Development of CYP3A4 Inhibition Models: Comparisons of Machine-Learning Techniques and Molecular Descriptors

RIEKO ARIMOTO, MADHU-ASHNI PRASAD, and ERIC M. GIFFORD

Computational models of cytochrome P450 3A4 inhibition were developed based on high-throughput screening data for 4470 proprietary compounds. Multiple models differentiating inhibitors (IC_{50} < 3 µM) and noninhibitors were generated using various machine-learning algorithms (recursive partitioning [RP], Bayesian classifier, logistic regression, k-nearest-neighbor, and support vector machine [SVM]) with structural fingerprints and topological indices. Nineteen models were evaluated by internal 10-fold cross-validation and also by an independent test set. Three most predictive models, Barnard Chemical Information (BCI)-fingerprint/SVM, MDL-keyset/SVM, and topological indices/RP, correctly classified 249, 248, and 236 compounds of 291 noninhibitors and 135, 137, and 147 compounds of 179 inhibitors in the validation set. Their overall accuracies were 82%, 82%, and 81%, respectively. Investigating applicability of the BCI/SVM model found a strong correlation between the predictive performance and the structural similarity to the training set. Using Tanimoto similarity index as a confidence measurement for the predictions, the limitation of the extrapolation was 0.7 in the case of the BCI/SVM model. Taking consensus of the 3 best models yielded a further improvement in predictive capability, kappa = 0.65 and accuracy = 83%. The consensus model could also be tuned to minimize either false positives or false negatives depending on the emphasis of the screening. (Journal of Biomolecular Screening 2005:197-205)

Key words: CYP3A4, BFC, in silico screening, machine learning, structural fingerprint, similarity index, kappa

Drug-drug interactions have resulted in costly late failures of drug development and withdrawal of drugs already on the market. Many of these drug-drug interactions can be traced back to cytochrome P450 3A4 (CYP3A4) inhibition. It is well known that CYP3A4 is the predominant enzyme involved in the oxidative metabolic pathways of many therapeutic agents and that inhibition of this enzyme in many cases leads to an undesired accumulation of the administered therapeutic agent. Accumulation of the unmetabolized drug creates an increased potential for causing various toxicities. It is obviously desirable to detect the potential of drug-drug interactions, as forecast by CYP3A4 inhibition, early in the drug discovery process. Fluorescent probe substrates enable high-throughput screening (HTS) of very large collections of compounds once they are synthesized. The data acquired from the in vitro assays have then been used to derive mathematical models. Such in silico screening tools are most useful for predicting the inhibitory potential of virtual compounds and of compounds yet to be screened.

CYP3A4 enzyme is known for its atypical interaction with its substrates in vitro as described in several reviews. Chemical inhibition of this enzyme is substrate dependent, and multiple-substrate experiments suggested that there are at least 3 classes of inhibitors, perhaps describing multiple modes of interactions with the enzyme. Past computational modeling efforts have mainly focused on ligand-based approaches such as pharmacophore mapping and quantitative structure activity relationship (QSAR) analysis. These modeling efforts were undertaken at a time when there was no publicly available crystal structure of CYP3A4. These approaches yielded predictive models for inhibition of various CYP3A4-mediated biotransformations. Ekins et al. reported Catalyst (Accelrys, San Diego, CA) pharmacophore models of CYP3A4 inhibitors for 2 metabolic reactions, midazolam 1′-hydroxylation (with 14 compounds), and cyclosporin A hydroxylation (with 32 compounds). Both models were able to predict Ki values for 7 of 8 test compounds within 1 log unit residual. A linear regression analysis of physicochemical properties of 30 diverse compounds found a strong correlation between IC_{50} for...
CYP3A4 catalyzed erythromycin N-demethylation and lipophilicity (logDpH).20 The study also pointed out that the presence of nitrogen-containing heterocycle in the structure increased the inhibitory potential, consistent with the property of prototypic CYP inhibitors such as ketoconazole, sulfaphenazole, and quinidine. Molnar and Keseru presented a neural network model differentiating inhibitor and substrate of CYP3A4 based on the literature data for 290 compounds.21 The data contained inhibitors for a number of different substrates, yet the model was very predictive, recognizing 89% of the inhibitors in the test set. This result might be suggesting that 1) the neural network was capable of formulating the complex SAR or 2) there are no distinct classes or substrate dependency for CYP3A4 inhibitors. To optimize the reliability and robustness of the result, it is necessary to model a large set of data, on the order of at least thousands of compounds, ideally measured using a uniform protocol executed under well-controlled experimental conditions. Recently, a recursive partitioning model trained with a commercially available database (Cerep, Redmond, WA) of 1750 compounds that predicted percentage inhibition of CYP2D6 and CYP3A4 was reported.22 The model was able to generate statistically significant rank ordering of the test set of 98 molecules.

It has been our experience that literature and commercially available models, in general, are of limited utility when applied to internal drug discovery projects. The published models provide an excellent source of information that can be used to focus modeling efforts on appropriate computational methods and chemical descriptors. Unfortunately, because of the differences in chemical compound collections and experimental methodologies between research institutions, models are usually not directly transferable. Because of these observations, we have embarked on building models to be more predictive for our own internal research efforts.

In the current study, we have exploited the data acquired from HTS for 4470 proprietary compounds to develop computational models of CYP3A4 inhibition. The experimental data used was percentage inhibition of 7-benzoyl-4-trifluoromethylcoumarin (BFC) metabolism, and the models were trained to differentiate inhibitor and noninhibitor. Combinations of various molecular descriptors and machine-learning methods were compared to determine the optimal descriptors and methods for generating the most predictive models. The limitations of the model’s predictive power outside the training set were also examined.

MATERIALS AND METHODS

Biological assay

All reagents were commercially obtained from Sigma-Aldrich (St. Louis, MO). Insect cell microsomes containing human NADPH cytochrome P450 reductase coexpressed with human CYP3A4 were purchased from Pan Vera LLC (Madison, WI). A single lot enzyme preparation was used to eliminate batch-to-batch variability.24 The fluorescent 7-BFC substrate was acquired from BD Biosciences (San Jose, CA). All test compounds were obtained from the Pfizer Global Research and Development compound collection.

The microsomal incubations were conducted in triplicate in a total volume of 75.75 µl in the presence of a freshly prepared NADPH regeneration system (1.3 mM NADP+, 3.3 mM Glc-6-P, 0.4 U/ml G6PDH, 3.3 mM magnesium chloride)25. A microsomal mixture of 0.368 micromolar protein, 50 µM 7-BFC, and 3 µM test compound26 in 84.33 mM potassium phosphate buffer, pH 7.4, was preincubated for 5 min at room temperature (28°C). Reactions were initiated by the addition of 10 µl of NADPH regeneration solution. Incubations without test compounds were used as high controls, and incubations without microsomes were used as low controls.

Fluorescence readings were taken every 2.07 min over a 25-min reaction period at λex/λem = 405/530 nM using a Tecan Farcyte plate reader (Amersham Pharmacia Biotech, Piscataway, NJ). Metabolite formation rates were determined from the slope of the time-dependent change in fluorescence intensity.24 The in vitro percentage inhibition of 7-BFC metabolism was calculated using the high and low controls as 0% and 100%, respectively.

Data with calculated values outside the 0% – 100% range were discarded. Test compounds with percentage inhibition greater than 55%, which is taken (with 5% margin) to imply an IC50 of less than 3 µM, were defined as CYP3A4 inhibitors. The 3-µM cutoff was chosen based on the implication that compounds with Ki values less than 1 µM are potent inhibitors likely to cause drug-drug interactions,26,27 not based on the distribution of the experimental data.

Descriptor generation

Molecular fingerprints and topological indices were used as molecular descriptors. With fingerprint systems, chemical structures can be represented by binary strings, with each bit indicating the presence or absence of certain structural features in the molecule. The 4470 molecular structures were represented in Daylight simplified molecular input line entry systems (SMILES) format,28 and MAKEBITS program (Barnard Chemical Information [BCI] Ltd, Sheffield, UK) was used to generate a BCI fingerprint of 4096 bits. Each is directly associated with a particular structural fragment, augmented atom, atom sequence, atom pair, ring composition, or ring fusion29 or a set of fragments, specified in the dictionary file supplied with the program. The SMILES strings were then converted into MDL SDfile format using CORINA30 to generate the following molecular descriptors. Molecular Operating Environment (Chemical Computing Group, Montreal, Canada) was used to generate MACCS fingerprints, consisting of 166 pre-defined fragments31 and Typed Graph Triangle (TGT) fingerprint of 13,608 bits. The TGT fingerprint encodes 3-point pharmacophores from 2-dimensional structures of the compounds. It assigns each atom a given molecular type in the set of donor, acceptor, polar, anion, cation, or hydrophobe, and all triplets were coded as features using the 3 graph distances (bond count) and 3 atom types.
in each triplet. MolconnZ (eduSoft, Ashland, VA) module within Sybyl (Tripos, St. Louis, MO) was used to calculate 156 topological and electrotopological indices. The topological indices, calculated from the 2-dimensional structures of the compounds, encode size, connectivity, branching, unsaturation, heteroatom content, and cyclicity of the molecules, and electrotopological indices represent an intrinsic electronic state with perturbations due to the topological state.

Model building

Multiple classification models were generated for all combinations of molecular descriptors (BCI, MACCS, TGT, MolconnZ) and machine-learning methods. The C5.0 program (RuleQuest, St. Ives, Australia) was used to construct decision trees by recursive partitioning (RP) with adaptive boosting. The boosting routine generates a sequence of shallow recursive partitions that are combined to form an overall prediction. It has been shown to significantly outperform a single tree. The misclassification cost for both inhibitor and noninhibitor were equally weighted. The RP tree-building algorithm chooses the best variable at every split; thus, variables that are less relevant to the inhibitory potency are omitted. In-house implementations of naive Bayesian classifier (BC), logistic regression, and k-nearest neighbor (k-NN) classification were also used to generate classification models. In k-NN, the default value of k = 9, number of neighbors, was used. The support vector machine (SVM) models were obtained using a publicly available program SVMlight. Radial basis kernel function and default values of capacity = 10 and cost-factor = 1 were used for the computations. Capacity represents the trade-off between training error and margin, and the cost factor is the number by which training errors on inhibitors outweigh errors on noninhibitors.

Models were evaluated in 2 ways: 1) 10-fold internal cross-validation using all the data (4470 compounds) and 2) validation with a fixed test set of 470 compounds. In the validation method (2), the test set (470) was randomly selected out of 4470 compounds, and the models were trained with the remaining 4000 compounds. In addition, selected combinations of methods and descriptors were tested with scrambled activities to investigate the likelihood of deriving models with reasonable fitness but that do not relate structure to inhibitory potential. The predictive abilities of the models were assessed by various statistics such as accuracy, precision, recall, and kappa. These statistics were calculated using the following formulae (1-4):

\[ \text{accuracy} = \frac{A + D}{A + B + C + D} \]  
\[ \text{precision (inhibitor)} = \frac{D}{B + D} \]  
\[ \text{recall (inhibitor)} = \frac{D}{C + D} \]  
\[ \text{kappa} = \frac{\text{accuracy} - E}{1 - E} \]  

where

\[ E = \text{expected agreement} = \frac{(A + C)(A + B) + (B + D)(C + D)}{(A + B + C + D)^2} \]

A, B, C, and D are the number of compounds in the confusion matrix as defined in Table 1. The accuracy is the overall correctness expressed by the fraction of the diagonal. Precision and recall represent the correctness of the prediction and retrieval rate in the certain class, “inhibitor” in equations (2) and (3). Kappa measures the degree of agreement between the classification of the model and the true classes on a scale from 0 to 1. Kappa = 0 indicates that the predictions are no more diagonal than expected from chance alone. Generally, models with kappa greater than 0.4 are considered to be predictive. Although accuracy, precision, and recall are affected by the composition of the data, that is, the distribution of the observations (inhibitor or noninhibitor), kappa uses expected agreement based on the ratio between the classes by chance (E in the equations above) and therefore is a better measure of a model’s predictive capability.

Consensus models with 3 selected classifiers were also tested. Given a chemical structure, each classifier voted for its result, inhibitor or noninhibitor, and the final prediction was based on the vote counts. The number of votes required for a compound to be predicted as an inhibitor varied from 1 to 3 to examine the effects on the model statistics.

All calculations were carried out on Silicon Graphics Octane R14000 and Origin R12000 workstations.

Similarity index

A standard similarity index was used to measure the structural similarity between the training set and a compound being evaluated. The similarity index used in this study was defined as the Tanimoto coefficient (overlap-bits/union-bits) of BCI fingerprint between the test compound and its nearest neighbor in the training set. To find the nearest neighbor, Tanimoto coefficients were calculated against all 4000 compounds, and the maximum value was assigned as the similarity index to the test compound.

RESULTS AND DISCUSSION

Figure 1 shows the experimental data used in this study. The training set (n = 4000) is in gray, and the hatched area on the bottom represents the test set (n = 470). The distribution between the classes (noninhibitor/inhibitor) with 55% cutoff is shown in the inset table. The ratios between the classes in the training set and in the test set are adequately close, 0.65/0.35 and 0.62/0.38, respectively.

When models were evaluated by 10-fold cross-validation, all 4470 compounds were used. The cross-validation routines left out
a set of 447 compounds in sequence for testing the models trained with the remaining 4023 compounds, repeating 10 times. The 10 models created during this process were not identical, but consistency in the classification rules among them was assumed for the validation to be reliable. In the validation with test set, 470 compounds were randomly selected and left out. The models were trained with the remaining 4000 compounds. Table 2 lists the kappa values of the test set validation and 10-fold cross-validation for the 19 models generated. The TGT fingerprint contains too many attributes (13,608 binary values) for the recursive partitioning approach to be investigated using this descriptor set. Kappa values of the 2 validation methods for each model were all quite similar as we expected. It indicates that the test set was properly representing the entire data and also that the 10 models generated by the cross-validation process were likely to share common classification rules. For comparison, single nearest-neighbor (k = 1) calculations were also carried out with BCI fingerprint. Kappa values were 0.56 for the test set and 0.54 for the cross-validation. These results were comparable to the k-NN (k = 9) model as we expected.

FIG. 1. Distribution of the experimental data, percentage inhibition of 7-benzyloxy-4-trifluoromethylcoumarin metabolism, used in the current study. The hatched area represents the test set of 470 compounds. The inset table lists the data set composition.

Table 1. Confusion Matrix of CYP3A4 Inhibition Model

<table>
<thead>
<tr>
<th></th>
<th>Predicted</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Noninhibitor</td>
<td>Inhibitor</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observed</td>
<td>A (true negative)</td>
<td>B (false positive)</td>
<td>C (false negative)</td>
<td>D (true positive)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>% inhibition</td>
<td>training set</td>
<td>test set</td>
<td>training set</td>
<td>test set</td>
<td>training set</td>
<td>test set</td>
</tr>
<tr>
<td>≤ 55 %</td>
<td>2,601</td>
<td>291</td>
<td>1,399</td>
<td>179</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt; 55 %</td>
<td>100</td>
<td>200</td>
<td>500</td>
<td>600</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>4,000</td>
<td>470</td>
<td>1,599</td>
<td>357</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

FIG. 2. Structures of common drugs in the data set.

Table 2. Kappa Values of the Models for the Test Set Predictions

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>Method</th>
<th>Information</th>
<th>MACCS</th>
<th>MolconnZ</th>
<th>Graph Triangle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barnard</td>
<td>Bayesian classifier</td>
<td>0.38 (0.44)</td>
<td>0.23 (0.32)</td>
<td>0.24 (0.24)</td>
<td>0.22 (0.27)</td>
</tr>
<tr>
<td></td>
<td>Recursive partitioning</td>
<td>0.53 (0.56)</td>
<td>0.59 (0.55)</td>
<td>0.62 (0.60)</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>k-nearest neighbor</td>
<td>0.56 (0.56)</td>
<td>0.48 (0.52)</td>
<td>0.58 (0.54)</td>
<td>0.51 (0.53)</td>
</tr>
<tr>
<td></td>
<td>Logistic regression</td>
<td>0.56 (0.58)</td>
<td>0.52 (0.50)</td>
<td>0.53 (0.51)</td>
<td>0.54 (0.54)</td>
</tr>
<tr>
<td></td>
<td>Support vector machine</td>
<td>0.61 (0.61)</td>
<td>0.62 (0.59)</td>
<td>0.54 (0.50)</td>
<td>0.53 (0.54)</td>
</tr>
</tbody>
</table>

Kappa values for 10-fold cross-validations are in parentheses.
Table 3. Predictions by Barnard Chemical Information/Support Vector Machine Model and IC_{50} for the Common Drugs in the Data Set

<table>
<thead>
<tr>
<th>Compound Name</th>
<th>+ Inhibitor</th>
<th>− Noninhibitor</th>
<th>IC_{50} (μM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Astemizole</td>
<td>+</td>
<td></td>
<td>1.2</td>
</tr>
<tr>
<td>Carbamazepine</td>
<td>−</td>
<td></td>
<td>175</td>
</tr>
<tr>
<td>Clotrimazole</td>
<td>−</td>
<td></td>
<td>0.002</td>
</tr>
<tr>
<td>Haloperidol</td>
<td>−</td>
<td></td>
<td>52</td>
</tr>
<tr>
<td>Ketoconazole</td>
<td>+</td>
<td></td>
<td>0.005</td>
</tr>
<tr>
<td>Miconazole</td>
<td>+</td>
<td></td>
<td>0.057</td>
</tr>
<tr>
<td>Nifedipine</td>
<td>−</td>
<td></td>
<td>8.4</td>
</tr>
<tr>
<td>Propofol</td>
<td>−</td>
<td></td>
<td>51</td>
</tr>
<tr>
<td>Quercetin</td>
<td>+</td>
<td></td>
<td>&gt;50</td>
</tr>
<tr>
<td>Quinidine</td>
<td>−</td>
<td></td>
<td>34</td>
</tr>
<tr>
<td>Tamoxifen</td>
<td>−</td>
<td></td>
<td>&gt;1000</td>
</tr>
<tr>
<td>Verapamil</td>
<td>+</td>
<td></td>
<td>0.36</td>
</tr>
</tbody>
</table>

FIG. 3. Comparisons of descriptors and methods. Graphs (a) and (b) show the same set of 19 classification models in different x-axes, descriptors, and learning methods, respectively. Triangles represent test set predictions, and squares are the results of 10-fold cross-validations. BCI = Barnard Chemical Information; TGT = Typed Graph Triangle; k-NN = k-nearest neighbor; LR = logistic regression; RP = recursive partitioning; SVM = support vector machine.
cross-validation kappa, and training kappa of these 3 models are listed in Table 4. The overall accuracies as well as kappa values were almost identical.

As one can see from the formulae (2) and (3) in the Materials and Methods section, the precision reflects false positives (B) and recall reflects false negatives (C). As the number of false positives (negatives) increases, precision (recall) value decreases. The MolconnZ/RP model was slightly overpredicting inhibitors compared to the other 2 models and resulted in more false positives (55 structures). Although this model identified more risky compounds correctly, the loss due to the false risky compounds was also greater. Considering that MolconnZ descriptors are distinctly different from BCI and MACCS fingerprints, they might have captured additional information about CYP3A4 inhibition SAR from the data, resulting in higher recall value and more false positives.

An additional test was conducted for these 3 descriptor/method combinations. The models were computed with scrambled experimental data. In training, recursive partitioning with MolconnZ descriptors classified all 4470 compounds as noninhibitors and did not carry out any boosting as the algorithm determined that no improvement would result from this operation. It can be explained that all compounds were classified into the default (dominant) class, noninhibitor, as no relationship between the descriptors and the dependent variables was found. BCI/SVM and MACCS/SVM combinations trained models with the randomized data just as well as with the real data (training kappa = 0.61, 0.78, respectively), but the validations yielded kappa values less than 0.1, proving that the resultant models did not have any predictive capability. These results illustrate examples of overtraining by SVM and stress the importance of thorough validations. The randomization test showed that the descriptors used for this study do indeed carry information that can be related to CYP3A4 inhibition and that the likelihood of identifying chance relationships with these methods is real but manageable using the aforementioned validation methods.

When new compounds are evaluated, it is necessary to know extrapolation limits of the models. Thus, the applicability of the model was investigated using the best model, BCI/SVM. If the test compound resides far outside of the area in the descriptor space that was covered by the model (training compounds), the classification of that compound would not be reliable. The similarity index described in the Materials and Methods section is one way to quantify how close a compound is to the training set on a scale from 0 to 1, with 1 being the closest. The model’s applicability was expected to correlate with the similarity index. Figure 4a shows the similarity indices of the test set. The median was 0.83, indicating that most compounds were in close neighborhoods of the training set. This was more or less expected because this test set and the training set were randomly separated from a single data set; the test set practically came from the same population as the training set. The model is clearly applicable for compounds with such high similarities, and the predictions indeed agreed well with the experimental measurement, as described above.

To further investigate the correlation between similarity and predictability for more diverse structures, additional 2195 compounds were tested with the BCI/SVM model. These new data were deposited into the corporate database as this study was in progress. The experimental data were acquired under the conditions described in the Materials and Methods section. Figure 4b shows the distribution of the similarity index of this additional test.
set of 2195 compounds, ranging from 0.125 to 1.0. This new data
set was divided into 5 groups with different similarities to the train-
ing set. The similarity index medians of the 5 groups were 0.3,
0.55, 0.7, 0.8, and 0.9. Each group contained ~400 compounds.
The BCI/SVM model was then applied to each group, and the
predictions and the experimental data were compared.

Figure 5 shows the plots between the similarity index and vari-
ous statistic measures related to the predictive power, kappa, over-
all accuracy, precision, and recall. The y-axis on the left is
for kappa values, and the y-axis on the right is for accuracy, precision, and recall in percentages. Each group is represented by its similarity index me-
dian in the graph (0.9, 0.8, 0.7, 0.55, and 0.3).

![Figure 5. Predictive ability versus similarity. Kappa, accuracy, precision, and recall for the 5 data sets (~400 compounds each) with different structural similarities to the training set are plotted. The y-axis on the left is for kappa values, and the y-axis on the right is for accuracy, precision, and recall in percentages. Each group is represented by its similarity index median in the graph (0.9, 0.8, 0.7, 0.55, and 0.3).](image-url)

131 compounds. Consequently, most of the compounds were as-
signed into the default (dominant) class, which was noninhibitor. This suggests that one must always be cautious when the model statistics are good due to an exceptional agreement in the default class. The large deviation between precision and recall in Figure 5 implies that the model is not actually predictive, and the results need to be examined in detail. Kappa alone could be misleading, particularly when the value is low.

Overall, the BCI/SVM model’s capability of extrapolation de-
clined sharply when the similarity index became lower than 0.8.
Following the kappa = 0.4 guideline,44 the reasonable similarity
threshold is 0.7. This implies that the compound being evaluated
must have a similarity index greater than 0.7 to obtain a reliable prediction. Although the correlation between similarity and the
model’s predictive power should hold for any type of computa-
tional model, the actual value of the similarity threshold depends
on the descriptors used and how the similarity index is defined.

It has been shown that using ensembles of models greatly im-
proves the predictions.50 Such ensemble learning methods include bagging,51,52 boosting,37 and random forest.53,54 Thus, the consen-
sus of the 3 best-performing individual models was considered.
Each of the 3 models voted for its prediction and various criteria to
determine the final predictions were compared. Table 5 lists pre-
dictions for the test set. The 1st column indicates the number of
votes required for a compound to be predicted as an inhibitor. The
1-vote model assigned a compound as a CYP3A4 inhibitor unless
all 3 models agreed that it was not and tended to overpredict inhibi-
tors (increase false positives, B). On the other end, the 3-vote
model determined a compound as an inhibitor only when all the
models agreed so, leading toward an underprediction of inhibitors
(increase false negatives, C). The 2-vote model was most balanced,
as expected, and in fact showed a slight improvement in kappa
value (0.65) and accuracy (83%) over any of the 3 individual mod-
els, demonstrating the benefit of the ensemble learning. This ap-
proach is particularly useful because a consensus model is tunable
depending on the purpose of the screening, minimizing false nega-
tives or false positives. In practice, some false negatives (a com-
 pound is in fact an inhibitor but predicted to be a noninhibitor) can
be tolerated in the early stage of drug discovery, but excluding too
many compounds, particularly false positives, should be avoided.
as much as possible. Therefore, tuning to minimize false positives while maintaining the ability to identify risky compounds is ideal in discovery projects.

CONCLUSIONS

Computational models were developed for the purpose of screening compounds against CYP3A4 inhibitory potential. The models are classifiers that determine if a compound is a CYP3A4 inhibitor, IC\textsubscript{50} less than 3 \mu M. BCI and MACCS fingerprints with SVM and MolconnZ descriptors with RP/boosting combinations resulted in highly predictive models. The overall accuracies of the models were 82\%, 82\%, and 81\%, and kappa values were 0.62, 0.61, and 0.62, respectively, for the test set not included in the model training. All models were trained with a large data set of 4000 proprietary compounds with percentage inhibition of 7-BFC metabolism by recombinant CYP3A4.

The limitation of the model’s extrapolation capability was also addressed using the BCI/SVM model. The reliability of the predictions was strongly correlated with the structural similarity between the test compounds and the training set. With additional 2195 compounds being tested, it was observed that the predictive power of the model dropped sharply when the similarity fell below 0.8. The limit that could sustain the applicability of the models, as defined by kappa > 0.4,\textsuperscript{44} was ~0.7.

The consensus learning was able to improve performance over any individual models, and the resultant model was tunable, minimizing false positives or false negatives, depending on the emphasis of the screening process.

The models are useful in identifying compounds with potential risk of inhibiting the CYP3A4 enzyme. This is valuable for drug discovery projects as drug-drug interaction, mostly caused by CYP3A4 inhibition, could be taken into consideration early in discovery, and the costly late failure could be avoided. The most attractive feature of such in silico tools is perhaps the ability to assess virtual compounds. The current models can rapidly screen a large number of structures, guide the chemical synthesis process, and help the discovery programs to allocate the resources efficiently.

ACKNOWLEDGMENTS

We thank Alain Calvet, Kjell Johnson, and Mark Snow for their comments, discussions, and support.

REFERENCES

CYP3A4 Inhibition Model