

RS-WebPredictor: a server for predicting CYP-mediated sites of metabolism on drug-like molecules

Jed Zaretski^{1,2}, Charles Bergeron³, Tao-wei Huang², Patrik Rydberg⁴, S. Joshua Swamidass¹ and Curt M. Breneman^{2,*}

¹Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO 63130, USA,

²Department of Chemistry and Chemical Biology, Rensselaer Polytechnic Institute, Troy, NY 12180, USA, ³Department of Basic and Social Sciences, Albany College of Pharmacy and Health Sciences, Albany, NY 12208, USA and

⁴Department of Drug Design and Pharmacology, University of Copenhagen, DK-2100 Copenhagen, Denmark

Associate Editor: Anna Tramontano

ABSTRACT

Summary: Regioselectivity-WebPredictor (RS-WebPredictor) is a server that predicts isozyme-specific cytochrome P450 (CYP)-mediated sites of metabolism (SOMs) on drug-like molecules. Predictions may be made for the promiscuous 2C9, 2D6 and 3A4 CYP isozymes, as well as CYPs 1A2, 2A6, 2B6, 2C8, 2C19 and 2E1. RS-WebPredictor is the first freely accessible server that predicts the regioselectivity of the last six isozymes. Server execution time is fast, taking on average 2s to encode a submitted molecule and 1s to apply a given model, allowing for high-throughput use in lead optimization projects.

Availability: RS-WebPredictor is accessible for free use at <http://reccr.chem.rpi.edu/Software/RS-WebPredictor/>

Contact: brenec@rpi.edu

Received on July 30, 2012; revised on November 19, 2012; accepted on December 8, 2012

1 INTRODUCTION

The cytochrome P450s (CYPs) are a family of heme-thiolate proteins that metabolize ~90% of FDA-approved drugs (Nebert and Russell, 2002). Most CYPs are 'regioselective', strongly favouring the oxidation of certain sites of metabolism (SOMs) over others. Knowledge of the SOMs, the specific atom(s) of a molecule that are oxidized by specific CYP isozymes, is valuable information for early-stage lead design and optimization. Armed with this knowledge, medicinal chemists can make rational modifications to a candidate lead in order to change its CYP-mediated metabolism. One way to use regioselectivity knowledge to perform rational lead optimization is to increase drug bioavailability based upon individual CYP-expression profile while retaining patient uniform end-target efficacy. Unfortunately, determining the isozyme-specific SOMs of early-stage lead candidates through experimental means is time- and resource-intensive, and not feasible for application to every candidate. Consequently, several groups in recent years developed *in silico* SOM prediction models (Kirchmair *et al.*, 2012).

Regioselectivity-Predictor (RS-Predictor) is an algorithm for creating accurate isozyme-specific SOM prediction models from any set of known substrates and metabolites (Zaretski *et al.*,

2011). In prior work, we manually curated the public literature to identify 680 CYP substrates and metabolites distributed across nine isozymes, the largest collection of cytochrome P450 metabolite data released to date. RS-Predictor was able to identify experimentally observed SOM(s) within the top two rank positions for substrate sets of each CYP isozyme with high levels of cross-validated accuracy: CYP isozyme (number of substrates, accuracy), 1A2 (271, 83.0%), 2A6 (105, 85.7%), 2B6 (151, 82.1%), 2C8 (142, 83.8%), 2C9 (226, 84.5%), 2C19 (218, 86.2%), 2D6 (270, 85.9%), 2E1 (145, 82.8%), 3A4 (475, 82.3%). These accuracies were significantly higher than those of commercial methods StarDrop (78.0%, 75.3%, 74.1%) from Optibrium and the P450 SOM Prediction workflow (72.1%, 68.1%, 76.4%) offered by Schrödinger for 2C9, 2D6 and 3A4 sets, respectively. Although there have been a number of SOM prediction models for these three isozymes, RS-Predictor offers the first ligand-based models for the remaining six isozymes.

The full details of the RS-Predictor algorithm have been described in prior work (Zaretski *et al.*, 2011) and are summarized here. RS-Predictor calibrates an isozyme-specific regioselectivity QSAR from a set of known isozyme substrates and metabolites by treating each substrate as an individual competition between candidate SOMs. Each SOM is represented by 148 topological descriptors—is the SOM in an aromatic ring, what is the size of the ring, what is the distribution of different atom types 1, 2, 3 and 4 bond lengths away from the given SOM, etc.—and a SMARTCyp-derived reactivity descriptor. SMARTCyp is an open-source ligand-based method that encoded density functional theory-derived transition state energies of molecule substructures with an isozyme non-specific CYP heme into a reactivity look-up table (Rydberg *et al.*, 2010). Next, SOM prediction models are created using MIRank (multiple-instance ranking), a generalization of support vector machines that is specifically designed to optimize the ranking of observed SOMs over non-observed SOMs on a substrate by substrate basis (Bergeron *et al.*, 2012). It is important to note that these models do not predict whether the given molecule is metabolized by a particular isozyme; they predict the exact location on that molecule that would be oxidized by that CYP if the molecule is a substrate for that particular isozyme. This limitation is shared by all other SOM prediction models, a comprehensive review of which was made by Kirchmair *et al.* (2012). Still, the high prediction accuracies of RS-Predictor

*To whom correspondence should be addressed.

models, and the fact that they encapsulate the largest collection of CYP substrate and metabolite information publicly available, make free access to predictions made by them a valuable contribution to the scientific and pharmaceutical communities.

2 RS-WEBPREDICTOR

Although the RS-Predictor algorithm has been described, no utility has been made available that lets users quickly and easily apply it to predict the CYP-mediated metabolism of candidate molecules. The main contribution of this application note is RS-WebPredictor, a public server that predicts sites of isozyme-specific CYP-mediated metabolism on any set of user-supplied molecules. Predictions may be made by SOM prediction models for the promiscuous 2C9, 2D6 and 3A4 CYPs, as well as CYPs 1A2, 2A6, 2B6, 2C8, 2C19 and 2E1. This is the first online tool to predict the regioselectivity of the last six isozymes. In addition, a combined model is made available that was calibrated on all available substrates and all CYP-mediated reactions they undergo, regardless of the metabolizing isozyme.

2.1 Input options

RS-WebPredictor lets the user submit candidate molecule(s) in one of two ways: (i) a single structure may be either drawn or copied from a MOL file, or (ii) a batch file of any number of compounds may be submitted in SDF or SMILES format, to allow for high-throughput isozyme-specific SOM predictions. The user can select which CYP model(s) they wish to apply and an optional email address where they wish to receive the results.

2.2 Server output

As output, the top three predicted SOMs of each chosen isozyme model are provided for each submitted molecule in three formats: (i) an SDF file is created with additional fields designated the primary, secondary and tertiary SOMs predicted by the selected model(s), (ii) a tabular result file containing the molecule name, and the atom IDs from the primary, secondary and tertiary predicted SOMs, and (iii) a web page is generated with graphic figures of each submitted molecule having numbers and circles drawn to designate the corresponding top three predicted SOMs by a given model. Links are provided in the web page that allow the user to see all predictions for the set of input compounds on a model-by-model basis. An example of the graphical output of 2B6, 2C9 and 2D6 model predictions on the substrate cinnarizine is illustrated in Figure 1. Graphical output is available on current versions of Mozilla Firefox, Google Chrome, Internet Explorer and Safari, or any browser that has the Google Chrome Frame plugin installed. It is important to note that predictions made by RS-WebPredictor on any compound that was contained within the initial calibration sets should be considered overtrained. These calibration sets may be found in the Supporting Information of Zaretski et al. (2012), as well as the release notes of this web server.

2.3 Execution time

When applied to a batch file of >100 random molecules, the average execution time is ~2s per molecule for descriptor generation and 1s per molecule for each model. So, execution time per substrate for an individual model is ~2s if executed on a large batch. There can be substantial increases in per-molecule

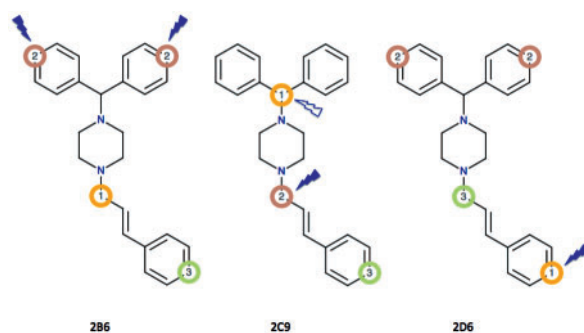


Fig. 1. The first, second and third predicted SOMs of 2B6, 2C9 and 2D6 RS-Predictor models when applied to cinnarizine. The graphical output of RS-WebPredictor is exactly as shown above, save for the lack of lightning bolts, which designate the experimentally observed SOM(s) for the given isozyme. Solid colouring denotes a primary metabolite while the outlined bolt designates a secondary metabolite of CYP 2C9

execution time on small batches. Applying the server to a single substrate takes 4.3s, 6.3s and 13s for molecules with 3, 21 and 88 heavy atoms, respectively. Nonetheless, this server is quick enough to feasibly analyse either individual molecules or small batches of several hundred molecules.

3 CONCLUSION

We release a publicly available server for fast web-based prediction of CYP-mediated metabolism on user-submitted molecules (<http://reccr.chem.rpi.edu/Software/RS-WebPredictor/>). This is the first server that makes site of metabolism predictions for nine isozymes, and uses models trained on the largest collection of CYP substrate and metabolite data publicly available. We expect it will be a valuable tool for lead optimization and academic investigations into CYP-mediated metabolism.

ACKNOWLEDGEMENT

The authors thank Dr Michael Krein for systems support and the Rensselaer Center for Biotechnology and Interdisciplinary Studies (CBIS).

Funding: NIH grants (1P20HG003899-01 and R01LM009731); ONR grant (N00014-06-1-0014); Lhasa Ltd; the Pathology and Immunology Department at Washington University; GlaxoSmithKline.

Conflict of Interest: none declared.

REFERENCES

- Bergeron, C. et al. (2012) Fast bundle algorithm for multiple-instance learning. *IEEE Trans Pattern Anal Mach Intell.*, **34**, 1068–1079.
- Kirchmair, J. et al. (2012) Computational prediction of metabolism: sites, products, SAR, P450 enzyme dynamics, and mechanisms. *J. Chem. Inf. Model.*, **52**, 617–648.
- Nebert, D.W. and Russell, D.W. (2002) Clinical importance of the cytochromes P450. *Lancet*, **360**, 1155–1162.
- Rydberg, P. et al. (2010) The SMARTCyp cytochrome P450 metabolism predictions server. *Bioinformatics*, **26**, 2988–2989.
- Zaretski, J. et al. (2011) RS-Predictor: a new tool for predicting sites of cytochrome P450-mediated metabolism applied to CYP 3A4. *J. Chem. Inf. Model.*, **51**, 1667–1689.
- Zaretski, J. et al. (2012) RS-Predictor models augmented with SMARTCyp reactivities: robust metabolic regioselectivity predictions for nine CYP isozymes. *J. Chem. Inf. Model.*, **52**, 1637–1659.