BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 3 2014, pages 398–405
doi:10.1093/bioinformatics/btt681

Systems biology

Advance Access publication November 22, 2013

In silico site of metabolism prediction for human UGT-catalyzed reactions

Jianlong Peng[1], Jing Lu[1], Qiancheng Shen[1], Mingyue Zheng[1,*], Xiaomin Luo[1,*], Weiliang Zhu[1], Hualiang Jiang[1,2] and Kaixian Chen[1]

[1]Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China and [2]School of Pharmacy, East China University of Science and Technology, Shanghai 200237, China

Associate Editor: Martin Bishop

ABSTRACT

Motivation: The human uridine diphosphate-glucuronosyltransferase enzyme family catalyzes the glucuronidation of the glycosyl group of a nucleotide sugar to an acceptor compound (substrate), which is the most common conjugation pathway that serves to protect the organism from the potential toxicity of xenobiotics. Moreover, it could affect the pharmacological profile of a drug. Therefore, it is important to identify the metabolically labile sites for glucuronidation.

Results: In the present study, we developed four *in silico* models to predict sites of glucuronidation, for four major sites of metabolism functional groups, i.e. aliphatic hydroxyl, aromatic hydroxyl, carboxylic acid or amino nitrogen, respectively. According to the mechanism of glucuronidation, a series of 'local' and 'global' molecular descriptors characterizing the atomic reactivity, bonding strength and physical–chemical properties were calculated and selected with a genetic algorithm-based feature selection approach. The constructed support vector machine classification models show good prediction performance, with the balanced accuracy ranging from 0.88 to 0.96 on test set. For further validation, our models can successfully identify 84% of experimentally observed sites of metabolisms for an external test set containing 54 molecules.

Availability and implementation: The software *somugt* based on our models is available at www.dddc.ac.cn/adme/jlpeng/somugt_win32.zip.

Contact: xmluo@simm.ac.cn or myzheng@mail.shcnc.ac.cn

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

Received on August 21, 2013; revised on October 20, 2013; accepted on November 18, 2013

1 INTRODUCTION

Human uridine diphosphate (UDP)-glucuronosyltransferases (UGTs) are major phase II drug-metabolizing enzymes that catalyze transfer of glucuronic acid from UDP-glucuronic acid to various substrates containing nucleophilic functional group, e.g. alcohols, phenols, carboxylic acids, amines, thiols and so forth. Up until now, 22 human UGT proteins have been identified, and they can be classified in four families: UGT1, UGT2, UGT3 and UGT8 (Rowland *et al.*, 2013). These isoforms are widely distributed and can be found in liver, intestine, kidney, lung and skin. Both endogenous (e.g. bilirubin) and xenobiotic (e.g. drug) chemicals can be metabolized by this family of enzymes, yielding water-soluble products that are easier to be excreted (Miners and Mackenzie, 1991). Most of those isoforms possess the overlapping substrate specificities as well as the capacity to metabolize a variety of structure unrelated substances. For example, UGT1A3 can catalyze glucuronidation of phenols, carboxylic acids and amines (Magdalou *et al.*, 2010).

In addition to an essential detoxification mechanism for structurally diverse drugs, the glucuronidation can also lead to short duration of action and the loss of pharmacological activity of the drugs. In such cases, medicinal chemist needs to optimize the lead compound to improve its pharmacokinetic properties. For example, to reduce potential for glucuronidation, Tanaka *et al.* (2007) designed metabolically more stable compounds by incorporating substituents at the ortho position of the phenolic hydroxyl group or exchanging the phenyl group. In addition, strategy like bioisosteric modifications of phenol moiety was also reported to overcome the issue of rapid glucuronidation (Kawada *et al.*, 2013; Wu *et al.*, 2005). In contrast, the property that glucuronidation may lead to a minimized systemic exposure can be used to reduce adverse events. For example, Millan *et al.* (2011) designed inhaled p38 inhibitors by introducing functional groups prone to be metabolized via cytochromes P450 (CYPs) or UGTs, to achieve high intrinsic clearance. Therefore, glucuronidation provides access to modifying the pharmacokinetic profile during lead optimization. It is necessary to identify whether a compound can be glucuronidated, as well as which functional group of the compound is able to undergo the reaction.

Normally the UGT-catalyzed biotransformation is determined experimentally by metabolite identification techniques. However, given that such experimental techniques are cost-intensive and time-consuming, there is growing interest to develop reliable *in silico* models to predict and deepen our understanding of the reaction. Till now, only the crystal structure of C-terminal domain of human UGT2B7 was determined (Miley *et al.*, 2007). Accordingly, current computational models for UGT-mediated reaction were based on the reported substrate profile. Early studies (Ethell *et al.*, 2002; Temellini *et al.*, 1991; Vashishtha *et al.*, 2001) mainly focused on analyzing kinetic constants for the glucuronidation of homologous series of substrates, suggesting that glucuronidation rate was related to

*To whom correspondence should be addressed.

molecular lipophilic, electronic and/or steric properties. Miners, Smith and Sorich extensively studied substrates and non-substrates of UGT isoforms ('reaction phenotyping') (Miners *et al.*, 2004; Smith *et al.*, 2003a, b; Sorich *et al.*, 2003, 2004a, b). By trying various machine learning methods, classification models for discriminating substrates of different UGT isoforms were developed. They also applied the strategy of pharmacophore fingerprint analysis to mine the structure signature related to glucuronidation, and found that local environment of nucleophilic site played an important role in identifying potential sites of metabolism (SOMs) (Smith *et al.*, 2003b; Sorich *et al.*, 2004b). Though many computational studies have been reported to simulate reaction rate and phenotyping, models predicting sites of glucuronidation were rare. Sorich *et al.* (2006) reported naïve Bayes models trained for eight UGT isoforms using three quantum chemical (QC) and constitutional descriptors. Their results indicated that the local environment of nucleophilic site is an important feature to characterize the SOM of glucuronidation, and general chemical properties may be of secondary importance. However, this study was based on a small dataset containing only dozens of glucuronidated sites, which limited its application to structurally diverse compounds outside the range of training data.

In the present study, we combined local QC descriptors and global physical–chemical properties to build four classification models to predict sites of glucuronidation using the learning method support vector machine (SVM) and a dataset covering a large chemical diversity space. First, human UGT-catalyzed reactions were retrieved from MDL Metabolite Database, and potential sites of glucuronidation were marked by matching predefined SMARTS strings. Then, both QC and geometrical descriptors were calculated to train SVM classifiers for each type of sites, i.e. aliphatic hydroxyl, aromatic hydroxyl, carboxylic and amino nitrogen. In the end, model performance was assessed by 10-fold cross-validation (10-fold CV) and test set. For further validation, another 54 molecules were collected and predicted using our constructed models.

2 METHODS

2.1 Datasets

Altogether 1377 *in vitro* human UGT-catalyzed reactions were retrieved from the MDL Metabolite Database. The following preparation procedures were automatically performed with an in-house C++ program using modules of OpenBabel 2.3.0 (O'Boyle *et al.*, 2011).

(1) *Extracting experimentally observed sites*. The experimentally observed SOMs were identified by applying a subgraph isomorphism algorithm to each reactant–product pair. Empirically, the substrate of a glucuronidation reaction is a subgraph of its product when ignoring hydrogen atoms. Figure 1a showed the glucuronidation of morphine as an example (Kilpatrick and Smith, 2005). By examining the 2D structures of morphine and its corresponding glucuronic acid conjugate, the SOM can be automatically determined at the OH group. Figure 1b and c presented two less common glucuronidation examples that were not considered in the current study.

(2) *Assigning class labels for potential SOMs*. The most common nucleophilic sites of glucuronidation include (Miners and Mackenzie, 1991): aliphatic or aromatic hydroxyl oxygen, singly bonded oxygen of carboxyl group, amino nitrogen, heterocyclic nitrogen, amide nitrogen and thiol sulfur. Table 1 summarized the corresponding SMARTS strings that were used to match potential SOMs. After removing the sites that appear too few times, we finally kept the following four types of sites: (i) aliphatic hydroxylic oxygen (*AlOH*), (ii) aromatic hydroxylic oxygen (*ArOH*), (iii) single bonded oxygen of a carboxylic acid group (*COOH*) and (iv) aromatic/aliphatic amino nitrogen and heterocyclic nitrogen (*Nitrogen*). Their SMARTS patterns and the number of reactants containing the corresponding groups are listed in Table 2. Among all the matched sites, those experimentally observed glucuronidation sites identified in step (1) were labeled as +1 (positive), and the others were labeled as −1 (negative).

(3) *Generating 3D structures and optimization*. The initial 3D structure of each reactant was constructed in Sybyl 6.8 (Sybyl, 2001) and further geometrically optimized using the semi-empirical AM1 method in the MOPAC 7.0 program (Stewart, 1990).

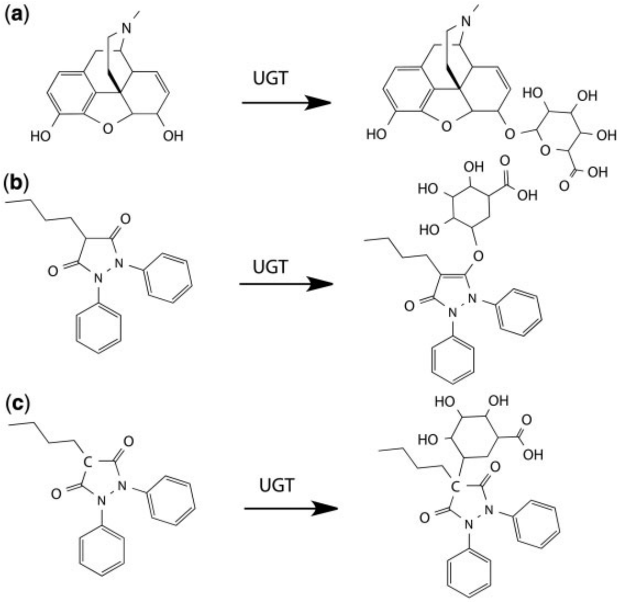

Fig. 1. Examples of glucuronidation reactions. (**a**) Reactant is a subgraph of the product when taking no account of hydrogen. (**b**) An example of uncommon reaction where keto-enol tautomerism occurs. (**c**) Phenylbutazone as an example of C-glucuronidation

To generate a representative test set for validation, the Kennard–Stone algorithm (Kennard and Stone, 1969) was used to split each dataset into a training and a test set in the ratio of 4:1. Moreover, to further evaluate the constructed models, an external test set containing 54 molecules was compiled from recently published studies using *in vitro* experiments to determine UGT-catalyzed SOMs. Each molecule contained at least one glucuronidation site. The same procedure described earlier in the text was applied to build and optimize 3D structures. The initial structures, their SOMs and original references are provided in Supplementary Material S1 and S2.

2.2 Descriptors

The mechanism of glucuronidation involves nucleophilic attack of the substrate on the cofactor (S_N2 reaction, a diagram was shown in Fig. 2) (Miners and Mackenzie, 1991). The descriptors capturing

Table 1. Definition of SMARTS of all potential sites[a]

Description	SMARTS	Description	SMARTS
Carboxyl oxygen	[OX2H][CX3] = O	Sulfonyl oxygen	[OX2H][SX4](=O)=O
Aromatic hydroxyl	[OX2H]a	Aliphatic Hydroxyl	[OX2H]A
Amide nitrogen	[#7X3][CX3] = O	Sulfonamide Nitrogen	[#7X3][SX4](=O)=O
Aromatic heterocyclic nitrogen	[n;!R0]	Aliphatic heterocyclic nitrogen	[n;!R0]
Aromatic amino nitrogen	[NX3,NX4]a	Aliphatic amino nitrogen	[NX3,NX4]A
Thiol sulfur	[SX2H]		

[a]For a given oxygen atom, SMARTS strings were tested in an order of carboxyl and sulfonyl > aromatic hydroxyl > aliphatic hydroxyl; for a given nitrogen atom, the order is amide, sulfonamide > N-heterocyclic > aromatic amine > aliphatic amine.

Table 2. Four types of glucuronidation sites to be modeled

Type	SMARTS pattern	Number of sites		Number of compounds
		Number of positive	Number of negative	
AlOH	[OX2H][A;!#1;!$ ([C,N,P,S] = O)]	210	254	301
ArOH	[OX2H]a	335	105	330
COOH	[OX2H][CX3] = O	117	42	154
Nitrogen[a]	[$([#7;!R0]), $([NX3,NX4]); !$([#7][C,S] = [O,S,N]); !$(N = O)]	125	506	311

[a]In the case of N-glucuronidation, one single model will be built on sites from heterocyclic nitrogen and aliphatic/aromatic amino nitrogen. And sites from amide were excluded in this study because the experimentally observed sites were rather rare.

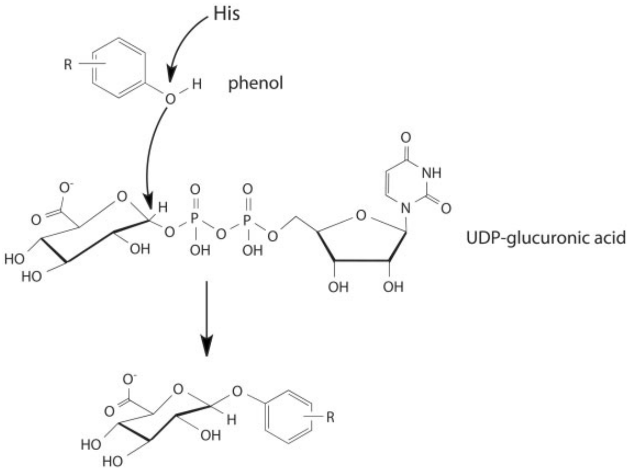

Fig. 2. Diagram showing glucuronidation of phenols by nucleophilic attacking on carbon of UDP-glucuronic acid. Catalytic base histidine helps proton abstraction from the phenol

information about the nature of nucleophilic atom (i.e. O, N) are supposed to contribute to discriminating whether a potential site can be glucuronidated. Previous studies (Cnubben *et al.*, 1994; Sorich *et al.*, 2004a, 2006; Zheng *et al.*, 2009) have shown that QC descriptors can well represent the atomic and molecular reactivity. Here 25 types of local QC features and 7 types of global QC features characterizing bond strength or reactivity of an atom were calculated (definition are available in Supplementary Material S3). Those one-center terms reflect the electrostatic behavior or reactivity ability of an atom toward the cofactor, whereas two-center terms are related to the bond strength between the atom and its bonded neighbors. More detailed information about these descriptors can be found elsewhere (Brown and Simas, 1982; Karelson *et al.*, 1996; Katritzky *et al.*, 1995–1997).

To calculate these QC features, the following key words were included in the input file of MOPAC program: AM1, MMOK, VECTORS, BONDS, PI, PRECISE, ENPART, EF, MULLIK and CHARGE = n, where n is the charge of the molecule studied. Then an in-house python program was used to calculate both local and global QC features by parsing the MOPAC output file. Because those two-center terms involve two atoms, the following transformations were made to use them as atomic descriptors. For the potential sites of *AlOH*, *ArOH* and *COOH*, both features of the bond O-H and O-C were separately calculated and used as atomic descriptors; for the sites of *Nitrogen*, all features of the bond N-X (where X can be any neighbor atom) were computed, of which the sum, max, min and mean values were used as atomic descriptors.

Besides, it has been shown that steric and lipophilic properties of the substrates were also responsible for the rate of glucuronidation (Cupid *et al.*, 1999; Ethell *et al.*, 2002; Kim, 1991; Mercier *et al.*, 1991; Vashishtha *et al.*, 2001). In this study, 13 molecular descriptors addressing geometrical and lipophilic properties were introduced (definitions are available in Supplementary Material S3). Geometrical descriptors were calculated via CODESSA 2.7.2 (CODESSA, 1995–2004), and logP (Wildman and Crippen, 1999) is computed using OpenBabel.

Overall, there are 56 descriptors calculated for the sites *AlOH*, *ArOH* and *COOH*, and 81 descriptors for the sites *Nitrogen*.

2.3 Support vector machine

SVM algorithm aims to find a hyperplane such that the sum of distances from the hyperplane to the nearest positive or negative training samples was maximized (Vapnik, 1999). The optimization problem can be described as follows:

$$\min\left\{\frac{\left\|\overrightarrow{w}\right\|^2}{2}+C_{+1}\sum_{\{i|y_i=+1\}}^{N_{+1}}\xi_i+C_{-1}\sum_{\{i|y_i=-1\}}^{N_{-1}}\xi_i\right\} \tag{1}$$

$$\text{subject to} \qquad \begin{array}{l} y_i*\left(\overrightarrow{w}*\overrightarrow{x}+b\right)\geq 1-\xi_i \\ \xi_i\geq 0 \end{array} \tag{2}$$

where ξ_i is the slack variable that allows training errors, whereas C_{+1} and C_{-1} are penalty of misclassifying positive and negative samples, respectively. Through those additional terms, SVM can handle non-linearly separable training data. Besides, kernel function is applied to avoid the dimensionality problems when constructing hyperplane in a high dimensional space. Details of the theory can be found elsewhere (Burges, 1998; Vapnik, 1999).

The SVM classifier for each reaction type was established using the LibSVM package (version 3.0) (Chang and Lin, 2001) with the radial basis function (RBF) kernel [$exp(-\gamma\|u\text{-}v\|^2$]. To avoid skewing to predicting majority class, a larger penalty is applied when misclassifying samples from minority class. In the package of LibSVM, the penalty of class i is calculated as $C_i = W_i{*}C$, where Wi is the weighting parameter. Here the weight of the majority class was set to be 1, and the weight of the minority class was determined according to the following relationship:

$$\frac{W_{+1}}{W_{-1}} = \frac{NEG}{POS} \tag{3}$$

where POS and NEG is the number of reactive and non-reactive sites, respectively. Finally, parameters C and γ were optimized simultaneously during genetic algorithm (GA)-based feature selection as described in the next section.

2.4 Feature selection and parameter optimization

Overfitting often occurs when a model uses more terms than necessary (Hawkins, 2004). Because the degree of degeneracy of our defined descriptors can be high, models built using all of them could be easily overfitting. Therefore, it is important to select a subset of relevant and non-redundant features. Up until now, there are a couple of frequently used feature selection techniques (Gonzalez *et al.*, 2008), such as forward and backward stepwise, GA, replacement method and so forth.

In the present study, for each pair of highly correlated features (Pearson correlation coefficient $r^2 \geq 0.9$), the one with smaller *F*-score was removed first. The F-score of *j*-th descriptor is defined as follows (Guyon *et al.*, 2002):

$$F_j = \frac{\left(\overline{X_j^{+1}} - \overline{X_j}\right)^2 + \left(\overline{X_j^{-1}} - \overline{X_j}\right)^2}{\frac{1}{n_{+1}-1}\sum_k \left(X_{k,j}^{+1} - \overline{X_j^{+1}}\right)^2 + \frac{1}{n_{-1}-1}\sum_k \left(X_{k,j}^{-1} - \overline{X_j^{-1}}\right)^2} \tag{4}$$

where, $\overline{X_j^{+1}}, \overline{X_j^{-1}}, \overline{X_j}$ are the average values of *j*-th descriptor of positive, negative and the overall samples, respectively; n_{+1}, n_{-1} are the number of positive and negative samples; $X_{k,j}^{+1}, X_{k,j}^{-1}$ are the *k*-th value of the *j*-th descriptor of positive and negative samples, respectively. Feature with a larger *F*-score was supposed to contribute more to the followed classification modeling.

Then, the framework of GA-based feature selection strategy as described previously (Wang *et al.*, 2010) was applied to select the descriptors and optimize the SVM parameters (i.e. C, γ) simultaneously. The fitness function of GA was defined as

$$Fitness = BACC_{cv} - weight * N_{desp} \tag{5}$$

where $BACC_{cv}$ is the balanced accuracy (defined in the next section) of 10-fold CV, N_{desp} is the number of selected descriptors to build SVM model and *weight* controls the number of selected descriptors. Different *weight* values were tested, and it was set to 0.0015 due to the associated $BACC_{cv}$ being consistently higher during the GA evolution.

2.5 Model performance

Given the quantities of true positive (TP), false negative (FN), true negative (TN) and false positive (FP), a classifier can usually be estimated by sensitivity [SE = TP/(TP + FN)], specificity [SP = TN/(TN + FP)] and accuracy [ACC = (TP + TN)/(TP + FN + TN + FP)]. SE is the measure of the rate correctly identifying a site as reactive given it is truly glucuronidated; SP, in contrast, is the measure of the rate correctly identifying a site as non-reactive given it is non-glucuronidated. Accuracy is the widely used measure of overall classification performance, but it is not appropriate when the dataset is imbalanced. Considering that the ratios of reactive to non-reactive sites for different reaction types varies from 0.25 to 3.19, three alternative metrics, balanced accuracy [BACC = (SE + SP)/2] (Lemnaru and Potolea, 2012), area under receiver operating characteristic curve (AUC) (Fawcett, 2006) and Matthews correlation coefficient (MCC) (Matthews *et al.*, 1975), were used to evaluate the performance of obtained models.

3 RESULTS

3.1 Overview of glucuronidation sites

To model the sites of glucuronidation, dataset containing four types of functional groups were collected and listed in Table 2. When looking at the fraction of SOMs of each functional group, we can see that the ability of glucuronidation varies over different atom species. In general, oxygen atoms have more tendency to be glucuronidated over nitrogen atoms, whereas the local structure of oxygen atoms can also affect the reaction. As shown in Table 2, both *ArOH* and *COOH* have much more reactive sites than non-reactive ones, whereas *AlOH* has nearly same number of reactive and non-reactive sites. As shown in previous study (Kerdpin *et al.*, 2009; Magdalou *et al.*, 2010), deprotonation of the substrate is one of the key steps during O-glucuronidation. Thus the observed different reactive SOM ratios could be partially explained by the deprotonation ability of different reaction types. If other substituent effects are neglected, the alkyl group of *AlOH* has electron donating effect, which increases the electron density of O-H bond. On the contrary, the electron density of O-H bond of *ArOH* and *COOH* is lower due to the withdrawing effect of the corresponding groups. Accordingly, the interaction between oxygen and hydrogen of *AlOH* could be in general stronger than that of *ArOH* and *COOH*, leading to less deprotonated oxygens and reactive sites. Therefore, the reactivity of a given atom could be altered by the local environment, and descriptors characterizing such influence were supposed to be a discriminating predictor. For example, of the calculated descriptors, Mulliken (1955) bond order, also known as Mulliken's overlap population, can characterize the accumulation of the electrons between two bonded atoms, and thus measures the strength of covalent bonding. Wilcoxon rank-sum test showed that *B_OH* (Mulliken bond order of O-H bond) of *AlOH* is on average greater than that of *ArOH* and *COOH* (both *P*-values are <0.001).

3.2 Validation of classifiers

A combination of selected descriptors and SVM parameters with the highest GA fitness score were used to model the final classifiers for each kind of glucuronidation reaction type. As being part of the fitness function of GA, BACC of 10-fold CV of each model was optimized during feature selection. After thousands of generations, the final result of 10-fold CV was listed in Table 3, from which one can see that a well and balanced performance can be obtained after GA-based feature selection. The statistics of each model on test set was also listed in Table 3. The results indicate that our classifiers have generally high performance of

Table 3. Statistics of model validation

Dataset	Type	SE	SP	BACC	AUC	MCC
Training set	*AlOH*	0.89	0.81	0.85	0.95	0.70
	ArOH	0.86	0.94	0.90	0.95	0.72
	COOH	0.94	0.94	0.94	0.98	0.84
	Nitrogen	0.91	0.92	0.92	0.97	0.78
Test set	*AlOH*	0.83	0.94	0.89	0.95	0.78
	ArOH	0.85	0.90	0.88	0.92	0.68
	COOH	0.92	0.89	0.90	0.94	0.78
	Nitrogen	0.96	0.96	0.96	0.96	0.88
10-fold CV	*AlOH*	0.84	0.79	0.82	0.85	0.63
	ArOH	0.82	0.80	0.81	0.84	0.57
	COOH	0.87	0.88	0.87	0.87	0.70
	Nitrogen	0.86	0.91	0.88	0.89	0.72

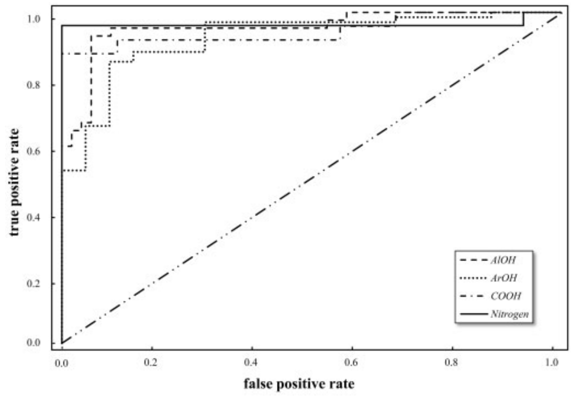

Fig. 3. ROC curves of prediction of four models on test set. The classifier is more discriminative if its ROC curve locates in the top left corner further away from the diagonal line

predicting both glucuronidated and non-glucuronidated sites, with BACC being 0.89, 0.88, 0.90 and 0.96, respectively. Besides, receiver operating characteristic (ROC) curve of each model on test set was shown in Figure 3, where the ROC curve for a random guess results was a diagonal line from bottom left to top right. It can be found that all curves were located far above the diagonal line, suggesting strong predictive performance of the obtained glucuronidation models.

To further evaluate our models, an external test set including 54 molecules was collected from reference. This set contains all four types of SOMs. By applying the built models to these molecules, 129 of 148 potential SOMs were correctly predicted with an overall accuracy of 0.87. Altogether 57 of 68 experimentally observed SOMs were successfully identified, yielding an SE value of 0.84. A summary of prediction results was listed in Table 4, from which we can see that the four models recognized most of experimentally observed SOMs. Some examples were shown in Figure 4, where potential SOMs were marked with circles. For example, dihydropyrrolopyrimidine is a PI3K inhibitor designed by Kawada *et al.* (2013) with IC50 value of 0.0086 μM. This compound was metabolically unstable because of rapid glucuronidation of the phenol moiety, as reflected by a UGT–glucuronidation activity assay and metabolite identification technique using LC/MS/MS. Using our models, the phenolic site can be successfully identified as reactive, which will provide useful information for structural modification to change the pharmacokinetic profile of the compound.

Table 4. Summary of prediction results of the external test set

Type	TP	FN	TN	FP
AlOH	17	4	2	1
ArOH	18	5	1	1
COOH	10	0	1	0
Nitrogen	12	2	68	6

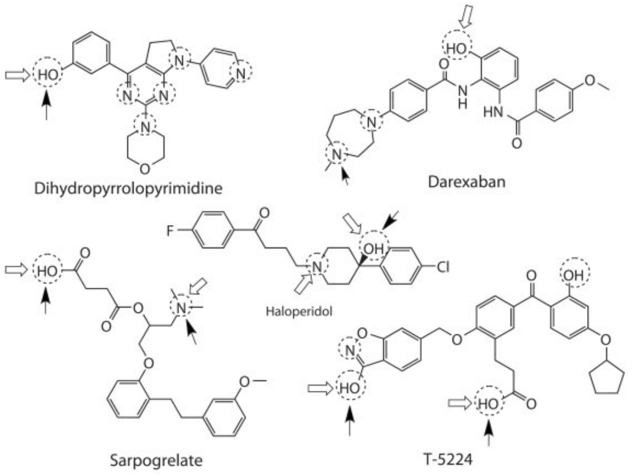

Fig. 4. Examples to show the UGT-catalyzed SOMs predicted by our models. Potential SOMs were marked with dashed circles; experimental observed SOMs were marked with hollow arrows; and predicted reactive SOMs were marked with black arrows. Amide nitrogens were excluded from potential SOMs

Another point of note is that there are 74 negative sites of *Nitrogen* type among those 54 molecules, and only 6 of them were misclassified as positive. It indicated that our weighted SVM classifiers can achieve excellent prediction performance for skewed data distribution. The detailed information about potential, experimentally observed and predicted SOMs of each molecule are provided in Supplementary Material S2.

3.3 Selected descriptors

To characterize the sites of glucuronidation, the initial descriptor set was reduced and optimized with a GA-based selection procedure. Table 5 summarized the selected descriptors for each model.

Among all those selected two-center QC descriptors, energy-related ones (e.g. *J*, *B*, *C*, *EE2*, *NN2*, *C2*) characterize the interaction energy between two bonded atoms (i.e. bonding strength), which is related to the deprotonation ability in the case of O-glucuronidation. The *polar_AB* describes how electronic charge of the studied reaction site is perturbed by its neighbor

Table 5. Descriptors and model parameters selected through GA

Type	Model parameters		Selected descriptors[a]
	C	γ	
AlOH	256.0	0.25	NA, EE2_OH, J_OH, B_OH, polar_AB_OH, B$\sigma\pi$_OH, B$\sigma\pi$_OX, HOMO, PMOIX, YZShadow, ZXShadow, MV, logP
ArOH	256.0	0.25	EA, SEA, SNA, C_OH, J_OH, EE2_OX, NN2_OX, B$\sigma\sigma$_OX, ABS_hard, XYShadow, ZXShadow/ZXR, logP
COOH	8.0	2.0	RA, SEA, EN1, E2_OH, E2_OX, B$\sigma\sigma$_OX, dipoleT, YZShadow/YZR, MV, MV/XYZBox, logP
Nitrogen	0.5	16.0	RA, EN1, C_sum, J_max, NN2_sum, B_sum, polar_AA, polar_AB_sum, B$\sigma\sigma$_mean, YZShadow

[a] Definition of each selected descriptor can be found in Supplementary Material S2.

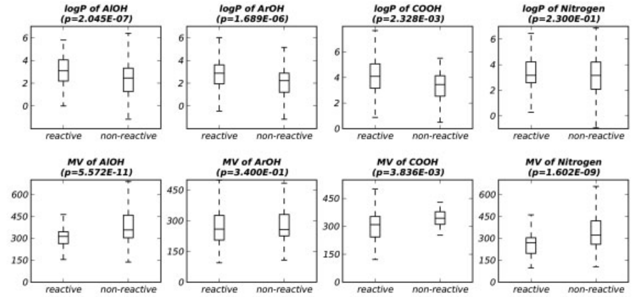

Fig. 5. Boxplots to show the distribution of descriptors *logP* and *MV*. The box edges are first and third quartiles, the band inside the box is the median. The values in parentheses are *P*-values of Wilcoxon rank-sum test. For *logP*, the alternative hypothesis is that values of molecules containing reactive sites are more likely larger than those of molecules containing non-reactive sites; for *MV*, the alternative hypothesis is that values of molecules containing reactive sites are more likely less than those of molecules containing non-reactive sites

(Karelson *et al.*, 1996). Of the selected one-center QC descriptors, *NA* (Fukui atomic nucleophilic reactivity index) and *SEA* (electrophilic superdelocalizability) describe the nucleophilicity of an atom toward glucuronic acid, and *RA* (Fukui atomic one-electron reactivity index) is a general indicator of the atomic susceptibility to electrophilic or nucleophilic attack (Brown and Simas, 1982; Karelson *et al.*, 1996). These three descriptors are representation of local reactivity of a molecule from different emphases. Accordingly, at least one of them was selected for each resulted models, which suggested that the GA-based approach is efficient in selecting relevant descriptors.

Apart from those local descriptors, we also calculated a couple of physical–chemical descriptors to characterize the overall property that a molecule possesses. These global descriptors, such as *logP* and molecular volume (*MV*), are useful to understand the favorable properties that a molecule should possess to be glucuronidated. As shown in Figure 5, for O-glucuronidation sites (i.e. *AlOH*, *ArOH* and *COOH*), the value of *logP* of the molecules containing reactive sites are more likely to be larger than those containing non-reactive sites. *P*-values of Wilcoxon rank-sum test were <0.005. The value of *MV* for the sites of *AlOH*, *COOH* and *Nitrogen*, instead, shows a reversed tendency. From these results, we may infer that molecules with higher lipophilicity and relatively small molecular size are generally more liable to glucuronidation, which is consistent with the intuitive perception of UGT substrates.

4 DISCUSSION

Sorich *et al.* (2006) reported *in silico* models for the first time to predict sites of glucuronidation. By analyzing the frequency of each functional group being metabolized, they found that the susceptibility of different nucleophilic types varies toward UGT-mediated glucuronidation. For example, UGTs prefer the atom sites attaching aromatic rings over those with aliphatic substituents. These observations are consistent with the distribution analysis result of our dataset. One of the major differences between Sorich's and our work is that their models predict isozyme-specific regioselectivity, whereas our models are reaction-specific that consider the overall effect of all UGT isoforms. Given the genetic polymorphism of the enzyme, it is expected that the isozyme-specific models may help understanding and predicting individual differences in drug response. However, after analyzing the metabolism data collected here, it is clear that there is a significant overlap between the isoforms. Excluding all substrates with overlapping specificity would not yield statistically meaningful results because the datasets were too small. In this study, we choose to construct separate SOM models for different functional groups, as they may possess different properties. For each of them, a GA-based approach was used to select the appropriate descriptors to characterize their reactivity. From the cross-validation results (only the results of 10-fold CV were provided in Sorich's work), our models showed improved and more balanced performance. Moreover, our models were evaluated by both internal and external test sets, and all these evaluations suggest that our models have excellent performances on predicting potential glucuronidation sites. Actually, at early stages of drug discovery, the UGT isoform responsible for a specific pathway is seldom experimentally determined, and medicinal chemists care more about the accurate 'soft' sites for structural protection and modification. The reaction-specific methods, due to a better consideration of local reactivity, would be more suitable for improving the prediction of UGT-mediated SOMs.

In general, the current models are capable of reproducing 84% of experimentally observed SOMs. Though reasonably successful, we noticed that they tend to give incorrect predictions sometimes, especially for some molecules containing a large number of potential sites. The reasons for this can be multifold, but primarily we think it is because the current models have not accounted for the kinetic issue. For a molecule containing many potential sites, those sites may compete for UGT with each other, and have different reaction rates. Consequently, there are only a few major routes, and some 'actually' active sites are not experimentally eminent. Because the current models cannot compare the priority of different sites within or across reactions, they do not match all of the experimentally observed SOMs. In addition, the current study only used the local reactivity descriptors and the overall physical–chemical properties, whereas there are other factors that should be modeled. For example, the local steric effect of ligand may hinder the nucleophilic attack of reaction site on the cofactor. Therefore, further improvement should focus on the reaction kinetic issue and try to take more factors related to the metabolism into consideration.

5 CONCLUSION

In the present work, we developed four *in silico* models to predict human UGT-catalyzed SOMs for the following potential sites: aliphatic hydroxyl, aromatic hydroxyl, carboxyl group, aliphatic/aromatic amino nitrogen and heterocyclic nitrogen. Extensive validations were carried out on the models, and the results demonstrated good predicting performances on both internal and external tests. Among the GA-selected descriptors, we found that local descriptors such as bond strength (between reactive site and its neighbor) and Fukui atomic reactivity indices are important to characterize the glucuronidation potential of a site. Moreover, the distribution analysis of global descriptors suggested that molecules with higher lipophilicity and relatively small molecular size are generally more liable to glucuronidation. Overall, the current study may provide insight for understanding human phase II metabolic profile of a chemical, and the resulted *in silico* models allow chemists to make rational decisions more quickly during pharmacokinetic lead optimization.

Funding: Hi-TECH Research and Development Program of China (2012AA020308), National S&T Major Project (2012ZX09301-001-002) and National Natural Science Foundation of China (21021063, 81001399, 81230076 and 21210003).

Conflict of Interest: None declared.

REFERENCES

Brown,R.E. and Simas,A.M. (1982) On the applicability of CNDO indices for the prediction of chemical reactivity. *Theor. Chim. Acta.*, **62**, 1–16.

Burges,C.J.C. (1998) A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Disc.*, **2**, 121–167.

Chang,C.C. and Lin,C.J. (2001) LIBSVM: a library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

Cnubben,N.H.P. *et al.* (1994) Molecular orbital-based quantitative structure-activity relationship for the cytochrome P450-catalyzed 4-hydroxylation of halogenated anilines. *Chem. Res. Toxicol.*, **7**, 590–598.

CODESSA, v 2.7.2. Semichem, Inc. 1995–2004. Shawnee, KS 66216.

Cupid,B.C. *et al.* (1999) Quantitative structure-metabolism relationships (QSMR) using computational chemistry: pattern recognition analysis and statistical prediction of phase II conjugation reactions of substituted benzoic acids in the rat. *Xenobiotica*, **29**, 27–42.

Ethell,B.T. *et al.* (2002) Quantitative structure activity relationships for the glucuronidation of simple phenols by expressed human UGT1A6 and UGT1A9. *Drug Metab. Dispos.*, **30**, 734–738.

Fawcett,T. (2006) An introduction to ROC analysis. *Pattern Recognit. Lett.*, **27**, 861–874.

Gonzalez,M.P. *et al.* (2008) Variable selection methods in QSAR: an overview. *Curr. Top. Med. Chem.*, **8**, 1606–1627.

Guyon,I. *et al.* (2002) Gene selection for cancer classification using support vector machines. *Mach. Learn.*, **46**, 389–422.

Hawkins,D.M. (2004) The problem of overfitting. *J. Chem. Inf. Comp. Sci.*, **44**, 1–12.

Karelson,M. *et al.* (1996) Quantum-chemical descriptors in QSAR/QSPR studies. *Chem. Rev.*, **96**, 1027–1043.

Katritzky,A.R. *et al.* (1995–1997) CODESSA Reference Manual, version 2.0. Semichem and the University of Florida, Gainesville, Florida.

Kawada,H. *et al.* (2013) Lead optimization of a dihydropyrrolopyrimidine inhibitor against phosphoinositide 3-kinase (PI3K) to improve the phenol glucuronic acid conjugation. *Bioorg. Med. Chem. Lett.*, **23**, 673–678.

Kennard,R.W. and Stone,L.A. (1969) Computer aided design of experiments. *Technometrics*, **11**, 137–148.

Kerdpin,O. *et al.* (2009) Influence of N-terminal domain histidine and proline residues on the substrate selectivities of human UDP-glucuronosyltransferase 1A1, 1A6, 1A9, 2B7, and 2B10. *Drug Metab. Dispos.*, **37**, 1948–1955.

Kilpatrick,G.J. and Smith,T.W. (2005) Morphine-6-glucuronide: actions and mechanisms. *Med. Res. Rev.*, **25**, 521–544.

Kim,K.H. (1991) Quantitative structure-activity relationships of the metabolism of drugs by uridine diphosphate glucuronosyltransferase, *J. Pharm. Sci.*, **80**, 966–970.

Lemnaru,C. and Potolea,R. (2012) Imbalanced classification problems: systematic study, issues and best practices. In: *Enterprise Information Systems*. Springer, Berlin Heidelberg, pp. 35–50.

Magdalou,J. *et al.* (2010) Insights on membrane topology and structure/function of UDP-glucuronosyltransferases. *Drug Metab. Rev.*, **42**, 159–166.

Matthews,B.W. *et al.* (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochem. Biophys. Acta*, **405**, s442–s452.

Mercier,C. *et al.* (1991) Modeling alcohol metabolism with the Darc Calphi system. *J. Med. Chem.*, **34**, 934–942.

Miley,M.J. *et al.* (2007) Crystal structure of the cofactor-binding domain of the human phase II drug-metabolism enzyme UDP-glucuronosyltransferase 2B7. *J. Mol. Biol.*, **369**, 498–511.

Millan,D.S. *et al.* (2011) Design and synthesis of inhaled p38 inhibitors for the treatment of chronic obstructive pulmonary disease. *J. Med. Chem.*, **54**, 7797–7814.

Miners,J.O. and Mackenzie,P.I. (1991) Drug glucuronidation in humans. *Pharmacol. Ther.*, **51**, 347–369.

Miners,J.O. *et al.* (2004) Predicting human drug glucuronidation parameters: application of *in vitro* and *in silico* modeling approaches. *Annu. Rev. Pharmacol.*, **44**, 1–25.

Mulliken,R.S. (1955) Electronic population analysis on LCAO-MO molecular wave functions. 1. *J. Chem. Phys.*, **23**, 1833–1840.

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

Rowland,A. *et al.* (2013) The UDP-glucuronosyltransferases: their role in drug metabolism and detoxification. *Int. J. Biochem. Cell Biol.*, **45**, 1121–1132.

Smith,P.A. *et al.* (2003a) *In silico* insights: Chemical and structural characteristics associated with uridine diphosphate-glucuronosyltransferase substrate selectivity. *Clin. Exp. Pharmacol. Physiol.*, **30**, 836–840.

Smith,P.A. *et al.* (2003b) Pharmacophore and QSAR modeling: complementary approaches for the rationalization and prediction of UDP-glucuronosyltransferase 1A4 substrate selectivity. *J. Med. Chem.*, **46**, 1617–1626.

Sorich,M.J. *et al.* (2003) Comparison of linear and nonlinear classification algorithms for the prediction of drug and chemical metabolism by human UDP-glucuronosyltransferase isoforms. *J. Chem. Inf. Comp. Sci.*, **43**, 2019–2024.

Sorich,M.J. *et al.* (2004a) Rapid prediction of chemical metabolism by human UDP-glucuronosyltransferase isoforms using quantum chemical descriptors derived with the electronegativity equalization method. *J. Med. Chem.*, **47**, 5311–5317.

Sorich,M.J. *et al.* (2004b) Multiple pharmacophores for the investigation of human UDP-glucuronosyltransferase isoform substrate selectivity. *Mol. Pharmaco.l*, **65**, 301–308.

Sorich,M.J. *et al.* (2006) The importance of local chemical structure for chemical metabolism by human uridine 5 '-diphosphate – glucuronosyltransferase. *J. Chem. Inf. Model.*, **46**, 2692–2697.

Stewart,J.J.P. (1990) Special issue - mopac - a semiempirical molecular-orbital program. *J. Comput. Aided Mol. Des.*, **4**, 1–45.

Sybyl, v 6.8. Tripos Inc. 2001. St. Louis, MO 63144-2913.

Tanaka,R. *et al.* (2007) Design and synthesis of piperidine farnesyltransferase inhibitors with reduced glucuronidation potential. *Bioorg. Med. Chem.*, **15**, 1363–1382.

Temellini,A. *et al.* (1991) Human liver sulphotransferase and UDP-glucuronosyltransferase: structure-activity relationship for phenolic substrates. *Xenobiotica*, **21**, 171–177.

Vapnik,V. (1999) *The Nature of Statistical Learning Theory (Information Science and Statistics)*. Springer, New York.

Vashishtha,S.C. *et al*. (2001) Quaternary ammonium-linked glucuronidation of 1-substituted imidazoles: studies of human UDP-glucuronosyltransferases involved and substrate specificities. *Drug Metab. Dispos*., **29**, 1290–1295.

Wang,Y. *et al*. (2010) Using support vector regression coupled with the genetic algorithm for predicting acute toxicity to the fathead minnow. *SAR QSAR Environ. Res*., **21**, 559–570.

Wildman,S.A. and Crippen,G.M. (1999) Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inf. Comp. Sci*., **39**, 868–873.

Wu,W.L. *et al*. (2005) Dopamine D1/D5 receptor antagonists with improved pharmacokinetics: design, synthesis, and biological evaluation of phenol bioisosteric analogues of benzazepine D1/D5 antagonists. *J. Med. Chem*., **48**, 680–693.

Zheng,M. *et al*. (2009) Site of metabolism prediction for six biotransformations mediated by cytochromes P450. *Bioinformatics*, **25**, 1251–1258.