Bioinformatics, 31(11), 2015, 1869–1871
doi: 10.1093/bioinformatics/btv043
Advance Access Publication Date: 22 January 2015
Applications Note

Data and text mining

CypRules: a rule-based P450 inhibition prediction server

Chi-Yu Shao[1,†], Bo-Han Su[2,†], Yi-Shu Tu[1], Chieh Lin[2], Olivia A. Lin[1] and Yufeng J. Tseng[1,2,*]

[1]Graduate Institute of Biomedical Electronics and Bioinformatics and [2]Department of Computer Science and Information Engineering, National Taiwan University, No.1 Sec.4, Roosevelt Road, Taipei, Taiwan 106

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Robert F. Murphy

Received on October 13, 2014; revised on January 16, 2015; accepted on January 18, 2015

Abstract

Summary: Cytochrome P450 (CYPs) are the major enzymes involved in drug metabolism and bioactivation. Inhibition models were constructed for five of the most popular enzymes from the CYP superfamily in human liver. The five enzymes chosen for this study, namely CYP1A2, CYP2D6, CYP2C19, CYP2C9 and CYP3A4, account for 90% of the xenobiotic and drug metabolism in human body. CYP enzymes can be inhibited or induced by various drugs or chemical compounds. In this work, a rule-based CYP inhibition prediction online server, *CypRules,* was created based on predictive models generated by the rule-based C5.0 algorithm. *CypRules* can predict and provide structural rulesets for CYP inhibition for each compound uploaded to the server. Capable of fast execution performance, it can be used for virtual high-throughput screening (VHTS) of a large set of testing compounds.

Availability and implementation: *CypRules* is freely accessible at http://cyprules.cmdm.tw/ and models, descriptor and program files for all compounds are publically available at http://cyprules.cmdm.tw/sources/sources.rar.

Contact: yjtseng@csie.ntu.edu.tw

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Among CYP enzymes, CYP1A2, CYP2C9, CYP2C19, CYP2D6 and CYP3A4 are responsible for 90% of drug oxidation or hydrolysis (Lynch and Price, 2007). More than 900 drugs have been identified in causing liver injury (Friedman *et al.*, 2003), making this the most common reason for a drug to be withdrawn from the market. Hepatotoxicity and drug-induced liver injury account for a substantial number of compounds failure, highlighting the need for drug screening assays capable of detecting potential compound toxicity early in the drug development processes. Previous CYP prediction models or prediction servers were: either based on a small number of compounds, screened with non-rule based approaches that could not suggest structural rules directly contributing to specific P450, or only gives yes/no classification results (Cheng *et al.*, 2011; Hammann *et al.*, 2009; Mishra *et al.*, 2010; Rostkowski *et al.*, 2013; Rydberg *et al.*, 2010; Sun *et al.*, 2011; Sushko *et al.*, 2011). In this study, we chose five of the most common P450 enzymes to build their corresponding inhibition prediction models with >16 000 compounds. The goals are: (i) to first build statistically high performance 2D and 3D QSAR models, utilizing rule-based C5.0 algorithm (Quinlan, 1993) as a prediction tool to provide structural information that contribute towards P450 inhibition endpoints in a faster and more direct manner, compared with other machine learning methods, (ii) and then provide a simple, easy to use, and freely

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

© The Author 2015. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

available web server for quick assessment of compound inhibitions towards the five major CYP enzymes.

2 Methods

The dataset used in this study was collected from the National Institutes of Health Chemical Genomics Center (NCGC) cytochrome panel assay's PubChem BioAssay database (AID1851), using the quantitative high-throughput screening (qHTS) technique. This dataset contains 16 561 compounds screened over five P450 endpoints, including CYP1A2, CYP2C19, CYP2C9, CYP2D6 and CYP3A4. Compound structures were pre-processed to remove any redundant ions and to convert them to 3D structures in preparation for the following 1D, 2D and 3D molecular descriptor calculations.

Molecular descriptors were calculated for each P450 endpoint, using three descriptor tools: PubChem 2D Fingerprint (ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.txt), PaDEL-Descriptor (Yap, 2011) and Mold^2 (Hong *et al.*, 2008). Different combinations of these descriptor sets were all tested in our works. For each P450 endpoint, corresponding trial descriptor set was randomly split into training set and testing set of 7:3 ratio (Akbani *et al.*, 2004). The best descriptor set was derived from the model with the highest G-mean, and the detailed results were shown in supplementary information.

Classification models were constructed using C5.0 algorithm. Rule-based C5.0 decision tree generating algorithm is an improved version of the well-known C4.5 (Quinlan, 1993). C4.5 is an algorithm used to generate a decision tree using the concept of information entropy. At each node on the decision tree, C5.0 chooses an attribute from the data that most effectively splits its initial set of samples into subsets enriched in one class or the other. An important/key feature of C5.0 is its ability to generate classifiers called 'rulesets' which consist of unordered collections of (relatively) simple if-then rules.

3 Results

The best C5.0 models for each CYP enzymes, after considering different combination of descriptor sets, are listed in Table 1, with classification accuracy ranging from 0.73 to 0.90 and G-mean ranging from 0.74 to 0.88. When compared with the other two freely available CYP prediction servers, *MetaPred* and *WhichCyp*, *CypRules* outperformed in most CYP isoforms. (Accuracy values of 0.77, 0.66, 0.68, 0.55 and 0.51 for *MetaPred* (Mishra *et al.*, 2010), and 0.88, 0.83, 0.85, 0.84 and 0.84 (Rostkowski *et al.*, 2013) for *WhichCyp* isoforms 1A2, 2C19, 2C9, 2D6 and 3A4, respectively). Most importantly, *MetaPred* and *WhichCYP* only give yes/no classification results, but *CypRules* can further provide detailed structural information that can be used as guidelines to refine drug candidates. Because the CYP2D6 dataset is highly imbalanced compared with other CYP datasets, CYP2D6 inhibitors were the most difficult to predict in all previous studies. It is worth noting that our reported CYP2D6 inhibition classification model outperformed all of the previous studies with 90% *accuracy*. The detailed description for CYP2D6 model construction was described in supplementary file.

Table 1. Statistics for the models applied to the test sets with different CYP endpoints

	1A2	2C19	2C9	2D6	3A4
Accuracy	0.80	0.86	0.77	0.90	0.73
Sensitivity	0.89	0.84	0.66	0.85	0.76
Specificity	0.72	0.86	0.82	0.91	0.72
G-mean	0.81	0.85	0.74	0.88	0.74

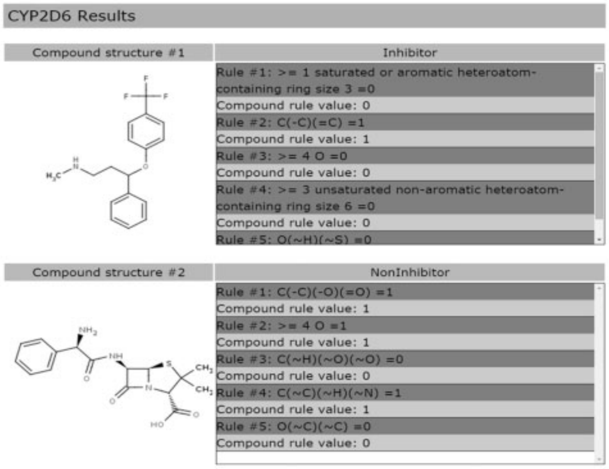
4. Web server

4.1 Interface features and implementation

CypRules is a web server built based on the models tested in this study, to provide a platform for virtually screening compound inhibition at each P450 endpoint. The web server not only predicts the inhibition of P450 endpoints, it also provides structural rulesets that contribute towards predicted inhibitions. Users can interpret the models based on these rules. Consequently, the prediction models embedded in *CypRule*s may offer some insights on how chemical structures correlate to P450 inhibition. Due to the nature of C5.0, which consumes less memory and executes faster than the well-know precursor C4.5, the web server provides a reliable executing speed for large data sets based on reasonable ruleset suggestions. Users can upload 2D or 3D SDF files containing thousands of chemical structures for screening simultaneously.

4.2 Example for inhibition prediction

For each P450 endpoint, *CypRules* deduced a prediction based on the best model reported. Figure 1 showed Fluoxetine, a CYP2D6 inhibitor, and Ampicillin, a CYP2D6 non-inhibitor, were correctly predicted, since the compounds' structural features corresponded to key descriptors, according to suggested rules. The suggested rules were listed sequentially with rule numbers in the figure. The rows containing 'Rule #' show the descriptor's meaning and its condition. The rows containing 'Compound rule value' show the compound's actual calculated value for each of the rules. For inhibitor Fluoxetine, there were five rules related to CYP2D6 inhibitory potency. Within these rules, Fluoxetine was predicted as an inhibitor because of its benzene ring, low oxygen count, and absence of heteroatom-containing rings. For non-inhibitor Ampicillin, there were five rules as well. Within these rules, Ampicillin was correctly predicted as a non-inhibitor because of its high oxygen count, carbonyl groups, amide groups, and absence of ether groups. With these informative rules that were specifically fitted for the five different CYP isoforms, users may utilize these rules/descriptors as a guideline for altering any inhibitor into non-inhibitor, or vice versa, by means of structural modifications.



CYP2D6 Results

Compound structure #1	Inhibitor
	Rule #1: >= 1 saturated or aromatic heteroatom-containing ring size 3 =0
	Compound rule value: 0
	Rule #2: C(-C)(=C) =1
	Compound rule value: 1
	Rule #3: >= 4 O =0
	Compound rule value: 0
	Rule #4: >= 3 unsaturated non-aromatic heteroatom-containing ring size 6 =0
	Compound rule value: 0
	Rule #5: O(~H)(~S) =0

Compound structure #2	NonInhibitor
	Rule #1: C(-C)(-O)(=O) =1
	Compound rule value: 1
	Rule #2: >= 4 O =1
	Compound rule value: 1
	Rule #3: C(~H)(~O)(~O) =0
	Compound rule value: 0
	Rule #4: C(~C)(~H)(~N) =1
	Compound rule value: 1
	Rule #5: O(~C)(~C) =0
	Compound rule value: 0

Fig. 1. Snapshot of CYP2D6 prediction results

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Funding

Ministry of Science and Technology, Taiwan (MOST 103-2325-B-002-048-, 103-2321-B-002-094-,103-2325-B-400-005); Computational Molecular Design and Metabolomics Laboratory, Department of Computer Science and Information Engineering at National Taiwan University.

Conflict of Interest: none declared.

References

Akbani,R. *et al*. (2004) Applying support vector machines to imbalanced datasets. *Proceedings of the 15th European Conference on Machine Learning*. pp. 39–50. Springer, Pisa, Italy.

Cheng,F. *et al*. (2011) Classification of cytochrome P450 inhibitors and noninhibitors using combined classifiers. *J. Chem. Inf. Model*., **51**, 996–1011.

Friedman,S. *et al*. (2003) *Current Diagnosis and Treatment in Gastroenterology*. McGraw-Hill, New York.

Hammann,F. *et al*. (2009) Classification of cytochrome P450 activities using machine learning methods. *Mol. Pharmaceutics*, **6**, 1920–1926.

Hong,H. *et al*. (2008) Mold2, molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics., *J. Chem. Inf. Model*., **48**, 1337–1344.

Lynch,T. and Price,A. (2007) The effect of cytochrome P450 metabolism on drug response, interactions, and adverse effects. *Am. Fam. Physician*, **76**, 391–396.

Mishra,N.K. *et al*. (2010) Prediction of cytochrome P450 isoform responsible for metabolizing a drug molecule. *BMC Pharmacology*, **10**, 8.

Quinlan,J.R. (1993) *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers. Morgan Kaufmann Publishers, Inc., San Francisco.

Rostkowski,M. *et al*. (2013) WhichCyp: prediction of cytochromes P450 inhibition. *Bioinformatics*, **29**, 2051–2052.

Rydberg,P. *et al*. (2010) SMARTCyp: a 2D method for prediction of cytochrome P450-mediated drug metabolism. *ACS Med. Chem. Lett*., **1**, 96–100.

Sun,H. *et al*. (2011) Predictive models for cytochrome P450 isozymes based on quantitative high throughput screening data. *J. Chem. Inf. Model*., **51**, 2474–2481.

Sushko,I. *et al*. (2011) Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. *J. Comput. Aided Mol. Des*., **25**, 553–554.

Yap,C.W. (2011) PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem*., **32**, 1466–1474.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016