Structural bioinformatics

Site of metabolism prediction for six biotransformations mediated by cytochromes P450

Mingyue Zheng, Xiaomin Luo, Qiancheng Shen, Yong Wang, Yun Du, Weiliang Zhu, and Hualiang Jiang

1Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, 555 Zuchongzhi Road, Shanghai 201203 and 2School of Pharmacy, East China University of Science and Technology, Shanghai 200237, China

Received on November 27, 2008; revised on February 9, 2009; accepted on March 9, 2009
Advance Access publication March 13, 2009

ABSTRACT

Motivation: One goal of metabolomics is to define and monitor the entire metabolite complement of a cell, while it is still far from reach since systematic and rapid approaches for determining the biotransformations of newly discovered metabolites are lacking. For drug development, such metabolic biotransformation of a new chemical entity (NCE) is of more interest because it may profoundly affect its bioavailability, activity and toxicity profile. The use of in silico methods to predict the site of metabolism (SOM) in phase I cytochromes P450-mediated reactions is usually a starting point in metabolic pathway studies, which may also assist in the process of drug/lead optimization.

Results: This article reports the Cytochromes P450 (CYP450)-mediated SOM prediction for the six most important metabolic reactions by incorporating the use of machine learning and semi-empirical quantum chemical calculations. Non-local models were developed on the basis of a large dataset comprising 1858 metabolic reactions extracted from 1034 heterogeneous chemicals. For validation, the overall accuracies of all six reaction types are higher than 0.81, four of which exceed 0.90. In further receiver operating characteristic (ROC) analyses, each of the SOM model gave a significant area under curve (AUC) value over 0.86, indicating a good predicting power. An external test was made on a previously published dataset, of which 80% of the experimentally observed SOMs can be correctly identified by applying the full set of our SOM models.

Availability: The program package SOME_v1.0 (Site Of Metabolism Estimator) developed based on our models is available at http://www.dddc.ac.cn/adme/myzheng/SOME_v1_0.tar.gz.

1 INTRODUCTION

The human cytochromes P450 (CYP450s) constitute a large family of thiolate-heme enzymes, and play a pivotal role in bio-transformation of both exogenous and endogenous compounds (Denisov et al., 2005). In drug development, the CYP-mediated metabolism of a new chemical entity (NCE) is of particular interest because it may profoundly affect the initial bioavailability, the desired activity and the safety profile of the compound. Any research leading to a deeper insight into the mechanistic aspects of metabolism will aid substantially in the rational design of drug candidates.

The site of metabolism (SOM) refers to the place in a molecule where the metabolic reaction occurs. Identification of phase I CYP-mediated SOM is usually a starting point in metabolic pathway investigation, and may assist in the process of lead optimization. If an active metabolite has improved pharmacological, pharmacokinetic and toxicological profiles compared with the parent, it can be conveniently used as a lead or even advanced to the clinic (Fura, 2006); while if a metabolite is unwanted, the SOM information can further guide the structure modification to a direction that will deactivate or eliminate the unstable sites to avoid such biotransformations. The experimental CYP-mediated SOM determination are highly time consuming and labor intensive: each candidate drug must be synthesized and assayed with identification of the site of metabolism will aid substantially in the rational design of drug candidates.

Some computational tools and models for predicting probable SOM have been developed, which mainly fall into two categories, ligand (substrate)-based and structure (enzyme)-based approaches. The former can further take two basic forms: expert systems that have libraries of molecular fragments reported to be involved in metabolic transformations, and computational models that attempt to explain reactivity utilizing molecular orbital calculations. Enzyme-based methods employ the structural information, in which a SOM is identified that is consistent with proximity to the enzyme reactive center. The differences between these methods as well as their individual capabilities have been reviewed elsewhere (Afzelius et al., 2007; de Graaf et al., 2005; Langowski and Long, 2002).

Intuitively, SOM can be roughly identified by the recognition of key functional groups within the complete chemical structure. The metabolic rules on O-, N-demethylation and aromatic hydroxylation have been established for about 30 years, some of which were extensively used in drug development (Bodor, 1997). However, the prediction accuracy of such methods is rather limited. To date, a

© The Author 2009. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org
AM1 molecular orbital calculations were developed to estimate the reactions reported for mammalian CYPs were initially exported from substrates. CYP450-mediated SOM, using only the 3D structure of the potential implement and computationally inexpensive method for predicting step. Both internal and external validations were performed to and the most discriminative model can be determined during this procedure was performed with the aid of various weighting schemes, supporting vector machines to classify whether a site undergoes a reactivity for the potential SOMs, which were used as input to train (QC)-based atomic features were calculated to represent the local data extracted from substrate structures. Then, quantum chemical type an individual dataset was prepared, containing entries of SOM by matching of characteristic functional groups. For each reaction with those from MetaSite (Cruciani 2013) and the yield field of this reaction is labeled as ‘minor’ or ‘trace’, which were deduced to minimize the risks brought by pooling data from different sources and measured by different assays. In the original database, all data were stored in the form of reactions and no SOMs were from different sources and measured by different assays. In the original database, all data were stored in the form of reactions and no SOMs were different colors correspond to different reaction centers.

Table 1. Six CYP450-mediated metabolic transformations, their SMARTS definitions and descriptive statistics of their corresponding SOM databases

<table>
<thead>
<tr>
<th>Type</th>
<th>SMARTS pattern</th>
<th>No. of reactions</th>
<th>No. of compounds</th>
<th>No. of SOM</th>
<th>Pos.</th>
<th>Neg.</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>[CH][O]</td>
<td>549</td>
<td>304</td>
<td>497</td>
<td>2344</td>
<td>2841</td>
<td></td>
</tr>
<tr>
<td>II</td>
<td>([H][O][F][Cl,Br,I])</td>
<td>476</td>
<td>305</td>
<td>503</td>
<td>1564</td>
<td>2067</td>
<td></td>
</tr>
<tr>
<td>III</td>
<td>([S][O][P][S])</td>
<td>483</td>
<td>355</td>
<td>503</td>
<td>1283</td>
<td>1788</td>
<td></td>
</tr>
<tr>
<td>IV</td>
<td>([S][O][S])</td>
<td>184</td>
<td>158</td>
<td>172</td>
<td>511</td>
<td>683</td>
<td></td>
</tr>
<tr>
<td>V</td>
<td>[N]</td>
<td>69</td>
<td>59</td>
<td>63</td>
<td>118</td>
<td>181</td>
<td></td>
</tr>
<tr>
<td>VI</td>
<td>[F]</td>
<td>97</td>
<td>74</td>
<td>81</td>
<td>6</td>
<td>87</td>
<td></td>
</tr>
</tbody>
</table>

These experimentally observed SOMs were marked by keeping a record of their atomic indices (for N- and O-dealkylations the carbon of the leaving group adjacent to the nitrogen or oxygen atom were marked). Topologically equivalent atoms were marked identically. Reactants of all the retrieved transformations were then categorized according to their reaction types, and filtered based on molecular weight (below 800) and atom-type content (only C, N, O, H, S, F, Cl, Br and I were permitted). Initial three-dimensional (3D) structures of these reactants were generated using the CORDINA program (Gasteiger et al., 1990), which were subsequently optimized using the semi-empirical AM1 method from the MOPAC package (Stewart, 1994). All the geometries had been fully optimized in XYZ mode with the convergence parameter GNORM = 0.01.

Within each of the six classes, reactions tend to involve a particular functional group transformation. For example, a compound that undergoes aliphatic hydroxylation reaction must have at least one aliphatic carbon atom that qualifies it as a substrate. These carbon atoms are in fact reaction centers indicating potential locations of aliphatic hydroxylation SOMs. Therefore, for any given reactant structures, an atom or group is a potential SOM only if it has structural features compatible with a class-specific reaction center, some examples of which were shown in Figure 1.

Like any molecular pattern, the reaction centers can be precisely encoded using SMARTS strings (James et al., 2004). In Table 1, we summarized all these predefined reaction center patterns. For instance, the SMARTS string \([\{N\}[O][O]]\) matches a carbon atom of an alkyl group single bonded to a nitrogen atom, which indicates a potential N-dealkylation SOM. As described above, the atomic indices of real SOMs (metabolic sites that have been experimentally observed) had been recorded, so we can divide the potential SOMs into positive (labeled as ‘+’) and negative (‘−’) examples. In the end, the resulted dataset of each reaction was randomly split into training and independent test sets in the ratio of 4:1.

The external test set compiled by Sheridan et al. (2007) was used for further model validation. For the external test set, the structure of Ketobemidone was found identical to MMTB00032411 of our training set, and thus removed. The final dataset contains 24 compounds with their SOM data. The procedures described above were followed to generate initial 3D structures, optimize geometries and mark SOMs.

number of human CYP450 crystal structures have been published, thereby there is a high expectation that this information might significantly improve the performance of current structure-based models for SOM prediction. For example, a recently reported docking procedure can correctly identify up to 80% of SOM of the CYP2D6 substrates (de Graaf et al., 2006). However, as having the evolutionary role in metabolizing diverse xenobiotics, many CYP450 isoforms (e.g. CYP3A4 and 2C9) show non-specific and low-affinity binding toward their substrates. Some researches proposed that preferred binding modes are seldom present (de Graaf et al., 2006).

Soley based on the substrate structural information, Singh et al. (2003) developed a rapid semi-quantitative model for evaluating the 3A4-mediated SOM, in which trend vector models based on AM1 molecular orbital calculations were developed to estimate H-abstraction energy and surface area exposure of the hydrogen atom. Recently, Sheridan et al. (2007, from the same laboratory, reported their further study for addressing the SOM by CYP3A4, 2C9 and 2D6. Purely empirical QSAR-models were constructed with only substructures and physical property descriptors, giving 20-fold cross-validated accuracies ranging from 72% to 77% for the investigated datasets. These results were found to be comparable with those from MetaSite (Cruciani et al., 2005)—a hybrid method combining both enzyme and substrate structure information.

In the present study, we report CYP450-mediated SOM prediction for the six most important metabolic reactions by incorporating the use of machine learning. First, potential SOMs are marked by matching of characteristic functional groups. For each reaction type an individual dataset was prepared, containing entries of SOM data extracted from substrate structures. Then, quantum chemical (QC)-based atomic features were calculated to represent the local reactivity for the potential SOMs, which were used as input to train supporting vector machines to classify whether a site undergoes a specific reaction or not. Finally, an embedded feature selection (FS) procedure was performed with the aid of various weighting schemes, and the most discriminative model can be determined during this step. Both internal and external validations were performed to test the predictive power and generalization ability of the models. The aim of the present study is to describe an accurate, easy to implement and computationally inexpensive method for predicting CYP450-mediated SOM, using only the 3D structure of the potential substrates.

2 METHODS

2.1 Datasets

Datasets used in the present study were collected from the MSD Metabolite Database (MMD, http://www.mdl.com). The dataset generation procedure is as follows. Six classes of CYP-catalyzed reactions were considered, including: (I) aliphatic C-hydroxylation, (II) aromatic C-hydroxylation, (III) N-dealkylation, (IV) O-dealkylation, (V) N-oxidation and (VI) S-oxidation. The reactions reported for mammalian CYPs were initially exported from the MMD unless the yield field of this reaction is labeled as ‘minor’ or ‘trace’, which were deduced to minimize the risks brought by pooling data from different sources and measured by different assays. In the original database, all data were stored in the form of reactions and no SOMs were explicitly reported, so the SOM information needs to be derived. Here, we manually compared each pair of reactant and product structures to determine SOMs, i.e. to locate atoms on the reactant where transformation occurs.
the Fukui reactivity index. Useful means to represent donor–acceptor interactions. Aside from these electrostatic interactions; atomic frontier orbital electron densities provide interactions between molecules and their chemical reactivity (Nakayama, valence-related and energy-related features.

Table 2. Overview of 23 QC feature types used in the study

<table>
<thead>
<tr>
<th>Category</th>
<th>ID</th>
<th>Name</th>
<th>Description</th>
<th>Atoms involved</th>
</tr>
</thead>
<tbody>
<tr>
<td>Charge-related</td>
<td>1</td>
<td>( R )</td>
<td>Fukui reactivity indices</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>( F )</td>
<td>Atomic orbital electron density</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>( N )</td>
<td>Atomic nucleophilicity indices</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>( E )</td>
<td>Atomic electrophilicity indices</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>( Q_s )</td>
<td>Self-charge, the charge kept by the atoms involved in the bond</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>( Q_e )</td>
<td>Active charge, the charge not involved in bonding</td>
<td>1</td>
</tr>
</tbody>
</table>

Valence-related

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Description</th>
<th>Atoms involved</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>( \delta^{-}_{\pi} )</td>
<td>( \pi \rightarrow \sigma ) component of atom-atom interactions from Mulliken population analysis</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>( \delta^{-}_{\sigma} )</td>
<td>( \sigma \rightarrow \pi ) component of atom-atom interactions from Mulliken population analysis</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>( p_{\sigma} )</td>
<td>( \pi \rightarrow \sigma ) component of atom-atom interactions from Mulliken population analysis</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>( p_{\rho} )</td>
<td>( \sigma \rightarrow \rho ) bond order</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>( p_{\delta} )</td>
<td>( \pi \rightarrow \delta ) bond order</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td>( P )</td>
<td>Bond degree</td>
<td>2</td>
</tr>
</tbody>
</table>

Energy-related

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Description</th>
<th>Atoms involved</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>( s_{\epsilon} )</td>
<td>One-center electron-electron repulsion energy</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>( s_{\epsilon^i} )</td>
<td>One-center electron-nuclear attraction energy</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>( s_1 )</td>
<td>Total one-center electronic energy</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>( s_{g} )</td>
<td>Electronic resonance energy</td>
<td>2</td>
</tr>
<tr>
<td>18</td>
<td>( s_{\epsilon^i} )</td>
<td>Electronic exchange energy</td>
<td>2</td>
</tr>
<tr>
<td>19</td>
<td>( s_{\epsilon^j} )</td>
<td>Electronic repulsion energy</td>
<td>2</td>
</tr>
<tr>
<td>20</td>
<td>( s_{\epsilon^i} )</td>
<td>Nuclear-electron attraction energy</td>
<td>2</td>
</tr>
<tr>
<td>21</td>
<td>( s_{\epsilon^j} )</td>
<td>Nuclear–nuclear repulsion energy</td>
<td>2</td>
</tr>
<tr>
<td>22</td>
<td>( s_{\epsilon^j} )</td>
<td>Coulomb interaction energy</td>
<td>2</td>
</tr>
<tr>
<td>23</td>
<td>( s_{\epsilon} )</td>
<td>Total of electronic and nuclear energy</td>
<td>2</td>
</tr>
</tbody>
</table>

FC indices (Fukui, 1988) provide us an indirect measure of chemical reactivity. Quantum mechanical energy-related features—these features characterize the intramolecular energy distribution using different partitioning schemes (Sklenar and Jäger, 1979; Strouf, 1986). The total energy, in eV, obtained by the addition of the electronic and nuclear terms, can be partitioned into mono- and bi-centric contributions, and these contributions in turn are divided into nuclear and one- and two-electron terms.

As mentioned above, for each of the substrates an energy minimization was performed. The optimized structures were then employed for the QC properties calculation with the AMI Hamiltonian of MOPAC. Keywords include: MMOK, VEKTORS, BONDS, PI, PRECISE, ENPART, EF, MULLIK and CHARGE=n, where n was determined by counting the total formal charges of the molecule studied. Generally, most features listed in Table 2 can be directly extracted from MOPAC output, except Fukui reactivity index (\( R \)), atomic nucleophilicity (\( N \)) and electrophilicity index (Katritzky et al., 1995–1997) (\( E \)). For a given atom A, values of these properties were calculated by the following functions:

\[
R_A = \sum_{i \in A} C_{\text{HOMO}} \cdot C_{\text{LUMO}} - \sum_{i \in A} C_{\text{LUMO}} \cdot C_{\text{HOMO}}
\]

\[
N_A = \sum_{i \in A} C_{\text{HOMO}} \cdot C_{\text{LUMO}} + \sum_{i \in A} C_{\text{LUMO}} \cdot C_{\text{HOMO}}
\]

\[
E_A = \sum_{j \in A} C_{\text{LUMO}} \cdot C_{\text{HOMO}} + D
\]

where the summations are performed over all atomic orbitals (AO), \( i \) at atom A. \( C_{\text{HOMO}} \) and \( C_{\text{LUMO}} \) denote the i-th and j-th AO coefficients on the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO), respectively. \( C_{\text{LUMO}} \) and \( C_{\text{HOMO}} \) are the energies of these orbitals, respectively. The reactivity indices estimate the relative reactivity of the atoms in the molecule for a given series of compounds and are related to the activation energy of the corresponding chemical reaction.

In this study, one reaction center pattern matches only one atom as a potential SOM, while the valence-related features and bi-centric energy contributions involve two atoms, which cannot be directly used as atomic descriptors. Hence, some straightforward transformations were made: suppose atom A of molecule M is under consideration, bi-centric terms of all A-related bonds in the M were collected as a set of L, and all A-related atom pairs in M were collected as another set of G. Then, sum, max, min, mean and norm values of L and G were calculated, respectively, and these parameters were used as descriptors of atom A. For mono-centric term, the same procedures were followed solely on L, which plus itself gives six descriptors. Therefore, altogether 194 (14 × 10 + 9 × 6) descriptors were calculated for each potential SOM studied. An in-house python program was used for parsing the MOPAC output and calculating these features.

2.2 Quantum chemical features

QC methods enable the definition of a large number of global and local quantities characterizing the reactivity and binding properties for molecules, which have been successfully used to predict ‘site of reactivity’ for CYP450 substrates (Chunben et al., 1994) and chemical metabolism by UDP-glucuronyltransferase (Sorich et al., 2004). To model the SOM data, we have considered 23 types of semi-empirical QC features (Table 2), which can be further divided into the following three categories: charge-related, valence-related and energy-related features.

Charge distribution-related features—these features describe the polar interactions between molecules and their chemical reactivity (Nakayama, 1993; Cizmadia, 1976; Fukui, 1970; Fukui et al., 1957; Mulliken, 1955). Among them, atomic partial charges constitute the major driving force of electrostatic interactions; atomic frontier orbital electron densities provide useful means to represent donor–acceptor interactions. Aside from these two interrelated concepts, another significant aspect for understanding the reactivity of a molecule is the chemical softness quantitatively measured by the Fukui reactivity index.

Valence-related features—these features relate to the strength of intramolecular bonding interactions and characterize the stability of the molecules, their conformational flexibility and other valence properties (Katritzky et al., 1995–1997; Sannigrahi, 1992). This type of property provides us an indirect measure of chemical reactivity.

2.3 Feature selection

FS is a prerequisite because the degree of redundancy and degeneracy of our defined indices may be high. In the current work, a backward elimination-based embedded model (Saray et al., 2007) was chosen due to its computational efficiency and effectiveness. The general procedures are briefly described as follows:

(1) Highly correlated features were first removed, with the filtering threshold in terms of correlation coefficient set to 0.95.

(2) The goodness of features was individually evaluated by six weighing schemes implemented in RapidMiner (Mierswa et al., 2006), including InfoGain (IG) (Quinlan, 1993), InfoGainRatio (IGR),...
The main idea of SVMs is to maximize the margin, which is defined as

\[
\frac{1}{2||w||^2} + C \sum_{i=1}^{N} \xi_i + C \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \sum_{j=1}^{N} x_i j
\]

(4)

subject to \(y_i (w \cdot x_i + b) \geq 1 - \xi_i \) and \( \xi_i \geq 0 \), \( i = 1, ..., N \), where \( w \) is a weight vector of the separating hyperplane, \( \xi \) is a slack variable that allows the margin constraints to be violated and \( C \) is a user parameter to be tuned. The first term of the objective function is about the margin maximization, and the second and third terms, respectively, control the penalties of positive and negative training samples. Geometrically, when positive data are in the minority, it gives more weight to the positive support vectors \( C \) is larger than \( C \), and pushes the hyperplane towards where the negative (majority) data exist.

In many supervised learning tasks, it is always more than necessary to convert the outputs of the classifier into well-calibrated posterior probabilities, particularly when the classification decision is cost-sensitive. To this end, Platt (1999) proposed a SVM + sigmoid method. For the estimated class membership \( f(x) \) of the observation \( x \), a sigmoid function is fitted to all estimated \( g(x) \) to derive probabilities of the form:

\[
P(Y=1|x) = \frac{1}{1+e^{-f(x)}}
\]

(5)

where \( A \) and \( B \) are estimated by minimizing the negative log-likelihood of the training data:

\[
\max_{A,B} \sum_{i=1}^{N} \left( y_i + \frac{1}{B} \log(p_i) + (1 - y_i) + \frac{1}{A} \log(1 - p_i) \right)
\]

(6)

Labels and decision values [estimated \( g(x) \)] are required to be independent, so a 5-CV was conducted to obtain decision values. In this study, the probability value of class membership might be interpreted as the metabolic priority of atoms having a similar environment, or the relative susceptibility of potential SOMs toward the investigated metabolic transformation.

The LibSVM package (Chang and Lin, 2001) (version 2.84) that implements both the cost-sensitive learning and probability estimation method was used in this study. Instead of directly using different penalty values, it provides a weighting parameter \( W \), and the penalty of class \( i \) is calculated as \( C_i = W_i \times C \). For our experiments we fix the weight \( W \) at 1, and determine the best ratio by varying \( W \). Therefore, the penalty parameter \( C \), the weighting parameter \( W \), and the radial basis function (RBF) kernel parameter \( y \) were tuned during the learning process to obtain SVM classifiers with optimal performance. The grid search strategy of LibSVM, originally determining the optimal \( C \) and \( y \), was tailored to perform a 3D (\( C \times W \times y \)) parameter scanning based on the performance measurements as will be described below.

### 2.5 Performance measurement

As in the case of all discriminative methods (Baldi et al., 2000; Roulston, 2002), the performance of statistical learning methods can be measured by the quantity of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). Accuracy \( [\text{ACC} = \frac{TP+TN}{TP+TN+FP+FN}] \) is a frequently used index for the overall classification performance, but it may be misleading as a result of highly unbalanced class distribution of used datasets. Sensitivity \( [\text{SE} = \frac{TP}{TP+FN}] \) and specificity \( [\text{SP} = \frac{TN}{TN+FP}] \) can assess a model’s ability to correctly identify TP and TN, respectively, while they are usually interpreted in combination with each other. Therefore, the mean value of SE and SP, or so-called balanced accuracy \( [\text{BACC} = \frac{\text{SE} + \text{SP}}{2}] \), is used as a major performance measurement during model training. For model validation and comparison, all these four indices were also calculated.

In addition to BACC, another method that evaluates a model in terms of the tradeoff between SE and SP is receiver operating characteristic (ROC) analysis (Fawcett, 2004), which is a performance graphing method becoming increasingly important for cost-sensitive learning. The ROC curve typically plots FP rate (1 - SP) versus TP rate (SE) while a decision threshold is being varied. The AUC of ROC plot provides a convenient way of comparing classifiers, where a random classifier has an area of 0.5, and an ideal classifier has an area of 1.0.

### 3 RESULTS

#### 3.1 Feature and model selection

The task of FS may not simply be to identify relevant characteristics to the target property, but also to provide a prediction system; this system can also be used to validate the quality of the selected features. In this way, one can get a better understanding about the interaction between FS and the classification method.

From Figure 2(a), one can clearly find that the initial QC property matrix contains a large number of highly correlated features that need to be removed at first. Figure 2(b) shows such a FS process for the case of aliphatic hydroxylation, the best performance was achieved by IGR when the feature number was reduced to 27, optimized parameter \( C \), \( y \) and \( W \) are 4.0, 8.0 and 5.0 (Table 3), respectively. For other reaction types, a common pattern was observed for those better performed weighting schemes that the decrease of prediction accuracy does not immediately follow the feature reduction. Instead, it undergoes a rise until the feature number decreased to some point, typically reaches peaks around 20, and then turns to descending. These results clearly indicate that there are indeed a substantial amount of redundant and unrelated
Table 3. Feature and model selection results

<table>
<thead>
<tr>
<th>Type</th>
<th>Model parameter</th>
<th>Feature related</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>γ</td>
<td>W</td>
</tr>
<tr>
<td>I</td>
<td>4.0</td>
<td>8.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>II</td>
<td>64.0</td>
<td>2.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>III</td>
<td>1024.0</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IV</td>
<td>64.0</td>
<td>2.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td>256.0</td>
<td>0.031</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VI</td>
<td>16.0</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1ID corresponds to the feature type ID in Table 2, a feature type is counted if any of its arithmetically derived forms is present in the selected feature subset.

The two valence-related features δ<sub>π→π</sub>-electrons of an investigated atom and π-electrons of its adjacent atom, which may be used to investigate whether such an electron flow along the direction induced by an external perturbation is easy. These two features were observed to be present in all cases, indicating the importance of valence-related features for developing CYP-mediated SOM models.

3.2 Performance measurements

As described in Section 2, SVM classifiers were developed to predict SOMs for CYP450-mediated biotransformation. In this study, a potential SOM is a substructure matching one of the reaction center patterns defined in Table 1, which is classified as reactive or non-reactive in the corresponding type of reaction based on the QC features calculated for the atoms in the putative reaction center. These features provide a characterization of the local reactivity and binding properties of the potential SOM. The prediction performance of each model was assessed by various indices (Table 4). For model construction, our method shows its superior efficiency for data fitting, with the average BACC obtained for training sets ranging from 0.912 to 0.978. The average ACC varies between 0.885 and 0.966; SE and SP are generally higher than 0.94 and 0.85, respectively. For validation, ACC of each reaction types exceeds 0.80, indicating a good predicting power of our method.

Among derived models, SOMs involving N- (III) and O-dealkylation (IV) are easier to predict because of their relatively simple mechanism mainly governed by the cleavage energy of the C–H bond in α-position to the N- or O-atom (Gasteiger, 2007). Accordingly, the best performances were observed for these two reaction types, of which all their statistical indices are over 0.9. In contrast, SOMs mediated by aromatic hydroxylation cannot be readily predicted. This type of reaction generally proceeds via an epoxide intermediate, but direct hydroxylation

![Fig. 2. (a) A Contour graph to show the inter-correlation matrix of QC features used for aliphatic hydroxylation, and (b) a plot to compare different feature weighting schemes for the SOM model of this biotransformation, where the average BACC derived from 5-CV is plotted as a function of the number of selected features.](image-url)
via an electrophilic aromatic substitution mechanism also occurs. Moreover, the formation of the epoxide intermediate sometimes undergoes the NIH shift (Guroff et al., 1967), which might result in multiple SOMs with relatively low selectivity and a variety of oxidized products. The difficulty to define such metabolism sites with high SE was also pointed out in Korolev’s work (Korolev et al., 2003). In our method, SE and SP for this transformation type are uniformly high, suggesting a well-balanced model that is capable of capturing the subtle differences between reactive and non-reactive hydroxylation site. Models for other reaction types are also the case with the exception of S-oxidation, where SE and SP vary considerably. However, the worst SP value observed in the S-oxidation model was partially due to the rare negative samples presented in the test set. In fact, only one non-reactive site in the test set was misclassified, meanwhile all the reactive sites were correctly identified.

Unlike SE or SP that reflect model performance at a single parameter (Hanley and McNeil, 1982), ROC curve analysis offers a global and unbiased measure of the effectiveness of classification algorithms. Figure 3 demonstrates ROC curves for six SOM prediction models on their respective test datasets, in which a completely random guess would give a point along the diagonal line (broken line in red), while the best possible prediction method would yield a point in the upper left corner of the ROC space. As can be seen, all the ROC curves are situated very high above the line of no-discrimination, indicating the success of these SOM models. The statistical results obtained from these ROC analyses are also given in Table 4. We may note that every SOM model has a significant AUC value, higher than 0.86, five of which are over 0.92. Here special attention should be put on the S-oxidation, the relatively large standard error of 0.129 suggests the model may be not as reliable as the statistic (AUC = 0.929) states. This model needs to be further tested with other qualified test sets in the future, especially one with sufficient negative samples.

4 DISCUSSION

To further validate the generalization ability of these SOM prediction models, the test set collected by Sheridan et al. was used. Aside from the non-congeneric structures and different literature sources, this external test is particularly challenging for at least the following two reasons: (i) the dataset contains metabolism information of CYP2D6, 2C9 and 3A4, while our models will predict all SOMs mediated by 11 CYP isotypes (data not shown). Thus, there might be some successfully recognized sites that were not considered in the original dataset, which may increase FP rates of our models; (ii) our models focus on six major reactions but Sheridan’s dataset compiled all possible SOMs, giving no special account of reaction types. The consequence of this disagreement, in contrast, may increase the FN rates of our models. Nevertheless, this external test is worth trying because CYP2D6, 2C9 and 3A4 are responsible for nearly 70% of all phase I metabolism (Lewis et al., 2002), and we assume that the six reactions modeled here constitute a major portion of metabolic pathways—such a test may just present us a practical idea of how these SOM models will behave in a real-world problem. Generally, combining the six models, 295 out of 373 potential SOMs were correctly classified with an overall accuracy of 0.79. In particular, among altogether 45 positive SOMs in the dataset 36 were identified, giving a fairly impressive SE value of 0.80. It should be mentioned that the results obtained by the Sheridan et al. model were reported in ‘molecule-scale’ (where the top two ranked atoms of a molecule containing a SOM was thought to be a correct prediction). We should not directly compare their results with ours. However, given the claim that neither their model nor MetaSite can explain more than about 70% of the regioslectivity data (Sheridan et al., 2007), we may speculate that the performance of our SOM prediction models is so far highly satisfactory.

Figure 4 shows some poorly and well-predicted example molecules in the external test set. For the cases of DRF4367 and Endosulfan, there was a full qualitative agreement between the in vivo experimental results and prediction results, either from our models or Sheridan’s. No FP predictions were made for the structure of DRF4367 where a large number of potential SOMs were marked. For Ellipticine, our models can successfully identify all the five SOMs and only introduce one FP prediction, while Sheridan’s model
retrieves four of them with the same FP rate. For other cases the results are not that optimistic: most misclassified SOMs are mediated by either aromatic-hydroxylation or S-oxidation, suggesting where further improvements not need be made to these classes. A common FN prediction can be found in the case of Phthalocaine, where an experimentally observed SOM by hydrolysis was missed on account of an undefined biotransformation. This example points to an issue discussed above—as long as a SOM in the test compound does not fall into any of the six reactions addressed, even the most sophisticated classifier must fail. This problem needs to be solved with the increase of high-quality data and the number of available training samples.

5 CONCLUSIONS
In this work, we developed SOM prediction models for six CYP450-mediated reactions of chemical compounds. A set of local QC properties were calculated with semi-empirical methods to represent the reactivity profile of a potential SOM; a cost-sensitive learning approach was selected to cope with the unbalanced nature of datasets; and an embedded FS procedure that incorporates a series of feature weighting schemes was followed to optimize the models. Both internal and external validation results suggest that the developed models are reasonably successful. Such a multistep approach, with derived models, would be highly desirable for the developed models are reasonably successful. Such a multistep approach, with derived models, would be highly desirable for the early assessment of human metabolic transformations of lead compounds and drug candidates at the preclinical stage of drug discovery.

ACKNOWLEDGEMENTS
The authors would like to thank Prof. Johann Gasteiger and Dr Niklas Blomberg from AstraZeneca R&D Mölndal for their insightful suggestions.

Funding: State Key Program of Basic Research of China (Grant 2009CB185502); Hi-TECH Research and Development Program of China (Grants 2006AA020402, 2006AA020236); the National Natural Science Foundation of China (Grants 20721033, 30672539); the Basic Research Project for Talent Research Group from the Shanghai Science and Technology Commission; the Key Project from the Shanghai Science and Technology Commission; the Key Project for New Drug Research from CAS.

Conflict of Interest: none declared.

REFERENCES