J Comput Aided Mol Des (2010) 24:23–35
DOI 10.1007/s10822-009-9309-9

Development of QSAR models for microsomal stability: identification of good and bad structural features for rat, human and mouse microsomal stability

Yongbo Hu · Ray Unwalla · R. Aldrin Denny · Jack Bikker · Li Di · Christine Humblet

Received: 28 July 2009 / Accepted: 11 November 2009 / Published online: 24 November 2009

Abstract High throughput microsomal stability assays have been widely implemented in drug discovery and many companies have accumulated experimental measurements for thousands of compounds. Such datasets have been used to develop *in silico* models to predict metabolic stability and guide the selection of promising candidates for synthesis. This approach has proven most effective when selecting compounds from proposed virtual libraries prior to synthesis. However, these models are not easily interpretable at the structural level, and thus provide little insight to guide traditional synthetic efforts. We have developed global classification models of rat, mouse and human liver microsomal stability using in-house data. These models were built with FCFP_6 fingerprints using a Naïve Bayesian classifier within Pipeline Pilot. The test sets were correctly classified as stable or unstable with satisfying accuracies of 78, 77 and 75% for rat, human and mouse models, respectively. The prediction confidence was assigned using the Bayesian score to assess the applicability of the models. Using the resulting models, we developed a novel data mining strategy to identify structural features associated with good and bad microsomal stability. We also used this approach to identify structural features which are good for one species but bad for another. With these findings, the structure-metabolism relationships are likely to be understood faster and earlier in drug discovery.

Y. Hu (✉) · R. Unwalla · R. A. Denny · J. Bikker · C. Humblet
Department of Computational Chemistry and Chemoinformatics,
Wyeth Research, Pearl River, NY 10965, USA
e-mail: huy2@wyeth.com

L. Di
Department of Chemical Profiling, Wyeth Research, Princeton,
NJ 08543, USA

Keywords Liver microsomal stability · Naïve bayesian · Prediction confidence · Chronological effect · Fragment mining · MetaSite prediction

Introduction

Metabolic stability is an important property of drug candidates. Several studies have implicated poor metabolism or PK as one of the major causes for drug candidates to fail in development [1]. Therefore, compounds are evaluated early in drug discovery to profile their ADME/TOX properties in an effort to reduce the attrition rate due to poor PK or bioavailability (“fail early fail cheap” strategy). High throughput metabolic stability assays have been widely implemented in many pharmaceutical companies to support early drug discovery efforts [2–4]. These assays are being used to predict in vivo performance, help lead selection, guide structural modification, and develop structure–metabolism relationships. After years of such efforts, many companies have generated large experimental datasets with thousands of compounds and have used them to develop *in silico* QSPR models to predict the metabolic stability of compounds [5–8]. Unlike datasets compiled from the literature, these datasets are usually collected from a single assay and have relatively low noise, which is essential to establish quality models. Recently, several predictive classification models of human or rat liver microsomal stability have been reported [5–8]. Such models can be applied to virtual compounds, permitting rapid and cost-effective bias of design toward desirable candidates prior to synthesis.

Most of the published models of microsomal stability have been proposed as ‘global’ models and are not restricted to a particular chemical class or structural series. As valuable as these models are, little has been done to detect most important variables and provide interpretations of the models. Thus, valuable information connecting structural features to metabolic stability or instability is lost [9]. A possible reason for lacking model interpretation might be the absence of the appropriate software to do such an analysis. SciTegic's Pipeline Pilot software offers the capability of identifying fragments associated with activity after a Bayesian model is built using fingerprints. The use of this capability to identify fragments that contribute to stability or instability of compounds in human liver microsomes has been reported [6]. Fragments associated with good and bad microsomal stability can be identified and highlighted in the molecules whose activity is predicted. In our hands, however, we find that the criteria used to score fragments are highly susceptible to bias due to the existence of close analogs in datasets.

In this paper, we will describe our efforts to develop and validate classification models of the metabolic stability of drug candidates in liver microsomes from different species. To the best of our knowledge, we are the first to report microsomal stability models for multiple species, although models for single species have been reported [5–8]. From a drug discovery perspective, ADME/Tox data for a drug candidate in the selected model species is very important for the compound to advance. The CYP450 enzyme family is the major drug-metabolizing enzyme family in liver. Although some CYP450 isoforms such as 3A4, 2C9, 2D6 are common to all mammalian species of interest to drug discovery, there is widespread variability in different species such as shown in Table 1 [10]. Thus, a compound's metabolic stability can vary when tested in microsomes from different species. Significant discrepancies between human metabolism and that in the species that serves as an in vivo model of disease presents a significant practical hurdle that must be overcome in a drug discovery program. We felt it was important to build microsomal stability models for other species besides human. We selected rat and mouse as the representative animal models because these two species are most frequently used as animal models for our discovery projects.

Table 1 Unique CYP450 enzymes for different species (common isoforms are not shown)

Human	Mouse	Rat	Dog	Monkey
3A4	3A11	3A1/3A23	3A12	3A8
3A5	3A13	3A2^{m}	3A26	
3A7	3A16	3A9^{f}		
3A43	3A25	3A18^{m}		
	3A41	3A62		
	3A44			

3A2 and 3A18 exist only in male rat, while 3A9 is found in female rat

In this study, a Naïve Bayesian classifier was applied to generate models of human, mouse, and rat metabolic stability based on data obtained for in-house compounds [6, 11]. The developed models were validated using test sets selected from the same pool of data used to train the models, followed by validation sets selected chronologically from subsequent quarterly metabolic testing. This approach also allowed us to estimate the prediction power drift in the chemical nature of the compounds being synthesized and the consequent interval after which the model must be retrained.

As mentioned above, a model would be more useful if the model results could be translated into practical ideas such as specific molecular features to modify in further synthetic effort during a lead optimization process. Naïve Bayesian classifier along with structure-based fingerprints allows us to develop a novel data mining strategy to identify structural features associated with good and bad microsomal stability. Structural features associated with a high incidence of rat, mouse, or human liver instability were mined from the available data. We also demonstrate that these fragments can be even further analyzed with MetaSite to provide insight in the source of compound instability in liver microsomes [12]. We also examined whether this approach could be used to identify features which are good for one species but bad for another. With these findings, we demonstrate how an integrated set of modeling tools can be used to establish structure-metabolism relationships that may move beyond the value of the classification model itself.

Methods

Microsomal stability assay procedure

The assay protocol has been described in detail elsewhere [13, 14]. Briefly, the samples were incubated with liver microsomes at 1 μM at microsomal protein concentration of 0.5 mg/mL at 37° C in the presence of NADPH cofactor. At both 0 and 15 min incubation time, cold acetonitrile was added to the wells to stop the reaction. The solution was centrifuged and the supernatants were analyzed using LC–MS. The microsomal stability assay was designed to minimize the effect of solubility on assay results by eliminating intermediate aqueous dilution steps, ensuring low test compound concentration, adding organic stock solution directly to microsomal proteins to avoid precipitation and pre-warming all the solutions. We found the method works very well for drug discovery compounds, even for very insoluble compounds with a Log P of 12 [13]. The microsomal stability assay concentration was very low to minimize any inhibitory effect. If a compound is a substrate and

Table 2 Size of training, test and validation sets for rat, human and mouse microsomal stability data

	Rat			Human			Mouse		
Dataset	Training	Test	Validation	Training	Test	Validation	Training	Test	Validation
Stable	5,933 (39%)	1,511 (39%)	2,085(43%)	2,410(57%)	603(58%)	825(69%)	681(42%)	191(47%)	260(60%)
Unstable	9,422 (61%)	2,329 (61%)	2,801(57%)	1,794(43%)	446(42%)	368(31%)	936(58%)	213(53%)	173(40%)
Total	15,355	3,840	4,886	4,184	1,049	1,193	1,617	404	433

the inhibitory effect is due to binding to the active site, metabolic stability is a precise measure of metabolism. If a compound is a substrate and the inhibitory effect is not due to binding to the active site, the actual metabolic rate might be higher. However, since the assay concentration is so low, the inhibitory effect is minimized.

Microsomal stability datasets

At Wyeth, half life is used as the end point for in vitro microsomal stability measurement. Considering the practical significance of microsomal stability in the context of drug design, we divided compounds into two classes: stable class ($t_{1/2} \geq 14$ min) and unstable class ($t_{1/2} < 14$ min). The selection of the cut-off was based on a previous study [14]. In that study, we examined the relationship between in vitro microsomal stability and in vivo clearance using 306 compounds from Wyeth drug discovery efforts. For compounds with low in vitro microsomal stability ($t_{1/2} < 14$ min), 87% showed high clearance in vivo (CL > 25 mL/min/kg). For compounds with high in vitro microsomal stability ($t_{1/2} \geq 14$ min), no significant differentiation was observed between high and low in vivo clearance compounds.

For the last decade, all of the requested tests of half life of compounds in different microsomes were performed by a dedicated central laboratory. To ensure the consistency of the half life measured at different time periods, we used only the data measured with the current standard high throughput assay [13]. We applied an internally-developed data retrieval and cleaning procedure to ensure the data integrity and accuracy. For the data cleaning process, problematic data with poor mass spectra or bad HPLC peaks were discarded. For compounds with multiple measurements, which account for about 5% of the total compounds, the median, maximum and minimum values were calculated. We discarded any data for which the calculated maximum was more than or equal to the half-life cutoff of 14 min and for which the minimum was less than 14 min. As a result, the compounds were consistently classified as either instable or unstable. We use mean value as the half - life for these compounds. To further minimize the noise of the data, we eliminated any data with a half-life value between 12 and 15 min.

The numbers of compounds tested for rat (SD, male), mouse (C56BL6, male), and human liver microsomes (Mixed) are shown in Table 2 (combination of training and test set). The rat dataset was the largest and consisted of 19,195 compounds from more than 60 discovery projects representing many different therapeutic areas. The human and mouse datasets were smaller than the rat set, and comprised 5,233 and 2,021 compounds, respectively. Some compounds were tested for their stability in the microsomes of two or three of the species. We partitioned each dataset into training (80%) and test (20%) sets. Table 2 also shows the number and percentage of stable and unstable compounds for the training and test sets using a half-life of 14 min as the threshold.

After the models were built and evaluated using the test sets, we evaluated their continuing performance using test results from compounds synthesized in the subsequent 3 months period. The distribution of stable/unstable compounds of these “validation” sets is shown in Table 2. As new compounds are synthesized, we hypothesized that the QSAR equation would become progressively less effective at predicting stability. Therefore, we assessed random subsets 1,000 compounds/quarter of the rat liver microsomal stability data for each 3 months for a period of 1 year. These 4,000 compounds were used to understand the chronological drift of the effectiveness of the models and thereby to determine how frequently the models should be updated.

Sampling for diversity

A proprietary algorithm was used to select diverse compound subsets. The similarity of two compounds was calculated using atom-pair similarity described by Carhart [15]. The algorithm conducts an exhaustive search to select a subset of compounds in which every compound has a pair-wise similarity less than a threshold. If all compounds in the full set are similar to at least one compound in this subset, a diverse set of compounds would return a subset of approximately equal size, whereas a group of close analogues could be represented by a small fraction of the full dataset. For this study, a similarity threshold of 0.7 (70%) was applied to select diverse subsets [13].

Model building

Models were built with a Naïve Bayesian classifier as implemented in Pipeline Pilot (version 7) [16]. The charges of the molecules were assigned with a Pipeline Pilot component at pH 7.4, which is the pH condition used for the microsomal stability assay. SciTegic's FCFP_6 fingerprints (Functional Class Finger Print of maximum diameter 6), ALogP and molecular weight were calculated in Pipeline Pilot and were used as the descriptors. A Bayesian classifier compares the frequency of occurrences of descriptors that are found in two or more groups that discriminate best between these groups. The 'Cross-validated Learn Good Molecules (Naïve Bayesian)' component in Pipeline Pilot builds models using samples from the training sets, and validates them using a leave-one-out cross-validation. In order to improve the capability of discriminating stable compounds from the rest, a post-processing step was performed to remove low-information FCFP_6 bins. Low-information bins are those that have normalized estimates in the range of −0.05 and 0.05.

After models were built, predictions were made for the test and validation sets. Statistical parameters were calculated to assess the accuracy and predictive power of the models. The parameters include concordance, sensitivity, specificity and Kappa values, as shown in Table 3. Concordance is calculated by the number of the correctly predicted compounds (A + D) divided by total compounds (*N*). Sensitivity is calculated by the number of the correctly predicted stable compounds (A) divided by the total number of the experimentally stable compounds (A + B). Specificity is the same for unstable. For a classification model, kappa is considered as true accuracy as the agreement by chance is corrected and therefore is a better parameter to estimate the prediction quality of the model [5, 6, 17]. The equations used to calculate Kappa are listed in Table 3. A kappa value of 0.4 or greater is often considered to be indicative of a model with useful predictive power [16].

Prediction confidence assignment

We used the calculated Bayesian probability scores of the test set compounds to estimate the confidence level of each prediction. The scores were binned using a bin size of 5. For each model, we calculated the actual/predicted concordances for each bin. Using a concordance of 0.8 as the cutoff (80% correctly predicted), the Bayesian score ranges with corresponding concordances above 0.8 were considered as 'High' confidence. If a prediction with the calculated Bayesian score falls into these ranges, it was labeled as 'High' Confidence. Similarly, if a concordance is between 0.8 and 0.7, the score range was considers as 'Medium' confidence. The rest was classified as 'Low' confidence.

Model interpretation

The use of the Naïve Bayesian classifier in Pipeline Pilot allowed us to apply their fragment mining method to identify substructures frequently found in stable and unstable compounds. To maximize the volume of the information, we combined the training, test and validation sets used in the modeling building process for each species. These combined datasets were used as new training sets for another round of Bayesian modeling. The top 10% FCFP_6

Table 3 Calculation of statistical parameters and the results for each model

		Predicted	
		Stable	Unstable
Observed	Stable	A	B
	Unstable	C	D

N =	A + B + C + D
Concordance =	(A + D)/N
Sensitivity =	A/(A + B)
Specificity =	D/(C + D)
E =	(A + C)(A + B) + (B + D)(C + D)/N x N
Kappa =	(concordance - E)/(1 - E)

	Rat		**Human**		**Mouse**	
Dataset	Test	Validation	Test	Validation	Test	Validation
Concordance (%)	78	71	77	72	75	77
Kappa	0.53	0.38	0.55	0.41	0.5	0.53
Sensitivity (%)	73	57	76	73	74	76
Specificity (%)	80	81	80	71	77	78

fingerprint features, which were most frequently observed in stable and unstable compound in each new training set, were saved. For each fingerprint feature retained, the compounds which contained this fingerprint were pulled out and saved as a compound set named by the fingerprint. To correct bias created by close analogues, a diverse subset was selected using atom-pair similarity of less than 0.7 [15]. The Bayesian probability scores were recalculated for each diverse subset and these top 10% FCFP_6 fingerprint features were re-ranked based on the adjusted scores. The FCFP_6 fingerprint features most associated with stable or unstable compounds were translated into 2D substructures within Pipeline Pilot.

To assess how each substructures might contribute to the instability of the compounds in human liver microsomes, the MetaSite software (version 3.0) was used to predict the metabolic soft spots of groups of unstable compounds. MetaSite is a computational procedure that predicts metabolic transformations related to cytochrome-mediated reactions in phase I metabolism in human [12]. The method uses flexible molecular interaction fields generated by GRID to determine the probability of an atom being the site of metabolism. Previous studies indicated that MetaSite's first three ranked predictions were experimentally confirmed for about 84% of the cases [12, 18–20]. The method not only predicts the metabolic soft spots of a compound, but also highlights the molecular moieties that help to direct the molecule in the cytochrome cavity such that the site of metabolism is in proximity to the catalytic center. We evaluated several fragments mined from our data above in MetaSite for consistency across predictive methods.

For each fragment associated with bad human microsomal stability, we visualized the first three ranked predicted soft spots for each compound. Based on this analysis, each bad fragment could be grouped into one of two categories. First, the fragments could contain a predicted soft spot; second, the fragments contained none of the first three ranked predicted soft spots, but one of their atoms was predicted as a major contributor for the CYP450 metabolism by helping the molecules to bind in the cytochrome cavity.

Results and discussion

The model development and validation procedure presented in this study provides a robust method for estimating the metabolic stability of compounds prior to synthesis. This includes an estimate of the accuracy of the prediction. By applying models developed on data of three species, the consistency of predicted metabolic behavior across species can be estimated. We report on our experience of the drift of model applicability over time, and propose a schedule for re-deriving such "global" models. We also report on our use of fragment mining to identify groups potentially leading to metabolic instability. By highlighting these groups as part of the prediction, these can be identified and modified at the design stage. In some cases, we demonstrate that these fragments can be further understood using the MetaSite software [12]. Taken together, these provide the medicinal chemists with an integrated assessment of the potential metabolic behavior of a compound prior to synthesis.

Diversity of the datasets

The chemical diversity of these datasets can be understood to some extent by applying a sampling algorithm to create diverse subsets of the training datasets, as shown in Table 4. These diverse subsets were generated with a proprietary algorithm using atom-pair similarity fingerprints. As a diverse set is defined by a similarity threshold of 0.7 [13], the diverse subset of the rat data was approximately 1/4 of the full dataset, whereas mouse and human returned subsets of 1/6 their parent size. The rat assay has been used as the primary in vitro microsomal stability assay at Wyeth. Any compound synthesized for a project is tested by this assay, as long as there is sufficient compound available. In contrast, human and mouse assays are only performed for more advanced compounds generated by later stage discovery projects. These compounds are usually derivatives of some specific scaffolds and are therefore less diverse. For this reason, the human and mouse data are less diverse than the rat data.

Overall accuracies of the models

Several studies have compared the performance of various methodologies, such as *k*-Nearest-Neighbor, Random Forest, and Naïve Bayesian [5–7]. It was found that the resulting differences between the various models in terms of prediction accuracies are relatively small. Based on this information, we selected a single method, Naïve Bayesian, and applied it across our datasets. The microsomal stabilities were classified into two classes: stable compounds with a half-life longer than or equal to 14 min, and unstable compounds with a half-life shorter than 14 min. By setting

Table 4 Size of diverse selection of the training sets with similarity thresholds

		Rat	Human	Mouse
Sim ≤	0.7	4,116	724	289
Sim ≤	0.8	7,071	1,465	590
Sim ≤	0.9	11,866	2,879	1,159
All		15,353	4,184	1,617

the stable compounds as "good" samples and the unstable compounds as "bad", the classifier derived from the training sets gave models with Receiver Operating Characteristic (ROC) accuracy of 0.84, 0.85, and 0.82 for rat, human and mouse liver microsomal stability models, respectively [21]. Table 3 lists the statistical parameters of the test and validation sets for the three resulting models. The concordance values show the overall prediction accuracy for each model. The rat model correctly classified 78% of the compounds, or 2,998 of 3,844 compounds in the rat test set. The human and rat Bayesian models were able to correctly classify 77 and 75% of the compounds in the test sets. For the validation sets, the rat, human and mouse models correctly predicted 71, 72 and 77% of the compounds in the corresponding test sets. The prediction accuracies for the validation sets slightly decreased, compared to those for the test sets, which is consistent with independent studies from other research groups [5–8].

The Kappa value is a measure of corrected accuracy which takes into account the agreement that may have occurred by chance, and therefore is a better assessment of true accuracy of a classification model [17]. A Kappa value could range from +1 to −1, with +1 as the perfect agreement and −1 as complete disagreement. A kappa value of 0.4 or greater is often considered to be indicative of a model with useful predictive power [5, 6, 17]. As shown in Table 3, all classifiers resulted in satisfactory Kappa values, except that for the validation set of the rat model which is slightly below. Sensitivity indicates the percentage of stable compounds correctly predicted while specificity reflects that of the unstable compounds correctly classified. In general, the models have higher values of specificity than sensitivity. These models are therefore somewhat biased in over-predicting unstable compounds.

Estimation of prediction confidence

For every prediction, it is vital to provide a user of the model an estimate of the confidence of that prediction. We used the calculated Bayesian probability scores to estimate this confidence. For each model, we analyzed the prediction accuracy for each range of Bayesian score for the test sets. Figure 1 shows that percentage of compounds correctly classified for each bin of Bayesian score for the models. As expected, the higher absolute value of the score, the higher percentage of compounds were correctly classified. For example, for the rat model, we assigned 'High' confidence level (overall 92% accuracy) for compounds with score of <−10 or >16, 'Low' confidence for a score ranging from −5 to 12, and 'Medium' confidence for the rest.

The predicted confidence is an indication of the applicability of the models in prospective design. For example, the microsomal stabilities of 228 Wyeth research compounds from a CNS project were predicted with the rat model (Data not shown). 221 compounds were predicted with High or Medium confidence. In this case, the experimental testing confirmed that 202 compounds were correctly predicted, with an accuracy of 91%.

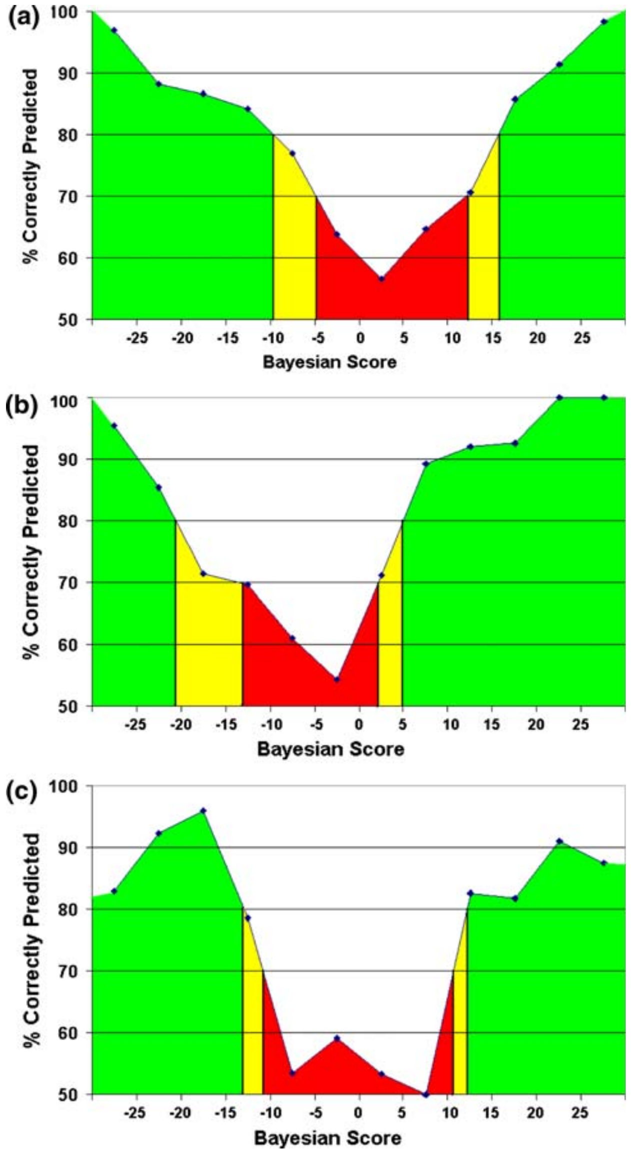

Fig. 1 Histogram representing the prediction accuracy of Bayesian scores in the testing sets of **a** rat, **b** human, **c** mouse. The score ranges for High, Medium and Low confidence predictions were colored as *green*, *yellow* and *red*, respectively

Chronological drift of prediction power

In order to understand the chronological drift of the effectiveness of the models, we randomly collected the rat liver microsomal stability data of 1,000 compounds tested for each 3-month period immediately after the rat model was built for 1 year. Thus, we obtained four datasets each consisting of 1,000 compounds. The trends of concordance, Kappa value, sensitivity and specificity are showed in Fig. 2. The statistical parameters for the datasets of the first two quarters are very similar to each other and are almost identical to those of the rat validation set shown in Table 3. However, the sensitivity values of the third and fourth quarter significantly dropped, although the specificity values did not change much. As a result, the concordance and Kappa values significantly decreased. As shown, the rat model lost effectiveness for the datasets of the third and fourth quarters. To better understand this result, a similarity analysis comparing the compounds of each quarter to the training set compounds was performed, using the FCFP_6 fingerprint. When the accuracy result was compared to the similarity analysis result (Fig. 3), a trend was observed. Overall, the similarities of the first and second quarter compounds with the training set are higher than those of the third and fourth quarter compounds. Models with kappa of ~0.4 were obtained with datasets in which 50% of the compounds had >60% similarity to some compound in the training set. This result suggests that these models are not "global", despite being developed using large datasets. The overall prediction power is related to the similarity between the compounds to be predicted and the training sets. Based on this analysis, we have chosen to update the model every quarter, which appears sufficient given our current discovery project evolution at Wyeth.

In order to understand how the performance of the model changes for the groups with High-, Medium- and Low-confidence predictions, we calculated the kappa values for the prediction groups of each validation set. Figure 4 (a) show the trend of the kappa values. Clearly, the chronological drift of prediction power is consistently observed for all of the High-, Medium- and Low-confidence prediction groups. In addition, the sample distribution of each group was analyzed. As shown at Fig. 4b, no obvious sample redistribution was observed among these three groups. Therefore, the lost of the prediction power for the third and fourth quarter validation sets is not because they have more compounds shifted from a higher confidence group to a lower confidence one.

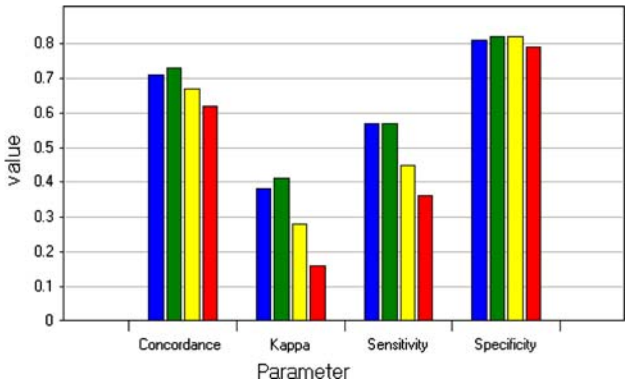

Fig. 2 Histogram of statistical parameters of four rat datasets collected for 1 year after the rat model was built. *Blue*, *green*, *yellow* and *red* are for the first, second, third and fourth quarters, respectively

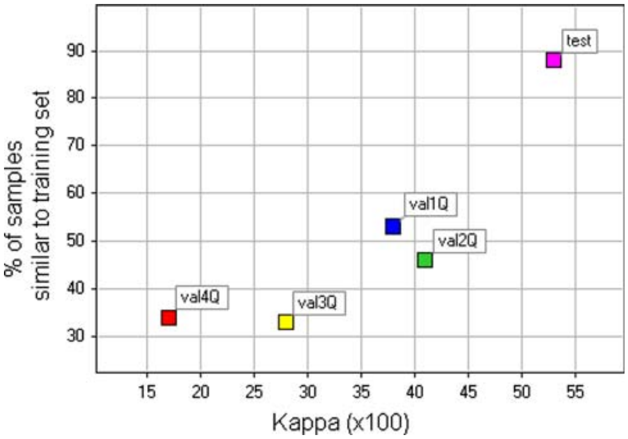

Fig. 3 The percentage of each validation dataset >60% similar to any compound in the training set, versus kappa. Models with kappa of ~0.4 were obtained with datasets in which 50% of the compounds had >60% similarity to at least one compound in the training set. Values for kappa are shown at ×100

Fragment mining

An advantage of using the Naïve Bayesian classifier based on the FCFP_6 fingerprints is that it identifies fingerprint features frequently found in the groups of microsomal stable or unstable compounds in the training set. The identified FCFP_6 fingerprint features can be then translated into 2D substructures within the Pipeline Pilot work frame. To maximize the volume of the information of the training set, we combined the training, test and validation sets for each species and conducted another round of Bayesian modeling, using the combined dataset. The statistical parameters of the resulting models are listed in Table 5. For example, the rat model was built using 24,081 compounds in the whole training set, which contained a total of 48,186 features of FCFP_6. The accuracy of the model for the training set was excellent, as indicated by the Kappa value of 0.61. Since the FCFP_6 fingerprints were ranked by normalized Bayesian probability, it was straightforward to identify the top fingerprints good or bad for microsomal stability for the compounds in the training set. However, this interpretation is totally dependent on the training set. As training sets are not perfectly diverse, one needs to be very cautious in drawing any conclusion from analysis of the statistics table to avoid over-interpretation. For example, all 54 compounds in our training set with dibenzothiophene substructure were stable for rat microsomes. However, all these compounds were from a single

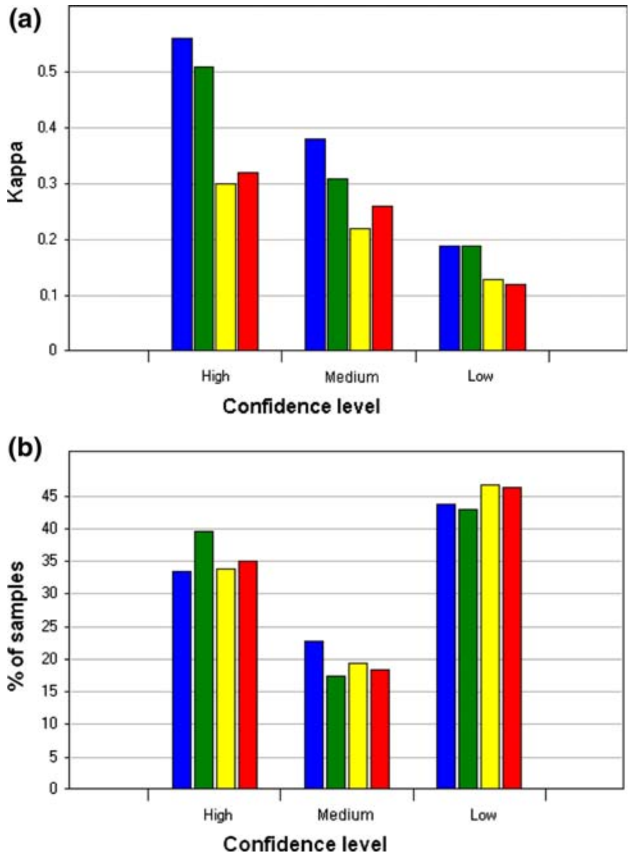

Fig. 4 Histograms representing **a** Kappa values and **b** compound distributions of the High-, Medium- and Low-confidence predictions for the four rat validation data sets. *Blue*, *green*, *yellow* and *red* are for the first, second, third and fourth quarters, respectively

Table 5 Statistical parameters of training sets for each model used to derive fragment information

	Rat	Human	Mouse
# of Compounds	24,081	6,426	2,454
# of FCFP_6	48,186	17,631	9,292
Concordance (%)	82	80	80
Kappa	0.61	0.6	0.58
Sensitivity (%)	79	76	76
Specificity (%)	83	85	82

chemical series and a vast majority of them were close analogues (Data not shown). Therefore, there was a possibility that these 54 similar compounds lacked a CYP liability because of a common scaffold. Therefore we decided not to rank the substructures within stable or unstable compounds using simple frequency analysis when the training set is not diverse.

To overcome this bias, we developed a novel method to mine good and bad substructures. Instead of considering the FCFP_6 fingerprint features most frequently observed using all stable and unstable compounds in the training set, we applied first a diversity filter to the compounds. Using an atom-pair similarity of less than 0.7 as the cutoff, we ensured that the compounds used to select substructures were structurally different from each other [15]. The Bayesian probability scores were then recalculated for each diverse subset and the FCFP_6 fingerprint features were re-ranked based on these adjusted scores. Based on the statistical significance determined by the adjusted Bayesian scores, the good and bad FCFP_6 features for each species were selected as following. First, we identified the score of a diverse subset containing five compounds where all five were classified as stable. Then, all FCFP_6 features with their Bayesian scores better than or equal to this score were considered as statistically significant good features. A similar approach was used to select the bad features. As a result, the good and bad FCFP_6 fingerprint features could be identified with a boosted confidence since this result is now based on diverse compound subsets and their Bayesian scores. Finally, the identified good and bad FCFP_6 fingerprint features were translated into 2D substructures within the Pipeline Pilot work frame. This analysis resulted in identifying 118, 53 and 12 good substructures for rat, human and mouse, respectively. Similarly, 720, 82 and 21 bad substructure features were chosen for rat, human and mouse, respectively. These structural features can be highlighted in molecules as to provide additional guidance to interpret what might lead to good features or to help avoid bad features, in order to improve stability of a compound.

Figures 5 and 6 show 10 examples of the top good and bad structural features for the stability of compounds in rat, human and mouse microsomes. By comparing the figures, it can be seen for example that substructures containing charged fragments are good for stability in microsomes of all three species. Since compounds with charged fragments usually have lower cLogP, this is consistent with our observation that the percentage of stable compounds increases with lower cLogP (Data not reported). It is interesting to note that similar structural features such as amine containing fragments observed by us were also identified by Pil H. Lee et al. [6] from their Bayesian human microsomal stability model. It is interesting that these two independent studies identified some similar substructures.

In order to gain a better understanding of why bad features cause the instability of compounds, we did substructure searches using the bad structural features for human liver microsomes. The metabolic soft spots of the resulting compounds were predicted with MetaSite [12, 22]. Although this method is not designed to predict the relative rate of metabolism of a compound, the identification of the metabolic soft spots could serve as a useful guide for designing

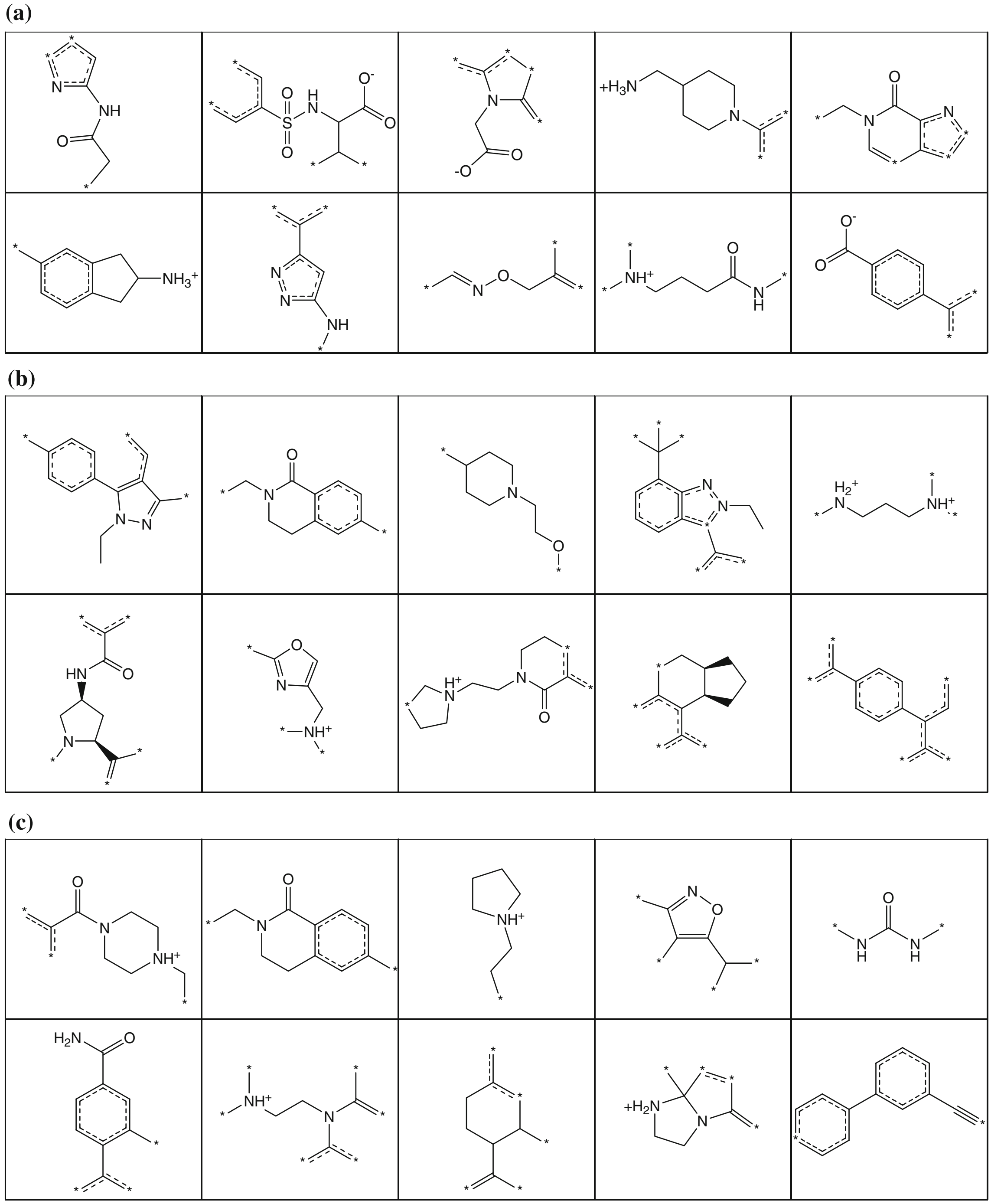

Fig. 5 Examples of top substructures good for stability in microsomes. **a** rat, **b** human, **c** mouse

more stable scaffolds. Cruciani et al. [12, 18–20] have described success rate of 84% in predicting the correct site of metabolism when its top three predictions are considered. Thus, we visualized the first three ranked predicted soft spots for each of the unstable compounds and identified some substructures that always contained predicted soft spots. Figure 7a shows the predicted primary sites of metabolism, located within the substructures identified as the bad features for human microsomes. The stability of this set of compounds might be improved by directly blocking the soft spots of metabolism. It was not true that all predicted soft spots were located inside the query substructure. In fact, in some cases, none of the first three ranked predicted was found directly in the substructure, but a structural moiety on the substructure was identified as a major contributor for the CYP450 recognition using the structure contribution feature

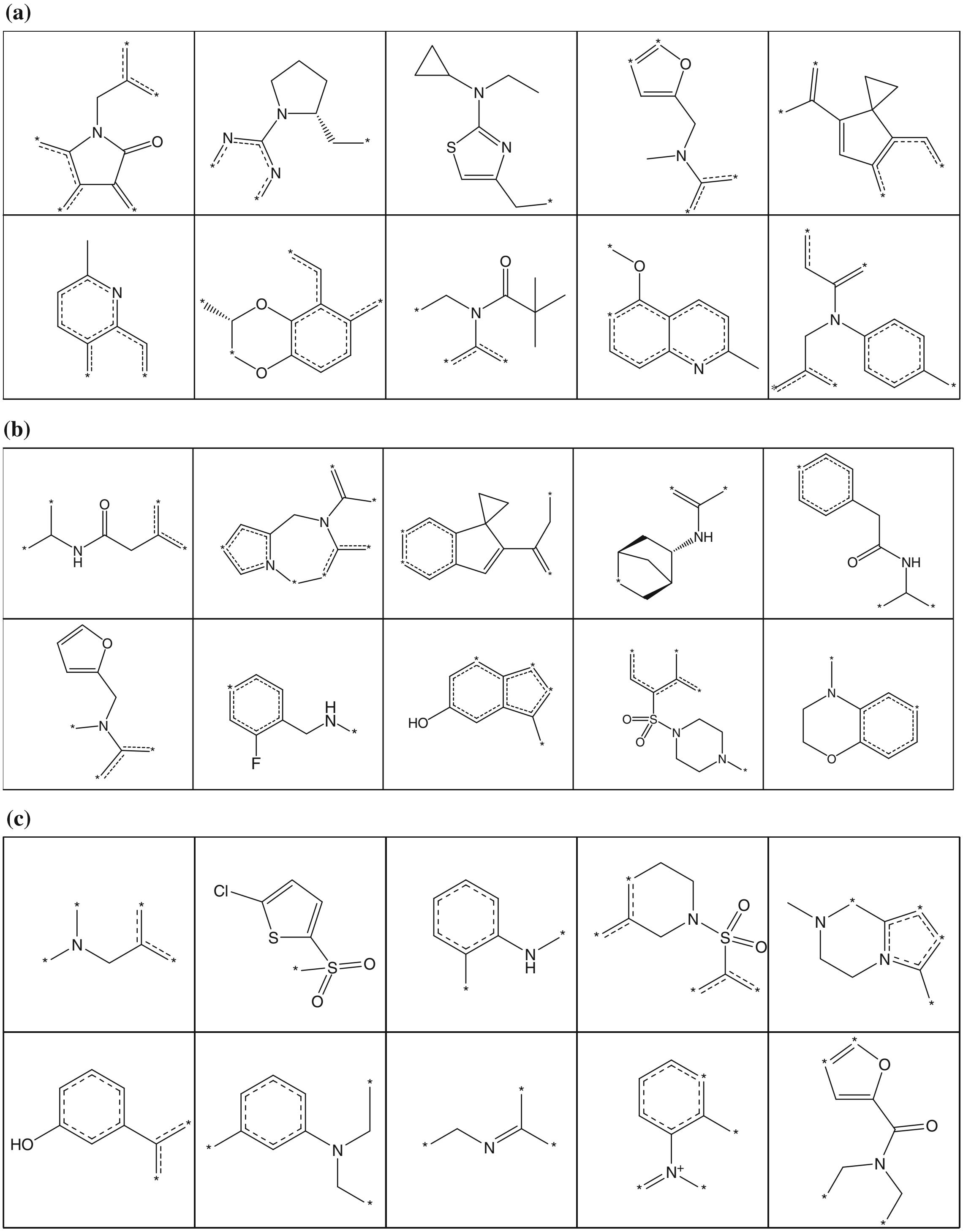

Fig. 6 Examples of top substructures bad for stability in liver microsomes. **a** rat, **b** human, **c** mouse

from MetaSite (Fig. 7b). Identifying this type of situation can be very useful, because directly blocking the primary site of metabolism can risk creating an inhibitor of the cytochromes. Moreover, it can negatively affect the activity or selectivity of the compound towards its therapeutic target. Modifying a contributing moiety that most influences the site of metabolism can bypass both of these potential problems.

Since MetaSite is based on only human CYP450-mediated reactions in phase I metabolism and this calculation has not been expanded to other species we did not apply this analysis to our rat and mouse data [12].

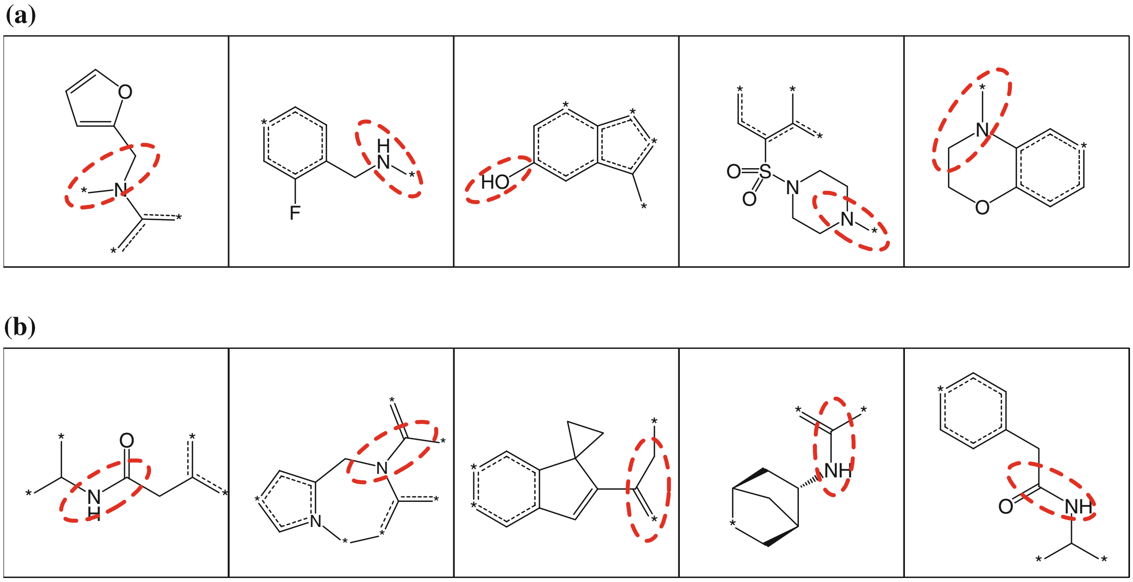

Fig. 7 MetaSite prediction. **a** Structure moieties, as indicted by *red dash ellipse*, contain predicted metabolic sites for human CYP450. **b** Structure moieties, as indicted by *red dash ellipse*, do not contain predicted metabolic sites for human CYP450, but are major contributors to CYP450 recognition as identified by MetaSite

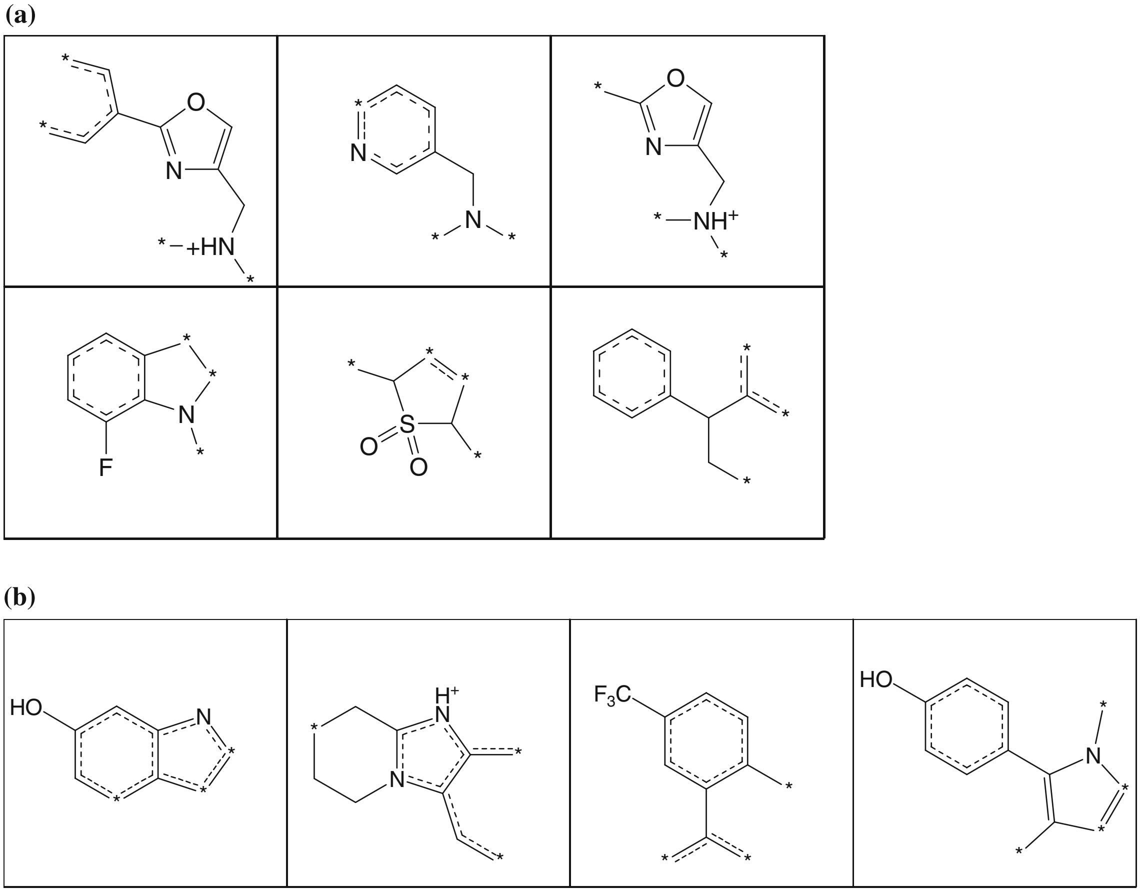

Fig. 8 **a** Structural features bad for rat microsomal stability but good for human microsomal stability. **b** Structural features good for rat microsomal stability but bad for human microsomal stability

Since we built models for three species, it was natural to extend this study to identify structural features which are good for one species but bad for another. It is known that different species have different CYP450 enzymes. For this reason, we often discover that some compounds are stable for one species but unstable for another. This presents a liability for these compounds advancing in our discovery programs. It would be of interest to find substructures that are contributors to this discrepancy, so we could avoid these structural features and optimize the use of our resources. As shown in Table 5, 24,081 and 6,426 compounds were used to develop the rat and human microsomal stability models. The size of the datasets was big enough to allow us to mine those substructures. Thus, the top 1,000 good FCFP_6 features for the stability in human and the top 1,000 bad ones for stability in rat were compared, after the diversity filter of 0.7 (atom-pair) was applied. The common features of these two sets of FCFP_6 features were saved. Figure 8a shows the 6 resulting statistically significant features that are bad for rat but good for human. Compounds with these 6 structural features increase the possibility that they bind to the rat specific CYP450 isoforms, although they are less likely to bind to the common CYP450 enzymes and the human specific isoforms. Similarly, the common features between the top 1,000 good FCPF_6 features for rat liver microsomes and the top 1,000 bad ones for human liver microsomes were identified. Figure 8b shows the four statistically meaningful structure features that are good for microsomal stability in rat liver, but bad in human liver. We suggest that all these 10 structural features should be avoided in order to have a higher likelihood of developing a stable compound for both species.

The same exercise was performed for the mouse data compared to human and rat data but no statistically meaningful structural features were found. We believe the main reason is that the size of the mouse data is too small to have a statistically meaningful result for this study.

Conclusion

We have developed three predictive classification models of stability of compound in rat, human and mouse liver microsomes using large in-house datasets. These models were built with FCFP_6 fingerprints using a Naïve Bayesian classifier and confidence level of the prediction could be assigned with Bayesian scores. These models are currently being used by discovery projects. To assist in interpretation model prediction, we developed a novel data mining strategy to identify structural features associated with good and bad microsomal stability. We also examined whether this approach could be used to identify structural features which are good for one species but bad for another. We reported examples of these structure features which can be used to better design stable compounds in liver microsomes.

Acknowledgments The authors thank Dr. Gary Walker for assistance with data retrieving. We thank Drs. Kristi Fan, Natasja Brooijmans, and Jason Cross for discussions on experimental design. Also we would like to mention our appreciation to Dr. David Diller for his help on preparing this manuscript. We thank Drs. Tarek Mansour and Will Somers for their support of this work.

References

1. Kola I, Landis J (2004) Can the pharmaceutical industry reduce attrition rates? Nat Rev Drug Discov 3(8):711–715
2. Masimirembwa CM, Thompson R, Andersson TB (2001) In vitro high throughput screening of compounds for favorable metabolic properties in drug discovery. Comb Chem High Throughput Screen 4(3):245–263
3. Di L et al (2005) Development and application of high throughput plasma stability assay for drug discovery. Int J Pharm 297(1–2):110–119
4. Di L et al (2003) Optimization of a higher throughput microsomal stability screening assay for profiling drug discovery candidates. J Biomol Screen 8(4):453–462
5. Sakiyama Y et al (2008) Predicting human liver microsomal stability with machine learning techniques. J Mol Graph Model 26(6):907–915
6. Lee PH et al (2007) Development of in silico models for human liver microsomal stability. J Comput Aided Mol Des 21(12):665–673
7. Shen M et al (2003) Development and validation of k-nearest-neighbor QSPR models of metabolic stability of drug candidates. J Med Chem 46(14):3013–3020
8. Chang C et al (2009) The development and validation of a computational model to predict rat liver microsomal clearance. J Pharm Sci 98(8):2857–2867
9. Fox T, Kriegl JM (2006) Machine learning techniques for in silico modeling of drug metabolism. Curr Top Med Chem 6(15):1579–1591
10. Martignoni M, Groothuis GM, de Kanter R (2006) Species differences between mouse, rat, dog, monkey and human CYP-mediated drug metabolism, inhibition and induction. Expert Opin Drug Metab Toxicol 2(6):875–894
11. Sun H (2006) An accurate and interpretable bayesian classification model for prediction of HERG liability. ChemMedChem 1(3):315–322
12. Cruciani G et al (2005) MetaSite: understanding metabolism in human cytochromes from the perspective of the chemist. J Med Chem 48(22):6970–6979
13. Di L et al (2006) High throughput microsomal stability assay for insoluble compounds. Int J Pharm 317(1):54–60
14. Di L et al (2008) Applications of high throughput microsomal stability assay in drug discovery. Comb Chem High Throughput Screen 11(6):469–476
15. Carhart RE, Smith DE, Venkataraghavan R (1985) Atom pairs as molecular features in structure-activity studies: definition and applications. J Chem Inf Comput Sci 25(2):64–73
16. www.accelrys.com. Accessed Mar 2009
17. Dunn G, Everitt B (1995) Clinical biostatistics: an introduction to evidence-based medicine. A Hodder Arnold Publication, London
18. Boyer D et al (2009) Utility of MetaSite in improving metabolic stability of the neutral indomethacin amide derivative and selective cyclooxygenase-2 inhibitor 2-(1-(4-chlorobenzoyl)-5-methoxy-

2-methyl-1H-indol-3-yl)-N-phenethyl-aceta mide. Drug Metab Dispos 37(5):999–1008

19. Caron G, Ermondi G, Testa B (2007) Predicting the oxidative metabolism of statins: an application of the MetaSite algorithm. Pharm Res 24(3):480–501
20. Trunzer M, Faller B, Zimmerlin A (2009) Metabolic soft spot identification and compound optimization in early discovery phases using MetaSite and LC-MS/MS validation. J Med Chem 52(2):329–335
21. Xia X et al (2004) Classification of kinase inhibitors using a Bayesian model. J Med Chem 47(18):4463–4470
22. http://www.moldiscovery.com. Accessed Mar 2009