BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 16 2013, pages 2051–2052
doi:10.1093/bioinformatics/btt325

Structural bioinformatics

Advance Access publication June 5, 2013

WhichCyp: prediction of cytochromes P450 inhibition

Michał Rostkowski[1], Ola Spjuth[2] and Patrik Rydberg[1,*]

[1]Department of Drug Design and Pharmacology, University of Copenhagen, Universitetsparken 2, DK-2100 Copenhagen, Denmark and [2]Department of Pharmaceutical Biosciences, Uppsala University, P.O. Box 591, SE-751-24 Uppsala, Sweden

Associate Editor: Anna Tramontano

ABSTRACT

Summary: In this work we present WhichCyp, a tool for prediction of which cytochromes P450 isoforms (among 1A2, 2C9, 2C19, 2D6 and 3A4) a given molecule is likely to inhibit. The models are built from experimental high-throughput data using support vector machines and molecular signatures.

Availability: The WhichCyp server is freely available for use on the web at http://drug.ku.dk/whichcyp, where the WhichCyp Java program and source code is also available for download.

Contact: pry@sund.ku.dk

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

Received on March 22, 2013; revised on May 13, 2013; accepted on May 30, 2013

1 INTRODUCTION

The cytochromes P450 (CYP) is a ubiquitous enzyme family found throughout the nature. These enzymes are involved in the metabolism of a large majority of the drugs in use today (Guengerich, 2006). Modeling of their interactions with drug-like compounds have been a hot topic during the past decade, stretching from quantum mechanical studies of reaction mechanisms to data mining approaches and QSAR studies (Carlsson *et al*., 2010; Kirchmair *et al*., 2012; Rydberg *et al*., 2012).

The five most important human CYP isoforms, with regard to drug metabolism, are 1A2, 2C9, 2C19, 2D6 and 3A4 (Guengerich, 2006). In 2009, a large high-throughput study of inhibition to these isoforms was performed (Veith *et al*., 2009), enabling the application of machine learning techniques to the classification of CYP inhibitors. Such classification models are important for the prediction of drug–drug interactions.

In this work, we present WhichCyp, a publically available software and web server for prediction of CYP inhibitors built from these data.

2 METHODS

Inhibitory data for 17 143 substances tested using quantitative high-throughput screening with *in vitro* bioluminescent assay against five major isoforms of cytochrome P450 was obtained from the PubChem BioAssay database, AID:1851 (Veith *et al*., 2009; Wang *et al*., 2012). Inorganic compounds, salts and mixtures, as well as entries classified as inconclusive were excluded from the dataset. For each of the five isoforms, 3000 compounds were extracted from the corresponding dataset to use as a test set, while the remaining compounds were used as a training set. The distribution of inhibitors/non-inhibitors for the 10 datasets is shown in the supporting information, Supplementary Table S1.

Classification models were constructed with support vector machines (SVM) (Cortes and Vapnik, 1995) using the libSVM library (Chang and Lin, 2011) together with R (http://www.r-project.org/). As features for the model building, we used molecular signatures (Faulon *et al*., 2003) of heights from 0 to 4, computed with the Bioclipse software (Spjuth *et al*., 2007, 2009, 2011). Molecular signatures are built up from atomic signatures, which in turn are a type of circular atomic fragments. The height of an atomic signature is the maximum number of bonds between the atom and the other atoms in its atomic signature. The SVM models were weighted by the relative number of positives and negatives in each dataset. The optimization of C and gamma parameters in the SVM was made through 5-fold cross validation to achieve a Matthews correlation coefficient (MCC) (Matthews, 1975) as high as possible, but with the constraint that the sensitivity and specificity should also be similar. To achieve this, the 1A2, 2C19 and 3A4 datasets could be optimized for MCC alone, whereas the datasets with the most uneven distributions (2C9 and 2D6) had to be optimized toward the highest sensitivity.

The constraint that sensitivity and specificity should be similar was added because there are multiple use cases for the models, which have opposite needs in this regard. First, a chemist might be interested in which CYP isoforms would be inhibited by a compound (to avoid drug–drug interactions), in which case sensitivity would be important. Second, the purpose could be to make sure a compound is not metabolized by CYPs (to enable a longer half-life), in which case specificity is important. Finally, after the models' performance was evaluated and best sets of signature heights were chosen, models for all isoforms were reoptimized and rebuilt using all experimental data—test sets used earlier in the studies were also included in the training set. These models are used in the publicly available code.

3 RESULTS

We investigated signatures with different heights, ranging from 0 to 4, and combinations thereof. The most successful combinations were combinations of heights 0–3, results for which are shown in Table 1. Combinations of signature heights for the ranges 0–2 and 0–4 gave similar results, but with slightly worse area under curve (AUC) values for almost all models (Supplementary Material). The final models (Table 1) give accurate predictions with AUC values ranging from 0.88 to 0.95, and even sensitivity and specificity for all models except the one for 2D6 (due to the unbalanced dataset). These results are improvements compared with earlier studies, in which combined classifiers using fingerprints resulted in AUC values of 0.81,

*To whom correspondence should be addressed.

Table 1. Statistics for the models applied to the test sets

	1A2	2C9	2C19	2D6	3A4
Sensitivity	0.87	0.84	0.86	0.75	0.84
Specificity	0.88	0.83	0.84	0.86	0.84
Prediction accuracy	0.88	0.83	0.85	0.84	0.84
AUC	0.95	0.90	0.91	0.88	0.92

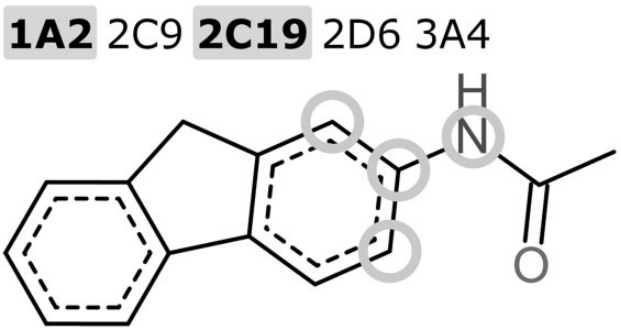

Fig. 1. The HTML output for 2-acetylaminoflourene. Isoform names with a dark background are predicted to be inhibited

0.86, 0.84, 0.88, 0.79 (Cheng *et al*., 2011) and 264 molecular descriptors in values of 0.93, 0.89, 0.89, 0.85, 0.87 (Sun *et al*., 2011) for isoforms 1A2, 2C9, 2C19, 2D6 and 3A4, respectively. Other approaches that have been used were discussed in a recent review (Kirchmair *et al*., 2012).

To investigate if we could determine an applicability domain type measure, we studied the relation between the number of signatures for the compounds in the test sets that did not exist in the training set, and the performance of the models. It was found that the number of missing signatures of height three (Supplementary Figs S1–S5) could be used as an estimation of the likelihood that the model performs well, and that the sensitivity was the first measure to drop (among sensitivity, specificity and prediction accuracy).

4 WEB SERVER AND SOFTWARE

4.1 Interface features

The WhichCyp web server offers the user three ways to submit molecules. The user can upload a file in any standard format, enter SMILES strings representing molecules or draw a molecule. The results are displayed directly in the browser and include the molecular structure and highlighting of the structural fragment that gave the most important contribution to the decision of the SVM model (see Figure 1), as well as a sensitivity warning if there are too many missing signatures of height three (defined as the number of missing signatures in the test set that resulted in a sensitivity of $\leq$0.6).

4.2 Implementation

The web server uses php code to run WhichCyp and the interface functionality. To support all standard formats, uploaded files are converted by Open Babel (O'Boyle *et al*., 2011) when necessary. The web server code was copied from the SMARTCyp web server (Rydberg *et al*., 2010). WhichCyp is implemented using the CDK Java library (Steinbeck *et al*., 2003, 2006).

4.3 Java program

The server runs a command line Java program that also is available for download as both an executable jar file and as source code. The Java program has several optional flags to enable the user flexibility in what output is generated.

ACKNOWLEDGEMENT

The authors thank Lars Carlsson for the Java code used to implement the SVM prediction in WhichCyp.

Funding: M.R. and P.R. acknowledge funding from Lhasa Ltd. O.S. acknowledges funding by the Swedish VR-2011-6129 and the Swedish strategic research program eSSENCE.

Conflict of Interest: none declared.

REFERENCES

Carlsson,L. *et al*. (2010) Use of historic metabolic biotransformation data as a means of anticipating metabolic sites using MetaPrint2D and Bioclipse. *BMC Bioinformatics*, **11**, 362.

Chang,C.C. and Lin,C.J. (2011) LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, **2**, 27.

Cheng,F. *et al*. (2011) Classification of cytochrome p450 inhibitors and noninhibitors using combined classifiers. *J. Chem. Inf. Model.*, **51**, 996–1011.

Cortes,C. and Vapnik,V. (1995) Support-vector networks. *Mach. Learn.*, **20**, 273–297.

Faulon,J.L. *et al*. (2003) The signature molecular descriptor. 1. Using extended valence sequences in QSAR and QSPR studies. *J. Chem. Inf. Comput. Sci.*, **43**, 707–720.

Guengerich,F.P. (2006) Cytochrome P450s and other enzymes in drug metabolism and toxicity. *AAPS J.*, **8**, E101–E111.

Kirchmair,J. *et al*. (2012) Computational prediction of metabolism: sites, products, SAR, P450 enzyme dynamics, and mechanisms. *J. Chem. Inf. Model.*, **52**, 617–648.

Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.

O'Boyle,N.M. *et al*. (2011) Open babel: an open chemical toolbox. *J. Cheminform.*, **3**, 33.

Rydberg,P. *et al*. (2010) The SMARTCyp cytochrome P450 metabolism prediction server. *Bioinformatics*, **26**, 2988–2989.

Rydberg,P. *et al*. (2012) Quantum-mechanical studies of reactions performed by cytochrome P450 enzymes. *Curr. Inorg. Chem.*, **2**, 292–315.

Spjuth,O. *et al*. (2007) Bioclipse: an open source workbench for chemo- and bioinformatics. *BMC Bioinformatics*, **8**, 59.

Spjuth,O. *et al*. (2009) Bioclipse 2: a scriptable integration platform for the life sciences. *BMC Bioinformatics*, **10**, 397.

Spjuth,O. *et al*. (2011) Integrated decision support for assessing chemical liabilities. *J. Chem. Inf. Model.*, **51**, 1840–1847.

Steinbeck,C. *et al*. (2003) The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. *J. Chem. Inf. Comp. Sci.*, **43**, 493–500.

Steinbeck,C. *et al*. (2006) Recent developments of the Chemistry Development Kit (CDK) - an open-source Java library for Chemo- and Bioinformatics. *Curr. Pharm. Des.*, **12**, 2111–2120.

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.

Veith,H. *et al*. (2009) Comprehensive characterization of cytochrome P450 isozyme selectivity across chemical libraries. *Nat. Biotechnol.*, **27**, 1050–1055.

Wang,Y. *et al*. (2012) PubChem's bioassay database. *Nucleic Acids Res.*, **40**, D400–D412.