
44. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han L, He J, He S, Shoemaker BA *et al.*: **PubChem Substance and Compound databases.** *Nucleic Acids Res* 2016, **44**(D1):D1202-D1213.
45. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M: **DrugBank: a knowledgebase for drugs, drug actions and drug targets.** *Nucleic Acids Res* 2008, **36** Suppl 1:D901-D906.
46. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J: **DrugBank: a comprehensive resource for in silico drug discovery and exploration.** *Nucleic Acids Res* 2006, **34**(Database issue):D668-672.
47. Halgren TA: **Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94.** *J Comput Chem* 1996, **17**(5/6):490-519.
48. Fox T, Kriegl JM: **Machine learning techniques for in silico modeling of drug metabolism.** *Curr Top Med Chem* 2006, **6**(15):1579-1591.
49. 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**:346-354.

50. 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.
51. Peng H, Long F, Ding C: **Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy.** *IEEE Trans Pattern Anal Mach Intell* 2005, **27**(8):1226-1238.
52. He Z, Zhang J, Shi XH, Hu LL, Kong X, Cai YD, Chou KC: **Predicting drug-target interaction networks based on functional groups and biological features.** *PLoS One* 2010, **5**(3):e9603.
53. Wu XD, Kumar V, Quinlan JR, Ghosh J: **Top 10 algorithms in data mining.** *Knowl Inf Syst* 2008, **14**:1-37.
54. Chang CC, Lin CJ: **LIBSVM : a library for support vector machines.** *ACM Trans Intell Syst Technol* 2011, **2**(27):1-27.
55. **CRC Handbook of Chemistry and Physics.** Boca Raton, FL: CRC Press; 1994.
56. Hall LH, Kier LB: **The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure-Property Modeling.** In: *Reviews in computational chemistry*. vol. 2. New York: VCH Publishers; 1991: 367-422.
57. Hall LH, Kier LB: **The Nature of Structure-Activity Relationships and Their Relation to Molecular Connectivity.** *Eur J Med Chem* 1997, **4**:307-312.
58. Balaban AT: **Five New Topological Indices for the Branching of Tree-Like Graphs.** *Theor Chim Acta* 1979, **53**:355-375.
59. Gasteiger J, Marsili M: **Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges.** *Tetrahedron* 1980, **36**(22):3219-3228.

Tables

Table 1. The features selected by most models and their descriptions

Name	Description
<i>a_heavy</i>	Number of heavy atoms
<i>a_acc</i>	Number of hydrogen bond acceptor atoms
<i>a_IC</i>	Atom information content (total). Let n_i be the number of occurrences of atomic number i in the molecule, $p_i = n_i / n$ where n is the sum of the n_i . Then $a_IC = -n \sum p_i \log p_i$
<i>b_single</i>	Number of single bonds (not including aromatic bonds).
<i>bpol</i>	Sum of the absolute value of the difference between atomic polarizabilities[55] of all bonded atoms in the molecule
<i>chi0</i>	Atomic connectivity index (order 0) [56, 57]. The sum of $1/\sqrt{d_i}$ over all heavy atoms i with $d_i > 0$, where d_i is the number of bonded neighbors of the atom i in the hydrogen suppressed graph
<i>VDistMa</i>	Sum of $\log_2 m - D_{ij} \log_2 D_{ij} / m$ over all i and j , where D_{ij} is the length of the shortest path from atoms i to j , and m is the sum of D_{ij}
<i>weinerPol</i>	Wiener polarity number: half the sum of all the distance matrix entries with a value of 3 as defined in [58]
<i>Kier1</i>	First kappa shape index: $(n-1)^2 / m^2$, where n and m denotes the number of atoms and the number of bonds in the hydrogen suppressed graph respectively[56]
<i>PEOE_VSA_PPOS</i>	Total positive polar van der Waals surface area. Sum of the v_i such that q_i is greater than 0.2. The v_i is calculated using a connection table approximation and q_i is the partial charge of atom i calculated by the Partial Equalization of Orbital Electronegativities (PEOE) method[59]
<i>vdw_area</i>	Area of van der Waals surface
<i>mr</i>	Molecular refractivity
<i>zagreb</i>	Zagreb index: the sum of d_i^2 over all heavy atoms i
<i>Std_dim3</i>	Standard dimension 3: the square root of the third largest eigenvalue of the covariance matrix of the atomic coordinates. A standard dimension is equivalent to the standard deviation along a principal component axis

Table 2. The validation result of ten CYP predictive models by four methods, represented by AUC.

	1A1	1A2	2A6	2B6	2C19	2C8	2C9	2D6	2E1	3A4
Improved Bayesian method	0.92	0.84	0.87	0.88	0.84	0.80	0.83	0.92	0.91	0.88
naïve Bayesian method	0.79	0.70	0.81	0.76	0.72	0.69	0.71	0.82	0.85	0.82
kNN	0.82	0.75	0.84	0.80	0.75	0.73	0.77	0.84	0.88	0.81
SVM	0.85	0.73	0.83	0.80	0.75	0.76	0.79	0.88	0.89	0.82

Table 3. The results of independent test set calculated by the improved Bayesian method

	1A1	1A2	2A6	2B6	2C19	2C8	2C9	2D6	2E1	3A4	Avg.
Specificity	0.95	0.93	0.95	0.94	0.87	0.91	0.90	0.86	0.95	0.72	0.90
Sensitivity	0.58	0.48	0.75	0.53	0.71	0.56	0.66	0.76	0.75	0.86	0.66
Accuracy	0.90	0.81	0.93	0.86	0.85	0.86	0.85	0.83	0.93	0.78	0.86

Table 4. The number of substrates metabolized by each CYP450 isoform

1A1	1A2	2A6	2B6	2C19	2C8	2C9	2D6	2E1	3A4	Total
87	167	83	109	144	115	166	225	127	423	1646

Figures

Figure 1

The proportion of the substrates metabolized by CYP2A6, 2D6, 2E1 or 3A4 in each group. All substrates are divided into 5 groups based on their molecular weight: very small (< 207), small (207-280), medium (280-326), large (326-405) and very large (>405). The number of substrates in each group is almost the same

Figure 2

The comparison of the validation results by four methods and ten CYP models

Figure 3

ROC curves of ten CYP models with four methods

Figure 4

Illustration of the two-steps training in Improved Bayesian method. The gray circle represents each feature and the black circle represents the target. The gray lines and the black arrows represent the relationship among features and the relationship between features and the target respectively

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplement1.doc](#)