*Bioinformatics*, 31(12), 2015, 2046–2048
doi: 10.1093/bioinformatics/btv087
Advance Access Publication Date: 11 February 2015
Applications Note

OXFORD

Structural bioinformatics

# SOMP: web server for *in silico* prediction of sites of metabolism for drug-like compounds

**Anastasia Rudik[1,*], Alexander Dmitriev[1], Alexey Lagunin[1,2], Dmitry Filimonov[1] and Vladimir Poroikov[1,2]**

[1]Laboratory for Structure-Function Based Drug Design, Institute of Biomedical Chemistry, 10/8 Pogodinskaya Str., Moscow, 119121, Russia and [2]Medicobiological Faculty, Pirogov Russian National Research Medical University, 1 Ostrovityanova Str., Moscow, 117997, Russia

*To whom correspondence should be addressed.

Associate Editor: Robert F. Murphy

Received on September 25, 2014; revised on January 29, 2015; accepted on February 7, 2015

## Abstract

**Summary:** A new freely available web server site of metabolism predictor to predict the sites of metabolism (SOM) based on the structural formula of chemicals has been developed. It is based on the analyses of 'structure-SOM' relationships using a Bayesian approach and labelled multilevel neighbourhoods of atoms descriptors to represent the structures of over 1000 metabolized xenobiotics. The server allows predicting SOMs that are catalysed by 1A2, 2C9, 2C19, 2D6 and 3A4 isoforms of cytochrome P450 and enzymes of the UDP-glucuronosyltransferase family. The average invariant accuracy of prediction that was calculated for the training sets (using leave-one-out cross-validation) and evaluation sets is 0.9 and 0.95, respectively.

**Availability and implementation:** Freely available on the web at http://www.way2drug.com/SOMP.

**Contact:** rudik_anastassia@mail.ru

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

## 1 Introduction

Drug metabolism affects the drug efficiency and toxicity and is divided into Phase I (oxidation, hydrolysis, reduction) and Phase II (conjugation) reactions. Cytochromes P450 (CYP) are the main enzymes of Phase I (Lewis and Ito, 2008; Williams *et al.*, 2004), which metabolize most drugs. Glucuronidation is the main reaction of Phase II, which is catalysed by UDP-glucuronosyltransferase (UGT) and serves as a clearance mechanism for drugs from many therapeutic classes (King *et al.*, 2000). Prediction of the sites of metabolism (SOM) is important during the drug discovery process to detect possible 'metabolic hot spots' in lead compounds that can be modified to obtain more metabolically stable compounds. The SOM prediction can be used for an integral estimation of activity and toxicity of a compound including its possible metabolites.

There are two freely available web servers, which allow predicting SOMs for xenobiotics that are metabolized by P450 isoenzymes. SmartCyp predicts SOMs for compounds that are metabolized by CYP2C and CYP2D6 (Rydberg *et al.*, 2010), and RS-WebPredictor (Zaretzki *et al.*, 2013) predicts SOMs for nine CYP isoforms (1A2, 2A6, 2B6, 2C8, 2C9, 2C19, 2D6, 2E1, 3A4).

We have previously developed a method to predict SOMs (Rudik *et al.*, 2014) by using only structural formulae, which distinguishes from traditional SOM-predicting methods that use molecular docking and molecular dynamics or quantum mechanical and semi-empirical calculations.

The detailed description of the algorithm was published (Filimonov *et al.*, 2014; Rudik *et al.*, 2014) and shown in the Supplementary text S1. In this article, we highlight the extension of

**Table 1.** Comparison of prediction accuracy for training and evaluation sets

| Enzyme | Nc | Np | Nn | $IAP_{LOO\ CV}$ | SOMP | | | | SMARTCyp | | | | RS-WebPredictor | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Top-1 | Top-2 | Top-3 | IAP | Top-1 | Top-2 | Top-3 | IAP | Top-1 | Top-2 | Top-3 | IAP |
| CYP3A4 | 960 | 1410 | 18621 | 0.89 | 0.69 | 0.77 | 0.8 | 0.85 | – | – | – | – | 0.31 | 0.34 | 0.4 | – |
| CYP2D6 | 588 | 726 | 8366 | 0.88 | 0.43 | 0.61 | 0.68 | 0.83 | 0.67 | 0.78 | 0.86 | 0.86 | 0.21 | 0.29 | 0.36 | – |
| CYP2C9 | 446 | 600 | 7471 | 0.89 | 0.58 | 0.75 | 0.88 | 0.87 | 0.63 | 0.75 | 0.92 | 0.91 | 0.17 | 0.21 | 0.25 | – |
| CYP2C19 | 388 | 517 | 6329 | 0.88 | 0.46 | 0.62 | 0.85 | 0.76 | – | – | – | – | 0.23 | 0.31 | 0.38 | – |
| CYP1A2 | 573 | 888 | 9375 | 0.89 | 0.47 | 0.68 | 0.68 | 0.83 | – | – | – | – | 0.37 | 0.37 | 0.37 | – |
| UGT (all isoforms) | 592 | 906 | 1136 | 0.98 | 0.9 | 0.95 | 0.95 | 0.96 | – | – | – | – | – | – | – | – |

Nc is the number of compounds in a training set; Np is the number of positive examples of SOMs; Nn is the number of negative examples of SOMs; Top-1, -2, -3 metrics: a molecule is considered as correctly predicted if any experimental SOM is ranked as first (Top-1), first or second (Top-2) or first or second or third (Top-3) within the SOM ranking (Rudik *et al.*, 2014).

our method and show a new freely available web server for SOM prediction for five major isoforms of P450 and UGT.

## 2 Materials and methods

We extended our previously described method with the prediction of glucuronidation sites. The interactions of xenobiotics with human enzymes were considered. The heterogeneous training sets were collected from the available *in vitro* and *in vivo* experimental studies in publications and the Biovia (Accelrys) Metabolite database. The collected data were used as the positive examples. The negative examples were generated as original structures with labelled atoms, which were not detected as SOMs in the experiments. The training set with the glucuronidation reaction (906 positive examples) and previously created training sets were used in the developed web server (Table 1).

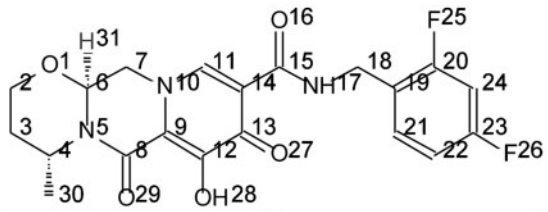
The leave-one-out cross-validation (LOO CV) procedure was performed during the training procedure, and two values (Pt and Pf) were calculated for every example structure from the training set. Pt is the probability that this example would be positive; Pf is the probability that this example would be negative. ΔP was calculated as Pt-Pf. The invariant accuracy of prediction (IAP) criterion, which is similar to the AUC [the area under the receiver operating characteristic curve (Swets, 1988)], was used to estimate the accuracy of the created method. Mathematically, the IAP (Filimonov *et al.*, 2008) value is equal to the probability that the estimated ΔP is higher for a randomly selected positive example than for a randomly selected negative example.

## 3 Site of metabolism predictor web server

The web server for the SOM prediction for drug-like compounds that are metabolized by the main CYP isoforms and UGT was created based on the aforementioned training sets and called the site of metabolism predictor (SOMP). The average IAP value that was calculated during the LOO CV procedure is approximately 0.9 (Table 1). The SOM prediction was validated based on the external evaluation sets with data on biotransformations for 68 cardiovascular drugs (full list in Supplementary text S1). The accuracy of the CYP validation was similar or higher than the results of SMARTCyp and RS-WebPredictor (Rudik *et al.*, 2014). The glucuronidation prediction results for the evaluation set (Table 1 and the Supplementary text S1) show high prediction accuracy.

### 3.1 Interface features

The 'JME Molecular Editor' (Bienfait and Ertl 2013) was used to draw and generate chemical structure. The user may submit InChI or SMILES strings (Weininger, 1988) [converted by Open Babel (Guha *et al.*, 2006)] or upload a file in 'MOL' format (Dalby *et al.*, 1992). Prediction is calculated only for single-component uncharged structures of low-molecular-weight organic molecules with at least three carbon atoms. The SOM prediction results include a tested structure with numbered atoms and tables, which include the atoms and their ranks according to the probability to be attacked by each enzyme. They can be saved as SDF or PDF files. The user obtains CYPs and UGT tables with the prediction results and can sort them by atom numbers or atom ranks.




| CYP3A4 | | | UGT | | |
|---|---|---|---|---|---|
| Atom number | Rank | DeltaP | Atom number | Rank | DeltaP |
| 2 | 1 | 0,533 | 28 | 1 | 0,088 |
| 18 | 2 | 0,079 | 19 | 2 | -0,476 |
| 6 | 3 | 0,013 | 18 | 3 | -0,362 |
| 17 | 4 | -0,709 | 22 | 4 | -0,360 |
| 19 | 5 | -0,681 | 20 | 5 | -0,352 |
| 14 | 6 | -0,570 | 21 | 6 | -0,346 |
| 26 | 7 | -0,549 | 23 | 7 | -0,318 |
| 20 | 8 | -0,532 | 4 | 8 | -0,311 |
| 1 | 9 | -0,507 | 14 | 9 | -0,306 |

**Fig. 1.** SOMP results for Dolutegravir (DTG; S/GSK1349572)

### 3.2 Implementation

The SOMP web server uses MySQL server to store the data and PHP and HTML codes to implement the main interface. The Python script is used to produce independent sub-processes to generate input to the prediction program and data processing. The prediction runs ~0.5 s on modern PC (requiring ~20 Mb RAM) independently for each user, and many users can simultaneously use it. For each non-hydrogen atom in the molecular structure that was sent to the prediction, ΔP was calculated, and the atoms are arranged in descending order of ΔP. The prediction for Dolutegravir (this drug was not included in the training set) is illustrated in Figure 1. The SOM that is known for UGT glucuronidation (atom number 28) was ranked as number one. The SOM that is known for being attacked by CYP3A4 (atom number 18) was correctly predicted according to

the top-ranking atom estimation (second position in the list) (Reese *et al.*, 2013). This result and the data in Table 1 show that the SOMP web server has reasonable prediction accuracy and may be used to study the drug metabolism.

## Funding

The study was supported by Russian Scientific Foundation grant 14-15-00449.

*Conflict of Interest*: none declared.

## References

Bienfait,B. and Ertl,P. (2013) JSME: a free molecule editor in JavaScript. *J. Cheminform.*, **5**, 24.

Dalby,A. *et al.* (1992) Description of several chemical structure file formats used by computer programs developed at molecular design limited. *J. Chem. Inf. Model.*, **32**, 244.

Filimonov,D.A. *et al.* (2008) In: Varnek,A. and Tropsha,A. (eds) *Chemoinformatics Approaches to Virtual Screening*. RSC Publishing: Cambridge, UK, pp. 182–216.

Filimonov,D.A. *et al.* (2014) Prediction of the biological activity spectra of organic compounds using the PASS Online web resource. *Chem. Heterocycl. Compd.*, **50**, 444–457.

Guha,R. *et al.* (2006) Blue Obelisk—interoperability in chemical informatics. *J. Chem. Inf. Model.*, **46**, 991–998.

King,C. *et al.* (2000) UDP-glucuronosyltransferases. *Curr. Drug Metab.*, **1**, 143–161.

Lewis,D.F. and Ito,Y. (2008) Human cytochromes P450 in the metabolism of drugs: new molecular models of enzyme-substrate interactions. *Expert Opin. Drug Metab. Toxicol.*, **4**, 1181–1186.

Reese,M.J. *et al.* (2013) In vitro investigations into the roles of drug transporters and metabolizing enzymes in the disposition and drug interactions of dolutegravir, a HIV integrase inhibitor. *Drug Metab. Dispos.*, **41**, 353–361.

Rudik,A.V. *et al.* (2014) Metabolism site prediction based on xenobiotic structural formulas and PASS prediction algorithm. *J. Chem. Inf. Model.*, **54**, 498–507.

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

Swets,J. (1988) Measuring the accuracy of diagnostic systems. *Rev. Sci.*, **240**, 1285–1293.

Weininger,D. (1988) SMILES, a chemical language and information system 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, **28**, 31–36.

Williams,J.A. *et al.* (2004) Drug-drug interactions for UDP-glucuronosyltransferase substrates: a pharmacokinetic explanation for typically observed low exposure (AUCi/AUC) ratios. *Drug Metab. Dispos.*, **32**, 1201–1208.

Zaretzki,J. *et al.* (2013) RS-WebPredictor: a server for predicting CYP-mediated sites of metabolism on drug-like molecules. *Bioinformatics*, **29**, 497–498.