Probabilistic Predictive Model of the Human Liver Microsomal Metabolism Regioselectivity

Justas Dapkunas[1,2], Andrius Sazonovas[1], Pranas Japertas[1]

[1] ACD/Labs, Inc., A.Mickeviciaus g. 29, LT-08117 Vilnius, Lithuania,
[2] Department of Biochemistry and Biophysics, Vilnius University, M.K.Ciurlionio g. 21/27, LT-03101 Vilnius, Lithuania.

INTRODUCTION

Analytical identification of metabolites for a drug candidate is usually a time consuming and low-throughput task which is performed only in late drug development phases. Therefore, the ability to predict possible sites of human liver microsomal metabolism using *in silico* techniques would be highly beneficial for any medicinal chemist. Moreover, *in silico* predictions of the most likely metabolism sites in a molecule could facilitate the analysis of spectroscopic data and thus ease the experimental identification of metabolites.

In this work, we present QSAR models for the prediction of metabolism regioselectivity. They provide the probability to be metabolized in human liver microsomes for every atom of the molecule and are based on a novel GALAS (**G**lobal, **A**djusted **L**ocally **A**ccording to **S**imilarity) methodology – an approach enabling the evaluation of Model Applicability Domain via the calculation of the prediction Reliability Index (RI).

DATA SET

Experimental data on metabolism in human liver microsomes for 873 compounds were collected from scientific publications dealing with analytical identification of the metabolites observed after the incubation of compound with human liver microsomes or recombinant cytochrome P450 enzymes. Every carbon atom with at least one hydrogen attached was marked as a site of metabolism, if hydroxylation at the atom was observed, or site of no metabolism otherwise. For dealkylation reactions, carbon atoms of the leaving groups were marked in the same manner. Similarly, every sulfur atom was marked whether oxidized or not. Some sites were marked as "inconclusive" and consequently not used in the modeling. The distribution of metabolized and not metabolized atoms for different enzymes is shown in Figure 1.

TABLE 1. The structure of the HLM metabolism regioselectivity dataset.

Subset	No. of compounds	No. of metabolism sites	Total No. of marked atoms
N-dealkylation	511	333	1173
O-dealkylation	488	260	1033
Aliphatic hydroxylation	723	318	2904
Aromatic hydroxylation	739	358	3341
S-oxidation	135	57	157
Total	873	1326	8606

The complete dataset of ca. 8500 marked atoms was divided into aromatic hydroxylation, N-dealkylation, O-dealkylation, S-oxidation, and aliphatic hydroxylation subsets according to the atom type [2]. The composition of obtained subsets in case of human liver microsomal metabolism is outlined in Table 1. Each subset was further divided into training (70%) and test (30%) sets and individual models were built for all types of reactions.

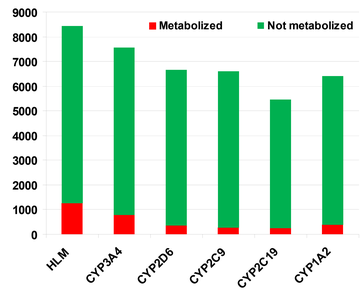


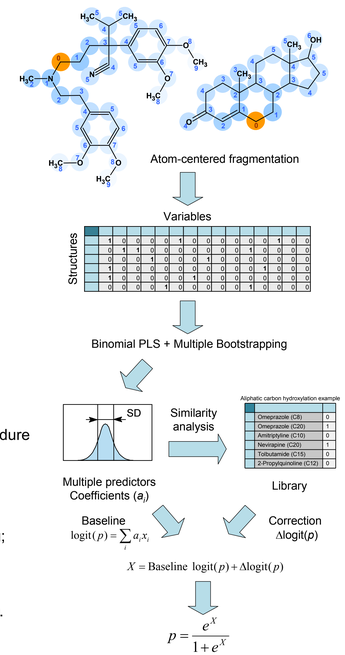
FIGURE 1. Distribution of metabolized and non-metabolized atoms in datasets for cytochrome P450 enzymes and overall HLM metabolism.

MODEL DEVELOPMENT

The modeling workflow is shown in Scheme 1. The procedure consists of the following steps:

- Atom centered fragmentation for every marked atom;
- Development of a baseline model predicting the probability to be metabolized for every atom in a molecule using PLS in combination with bootstrapping;
- Correction of baseline predictions according to experimental data for 5 most similar atoms in the training set accompanied by the estimation of prediction reliability in the form of Reliability Index (RI).

The entire process of the molecule fragmentation and subsequent statistical analysis was realized using Algorithm Builder application [2]. More details on GALAS modeling method can be found in our recent publications [3,4].



SCHEME 1: Outline of modeling methodology

MODEL VALIDATION

TABLE 2. Detailed results of the HLM metabolism regioselectivity algorithms' validation using the test sets constituting 30% of the initial HLM databases for each reaction.

			Predicted value					
			Baseline*		Similarity Corrected*		Similarity Corrected (RI>0.5)	
			Positive	Negative	Positive	Negative	Positive	Negative
Observed value	N-dealkylation	Positive	**48**	31	**60**	19	**33**	5
		Negative	10	**190**	10	**190**	3	**105**
	O-dealkylation	Positive	**44**	11	**46**	9	**23**	1
		Negative	10	**179**	7	**182**	3	**118**
	Aliphatic hydroxylation	Positive	**41**	21	**35**	27	**27**	11
		Negative	149	**574**	31	**692**	13	**498**
	Aromatic hydroxylation	Positive	**72**	25	**56**	41	**33**	17
		Negative	191	**669**	46	**814**	22	**589**
	S-oxidation	Positive	**7**	0	**6**	1	**3**	0
		Negative	1	**20**	1	**20**	0	**16**

* - Unreliable predictions (RI<0.3) were not considered which led to exclusion of 73 marked atoms (21 metabolism sites) from the initial test set in case of N-dealkylation, 66 marked atoms (26 metabolism sites) in case of O-dealkylation, 111 marked atoms (27 metabolism sites) in case of aliphatic hydroxylation, and 84 marked atoms (14 metabolism sites) in case of the aromatic hydroxylation, and 20 marked atoms (13 metabolism sites) in case of S-oxidation.

The baseline models already provide satisfactory prediction results, however some problems are also evident. For example, in case of aliphatic and aromatic hydroxylations, a large number of false positive predictions is observed (~20% of all predictions). This situation noticeably improves after the application of similarity correction. Analogous improvement of baseline model predictions is also obtained in case of N- and O-dealkylation models. The percent of correct classifications produced by these models approaches 100 if only predictions of high reliability (RI>0.5), which still constitute ca. 50% of the test set, are taken into account. These results prove the effectiveness of local similarity correction and the usefulness of Reliability Indices in assessing the Model Applicability Domain as the statistical parameters of the model clearly correlate with the reliability of considered predictions.

MODEL ADAPTATION FOR INDIVIDUAL CYTOCHROME P450 ENZYMES REGIOSELECTIVITY PREDICTION

Table 3 below presents results for a few selected examples of a GALAS model training with separate enzyme data. The baseline model for the prediction of each particular reaction regioselectivity in HLM was separately trained with a corresponding reaction data for each of the five considered individual P450 enzymes (e.g., in Table 3 HLM N-dealkylation baseline model was trained using CYP1A2 N-dealkylation data, HLM baseline model for the prediction of aromatic hydroxylation – using CYP2D6 aromatic hydroxylation data, etc.).

TABLE 3. Results for a few selected examples of an HLM model adaptation using separate enzyme data.

			Predicted value					
			Baseline*		Similarity Corrected*		Similarity Corrected (RI>0.5)	
			Positive	Negative	Positive	Negative	Positive	Negative
Observed value	CYP1A2 N-dealkylation	Positive	**14**	5	**13**	6	**4**	3
		Negative	25	**145**	7	**162**	3	**73**
	CYP2C9 O-dealkylation	Positive	**6**	1	**5**	2	**2**	0
		Negative	30	**135**	8	**157**	1	**77**
	CYP3A4 Aliphatic hydroxylation	Positive	**31**	17	**28**	20	**18**	8
		Negative	136	**538**	23	**651**	5	**453**
	CYP2D6 Aromatic hydroxylation	Positive	**21**	1	**17**	5	**11**	1
		Negative	176	**512**	15	**673**	5	**446**

* - Unreliable predictions (RI<0.3) were not considered which led to exclusion of 74 marked atoms (8 metabolism sites) from the initial test set in case of CYP1A2 N-dealkylation, 56 marked atoms (11 metabolism sites) in case of CYP2C9 O-dealkylation, 84 marked atoms (16 metabolism sites) in case of CYP3A4 aliphatic hydroxylation and 65 marked atoms (10 metabolism sites) in case of the CYP2D6 aromatic hydroxylation.

As it could have been expected prior to this test, the baseline HLM models produce large numbers of false positive predictions on data for separate enzymes. This is due to the fact that HLM models take into account metabolites produced by all enzymes, yet a particular enzyme of interest in each case can be only responsible for some of them. As a result, the sensitivity of HLM baseline models applied to the individual enzyme data is comparable to that of the HLM baseline model applied to the HLM test set, while the specificity is generally much lower (see Figure 2A).

After the application of similarity corrections based on the Self-training Libraries, containing corresponding individual enzyme data, the number of false positive atoms decreases considerably. Within the Model Applicability Domain (RI>0.3) the specificity parameters for individual enzyme regioselectivity predictions become comparable to those of the corresponding HLM data based models of metabolism in human liver microsomes (see Figure 2B).

Such results clearly illustrate that the GALAS modeling methodology allows for an easy and straightforward adaptation of the resulting HLM baseline models for predicting regioselectivity of any drug metabolizing enzyme.

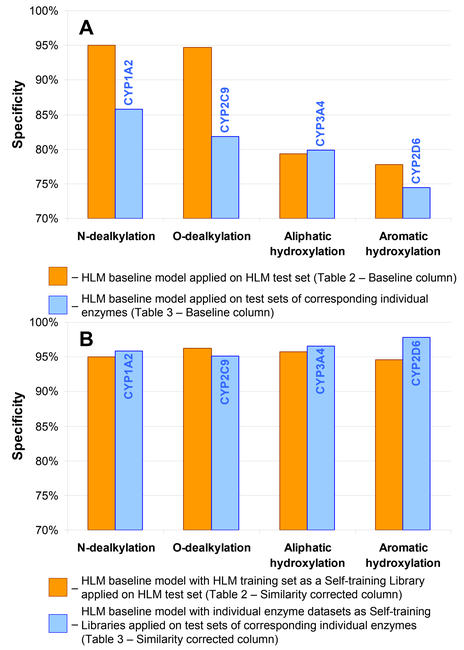


FIGURE 2. Specificity of metabolism regioselectivity models. A – HLM baseline models, B – Similarity corrected models with corresponding Self-training Libraries.

PREDICTION EXAMPLES

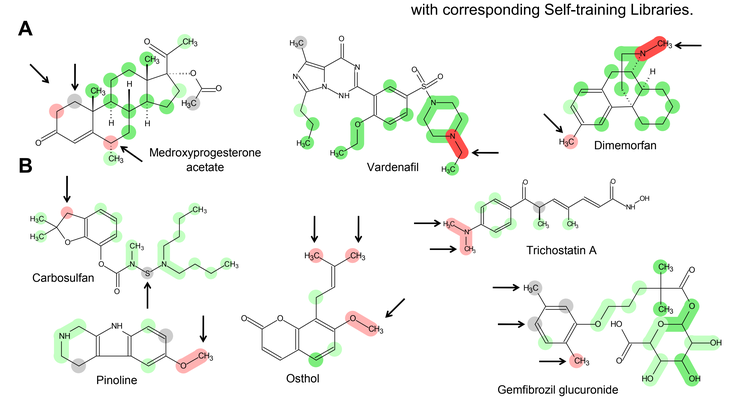


FIGURE 3. Examples of regioselectivity predictions for the compounds not used in the model development: A – compounds from well-known drug classes, B – novel drug-like molecules having nothing similar in the training set of the model. Arrows indicate experimentally identified metabolism sites.

REFERENCES

[1] Testa B. 'The Metabolism of Drugs and Other Xenobiotics: Biochemistry of Redox Reactions', Academic Press, San Diego, **1995**.
[2] Japertas P et al. *Quant. Struct.-Act. Relat.* **2002**, 21, 23.
[3] Sazonovas A et al. *SAR QSAR Environ. Res.* **2010**, 21, 127.
[4] Didziapetris R et al. *J. Comput. Aid. Mol. Des.* **2010,** 24, 891.

Advanced Chemistry Development, Inc.
Tel: (416) 368-3435
Fax: (416) 368-5596
Toll Free: 1-800-304-3988
Email: info@acdlabs.com
www.acdlabs.com