



Original article

Classification models for CYP450 3A4 inhibitors and non-inhibitors

Inhee Choi^a, Sun Young Kim^b, Hanjo Kim^b, Nam Sook Kang^c, Myung Ae Bae^c, Seung-Eun Yoo^c,
Jihoon Jung^d, Kyoung Tai No^{a,b,d,*}

^a Institute of Life Science and Biotechnology, Yonsei University, Seoul 120-749, Republic of Korea

^b Bioinformatics and Molecular Design Research Center, Seoul 120-749, Republic of Korea

^c Korea Research Institute of Chemical Technology, Yuseong-Gu, Daejeon 305-600, Republic of Korea

^d Department of Biotechnology, College of Engineering, Yonsei University, Seoul 120-749, Republic of Korea

ARTICLE INFO

Article history:

Received 4 February 2008

Received in revised form 21 August 2008

Accepted 26 August 2008

Available online 18 September 2008

Keywords:

Cytochrome P450 3A4

Inhibitors

Classification

2D molecular descriptors

Recursive partitioning

Random forest

ABSTRACT

Cytochrome P450 3A4 (CYP3A4) is the predominant enzyme involved in the oxidative metabolic pathways of many drugs. The inhibition of this enzyme in many cases leads to an undesired accumulation of the administered therapeutic agent. The purpose of this study is to develop *in silico* model that can effectively distinguish human CYP3A4 inhibitors from non-inhibitors. Structural diversity of the drug-like compounds CYP3A4 inhibitors and non-inhibitors was obtained from Fujitsu Database and Korea Research Institute of Chemical Technology (KRICT) as training and test sets, respectively. Recursive Partitioning (RP) method was introduced for the classification of inhibitor and non-inhibitor of CYP3A4 because it is an easy and quick classification method to implement. The 2D molecular descriptors were used to classify the compounds into respective inhibitors and non-inhibitors by calculation of the physicochemical properties of CYP3A4 inhibitors such as molecular weights and fractions of 2D VSA chargeable groups. The RP tree model reached 72.33% of accuracy and exceeded this percentage for the sensitivity (75.82%) parameter. This model is further validated by the test set where both accuracy and sensitivity were 72.58% and 82.64%, respectively. The accuracy of the random forest model was increased to 73.8%. The 2D descriptors sufficiently represented the molecular features of CYP3A4 inhibitors. Our model can be used for the prediction of either CYP3A4 inhibitors or non-inhibitors in the early stages of the drug discovery process.

© 2008 Elsevier Masson SAS. All rights reserved.

1. Introduction

Cytochrome P450s (CYPs) play a crucial role in metabolism [1]. P450-mediated metabolism is an oxidation reaction that is a part of Phase I metabolic cycle. P450s are the strongest oxidizing agents known in living systems, consequently many drugs can be oxidized by more than one P450 enzyme [2]. These enzymes metabolize a vast array of structurally diverse drugs in market – approximately 90% of all marketed drugs – and are responsible for major routes of drug clearance [2,3]. Members of the CYP family are particularly prone to inhibition because of their broad substrate specificity, which is amenable to competitive inhibition arising from different types of structurally diverse drugs that can be metabolized by the same enzymes [4]. Thus, it has become a widely accepted practice that potent inhibition of these enzymes should be avoided where possible [2]. In drug therapy, CYP inhibition may result in undesirable consequences: (1) an increase of toxicity caused by

decreased drug metabolism rate, (2) a decrease in pharmacological effects due to the decreased formation of reactive metabolites of pro-drugs, and (3) drug–drug interactions (DDI) by double medication [4] which lead to a decreased clearance of one of the drugs when two or more drugs are administered simultaneously. It is particularly important to consider the biotransformation and elimination of drugs during their development to determine their potential interactions with CYPs as early as possible [1].

CYP3A4, one of the key P450s, has been identified and modeled extensively. CYP3A4 accounts for 50–60% of drug metabolisms [2]. Now, it is a common practice in drug discovery to screen out chemicals that possess inhibitory activity against CYP3A4-mediated metabolism in order to predict and manage DDI [5]. In this sense, the high quality *in silico* models are very useful in ADME/DDI predictions.

The machine-learning algorithm used primarily in this work is recursive partitioning (RP) which is generally known to be fast compared to other methods and provides easily interpretable results. It is a simple, yet powerful, statistical method that seeks to uncover relationships in large complex data sets. It is a method for determining statistically meaningful rules that classify objects into similar categories [6]. It is known to be sensitive to the descriptors

* Corresponding author. Bioinformatics and Molecular Design Research Center, Seoul 120-749, Republic of Korea. Tel.: +82 2 393 9550; fax: +82 2 393 9554.

E-mail address: ktno@bmdrc.org (K.T. No).

Table 1
Number of CYP450 3A4 inhibitors and non-inhibitors in training and test sets

Data set	Inhibitors	Non-inhibitors	Total
Training	273	468	741
Test	121	65	186

used, and to the composition of the data sets, that can radically change the decision tree. In fact, RP methods have been used extensively with large sets of molecules and either continuous or binary data, for therapeutic target end points as well as CYP inhibition and toxicity properties such as AMES mutagenicity status [7]. In addition, we have used an ensemble method, random forest (RF). The purpose of this study was to present models which predict CYP3A4 inhibitors and non-inhibitors with molecular descriptors that well represent and easily explain their physicochemical properties with the aim of early identification during the drug development process.

2. Results and discussion

In order to build efficient models, access to a sufficient number of structure data as well as the diversity of the data sets is important. The quantity of data was sufficient for both training and test sets (~100 compounds or more) [1] in this study. The number of inhibitors and non-inhibitors in the respective sets is reported in Table 1. The inhibitors and non-inhibitors of both training set (Fig. 1A) and test set (B) covered a well-defined region in the diversity space described by the three principal components (PCs). A clearer view of the diversity between training and test sets was elucidated from an additional principal component analysis (PCA) using “predefined set” of molecular descriptors in SciTegic Pipeline Pilot (Fig. 2) [8]. Here, PCA gives three significant PCs, which explains 83.3% of the variation in the data (44.5%, 25.3%, and 13.4%, respectively). The values of each molecular descriptor to respective PCs are shown in Table 2.

The decision tree was trained in order to classify the compounds into two active classes, inhibitors and non-inhibitors. The classes of inhibitors and non-inhibitors of CYP3A4 training set could be differentiated with the correct recognition rate of 72.33%. To confirm the performance of our model, the external test set of 186 compounds was classified by use of the generated model. The accuracy was 72.58% similar to that of training set. Thus, the similar prediction rate with that of the training set validated the robustness of our decision tree. There were ten compounds in the test set with ambiguity in biological data where their percentages for CYP3A4

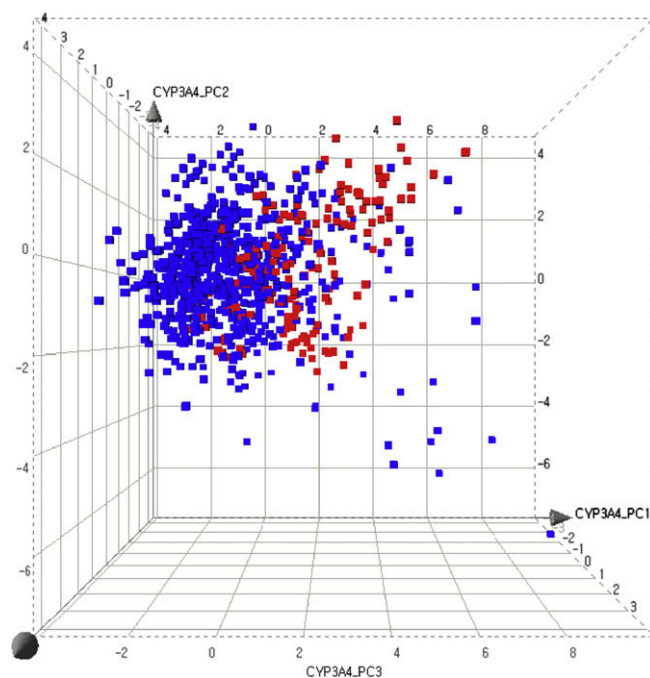


Fig. 2. Principal components analysis of CYP3A4 inhibitors and non-inhibitors in the training (blue dots) and test (red dots) sets (Spotfire Decision Site 9.1.1) [32]. The comparison is based on the first three principal components calculated from the “predefined set” descriptors of SciTegic Pipeline Pilot. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

inhibition were above 40% and less than 60%. When these compounds were not included in the test set, the predictability increased slightly (75.2%). The sensitivity (recall rate) of the training set was 75.82%. In case of the CYP3A4 test set, 100 out of 121 inhibitors were predicted to the correct class (82.64%). Overall, the RP model showed that the inhibitors of both training and test sets were predicted with higher certainty than the non-inhibitors (Fig. 3).

Table 3 summarizes the performance parameters – accuracy, sensitivity, specificity, kappa, and MCC – involving the training and test sets. Matthews Correlation Coefficient (MCC) values were about 0.38 for both sets where they were above 0, indicating improved prediction compared to random prediction [9,10]. Kappa values for the training and test sets were 0.44 and 0.38, respectively. According to the guideline for interpreting kappa values,

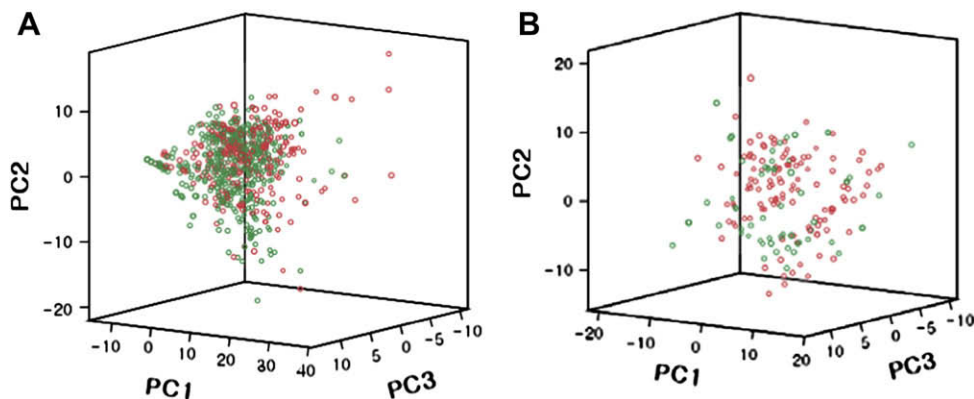


Fig. 1. Distribution of CYP3A4 inhibitors (red dots) and non-inhibitors (green dots) in the training (A) and test (B) sets. The comparison is based on the first three principal components calculated from 240 PreADMET 2D descriptors. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2

PCA loadings obtained for selected descriptors with three principal components for training and test sets

Descriptors	Training set			Test set		
	PC1	PC2	PC3	PC1	PC2	PC3
Constant	−5.742523	−2.651090	−0.485238	−9.381893	−4.479217	0.221509
ALogP	−0.010377	0.318532	0.260999	0.132020	−0.180071	−0.000856
Molecular_Weight	0.005385	0.001106	0.001437	0.004917	0.003871	−0.000856
Num_H_Donors	0.217122	−0.234540	−0.169823	−0.082838	0.696758	−0.760364
Num_H_Acceptors	0.183318	−0.136504	−0.052674	0.192140	0.355695	0.536200
Num_RotatableBonds	0.090299	−0.039291	0.204295	0.153406	0.016335	0.026038
Num_Atoms	0.077984	0.022011	0.007230	0.070237	0.042052	−0.016883
Num_Rings	0.208034	0.305987	−0.357279	0.373853	−0.164702	−0.087517
Num_Aromatic_Rings	0.080250	0.472039	−0.177699	0.396609	−0.209780	−0.233588

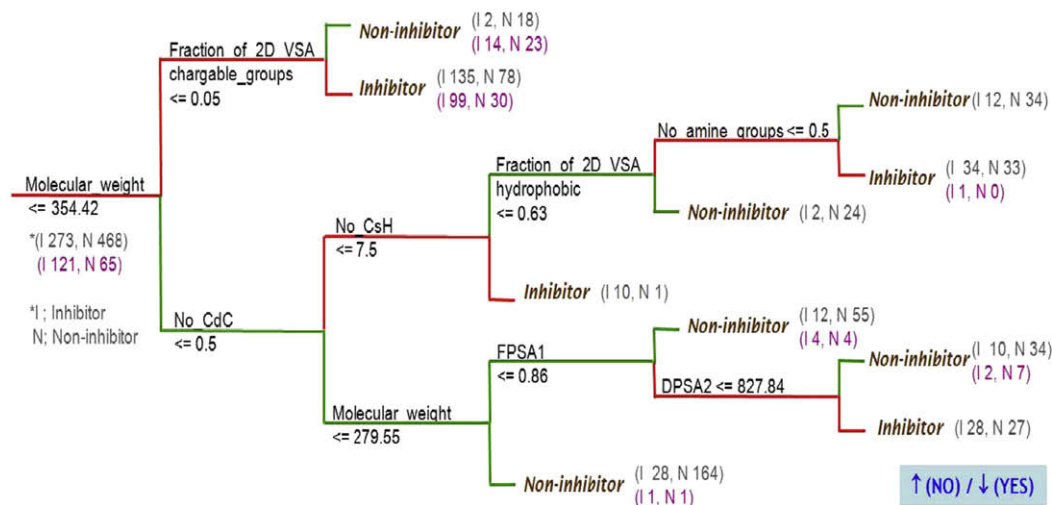


Fig. 3. Decision tree and 2D descriptors built with 741 compounds of CYP3A4 inhibitors and non-inhibitors in the training set. Two classes were used: non-inhibitors (N; green) and inhibitors (I; red). 2D descriptors from PreADMET were used. The classified number of inhibitors and non-inhibitors for training set (black) and test set (purple) is shown in parentheses. The distribution of molecules of the training set (black) and the test set (purple) in each leaf is positioned between brackets (number of correctly classified compounds/ number of misclassified compounds). The arrows show the direction of the branches. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

kappa of 0.41–0.60 is considered to be of moderate agreement [11], and thus, we concluded that our RP model is a predictive one.

Meaningful 2D descriptors affect the differentiation of inhibitors and non-inhibitors of CYP3A4. We have used the following well-chosen descriptors such as: (1) constitutional descriptors like

formal charges, fraction of rotatable bonds, number of rigid bonds, number of rings, number of charged groups, etc., (2) electrostatic descriptors that include charge polarization, relative charge, etc., (3) physicochemical descriptors like AlogP98 value, and (4) geometrical descriptors containing information from the 3D

Table 3

Performance parameters – accuracy, sensitivity, specificity, kappa, Matthews correlation coefficient – for RP model corresponding to CYP3A4 training and test sets

Data set	Accuracy, %	Sensitivity, %	Specificity, %	Kappa ^a	Matthews correlation coefficient ^b (MCC)
Training	72.33 (536/741)	75.82 (207/273)	70.3 (329/468)	0.44	0.383
Test	72.58 (135/186)	82.64 (100/121)	53.85 (35/65)	0.38	0.379

^a kappa = accuracy – $E/1 - E$, where E = expected agreement = $(TN + FN)(TN + FP)(FP + TP)(FN + TP)/(TP + FP + FN + TN)^2$.

^b MCC = $(TP \times TN) - (FP \times FN) / \sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}$.

Table 4

Summary definition of all descriptors that were found to be important by the CYP3A4 RP decision tree with the best predictive ability

Descriptor group	Name	Representation
Constitutional	Molecular_weight	Molecular weight
	No_amine_groups	The number of amine groups
	No_CdC	The number of double bonds between C atoms and C atoms
	No_CsH	The number of single bonds between C atoms and C atoms
Electrostatic	DPSA2	The difference between total charge weighted partial positive and negative surface areas
	FPSA1	The partial positive VDW surface area divided by the total VDW surface area
Geometric	Fraction of 2D VSA chargeable groups	Fraction of 2D van der Waals chargeable groups of surface area
	Fraction of 2D_VSA hydrophobic	Fraction of 2D van der Waals hydrophobic of surface area

structure of a compound, such as the polar surface area [12]. Table 4 lists the description of 2D descriptors where the combination of them was necessary to develop the model. The 2D descriptors of our RP tree model provide insight of molecular properties which are important for CYP3A4 inhibitors and can aid in the prediction of CYP3A4 inhibition activity of unknown compounds.

Among various descriptors, molecular weight is the first decisive factor for distinguishing inhibitors from non-inhibitors. In general, CYP3A4 is known to be able to accommodate bulkier compounds and metabolize very large molecules than most other P450 enzymes [2,13]. It explains the preference for inhibitors with higher molecular weight (above 354.42) in our model. It is consistent with other CYP3A4 inhibition models which identified the size of the compounds as one of the important features promoting CYP3A4 inhibition [14].

The next decisive descriptor to classify inhibitors is the geometrical descriptor, Fraction of 2D VSA chargeable groups,

which considers the delocalization of all charged atoms. Most of CYP3A4 inhibitors are hydrophobic and a hydrophobic domain in CYP3A4 is important for enzyme activation [15]. The binding site of CYP3A4 is predominantly consisted of hydrophobic residues (such as Phe), as shown by recently published X-ray structures of CYP3A4 [16,17]. It is the reason that compounds having less than 5% chargeable groups were classified as inhibitors. It is known that key molecular properties which favor CYP3A4 inhibition are the size of a molecule and the hydrophobic nature of a compound [14]. These two descriptors are meaningful in classifying CYP3A4 inhibitors for the sensitivity was higher than accuracy. To further illustrate the relation between key molecular properties and the CYP3A4 inhibition, the values of these selected descriptors for drugs that are known as strong CYP3A4 inhibitors [18] are listed in Table 5.

Other descriptors which favored CYP3A4 inhibition were No_amine groups (the number of amine groups), No_CsH (the number of single bonds between C atoms and C atoms) and DPSA2

Table 5

Strong CYP3A4 inhibitors from the training set and the corresponding values of descriptors of the model

Molecule	Structure	Fraction_of_2D_VSA_chargeable_groups	Molecular_weight
Doxycycline		0	444.4402
Itraconazole		0	705.6424
Ketoconazole		0	531.4376
Methylprednisolone		0	374.476
Miconazole		0	416.1334

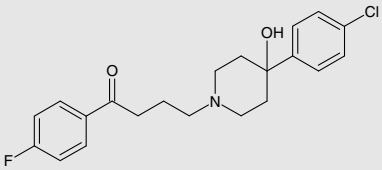
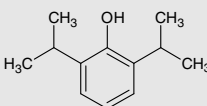
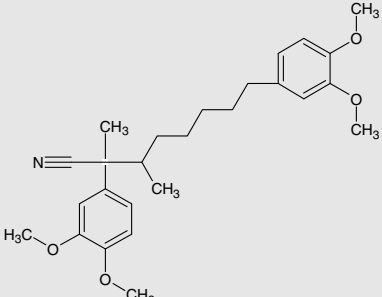
(difference between total charge weighted partial positive and negative surface areas). In our model, compounds that were classified as CYP3A4 non-inhibitors generally had more than one amine group. Among those compounds classified as non-inhibitors, there might be compounds binding to CYP2D6 active site. It is known that all drugs metabolized by CYP2D6 contain a basic nitrogen atom; CYP2D6 inhibitors and substrates generally fit a pharmacophore that contains a basic amine [2,19,20]. In fact, there were CYP2D6 substrates classified as non-inhibitors such as zuclopenthixol and haloperidol which have a piperazine moiety and a piperidine moiety, respectively. On the other hand, one of the strong CYP3A4 inhibitors, bergamottin, which has no nitrogen atom in the structure, was correctly classified as an inhibitor.

The number of single bonds descriptor was also used to classify inhibitors. In the study done by Zuegge et al., the first principal component to cluster compounds with low CYP3A4 IC_{50} values showed some correlation with the total number of single bonds [21]. DPSA2, one of the partial surface area descriptors introduced by Stanton and Jurs [22], is defined as the difference between total charge weighted partial positive and negative surface areas in the molecule. This descriptor represents the charge distribution on the surface. Thus, compounds with lesser DPSA2 value could mean that their surface charge distribution is small. The fact that compounds with lesser DPSA2 value were accounted as inhibitors reveals that the hydrophobic character of compounds is the major contributing factor to the CYP3A4 inhibitory activity. The majority of non-inhibitors of CYP3A4 were classified by other descriptors starting from No_CdC (the number of double bonds between C atoms and C atoms) as the second descriptor following molecular weight descriptor.

Due to the differences in chemical compound collections such as commercial or proprietary databases, and experimental methodologies between research institutions, all published models up to date are usually not directly transferable [23]. Therefore, the current model generated in this work would be only as good as the data used like other models. Although the quality of a model depends on the used data sets, generally a model should predict well for clinically important drugs. Nine out of 12 common drugs in the data set of Arimoto's et al. (Support Vector Machine (SVM) model with Barnard chemical information (BCI) fingerprints) [23] are included in our training set. Three other drugs – haloperidol, propofol, and verapamil – have been predicted with our RP model as an additional external validation compounds. Based on the recent review of clinically important drugs as CYP3A4 inhibitor and/or substrate [18], the RP model predicted these three compounds well (Table 6). Haloperidol is correctly predicted with our RP model as a CYP3A4 inhibitor according to the review. However, according to the IC_{50} of 7-benzyloxy-4-trifluoromethylcoumarin (BFC) metabolism result which was used to collate with Arimoto's SVM prediction result [23], haloperidol is a non-inhibitor. The discrepancy among inhibition assay results as such is inevitable since CYP3A4 the human P450 metabolism related assays are substrate dependent. Propofol and verapamil were predicted as non-inhibitor and inhibitor, respectively, just like the SVM model.

There are some limitations to the model that could have affected the quality of the model although classification results are significantly accurate, particularly in the case of inhibitors. First, the training set used in our model was classified based on results from literature. Therefore, the data set lacks consistency in the assay methods. Thus, the future predictive power of our model will be

Table 6
Predictions of common drugs by RP model

Molecule	Structure	Prediction RP model	CYP3A4 ^a		Prediction BCI/SVM model ^b	IC_{50} (μ M) ^c
			Inhibitor	Substrate		
Haloperidol		I	▲	▲	NI	52
Propofol		NI	▲	△	NI	51
Verapamil		I	▲▲	▲	I	0.36

I = Inhibitor; NI = Non-inhibitor.

^a Classification data of clinically important drugs as CYP3A4 inhibitor and/or substrate [18]. ▲, major role or significant inhibitory and/or inducing effect on CYP3A4; △, minor role or weak inhibitory and/or inducing effect on CYP3A4; ▲▲, mechanism-based inhibitor of CYP3A4.

^b Prediction results of BCI-fingerprint/SVM model from Arimoto et al. [23].

^c IC_{50} values of CYP3A4-mediated BFC metabolism [23].

enhanced if homogenous assay methods become available for CYP3A4 inhibitors and non-inhibitors in the training set.

Second, the limitations in the predictive power of our generated model appear to be related to the training data set. The compound ratio between two classes in the training set was not adequately close; this set was inadvertently heavily skewed toward the great number of non-inhibitors. Therefore, the unbalanced number of compounds in the training set might have affected the accuracy value of the model. Future improvement of the model with the focus on data quality is being considered. Despite these two limitations, high sensitivity value of the model indicated that the selected 2D descriptors in the model well portrayed chemical aspects of CYP3A4 inhibition.

Next, we have built an additional model using the Random Forest (RF) algorithm [24]. Recent studies regarding the random forest have shown that prediction results can be improved by growing a set of decision trees and letting them to vote. Because of its ability to increase predictive accuracy over the individual model established, ensemble methods are investigated for their ability to provide insight into the confidence associated with each prediction [25]. We observed an increase of the prediction quality from RF model. When compared with RP model result, the accuracy of the RF model was 73.8%. The out-of-bag (OOB) estimate of error rate was 26.18% and the receiving operating characteristics (ROC) score for OOB data was 0.749.

3. Conclusions

In conclusion, we have developed the RP and RF models to predict CYP3A4 inhibitors and non-inhibitors. CYP3A4 is known to metabolize structurally diverse, large lipophilic molecules. The choice of descriptors in our model was suitable because they represented those well-known properties like the lipophilicity and the size of a molecule in addition to the charge aspect that is concerned with the inhibition of drugs (or candidates) against CYP3A4 enzyme. Our PreADMET descriptors – especially, molecular weight and fraction of 2D VSA chargeable groups – were satisfactory for classifying CYP3A4 inhibitors as majority of inhibitors were reasonably classified. The accuracy (72.33%) and sensitivity (75.82%) of the classification algorithm proved that our RP model is valid for the classification and prediction of inhibition and non-inhibition against CYP3A4 of new drug candidates. The RP model predicted an external validation set with the accuracy of 72.58% and sensitivity of 82.64%. The accuracy was increased to 73.8% in the RF model derived from using the ensemble method.

Although there are few limitations on the derived model, the fact that the sensitivity is quite higher than accuracy shows that the selected molecular descriptors appropriately represented the features of CYP3A4 inhibition. Hence, these RP and RF models could be applied to predict compounds in order to assist earlier identification of either CYP3A4 inhibitors or non-inhibitors during the preliminary step of drug discovery.

4. Experimental protocols

4.1. Data set

In order to assure an adequate extrapolation power for the models, both training and test sets with the maximum molecular diversity were chosen. The inhibitors and non-inhibitors of CYP450 3A4 were collected from commercially available Fujitsu database as the training set [26]. The database was constructed by retrieving inhibition data from a great number of publications. Any compound that exhibits inhibition (strong, moderate and/or weak) is classified as an inhibitor. Compounds that have been determined to be very weak inhibitors were excluded from the data set in order to remove

the bias in the compound distribution. Non-inhibitors were considered as compounds that are known to be inhibitors/substrates of other CYP450 isoforms except 3A4 [26]. Selenium and few metal containing compounds were excluded from the training set. Test set was generated using the Korea Research Institute of Chemical Technology (KRICT) proprietary drug candidate database.

4.2. Evaluation of effect of test set compounds on CYP3A4 activity

As for the test set, a number of synthesized compounds were experimentally determined as either CYP3A4 inhibitors or non-inhibitors. The CYP3A4 enzyme assay for the compounds in the test set was carried out using fluorometric enzyme assays with Vivid CYP3A4 assay kit (PanVera, USA, CA) in a 96-well microtiter plate following the manufacturer's instruction with some modification. Test compounds including ketoconazole known as CYP3A4 inhibitor were prepared in acetonitrile to give final concentrations of 10 μ M. NADP generating solution (1.0 mM NADP⁺, 3.3 mM glucose-6-phosphate, 3.3 mM MgCl₂·6H₂O, and 0.4 U/mL glucose-6-phosphate dehydrogenase in 10 mM KPO₄, pH 8.0) was added to each well of the microtiter plate followed by the vehicle acetonitrile (control) and the test samples. The plate was covered and then incubated at 37 °C for 20 min. Enzyme reaction was initiated by the addition of enzyme/substrate (E/S) mixture (0.5 pmol recombinant human CYP3A4 enzyme and 5 μ M dibenzylfluorescein, DBF). The plate was further incubated for 20 min, followed by the addition of the stop solution to terminate the enzyme activity. Background reading was measured in a similar manner except for the E/S mixture which was added after the enzyme reaction was terminated. The fluorescence of DBF metabolite fluorescein was measured on a fluorescence plate reader with an excitation wavelength of 485 nm and an emission wavelength of 530 nm. The effect of test compounds on CYP3A4 enzyme was calculated as the percentage of the enzyme activity. Test compounds with percentage inhibition less than 50% were defined as CYP3A4 non-inhibitors, and those above 50% were CYP3A4 inhibitors.

4.3. Drug-likeness factor analysis for training and test sets

The drug-likeness factors were analyzed for CYP3A4 inhibitors and non-inhibitors in training as well as test sets prior to building the models. The statistical analysis of drug-likeness properties such as molecular weight, calculated AlogP98 and topological PSA was done with SPSS (version 12.0, SPSS Inc., Chicago, IL). The compounds in the training set were distributed: (1) within the wide range of molecular weight from 200 AU to 700 AU where most of them were populated around 200–500 AU, (2) between the value of 1 and 5 AlogP98, and (3) between 0 and 150 topological PSA. In the test set, the compounds had: (1) molecular weight within the range of 300–600 AU, (2) AlogP98 value between 1 and 7, and (3) topological PSA between 50 and 125. Thus, the majority of the compounds in both data sets satisfied Lipinski's rule [27].

4.4. PCA analysis

We calculated each molecule in the training set with 2D molecular descriptors implemented in PreADMET software developed by BMDRC [12]. The PreADMET program provides rapid and reliable data of drug-likeness and ADME properties. It can also calculate constitutional, electrostatic, physicochemical, geometrical and topological descriptors, which have been developed in response to need for rapid prediction of drug likeness and ADME/Toxicity data. A total number of 240 molecular descriptors were calculated for our present study excluding topological descriptors. To visualize the chemical diversity of inhibitors and non-inhibitors within each training and test set, a statistical approach known as

PCA was used with SPSS. The variability of the descriptors was redistributed across a set of orthogonal indices of major three PCs. Next, the chemical diversity between training and test sets was assessed by PCA conducted on the default set of calculable molecular properties termed as the “predefined set (ALogP, MW, No of H donors, No of H acceptors, No of rotatable bonds, No of atoms, no of rings, no of aromatic rings and no of fragments)” in SciTegic Pipeline Pilot (version 6.1) [8].

4.5. Recursive partitioning and random forest models' development

The calculated molecular descriptors were used to configure a RP model. The RP training was performed using the CART (Classification and Regression Trees) algorithm implemented in Cerius2 using default settings [28]. The splits were scored using the Gini Impurity scoring function. The value of 1/100 of samples was considered as the minimum number of samples in any node. The maximum tree depth was 5 and the default values were accepted for the maximum number of generic splits and the number of knots per variable.

In addition, the RF algorithm in R [29] which is implemented in SciTegic Pipeline Pilot (version 6.1) [8] has been used for multi-class classification. In a RF, each tree is created from a bootstrapped sample of the training data, but at each node, only *mtry* variables are searched for possible split values [24,25]. We have trained a RF classifier with 500 trees. The default value of the square root of the number of descriptors was used to reflect the optimal degree of randomness [24,25].

4.6. Evaluation criteria

To evaluate the performance of the RP model, several measures were used. Accuracy is the overall classification accuracy of a prediction model; it corresponds to the ratio of correctly classified compounds to the total compounds. Sensitivity is the ratio of inhibitors correctly predicted, whereas specificity is the ratio of non-inhibitors correctly predicted. The performance of each prediction method is further evaluated by MCC [9,10]. It provides more detailed understanding for the prediction power by taking into account true and false positives and negatives. Its strength is that it could be used to measure the quality of binary (two class) classifications even for the imbalanced data sets. Kappa statistics is an index which compares the agreement against that which might be expected by chance. Kappa can be thought of as the chance-corrected proportional agreement [30]. Although accuracy, precision, and recall are affected by the composition of the data, that is, the distribution of the observations (inhibitor or non-inhibitor) [23], kappa uses expected agreement based on the ratio between the classes by chance [31] and therefore is a better measure of a model's predictive capability. Kappa measures the degree of agreement between the classification of the model and the true classes on a scale from 1 (perfect agreement) via 0 (no agreement above that expected by chance) to –1 (complete disagreement) [30].

Since there is no need for cross-validation or a separate test set to get an unbiased estimate of the test set error in RF [24], the OOB error estimate and ROC score determine the performance of the RF model.

Acknowledgements

This work was supported by Korea Research Foundation Grant (KRF-2006-005-J04501) of the Ministry of Education and Human Resource Development (MOEHRD), Korea and the 21st Century Frontier R&D program (CBM31-B2000-01-00-00) of the Center for Biological Modulators funded by the Ministry of Science and Technology, Korea.

References

- [1] J. Burton, I. Ijjaali, O. Barberan, F. Petitet, D.P. Vercauteren, A. Michel, J. Med. Chem. 49 (2006) 6231–6240.
- [2] N. Manga, J.C. Duffy, P.H. Rowe, M.T. Cronin, SAR QSAR Environ. Res. 16 (2005) 43–61.
- [3] S. Ekins, J. Berbaum, R.K. Harrison, Drug Metab. Dispos. 31 (2003) 1077–1080.
- [4] L.C. Wienkers, T.G. Heath, Nat. Rev. Drug Discovery 4 (2005) 825–833.
- [5] H. Park, S. Lee, J. Suh, J. Am. Chem. Soc. 127 (2005) 13634–13642.
- [6] A. Rusinko, M.W. Farmen, C.G. Lamber, P.L. Brown, S.S. Young, J. Chem. Inf. Comput. Sci. 39 (1999) 1017–1026.
- [7] D.R. Jones, S. Ekins, L. Li, S.D. Hall, Drug Metab. Dispos. 35 (2007) 1466–1475.
- [8] SciTegic Inc., 9665 Chesapeake Dr., Suite 401, San Diego, CA 92123, USA, Pipeline Pilot 6.1.5, version 6.1.5, 2007.
- [9] B.W. Matthews, Biochim. Biophys. Acta 405 (1975) 442–451.
- [10] P. Baldi, S. Brunak, Y. Chauvin, C.A.F. Andersen, H. Nielsen, Bioinformatics 16 (2000) 412–424.
- [11] J.R. Landis, G.G. Koch, Biometrics 33 (1977) 159–174.
- [12] BMDRC, PreADMET 2.0, Bmdrc, Seoul, Korea, 2007.
- [13] M.G. Hudelson, J.P. Jones, J. Med. Chem. 49 (2006) 4367–4373.
- [14] J.M. Kriegl, L. Eriksson, T. Arnhold, B. Beck, E. Johansson, T. Fox, Eur. J. Pharm. Sci. 24 (2005) 451–463.
- [15] S. Ekins, D.M. Stresser, J.A. Williams, Trends Pharmacol. Sci. 24 (2003) 161–166.
- [16] P.A. Williams, J. Cosme, D.M. Vinkovic, A. Ward, H.C. Angove, P.J. Day, C. Vonrhein, I.J. Tickle, H. Jhoti, Science 305 (2004) 683–686.
- [17] J.K. Yano, M.R. Wester, G.A. Schoch, K.J. Griffin, C.D. Stout, E.F. Johnson, J. Biol. Chem. 279 (2004) 38091–38094.
- [18] S.F. Zhou, C.C. Xue, X.Q. Yu, C. Li, G. Wang, Ther. Drug Monit. 29 (2007) 687–710.
- [19] M.J. de Groot, N.P. Vermeulen, J.D. Kramer, F.A. van Acker, G.M. Donné-Op den Kelder, Chem. Res. Toxicol. 9 (1996) 1079–1091.
- [20] D.R. Armour, M.J. de Groot, D.A. Price, B.L. Stammen, A. Wood, M. Perros, C. Burt, Chem. Biol. Drug Des. 67 (2006) 305–308.
- [21] J. Zuegge, U. Fechner, O. Roche, N.J. Parrott, O. Engkvist, G. Schneider, Quant. Struct. Act. Relat. 21 (2002) 249–256.
- [22] D.T. Stanton, P.C. Jurs, Anal. Chem. 62 (1990) 2323–2329.
- [23] R. Arimoto, M.A. Prasad, E.M. Gifford, J. Biomol. Screening 10 (2005) 197–205.
- [24] L. Breiman, Machine Learning 45 (2001) 5–32.
- [25] C.E. Keefer, N.A. Woody, Chemom. Intell. Lab. Syst. 84 (2006) 40–45.
- [26] S. Rendic, ADME Database – A Database of Substrates, Inhibitors, Inducers and Activators of Cytochrome P450 and Drug Transporters: URL: <http://jp.fujitsu.com/group/fqs/services/lifescience/english/asp/admedb/index.html>.
- [27] C.A. Lipinski, F. Lombardo, B.W. Dominy, P.J. Feeney, Adv. Drug Delivery Rev. 46 (2001) 3–26.
- [28] Cerius2 (Version 4.9), Accelrys, Inc., San Diego, CA, USA, 2003.
- [29] R Development Core Team, R: A Language and Environment for Statistical Computing ISBN 3-900051-07-0, R Foundation for Statistical Computing, Vienna, Austria, 2005, <http://www.R-project.org>.
- [30] <http://www.dmi.columbia.edu/homepages/chuangj/kappa/>.
- [31] J. Cohen, Educ. Psychol. Meas. 20 (1960) 37–46.
- [32] Spotfire® DecisionSite® 9.1.1, <http://spotfire.tibco.com>, Somerville, MA, USA, 2008.