J Comput Aided Mol Des (2008) 22:385–392
DOI 10.1007/s10822-008-9176-9

Evaluation of descriptors and classification schemes to predict cytochrome substrates in terms of chemical information

John H. Block · Douglas R. Henry

Received: 10 October 2007 / Accepted: 9 January 2008 / Published online: 23 January 2008

Abstract Using a small database of defined substrates in humans for cytochrome P450 mixed function oxidases, a series of descriptors and classification methods were evaluated with respect to how well they correctly classified substrates. The descriptors ranged from structural keys to topological to electronic. A variety of classification schemes were examined in terms of their ability to point out which descriptors are important for predicting the cytochrome P450 specificity for a substrate. Results illustrate the relative effectiveness of the various kinds of descriptors and classification methods, as well as the value of using as well-defined data set as possible.

Keywords Cytochrome substrates · Data mining · E-state indices · Simple decision tree · K-nearest neighbor (K-NN) · Logistic regression · Molecular properties · Naïve-Bayes · Ripper · Structural keys · Support vector machine (SVM) · Topological indices

Introduction

Predicting a drug's fate has been necessary and a challenge in the drug development process. It is obvious that a biologically active xenobiotic must leave the patient's body within a reasonable time following administration. Knowing its metabolic fate is important for at least three reasons: (1) the metabolites might be beneficially active or toxic; (2) the drug's metabolism might be influenced by other xenobiotics taken by the patient (drug–drug interaction); and (3) the patient's genetic make-up might affect the rate of the xenobiotic's metabolism. Approaches to predicting a drug's metabolic fate are varied. As will be seen in the following very brief overview of the prior literature, the amount of chemical information provided by the predictors or descriptors vary. Some descriptors are useful for the medicinal chemist and others, while being good predictors, provide little information that the medicinal chemist can use when altering the molecule.

Prior to easy access to computers and a variety of software, classical SAR using homologous series of ethers showed a relationship between rate of O-dealkylation to the length and type of the ether side chain [1]. Rules began to be developed that would match the chemical characteristics of the substrate with specific cytochromes. These included the degree of steric hindrance where the cytochrome iron–oxygen complex carries out the oxidation on the substrate and the actual ease of electron or hydrogen abstraction from the substrate's various carbons or heteroatoms [2]. An early analysis showed that combinations of descriptors such as molecular dimensions, surface area, ΔE and dipole moment could help predict cytochrome substrate specificity. However, the authors used the information to discuss the interactions at the cytochrome active site and did not report on the accuracy of specific classification methods using these descriptors [3]. An early approach to determining structural requirements of cytochrome substrates was based on the different amino acid sequences of each cytochrome group and the enzymes' X-ray crystallographic

Electronic supplementary material The online version of this article (doi:10.1007/s10822-008-9176-9) contains supplementary material, which is available to authorized users.

J. H. Block (✉)
Department of Pharmaceutical Sciences, College of Pharmacy, Oregon State University, Corvallis, OR 97330, USA
e-mail: John.Block@oregonstate.edu; blockj@onid.orst.edu

D. R. Henry
BIOSAR Research, San Leandro, CA 94577, USA

structures [4, 5]. While the authors provided rules for matching substrate with the specific cytochome, they did not indicate the accuracy of these rules using classification schemes. As computing power reached the desktop, pharmacophore modeling including 3D- and 4D-QSAR have been used to focus on the specific interactions at the cytochrome active site [6]. Results from the probes indicated that the significant pharmacophore factors were hydrogen bond donor, hydrogen bond acceptor and hydrophobe [7]. The 3D-QSAR methodolgy has been used to develop a models of cytochrome active sites [8–12]. An example is the CYP2C9 active site where the 3D-QSAR analysis indicated the importance of charge and steric properties of the substrate [13]. Two pharmacophores were determined for CYP2B6 substrates, both of which include two hydrophobic regions and one hydrogen bond acceptor [14]. A 3D-QSAR analysis of CYP2D6 substrates indicated that the substrates should have two areas of positive (I+) charge, one near the site of metabolism. Between the two I+ regions is an area of negative (I−) charge. Interestingly, this study indicated that neither pKa nor lipophilicity were important for binding [15]. The latter observation is in contrast to a report showing compound lipophilicity being of key importance to CYP binding affinity and enzyme selectivity [16].

Paralleling this increase in computing power and the increasing number of chemical descriptors, there have been several packaged software packages for predicting xenobiotic metabolism [17–20]. In addition to predicting substrate and specific cytochrome, predicting the site where the transformation occurs on the xenobiotic has been a goal [21]. The regioselectivity model is a QSAR-based approach where the focus is on the atom where oxidation might occur and the influence of its immediate neighbors and was evaluated on substrates for CYP2D6, CYP2C9 and CYP3A4 [22].

Finally, because many drug–drug interactions are based on drug A inhibiting the metabolism of drug B, being able to identify the structural properties of a cytochrome inhibitor would be very useful. Using methodologies similar to that used for predicting cytochrome substrate specificity, studies predicting inhibition of specific cytochromes have been published. A series of topological descriptors calculated from the molecules' 2D structures were able to predict which compounds would inhibit CYP2D6 with 75–80 percent accuracy [23]. A 3D-QSAR analysis of CYP1A2 inhibitors showed that both steric and electrostatic interactions were the most important properties whereas as a series of 2D statistical methods showed lipophilicity, aromaticity, charge and the HOMO/LUMO energies as important descriptors [24, 25]. Using 3D-QSAR, the LUMO aspect of CYP2A inhibitors was important, possibly because of π–π stacking between the inhibitors and aromatic acid residues in the binding cavity [26]. Just as the CYP3A4 substrates show wide structural variation, the same is observed for inhibitors of this isozyme. One approach that gave fair results was invoking the multiple pharmacophore hypthesis method [27]. Using recursive partitioning and fairly standard descriptors (density, SlogP, number of double/triple/rotatable bonds, number of halogen atoms, number of H-bonds donor/acceptor, etc), 80 percent accuracy was obtained for CYP2D6 and CYP1A2 inhibitors.[28] Using K-NN and descriptors generated from structural fragments 82 percent of the CYP2D6 and 88 percent of the CYP3A4 inhibitors were correctly classified [29]. This paper also advocated the use of the κ-index a measure of agreement in evaluating classification methods. In contrast with the "QSAR" and data mining methods, a study using dose–response curves showed a clear relationship between polar surface area and charge for CYP2D6 inhibitors but no clear trend between activity and physicochemical properties for CYP2C9 inhibitors [30].

Outline of this study

Initially, we wanted to evaluate the ability of the 166 bit and 324 bit MDL structural keysets to classify xenobiotics according to which cytochrome mixed function oxidase would oxidize them. We started with the Elsevier-MDL Metabolite (2003.1) database, but could not develop queries that would restrict the metabolism to intact humans. Results obtained from human microsomal and cell culture preparations also were included when specifying "human" in the search queries. We also were not able to eliminate xenobiotics that were metabolized by more than one cytochrome. In other words, classifying CYP 1A2 versus 2C9 would include those xenobiotics metabolized by both isozymes.

This led us to focus on a more defined database of drugs. We selected the 2006 Cytochrome P450 Drug Interactions table published by Pharmacist's Letter [31]. This list is based on one publicly available on the Web [32]. The list of drugs and their substrate classification by cytochrome is located in the supplemental material. Nevertheless, there is little indication of the degree to which a particular drug is metabolized by a mixed function oxidase. The substrate classification tables denote those substrates that are involved in a drug interaction of clinical relevance and/or associated with strong drug interaction warnings. Those drugs are indicated in the column titled "Clinically Significant". In addition, we included drugs that are not substrates for mixed function oxidases. In Table 1 these are in the column titled "Not a Substrate". Their route of metabolism was obtained from the drug's package inserts. In other words, xenobiotics can be metabolized by

Table 1 Principal components analyses on correlation matrices

	Molecular connectivity	E-state indices	Spartan descriptors	166 structural keys	324 structural keys
No. of descriptors	79	137	14	166	324
No. of eigenvalues required for 95% of variance	10	59	7	74	115
No. of eigenvalues required for 99% of variance	19	78	9	104	174
X of descriptors with eigenvalues greater than Y (=1/Z)	X = 8	X = 100	X = 9	X = 137	X = 262
	Y = 0.013	Y = 0.0072	Y = 0.071	Y = 0.006	Y = 0.003
	Z = 1/79	Z = 1/137	Z = 1/14	Z = 1/166	Z = 1/324
Largest eigenvalue	43.3	16.83	4.67	16.31	29.19

non-cytochrome routes. Thus, it should not be concluded that the drugs in the “Not a Substrate” column are metabolically inert.

Experimental

We expanded the descriptors from our previous study used to describe a molecule. These included Kier and Hall χ, χ^v and the newer S topological indices calculated using MOLCONN-Z (Hall Associates). Chemical properties estimated using Advanced Chemistry Development Software (8.14) from ACD Labs were obtained from Chemical Abstracts [33] and calculated by Spartan ‘06’ (Wavefunction). For the latter, the structures were minimized using the MMFF force field obtaining the lowest energy conformation in water. The properties calculated by the ACD Labs software were log P, log D (at pH 7), number of freely rotatable bonds, number of hydrogen acceptors, number of hydrogen donors, the sum of hydrogen donors and hydrogen acceptors, molar volume, pKa and polar surface area. The properties calculated by the Wavefunction software were conformational energies, dipole moment, molar area, molar volume and polar surface area. The reason for using molar volume and molar area from both software packages was to see if one software package might provide a more accurate number for purposes of classification. Using the pKa values the percent of the molecule having a I− or I+ charge at pH 7.4 was obtained.

When using large numbers of descriptors, there is the risk of high degree of correlation among descriptors. In lieu of printing the large correlation matrices, we ran principal components analysis on the correlation matrices. The results give an indication of the degree of correlation among the descriptors, which, all things considered, is not particularly high. The measures we report are the number of eigenvalues needed to account for 95 and 99% of the variance; the number of eigenvalues greater than (1/number of descriptors)—for a perfectly uncorrelated data set, the eigenvalues will all be equal to this value; and finally, the largest eigenvalue, which for most of the data sets was around 1/10 the number of descriptors. The results from the principal components analyses are summarized in Table 1.

Several classification methods were tried using WEKA 3.5.6, an open source data mining software package issued under the GNU General Public License [34]. Six different classification schemes were evaluated. Naïve-Bayes, K-Nearest Neighbor (K-NN), Logistic Regression, Support Vector Machine (SVM), Simple Decision Tree and Repeated Incremental Pruning to Produce Error Reduction (Ripper).

In classification problems of the type reported in this paper, better results are usually obtained when class sizes are equal. This applies to both original sample (all the data) and to cross validation (holdout) subsets. Otherwise, there is the risk that the smaller set of data will be classified incorrectly and the larger data set correctly producing an acceptable overall percent correct. When class sizes are disparate, we need to use a sampling technique that will provide holdout samples for training and test sets in which each class is represented in about the same proportion as in the original data. This is crucial in methods that rely on prior probabilities, and it is generally good practice. The process of representing classes properly in samples is called stratification. In Weka, the standard cross validation procedure is a stratified ten-fold cross-validation. The data is divided randomly into ten parts in which each class is represented in approximately the same proportions as in the full dataset. Each part is held out in turn and the learning schemes trained on the remaining nine-tenths; then its error rate is calculated on the holdout set. Thus the learning is executed a total of 10 times on different training sets. The 10 error estimates are averaged to yield an overall error estimate.

There are many ways to evaluate the degree of effectiveness of classification methods. Following a previously published procedure, we used the κ-index. (This should not be confused with the topological κ-shape indices.) Its definition is shown in Eq. 1, and its calculation from a confusion index is described in Fig. 1 [25]. Its validity is best when the ratios between two sets of observations are

		Actual	
		+	−
Predicted	+	a	b
	−	c	d

Fig. 1 A confusion matrix. $\kappa = \frac{(a+d)-[((a+c)(a+b)+(b+d)(c+d))/N]}{N-[((a+c)(a+b)+(b+d)(c+d))/N]}$ a = true positive; b = false positive; c = true positive; d = true negative

approximately equal [35]. In Table 2, we have indicated with italics those results with a κ-index greater than 0.6 on the test sets.

$$\kappa = \frac{\text{observed agreement} - \text{chance agreement}}{\text{total observed - chance agreement}} \tag{1}$$

Naïve-Bayes can be considered analogous to linear discriminant analysis in that the descriptors and classes can be thought of as independent and dependent variables, respectively. Bayes' theorem is only exact in the case where the descriptors are statistically independent of one another. This ideal situation is almost never seen in practice, but minimizing redundancies often provides a useful model nonetheless. A drawback to this classification method is that there is no built in weighting or selection of the descriptors most contributing to the differentiation of classes. In its basic use, it does not tell the chemist which descriptors are important for the xenobiotic to be metabolized by a particular mixed function oxidase. Using feature selection procedures, Naïve-Bayes can be used to evaluate descriptors [36].

K-NN is a form of classification based on similarities among the members of a class. Classification is done according to a "vote" among nearest neighbors. It does not provide a set of global or generic rules for predicting the classification of new objects, and it does not provide structural information useful to the chemist. A unique feature of nearest-neighbor methods is that they allow the construction of irregular boundaries between classes. Also, this method has been used to optimize variable selection when developing quantitative structure-property relationships (QPSR) [37].

Logistic regression is replacing 2-way discriminant analysis as a means to carry out a classification. An advantage is that it does not require continuous variables. Instead, a mixture of continuous and categorical variables can be used. Another advantage is that it does show which descriptors are important for carrying out the classification.

SVMs can be considered analogous to quadratic discriminant analysis in that linear models are used to form nonlinear class boundaries. Many times this method will produce the best results, but, because of the large number of variables produced in the transformation into nonlinear space, overfitting can be a possibility. Stratified cross validation can reduce the impact of overfitting. Finally, there is no explicit selection of descriptors.

A decision tree searches for those descriptors that distinguish between classes and then uses a series of yes and no answers to decide if a descriptor can provide a rule for classifying objects into the correct class. Decision tree analysis is most robust with a relatively similar number of objects in each class. It is important to note that the shape of the decision tree can depend on which descriptor is the starting point or establishing a set of rules. All decision tree methods represent a suboptimal solution to an np-complete problem [38]. One approach to dealing with this is the generation of multiple decision trees (the random forest approach). Another is to prune the complete decision tree using parameters. We have examined both approaches, and we found the pruning approach worked best for these data sets. Even with these limitations, only those descriptors useful to carry out the classification are in the decision tree, this method can produce useful information to the chemist. Weka's simple decision tree (J48) is a Java implementation of Ross Quinlan's C4.5 decision tree [39], with some improvements as described in the Weka source code.

The Ripper method of classification first establishes a set of rules and then "prunes" the tree to reduce the number of mis-classifications and the number of rules. Class size differences can determine the default classification. Compared to a decision tree, Ripper produces fewer rules. Like a decision tree, the descriptors that are used to form the rules are clearly identified.

A wide variety of descriptors were evaluated. Because of our experience with the MDL structural keys, we used them again with the Pharmacist's Letter set of cytochrome substrates. A keyset contains information on the atom(s), bond(s), atom type(s), local environments and number of occurrences. There is the 166 bit keyset and the larger 324 bit keyset. The latter was optimized for drug discovery. It was postulated that this keyset might provide descriptors suitable for classifying cytochrome substrates.

The key sets are not strictly independent as required by a Bayesian classification, but it appears that the 324 keyset is more uncorrelated than the 166 keyset. Complicating this observation is that the problems with covariance vary with structures [40]. The 166 keyset contains mostly specific sequences of atoms. For example, key #50 is C = C(C)C. In contrast, the keys in the 324 keyset can be more difficult to interpret. Key #36 is any atom in a 5-membered ring, whereas key #250 is C=N.

Topological descriptors have been widely used in data mining, classification and QSAR studies. There are a wide

Table 2 Comparison of various classification methods to differentiate between sets of cytochrome substrates using different types of descriptors

	Molecular connectivity			E-state connectivity			Chemical properties			166 MDL key set			324 MDL KEY set		
	(Train)	Test	RMSE	(Train)	Test	RMSE	(Train)	Test	RMSE	(Train)	Test	RMSE	(Train)	Test	RMSE
1A2 (30 drugs) versus 2C9 (34 drugs)															
Naïve-Bayes	(71.9)	67.2	0.5702	(89.1)	78.1	0.4588	(93.8)	*84.4*	0.3872	(82.8)	70.3	0.5164	(84.4)	75.0	0.4880
K-nearest neighbor	(100)	73.4	0.5069	(100)	71.9	0.5216	(100)	*85.9*	0.3690	(100)	71.9	0.5254	(100)	*84.4*	0.3692
Logistic regression	(100)	76.0	0.5010	(100)	65.6	0.5863	(96.9)	71.9	0.5285	(100)	71.9	0.5111	(100)	76.6	0.4756
Support vector machine	(84.4)	73.4	0.5154	(93.8)	73.4	0.5154	(92.2)	78.1	0.4677	(100)	78.1	0.4677	(100)	*81.3*	0.4330
Simple decision tree	(95.3)	*81.3*	0.4336	(96.9)	76.6	0.4787	(95.3)	*81.2*	0.4150	(98.4)	78.1	0.4483	(98.4)	75.0	0.4822
Ripper	(84.4)	71.8	0.4833	(89.1)	79.7	0.4209	(93.8)	67.2	0.4932	(85.9)	70.3	0.4570	(90.6)	75.0	0.4507
1A2 (21 drugs) versus 2D6 (46 drugs)															
Naïve-Bayes	(59.7)	74.6	0.4958	(91.0)	65.7	0.5642	(86.6)	83.6	0.3971	(82.1)	74.6	0.4970	(86.6)	71.6	0.4674
K-nearest neighbor	(100)	74.6	0.4958	(100)	77.6	0.4660	(100)	77.6	0.4658	(100)	79.1	0.4646	(100)	77.6	0.4544
Logistic regression	(100)	80.6	0.4387	(100)	70.1	0.5451	(91.0)	62.7	0.5906	(100)	73.1	0.5060	(100)	71.6	0.5192
Support vector machine	(82.1)	71.6	0.5325	(94.0)	79.1	0.4571	(86.6)	85.1	0.3863	(100)	74.6	0.5037	(100)	82.1	0.4232
Simple decision tree	(98.5)	74.6	0.4787	(94.0)	68.7	0.4820	(94.0)	68.6	0.5057	(94.0)	62.7	0.5842	(95.5)	70.1	0.5387
Ripper	(85.1)	73.1	0.4782	(94.0)	74.6	0.4510	(94.0)	83.6	0.3730	(85.1)	74.6	0.4761	(92.5)	74.6	0.4829
1A2 (31 drugs) versus 2C19 (23 drugs)															
Naïve-Bayes	(85.2)	61.1	0.6253	(74.1)	61.1	0.6081	(77.8)	75.9	0.4978	(81.5)	70.4	0.5176	(81.5)	66.7	0.5716
K-nearest neighbor	(100)	79.6	0.4428	(100)	55.6	0.6537	(100)	70.4	0.5338	(100)	59.3	0.6187	(100)	70.4	0.5341
Logistic regression	(100)	64.8	0.5842	(100)	53.7	0.6697	(100)	68.5	0.5610	(100)	64.8	0.5835	(100)	74.1	0.5004
Support vector machine	(85.2)	55.6	0.6667	(88.9)	51.9	0.6939	(87.0)	70.4	0.5443	(100)	63.0	0.6086	(100)	72.2	0.5270
Simple decision tree	(96.3)	46.3	0.7160	(98.1)	66.7	0.5333	(83.3)	64.8	0.5382	(94.4)	72.2	0.4787	(96.3)	63.0	0.5776
Ripper	(90.7)	61.1	0.5410	(79.6)	68.5	0.4892	(81.5)	68.5	0.5047	(70.4)	55.6	0.5606	(79.6)	63.0	0.5544
2C9 (35 drugs) versus 2D6 (56 drugs)															
Naïve-Bayes	(72.5)	67.0	0.5410	(91.2)	*83.5*	0.4067	(96.7)	*94.5*	0.2238	(85.7)	80.2	0.4333	(87.9)	*83.5*	0.3917
K-nearest neighbor	(100)	74.7	0.4969	(100)	*86.8*	0.3590	(100)	*94.5*	0.2319	(100)	*91.2*	0.2792	(100)	*86.8*	0.3555
Logistic regression	(100)	*82.4*	0.4211	(100)	80.2	0.4403	(100)	*85.7*	0.3766	(100)	78.0	0.4519	(100)	73.6	0.4808
Support vector machine	(91.2)	80.2	0.4447	(97.8)	*86.8*	0.3631	(94.5)	*93.4*	0.2568	(100)	*86.8*	0.3631	(100)	*84.5*	0.3922
Simple decision tree	(96.7)	70.3	0.5296	(95.6)	*89.0*	0.3196	(94.5)	*90.1*	0.3086	(97.8)	79.1	0.4416	(96.7)	*86.8*	0.3554
Ripper	(92.3)	76.9	0.4340	(91.2)	*82.4*	0.4062	(94.5)	*91.2*	0.2884	(90.1)	78.0	0.4352	(89.0)	*89.0*	0.3137
1A2 + 2C9 + 2C19 (41 drugs) versus 2D6 (57 drugs)															
Naïve-Bayes	(71.4)	61.2	0.6015	(81.6)	71.4	0.5245	(86.7)	*85.7*	0.3646	(79.6)	74.5	0.4855	(77.6)	75.5	0.4896
K-nearest neighbor	(100)	70.4	0.5380	(100)	74.5	0.4998	(100)	*83.7*	0.3997	(100)	77.6	0.4529	(100)	76.5	0.4794
Logistic regression	(100)	68.4	0.5537	(100)	72.4	0.5209	(90.8)	75.5	0.4574	(100)	72.4	0.5184	(100)	*81.6*	0.4131

Table 2 continued

	Molecular connectivity			E-state connectivity			Chemical properties			166 MDL key set			324 MDL KEY set		
	(Train)	Test	RMSE	(Train)	Test	RMSE	(Train)	Test	RMSE	(Train)	Test	RMSE	(Train)	Test	RMSE
Support vector machine	(81.6)	74.5	0.5051	(94.9)	*80.6*	0.4403	(87.8)	*84.7*	0.3912	(100)	77.6	0.4738	(100)	*80.6*	0.4403
Simple decision tree	(95.9)	61.2	0.6032	(95.9)	66.3	0.5654	(88.8)	79.6	0.4261	(95.9)	75.5	0.4824	(96.9)	*80.6*	0.4271
Ripper	(72.4)	67.3	0.4885	(93.9)	73.5	0.4556	(93.9)	77.6	0.4188	(86.7)	72.4	0.4606	(90.8)	78.6	0.4281
1A2 + 2C9 + 2C19 + 2C8 + 2D6 (107 drugs) versus 3A4 (117 drugs)															
Naive-Bayes	(68.3)	66.7	0.5724	(72.8)	69.6	0.5421	(74.1)	74.1	0.4611	(75.0)	70.5	0.5188	(73.7)	71.4	0.5300
K-nearest neighbor	(99.6)	71.0	0.5361	(99.6)	69.6	0.5483	(99.6)	69.6	0.5483	(99.6)	71.0	0.5278	(99.6)	68.8	0.5472
Logistic Regression	(82.6)	61.2	0.5557	(89.7)	63.3	0.5847	(81.7)	76.3	0.4238	(99.6)	59.4	0.6350	(99.6)	62.1	0.6081
Support vector machine	(77.7)	71.0	0.5387	(83.9)	72.7	0.5218	(78.1)	73.2	0.5175	(92.9)	63.8	0.6013	(99.6)	66.1	0.5825
Simple decision tree	(97.8)	71.0	0.5053	(95.1)	67.9	0.5400	(84.8)	72.7	0.4776	(91.5)	62.1	0.5809	(94.2)	64.7	0.5746
Ripper	(84.8)	69.2	0.4797	(91.5)	71.4	0.4757	(80.4)	73.7	0.4508	(67.4)	65.2	0.4869	(79.0)	71.0	0.4633

All drugs in each comparison are in the training set. The test set is based on a 10-fold stratified cross-validation. The number in each column is the percent correct for that pair of substrates. RMSE is the root mean squared error. An *italicized percent* indicates a κ-index of 0.60 or greater [25]

variety of this class of descriptors. While excellent results can be obtained using topological descriptors, criticism has focused on how to interpret them and translate the results to the "bench" in designing new compounds. For this study, we used topological descriptors that can be calculated using low-cost software that runs on the desktop. Thus, we used the original Kier and Hall chi, valence chi (χ^v), kappa shape (κ) and E-state (S) indices [41–44]. In selecting the specific descriptors for the classification analysis, each descriptor had to have a value for at least a third of the molecules. In other words, descriptors specific for a guanidium, bromide or other moieties that were part of only a few molecules were omitted.

The drug list in the supplemental table consists of 247 xenobiotics metabolized by one or more cytochrome isozymes and 53 that are not substrates for this enzyme system. Because many of these xenobiotics are metabolized by more than one cytochrome isozyme, the totals will exceed 247. Duplicate xenobiotics were removed before running a classification. The 100 "clinically significant" drugs were used to see if it were possible to distinguish non-cytochrome substrates from substrates.

Results and discussion

We did try a cluster analysis. Weka provides the capability to compare cluster analysis results to known class memberships (a cluster is "assigned" the class to which a majority of its observations belong). We tested this using the K-means cluster technique with a variety of the descriptor sets and activity classes. In the majority of cases, the overall percent correct in the class-to-cluster comparisons was around 50% and that would correspond to a κ of zero. When a subset of descriptors was selected, corresponding to those chosen during, say, a decision tree analysis, the class-to-cluster results generally improved. They did not generally approach, and they never exceeded the best classification results. We conclude that having a large number of irrelevant descriptors in a cluster analysis can give correspondingly poor classification results.

We also ran Kohonen self-organizing map (SOM) analysis using the JavaSOM package ver. 1.2 [45]. Analyses were run using the 166 and 324-key data sets, coding the compounds by either the substrate or the inhibitor tendency. Thus, if a compound is a substrate for three of the Cyp enzymes, its label in the SOM would show this. The SOM maps, created using a 30 × 30 2D hexagonal geometry, were examined to see if compounds of similar substrate or inhibitor properties tended to cluster together. Some clustering was evident, but no tight clusters were found in the experiments we ran, and we concluded that this approach was only qualitatively useful, especially when all the descriptors are used.

Table 2 contains the results from a variety of substrate pairings and combinations. In general, if the number of drugs in one class of substrates exceeded another by 2:1, the analysis was not run. Thus, there are no results reported for CYP 3A4 versus CYP 1A2, 2C9, 2C19, 2D6 or 2C8. Combinations of these substrates versus 3A4 were run. It is obvious from an examination of Table 2 that good results are dependant on the substrate class and classification method. In general, the chemical properties obtained from the ACD and Wavefunction calculations produced the best overall results.

For CYP 1A2 versus 2C9, the simple decision tree indicated that the percent I− charge, Wavefunction calculated polar surface area and ACD calculated Log P were important properties for differentiating 2C9 substrates from 1A2 substrates. Using different descriptors, the simple decision tree indicated that the simple three atom cluster, number of hydrogen-bonds, count of oxygen atoms, simple $^{6}\chi^{p}$ and site of nucleophilic attack were important for differentiating 1A2 substrates from 2C9.

Classification of 1A2 versus 2D6 illustrates how the results are dependent on which substrate pair is being compared and the importance of trying to keep the number of drugs in each group approximately equal. The fact that 69 percent of the substrates being compared were 2D6 explains why the overall results were better for this group than the 1A2 substrates. Nevertheless, note that this time the simple decision tree gave inferior results relative to the SVM and Ripper. The latter indicated that the calculated percent I+ charge, dipole moment and log P were important discriminators in the chemical properties group. The only simple decision tree descriptor in common with the SVM was the calculated percent I+ charge. The latter method indicated that the pKa used to calculate the I+ charge also was an important discriminator.

CYP 1A2 versus 2C19 were poorly classified. The best results for single pairs was CYP 2C9 substrates versus 2D6 with chemical properties having the significant descriptors. This may not be surprising considering that 2C9 and 2D6 each showed better result when compared with 1A2. It is interesting to note which descriptors are picked from the different groupings by those classification methods with higher κ-indices. For logistic regression (κ = 0.6286), the important molecular connectivity indices included number of non-hydrogen elements, $^{4}\chi^{p}$, $^{6}\chi^{p}$, $^{3}\chi^{vc}$, $^{5}\chi^{vc}$, and $^{6}\chi^{vc}$. It should be kept in mind that there is a higher degree of intercorrelation among connectivity terms relative to the other descriptors used in this study. In contrast, the significant chemical properties identified by logistic regression were log D and number of hydrogen donors. The significant descriptors for both 2C9 and 2D6 substrates used by the SVM were the presence of a I+ charge, the percent I+ charge at pH 7 and the pKa determining that charge, but only the percent I+ charge was used by the simple decision tree and Ripper.

In an attempt to more equalize numbers of substrates between classes, CYPs 1A2, 2C9 and 2C19 were combined into one class and compared against 2D6. According to the SVM, all of the chemical properties contributed to the classification. It must be kept in mind that the SVM can over compensate, and these results should be interpreted with caution. Finally, combining all but CYP 3A4 substrates into one class and comparing it against 3A4 produced poor results overall.

Conclusion

Data mining xenobiotic databases using a variety of descriptors and classification methods can produce inconsistent results. How the results are presented also exacerbate these inconsistencies. For example, compare the Naïve-Bayes chemical properties results for 1A2 versus 2C9 and 1A2 versus 2D6. Within the 1A2–2C9 pair, 83.3 percent of 1A2 and 85.3 percent of 2C9 were correctly classified for an overall 84.4 percent correctly classified ($\kappa = 0.6863$). In contrast, for the 1A2–2D6 pair, only 61.9 percent of 1A2 were correctly classified, but 93.5 percent of 2D6 were correctly classified giving an overall 83.6 percent correctly classified ($\kappa = 0.5921$), nearly the same overall accuracy as the 1A2–2C9 pair. Examining the 1A–2D6 pair using the chemical properties descriptors further, SVM and Ripper show the same unequal percent correct pattern as Naïve-Bayes, while K-NN, logistic regression and simple decision tree produce more consistent percent correct for 1A2 and 2D6. This illustration highlights the fact that results from studies of this type are dependent on the chemical structures in the substrate sets and the classification methods used.

The biggest use of QSAR analyses on large datasets has been for screening prior to more detailed analysis or modeling. The results are highly dependent on the composition of the database and the classification method. QSAR can generally be counted on to work “better” with homologous structures than with diverse ones. Small changes in structure cause predictable changes in activity, and the most precise models are usually obtained with a series of similar structures. Encoding small, relevant structural changes for such series using substructural keys is straightforward. As the collection of structures becomes more diverse, encoding relevant structural changes with substructural keys becomes problematic, since many of the changes may not be relevant to binding to the enzyme. In this case, whole-molecule properties, counts, and min/max descriptors like those generated in Spartan and ACD Labs, although they express more isotropic properties, may relate to bulk phenomena like transport and partitioning, which can also be important in the overall activity. We have "averaged" the activity by using a binary descriptor (rather than using more quantitative measures like Ki or I_{50}) and perhaps "averaged" descriptors are better able to predict this in our case. It should be kept in mind that obtaining false positives is preferable to obtaining false negatives to reduce the risk of missing possible hits. It is possible to "tweak" most of the methods in favor of false positives. Refining these analyses to obtain chemically consistent descriptors likely will require docking into active sites or other alignment methods [46].

References

1. Burke MD, Thompson S, Elcombe CR, Halpert J, Haaparanta T, Mayer RT (1985) Biochem Pharmacol 34:3337
2. Smith DA (1994) Eur J Pharm Sci 2:69
3. Lewis DFV, Eddershaw PJ, Dickins M, Tarbit MH, Goldfarb PS (1998) Chem Biol Interact 115:175
4. De Rienzo F, Fanelli F, Menziani MC, De Benedetti PG (2000) J Comput Aided Mol Des 14:93
5. Kirton SB, Baxter CA, Sutcliffe MJ (2002) Adv Drug Deliv Rev 54:385
6. Ekins S, Waller CL, Swaan PW, Cruciani G, Wrighton SA, Wikel JH (2000) J Pharmacol Toxicol Methods 44:251
7. Ekins S, Bravi G, Binkley S, Gillespie JS Ring BJ, Wikel JH, Wrighton SA (2000) Drug Metab Dispos 28:994
8. Ekins S, de Groot MJ, Jones JP (2001) Drug Metabol Dispos 29:936
9. de Groot MJ, Ekins S (2002) Adv Drug Deliv Rev 54:367
10. Boyer S, Zamora I (2002) J Comput Aided Mol Des 16:403
11. de Groot MJ, Kirton SB, Sutcliffe MJ (2004) Curr Top Med Chem 4:1803
12. de Groot MJ (2006) Drug Discov Today 11:601
13. Rao S, Aoyama R, Schrag M, Trager WF, Rettie A, Jones JP (2000) J Med Chem 43:2789
14. Wang Q, Halpert JR (2002) Drug Metab Dispos 30:86
15. Haji-Momenian S, Rieger JM, Macdonald TL, Brown ML (2003) Bioorg Med Chem 11:5545
16. Lewis DFV, Jacobs MN, Dickins M (2004) Drug Discov Today 9:530
17. Langowski J, Long A (2002) Adv Drug Deliv Rev 54:407
18. Korolev D, Balakin KV, Nikolsky Y, Kirillov E, Ivanenkov YA, Savchuk NP, Ivashchenko AA, Nikolskaya T (2003) J Med Chem 46:3631
19. Borodina Y, Sadym A, Filimonov D, Blinova V, Dmitriev A, Poroikov V (2003) J Chem Inf Comput Sci 43:1636
20. Van de Waterbeemd H, Gifford E (2003) Nat Rev Drug Discov 2:192
21. Zamora I, Afzelius L, Cruciani G (2003) J Med Chem 46:2313
22. Sheridan RP, Korzekwa KR, Torres RA, Walker MJ (2007) J Med Chem 50:3173
23. Susnow RG, Dixon SL (2003) J Chem Inf Comput Sci 43:1308
24. Korhonen LE, Rahnasto M, Mähönen NJ, Wittekindt C, Poso A, Juvonen RO, Raunio H (2005) J Med Chem 48:3808
25. Chohan KK, Paine SW, Mistry J, Barton P, Davis AM (2005) J Med Chem 48:5154
26. Rahnasto, Raunio H, Poso A, Wittekindt C, Juvonen RO (2005) J Med Chem 48:440
27. Mao B, Gozalbes R, Barbosa F, Migeon J, Merrick S, Kamm K, Wong E, Costales C, Shi W, Wu C And Froloff N (2006) J Chem Inf Model 46:2125
28. Burton J, Ijjaali I, Barberan O, Petitet F, Vercauteren DP, Michel A (2006) J Med Chem 49:6231
29. Jensen BF, Vind C, Padkjaer SB, Brockhoff PB, Refsgaard HF (2007) J Med Chem 50:501
30. McMasters DR, Torres RA, Crathern SJ, Dooney DL, Nachbar RB, Sheridan RP, Korzekawa KR (2007) J Med Chem 50:3205
31. Pharmacist's letter, detail-document #220233, 2006. (http://www.pharmacistsletter.com)
32. http://www.medicine.iupui.edu/Flockhart/table.htm
33. The chemical properties calculated by the ACD software were obtained from Chemical Abstracts online using its SciFinder Scholar 2007 interface
34. Witten IH, Frank E (2005) Data mining: practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco. (http://www.cs.waikato.ac.nz/ml/weka)
35. http://ourworld.compuserve.com/homepages/jsuebersax/kappa.htm
36. Bender A, Mussa HY, Glen RC, Reiling S (2004) J Chem Inf Comput Sci 44:170
37. Shen M, Xiao Y, Golbraikh A, Gombar VK, Tropsha A (2003) J Med Chem 46:3013
38. Hyal L, Rivest R (1976) Inform Process Lett 35:15
39. Quinlan JR (1993) C4.5: programs for machine learning. Morgan Kaufmann Publishers, San Mateo, CA
40. Durant JL, Leland BA, Henry DR, Nourse JG (2002) J Chem Inf Comput Sci 42:1273
41. Kier LB, Hall LH (1976) Molecular connectivity in chemistry and drug research. Academic Press, New York
42. Kier LB, Hall LH (1986) Molecular connectivity in structure–activity analysis. Research Studies Press, Letchworth, England
43. Kier LB, Hall LH (1999) Molecular structure description: the electrotopological state. Academic Press, New York
44. Kier LB (1989) Quant Struct Act Relat 8:221
45. http://javasom.sourceforge.net/
46. Fox, T, Kriegl, JM (2007) In: Spellmeyer, DC, Wheeler RA (eds) Annual reports in computational chemistry. Elsevier, New York, 3, pp 63–81