molecular pharmaceutics

ARTICLE

pubs.acs.org/molecularpharmaceutics

ADME Evaluation in Drug Discovery. 10. Predictions of P-Glycoprotein Inhibitors Using Recursive Partitioning and Naive Bayesian Classification Techniques

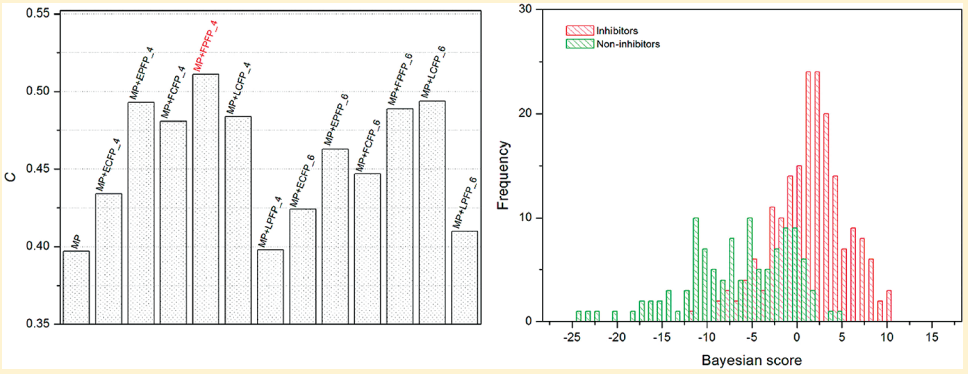
Lei Chen,[†] Youyong Li,[†] Qing Zhao,[‡] Hui Peng,*[,‡] and Tingjun Hou*[,†]

[†]Institute of Functional Nano & Soft Materials (FUNSOM) and Jiangsu Key Laboratory for Carbon-Based Functional Materials & Devices, Soochow University, Suzhou, Jiangsu 215123, China

[‡]Department of Molecular Immunology, Institute of Basic Medical Sciences, Beijing 100850, China

Supporting Information

ABSTRACT:



P-Glycoprotein (P-gp), an efflux transporter, plays a crucial role in drug pharmacokinetic properties (ADME), and is critical for multidrug resistance (MDR) by mediating the active transport of anticancer drugs from the intracellular to the extracellular compartment. Here we reported an original database of 1273 molecules that are categorized into P-gp inhibitors and noninhibitors. The impact of various physicochemical properties on P-gp inhibition was examined. We then built the decision trees from a training set of 973 compounds using the recursive partitioning (RP) technique and validated by an external test set of 300 compounds. The best decision tree correctly predicted 83.5% of the inhibitors and 67.0% of the noninhibitors in the test set. Finally, we applied naive Bayesian categorization modeling to establish classifiers for P-gp inhibitors. The Bayesian classifier gave average correct prediction for 81.7% of 973 compounds in the training set with leave-one-out cross-validation procedure and 81.2% of 300 compounds in the test set. By establishing multiple decision trees and Bayesian classifiers, we evaluated the impact of molecular fingerprints on classification by the prediction accuracy for the test set, and we found that the inclusion of molecular fingerprints improves the prediction obviously. As an unsupervised learner without tuning parameters, the Bayesian classifier employing fingerprints highlights the important structural fragments favorable or unfavorable for P-gp transport, which provides critical information for designing new efficient P-gp inhibitors.

KEYWORDS: P-glycoprotein, naive Bayesian classification, recursive partitioning, ADME/T, ABCB1, ABC transporter, multidrug resistance (MDR)

INTRODUCTION

P-Glycoprotein (MDR1/ABCB1), the product of the mdr-1 gene, is a member of the ATP-binding cassette (ABC) family that is extensively distributed and expressed in many human organs with secretory or barrier functions, such as intestinal epithelium, hepatocytes, renal proximal tubular cells, adrenal gland and endothelial capillaries of the brain comprising the blood−brain barrier (BBB).[1] P-gp functions as an energy-dependent efflux pump that exports environmental xenobiotics out of cells.[2,3] P-gp has broad poly specificity, recognizing hundreds of compounds or drugs as small as 330 up to 4000 Da; therefore, it plays a crucial role in drug bioavailability, metabolism and toxicity.[2,4] Increasing attention has been given to explore the important biological roles of P-gp in pharmacokinetic properties of many clinical therapeutic agents.[5−9]

Received: December 31, 2010
Accepted: March 17, 2011
Revised: March 12, 2011
Published: March 17, 2011

Overexpression of P-gp has been demonstrated to generate resistant phenotypes in cultured cancer cell lines and various tumor models.[10] It is believed that P-gp recognizes and transports a wide range of hydrophobic compounds or drugs, leading to a phenomenon known as multidrug resistance, which is one of the major reasons for the failure of chemotherapeutic treatment of cancers. Therefore, the development of potent P-gp inhibitors may be an efficient way of sensitizing drug resistance in cancer chemotherapy.[11,12]

The transport activity assessment of MDR reversal (MDRR) agents (i.e., P-gp inhibitors) or P-gp substrates can be achieved experimentally through *in vitro* or *in vivo* assays, such as the transepithelial transport assays on confluent cell monolayer accumulation (e.g., monolayer efflux assay), the efflux assays using fluorescent probes (e.g., calcein-AM assay), and the ATPase assays that monitor the ATPase activity of P-gp.[13] However, these experimental assays are expensive, laborious, and time-consuming. They are usually used in the later stages of drug design or optimization when the drug candidates exhibit adequate potency and acceptable pharmacokinetic properties. Thus the development of *in silico* models that provide rapid and efficient screening platform for identifying P-gp inhibitors becomes the critical tool in the early stage of drug design or optimization.

To date, numerous computational models, which were generated by quantitative structure−activity relationship (QSAR) studies and pharmacophore modeling, have been developed to predict P-gp inhibitors. For example, in 2002, Ekins and co-workers built a series of phamacophore models based on structurally diverse inhibitors of P-gp. The representative pharmacophore model includes four features: a hydrogen bond acceptor, two hydrophobes, and a ring aromatic center. The models show potentials to predict inhibitors, which are not included in the training set.[14,15] In 2005, Wang and co-workers applied an unsupervised machine learning approach based on the Kohonen self-organizing maps (SOM) and a preselected set of molecular descriptors to construct classification models for inhibitors and substrates of P-gp.[16] The SOM models can successfully predict 83.3% of the substrates and 80.8% of the inhibitors. However, the prediction accuracy of Wang's model was not fully validated against an extensive test set, and the reported results are not reliable enough. In 2005, Sun developed a naive Bayesian classifier to categorize MDRR agents into active and inactive classes based on atom-type-based molecular descriptors,[17] and the model correctly predicted MDR reversal activities for 82.2% of 185 compounds in a test set. In 2006, Chang and co-workers generated a pharmacophore model of P-gp inhibitors, which includes four hydrophobic features and one hydrogen bond acceptor feature. Seven high-scoring molecules identified by the pharmacophore model were selected for *in vitro* testing, and six of them were confirmed as P-gp inhibitors. Chang's work showed the potential of pharmacophore modeling for identifing the P-gp inhibitors.[18]

Overall, the studies on *in silico* predictions of P-gp inhibitors are limited. To our knowledge, there is no *in silico* model to classify P-gp inhibitors and noninhibitors involving a large amount of molecules (larger than 1000). A disadvantage of QSAR and SAR models for P-gp inhibitors is the use of congeneric series of compounds in the data set. The broad multispecificity of P-gp and the lack of an extensive database of P-gp inhibitors have proved two almost insurmountable obstacles to establish accurate prediction models. Here we first present a large data set of 1273 molecules which are categorized into inhibitor and noninhibitor classes. We examined the impact of six important molecular properties on P-gp transport, which were widely used in the predictions of ADME. We then used the recursive partitioning technique to create decision trees to classify P-gp inhibitors and noninhibitors. Then the impact of different molecular fingerprints, as well as several essential parameters to generate the decision trees, on classification accuracy was studied. Finally, naive Bayesian classifiers based on molecular properties and fingerprints were developed and compared with the decision trees. Particularly, we expect that the Bayesian classifiers are predictive and easily interpretable. The naive Bayesian classifiers can identify the important structural features necessary for differentiating P-gp inhibitors and noninhibitors, which are promising for the design of potent P-gp inhibitors. Both decision trees and Bayesian classifiers, validated by the external test set, can be implemented as virtual screening tools in early phase of drug discovery.

METHODS AND MATERIALS

1. Data Set of P-gp Inhibitors and Noninhibitors. The entire data set contains 1273 structurally diverse molecules, among which 797 molecules are P-gp inhibitors and 476 molecules are P-gp noninhibitors. The whole data set was collected from 105 literatures. The data set reported by Bakken and Jurs is an important source of our data set. The Bakken's data set has 609 compounds with MDRR activity values measured using adriamycin-resistant P388 murine leukemia cells.[19] The experimental data for 347 compounds reported by Ramu and co-worker[20,21] is another important source of our data set. The experimental MDRR ratio was used as the criterion to determine whether a compound is an inhibitor or not: If MDRR ratio is less than 4.0, the compound was categorized into the noninhibitor class; if MDRR ratio is larger than 5.0, the compound was categorized into the inhibitor class; if MDRR ratio is ≥ 4 and ≤ 5.0, this compound was considered as moderately active and not included in the data set for classification. It should be noted that the MDRR ratio reported by Ramu and co-worker is not a direct measure of P-gp inhibition, and the inhibition class (inhibitor or noninhibitor) of a compound needs to be checked carefully if it was found in multiple publications. For example, according to Ramu's data, nifedipine is a noninhibitor (MDRR ratio = 1.4); however, according to the work reported by Ford et al. and that reported by Polli et al., nifedipine is a P-gp inhibitor.[22,23]

The structures of the compounds were built using the Sybyl molecular simulation package.[24] Structures were cross-checked in a search of the Beilstein database. Each molecule in the data set was optimized by using molecular mechanics (MM) with the MMFF94 force field.[24] All molecules were saved to a MACCS sdf file. Two data sets were collected for the following analysis: a data set with 1273 compounds that are categorized into inhibitors or noninhibitors, and a data set with 557 compounds that have quantitative MDRR activity. The 3-D structures, the indicators of P-gp inhibitors and noninhibitors, and the MDRR activity for all molecules are available online: http://modem.ucsd.edu/adme.

2. Calculation of Molecular Descriptors. Here, thirteen molecular descriptors widely adopted in ADME predictions were used in our analysis. These descriptors include octanol−water partitioning coefficient (AlogP) based on Ghose and Crippen's

method,[25] the apparent partition coefficient at pH = 7.4 (log D) based on the Csizmadia's method, molecular solubility (log S) based on the multiple linear regression model developed by Tetko et al.,[26] molecular weight (MW), the number of hydrogen bond donors (n_{HBD}), the number of hydrogen bond acceptors (n_{HBA}), the number of rotatable bonds (n_{rot}), the number of rings (n_R), the number of aromatic rings (n_{AR}), the sum of oxygen and nitrogen atoms (n_{O+N}), polar surface area (PSA), molecular fractional polar surface area (MFPSA), and molecular surface area (MSA). All the descriptors were calculated using the Discovery Studio molecular simulation package (version 2.5).[27]

3. Calculation of Molecular Fingerprints. Here, we used the SciTegic extended-connectivity fingerprints (FCFP, ECFP and LCFP) and Daylight-style path-based fingerprints (FPFP, EPFP and LEFP). The first character, F, E or L, in the fingerprint name denotes the atom abstraction method used to assign initial atom code: F represents functional role code, E represents atom type code and L represents AlogP atom type code. Functional role code is a combination of a hydrogen-bond acceptor, hydrogen-bond donor, positively ionized or positively ionizable, negatively ionized or negatively ionizable, aromatic, and halogen. Atom type code is derived from the number of connections to an atom, the element type, the charge, and the atomic mass. AlogP atom type code uses the 120 atom types used in the calculation of AlogP. The second character, C or P, in the fingerprint name denotes the type of fingerprint: C represents extended-connectivity fingerprints, and P represents path-based fingerprints. The fourth character of a fingerprint class is followed by an underscore and the maximum distance. For example, a functional class extended-connectivity fingerprint of maximum diameter 6 generates a property named FCFP_6. Here for each fingerprint class, two diameters, 4 and 6, were considered in our analysis. The smaller diameter, 2, is not used because the structural fragments based on the diameter of 2 are small and general. The detailed descriptions of these fingerprints were reported.[28,29]

These fingerprints we used here are different from the substructures in the prediction of ADME properties reported previously.[30,31] They represent a much larger set of features than predefined substructures. Furthermore, these fingerprints do not need to be preselected or predefined because they are generated directly from the molecules. Therefore, novel molecular classes are so easily handled as the common classes. Here we generated the structural fingerprints by using Discovery Studio molecular simulation package.[27]

4. Recursive Partitioning Models. Recursive partitioning (RP) in Discovery Studio molecular simulation package was used to develop decision trees to categorize the compounds into P-gp inhibitors and noninhibitors. RP is a classification method that constructs one or more decision trees to decipher the relationship between a dependent property Y (inhibition class) and a set of independent properties X (molecular properties and fingerprints), and it classifies data by using a set of hierarchical rules to split a data set into smaller and smaller subsets. The result of RP is a "decision tree" or "graph", which is constructed through a recursive partitioning process that divides the study samples into smaller and smaller samples (every subsample is called a node) according to whether a particular selected predictor is above a chosen cutoff value or not. At each step, all of the molecular descriptors are sequentially analyzed to find the best criterion for subdividing compounds. Once the best criterion is confirmed, the procedure is repeated for each of the obtained classes of compounds. RP is not designed to stop splitting at the right moment; instead, it is designed to "over-split" and then prune the tree backward. In our study, 5-fold cross-validation was used to determine the degree of pruning required for the best predictive performance. The minimum number of samples at each node was set to 8, and the maximum tree depth was changed from 3 to 10 systematically in order to evaluate the effect of tree depth on the predictions and choose the best value. The whole data set was randomly split into a training set of 973 compounds and a test set of 300 compounds, respectively. The training set was used to create the decision trees, and the 300 compounds of the external test set were used to evaluate the true prediction power of the models.

5. Naive Bayesian Classifiers. The naive Bayesian categorization models were developed using Discovery Studio.[27] In our study, each compound can be categorized into inhibitor (+) or noninhibitor (−) class, and a vector $\mathbf{f} = \langle f_1, f_2, ... f_n \rangle$, where $f_1, f_2, ..., f_n$ are the calculated values for the feature variables (molecular properties and fingerprints) $F_1, F_2, ..., F_n$. Then using Bayes's theorem, we get:

$$p(C|F_1, F_2, ..., F_n) = \frac{p(C)p(F_1, ...F_n|C)}{p(F_1, ..., F_n)} \quad (1)$$

where C refers to a compound's class, $p(C|F_1,F_2,...,F_n)$ is the posterior probability of the compound class, $p(C)$ is the prior probability, a probability induced from the training set, $p(F_1,...F_n|C)$ is the probability that a compound has certain descriptors given that it is an inhibitor or noninhibitor, and $p(F_1,...,F_n)$ is the marginal probability that given the descriptors will occur in the data set.

In naive Bayesian classifier, each descriptor is conditionally independent of every other descriptor. From this assumption, we can get

$$p(F_1, ...F_n|C) = p(F_1|C)...p(F_n|C) = \prod_{i=1}^{n} p(F_i|C) \quad (2)$$

Each factor in eq 2 can be estimated from the training set:

$$p(F_i = f_i| +) = \frac{\text{count}(F_i = f_i \cap C = +)}{\text{count}(C = +)}$$
$$p(F_i = f_i| -) = \frac{\text{count}(F_i = f_i \cap C = -)}{\text{count}(C = -)} \quad (3)$$

In naive Bayesian classification, the following strategy is used to avoid evaluating the marginal probability. In the current work, all the compounds are categorized into two classes: inhibitor class or noninhibitor class. Here $p(+)$ is the prior probability of the inhibitor class and $p(-)$ is the prior probability of the noninhibitor class, and then the posterior probability of a compound to be an inhibitor (p) or a noninhibitor (q) is

$$p = \frac{p(+)}{p(F_1 = f_i, ..., F_n = f_n)} \prod_{i=1}^{n} p(F_i = f_i| +) \quad (4)$$

$$q = \frac{p(-)}{p(F_1 = f_i, ..., F_n = f_n)} \prod_{i=1}^{n} p(F_i = f_i| -) \quad (5)$$

where the marginal probability $p(F_1=f_1,...,F_n=f_n)$ is a constant. Since $p + q = 1$, then we have

$$\log\frac{p}{q} = \log\left(\frac{1-q}{q}\right) = \sum_{i=1}^{n}(\log p(F_i = f_i|+) - \log p(F_i = f_i|-)) + (p(+) - p(-)) \quad (6)$$

In the naive Bayesian learning process, we need to solve two problems, zero counts and missing values. Zero counts refers to the situation where a certain descriptor never occurs for a given class in the training set. A related issue, missing values, may occur when evaluating a tested compound. For example, a fingerprint can be found in a compound from the test set but not in any compounds from the training set. The problem of zero counts is usually solved by using Laplace correction:

$$p(F_i = f_i|+) = \frac{(\text{count}(F_i = f_i \cap C = +) + 0.5)}{(\text{count}(C = +) + 1)}$$
$$p(F_i = f_i|-) = \frac{(\text{count}(F_i = f_i \cap C = -) + 0.5)}{(\text{count}(C = -) + 1)} \quad (7)$$

Missing values, or missing fingerprint, are usually ignored to avoid introducing unproved information. An advantage of the naive Bayesian classifier is that it only requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Moreover, Bayesian classification can process large amounts of data, can learn fast, and is tolerant of random noise. The naive Bayesian classifiers were developed using the Create Bayesian Model protocol in Discovery Studio.[27,28,32]

6. Validating the Prediction Accuracies of the RP and Bayesian Models. The quality of the Bayesian and RP models was measured by the quantity of true positives (TP), true negatives (TN), false positives (FP), false negatives (FN), sensitivity (SE), and specificity (SP), the prediction accuracy for inhibitors (Q_+), the prediction accuracy for noninhibitors (Q_-), and the Matthews correlation coefficient (C).[33]

$$\text{SE} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (8)$$

$$\text{SP} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (9)$$

$$\text{PRE1} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (10)$$

$$\text{PRE2} = \frac{\text{TN}}{\text{TN} + \text{FN}} \quad (11)$$

$$C = \frac{\text{TP} \times \text{TN} - \text{FN} \times \text{FP}}{\sqrt{(\text{TP} + \text{FN})(\text{TP} + \text{FP})(\text{TN} + \text{FN})(\text{TN} + \text{FP})}} \quad (12)$$

The value of C is the most important indicator for the classification accuracy of the models. The above quantities were calculated for both training and test sets.

RESULTS AND DISCUSSIONS

1. The Roles of Simple Molecular Properties on P-gp Transport. A lot of molecular descriptors have been proven to be useful for the predictions of ADME properties,[2,33−49] and they describe a variety of molecular properties, such as lipophilicity, hydrogen bonding ability, molecular flexibility, molecular bulkiness, etc. Here, the relationships between eight molecular properties widely used in ADME predictions and P-gp inhibition were examined systematically. These eight molecular properties include MW, log S, log $D_{7.4}$, log P, PSA, n_{HBA}, n_{HBD} and n_{rot}. The distributions of these eight molecular properties for the inhibitor and noninhibitor classes are shown in Figure 1. Student's t-test was employed to evaluate the significance of the difference between the means. As a complementary test, the linear correlations between each of these eight molecular properties and the MDR-reversal activities of 557 compounds are shown in Figure 2.

For the eight molecular properties studied here, three of them are related to hydrophobicity of a molecule: log P, log D and log S. log P is distributed between −9.92 and 15.72, with a mean of 2.64, log D between −9.01 and 15.72, with a mean of 3.04, and log S between −13.48 and 2.51, with a mean of −6.19. The distributions of log P of the inhibitor class and the noninhibitor class are shown in Figure 1. The mean values of log P for the 797 inhibitors versus 476 noninhibitors are 3.28 and 1.58, respectively; that is to say, P-gp prefers to transport hydrophobic molecules. We believe that the inhibitors with higher hydrophobicity may form stronger hydrophobic interactions with the hydrophobic binding region of membrane-embedded transmembrane (TM) domains. Student's t-test was employed to evaluate the significance of the difference between the means. The p-value associated with the difference in the mean log P values of the inhibitors versus those of the noninhibitors was 6.97 e^{-50} at the 95% confidence level, indicating that the two distributions are significantly different. Compared with log P, log D shows better classification capability for inhibitors and noninhibitors because the p-value of log D shown in Figure 1 is 5.97 e^{-58}. In Figure 1, log S shows better classification capability than log P and log D because the p-value associated with the difference in the mean log S values of the inhibitors versus those of the noninhibitors was 6.48 e^{-76}. The mean values of log S for the 797 inhibitors versus 476 noninhibitors are −7.71 and −4.93, respectively. Similar results are obtained in Figure 2. log S shows a better linear correlation ($r = -0.34$) with MDR-reversal activities than log D ($r = 0.31$) and log P ($r = 0.23$). Indubitably, log S is more effective to predict P-gp inhibition than log D and log P. However, the two distributions of log S for inhibitors and noninhibitors are still strongly overlapped.

As shown in Figure 1, molecular weight has relatively high impact on P-gp inhibition. The mean values for inhibitors and noninhibitors are 490.5 and 362.7, respectively. The p-value associated with the difference in the mean MW values of the two classes is 2.92 e^{-31} at the 95% confidence level, indicating that MW also has high impact on P-gp inhibition. But it is worse to discriminate the inhibitor class from the noninhibitor class than log D. Moreover, our finding is consistent with Gleenson's results.[50] According to Gleeson's analysis, molecules having MW > 400 and/or log P value >4 are more likely to be transported by P-gp.[50] Certainly, according to our analysis shown in Figure 1, MW > 400 is not a good indicator for P-gp transport either. In Figure 1, we can observe that the descriptor, n_{rot}, shows high impact on P-gp transport because the difference in the mean n_{rot} values of the two classes is 4.41 e^{-31}. In many cases, n_{rot} can be considered as a descriptor to characterize the bulkiness of a molecule because a larger molecule usually has more rotatable

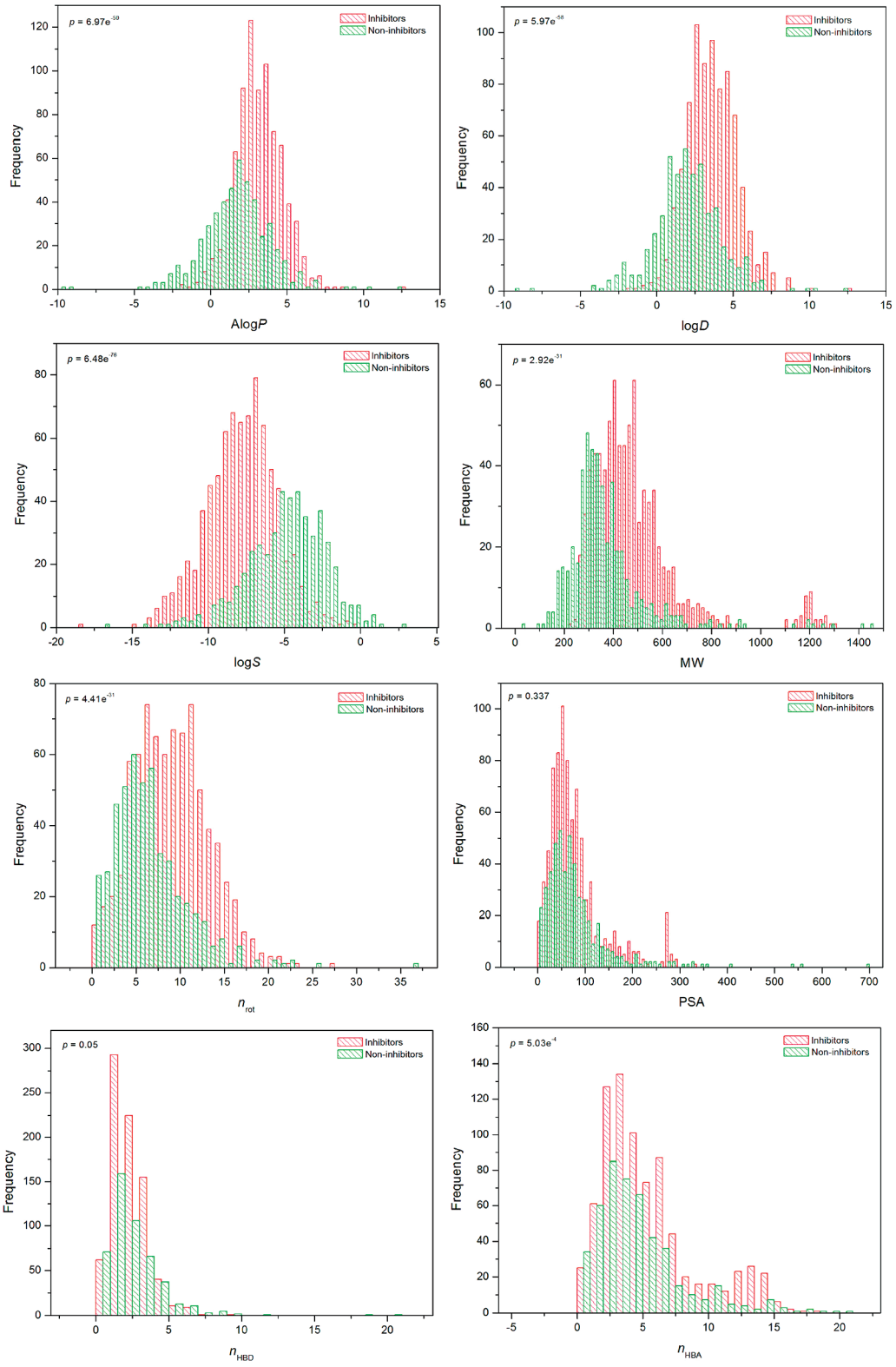


Figure 1. Distributions of eight molecular properties, including AlogP, log *D*, log *S*, MW, PSA, n_{rot}, n_{HBD} and n_{HDA}, for the inhibitor and noninhibitor classes.

bonds. In fact, n_{rot} has relatively obvious correlation with molecular weight (r = 0.64).

The other three descriptors (PSA, n_{HBD} and n_{HDA}) representing electrostatic or H-bond features do not show any capability to discriminate inhibitors from noninhibitors. The p-values associated with the difference in the mean values of the two classes for PSA, n_{HBD} and n_{HDA} are 0.337, 0.05 and 5.03 e^{-4}. In Figure 2, it can be observed that MDR-reversal activities have very low

dx.doi.org/10.1021/mp100465q |*Mol. Pharmaceutics* 2011, 8, 889–900

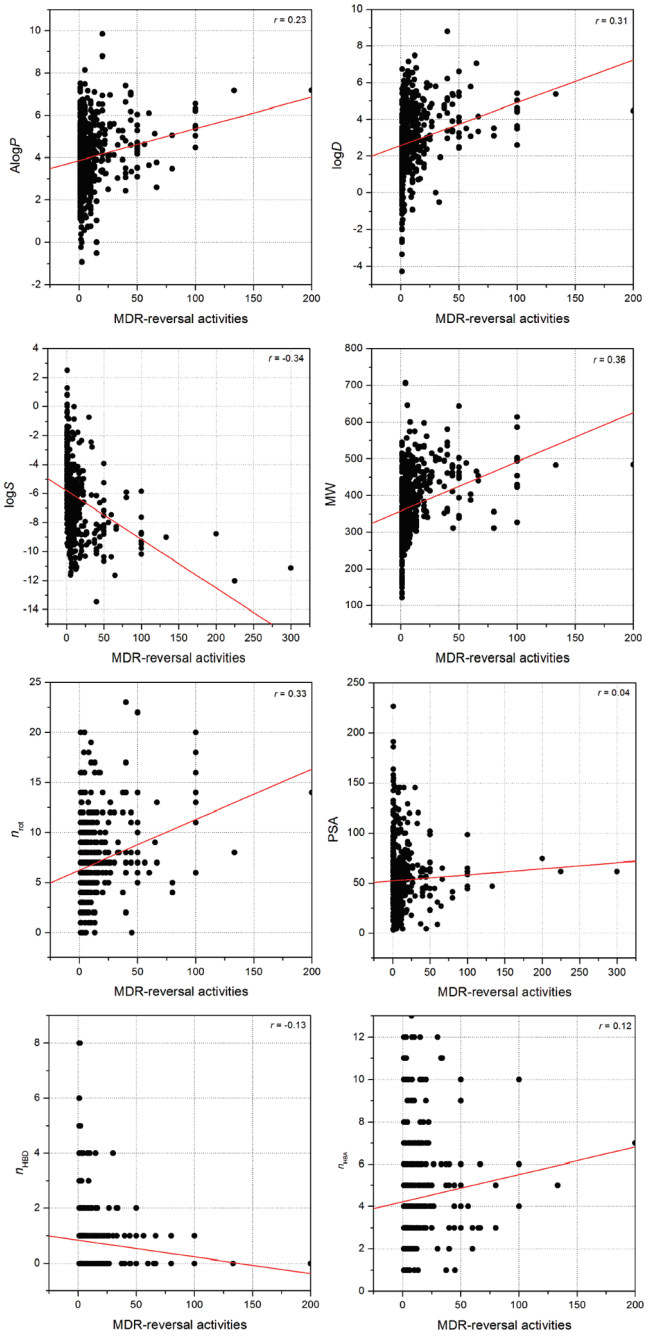


Figure 2. Correlations between eight molecular properties, including AlogP, log *D*, log *S*, MW, PSA, n_{rot}, n_{HBD} and n_{HDA}, and MDR-reversal activities.

correlations with PSA ($r = 0.04$), n_{HBD} ($r = -0.13$) and n_{HDA} ($r = 0.12$). Our findings are not consistent with the reported pharmacophore models of P-gp inhibitors because all these pharmacophore models have H-bond acceptor features. The basic reason for the inconsistence is that the three descriptors, PBSA, n_{HBD} and n_{HDA}, are too general to characterize the spatial constraint among these H-bond elements while the pharmacophore model is able to characterize.

2. Recursive Partitioning Models. According to our analysis above, it is obvious that a single molecular property is not a good criterion for classification. In order to develop more precise and understandable classification models, we applied the RP technique to establish decision trees to classify molecules into different groups. Compared with "the blind operations" of ANNs and SVMs, the results of RP can be converted to simple hierarchical rules, which are easily interpreted. First, we constructed a decision tree based on the thirteen molecular properties, and the tree depth was set to 7. The performance of the RP model is shown in Table S1 in the Supporting Information. For the training set, the sensitivity and specificity are 87.8% and 84.9%, respectively. However, the prediction accuracy of the model for the test set ($C = 0.397$, $SE_{test} = 75.0\%$ and $SP_{test} = 65.2\%$) is much worse than that for the training set. The molecular properties we used can depict whole-molecule properties, but they cannot characterize the important substructures or molecular fragments that are important to P-gp transport. Then, molecular fingerprints, together with molecular properties, were used simultaneously as the descriptors in RP analysis. Here we compared the performance of 12 sets of fingerprints for classification. Obviously, the addition of fingerprint can improve the performance of the RP models because the *C* values of the RP models after considering fingerprints are higher than that of the PR model only based on thirteen molecular properties (Figure 3a). According to the *C* values shown in Figure 3a ($C = 0.730$), the fingerprint set ECFP_4 performs better than the others. Moreover, in order to compare the actual prediction capability of fingerprints, all models shown in Table 1 were used to make

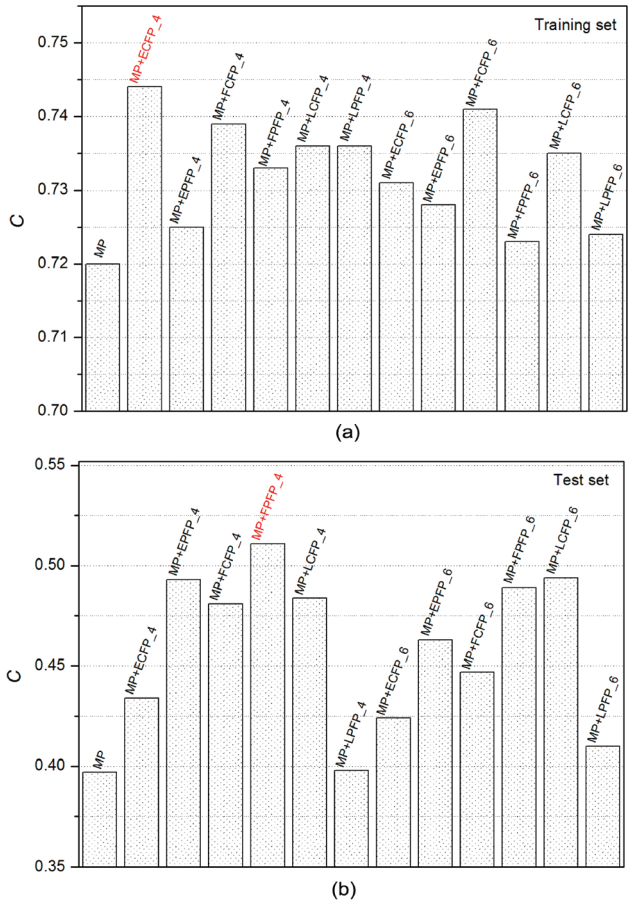


Figure 3. The Matthews correlation coefficient (*C*) of the thirteen decision trees based on molecular properties and twelve sets of fingerprints for (a) the training set and (b) the test set.

Table 1. The Performance of the Bayesian Classifiers for the Training and Test Sets

	training set									test set								
descriptors	TP	FN	TN	FP	SE	SP	Q_+	Q_-	C	TP	FN	TN	FP	SE	SP	Q_+	Q_-	C
MP[a]	510	99	281	83	0.837	0.772	0.860	0.739	0.604	145	43	78	34	0.771	0.696	0.810	0.645	0.461
MP+ECFP_4	511	98	302	62	0.839	0.830	0.892	0.755	0.658	155	33	81	31	0.824	0.723	0.833	0.711	0.546
MP+EPFP_4	444	165	304	60	0.729	0.835	0.881	0.648	0.546	129	59	85	27	0.686	0.759	0.827	0.590	0.431
MP+FCFP_4	515	94	299	65	0.846	0.821	0.888	0.761	0.658	157	31	82	30	0.835	0.732	0.840	0.726	0.566
MP+FPFP_4	418	191	337	27	0.686	0.926	0.939	0.638	0.595	129	59	97	15	0.686	0.866	0.896	0.622	0.535
MP+LCFP_4	505	104	302	62	0.829	0.830	0.891	0.744	0.647	151	37	83	29	0.803	0.741	0.839	0.692	0.537
MP+LPFP_4	461	148	295	69	0.757	0.810	0.870	0.666	0.551	142	46	81	31	0.755	0.723	0.821	0.638	0.468
MP+ECFP_6	490	119	312	52	0.805	0.857	0.904	0.724	0.645	148	40	90	22	0.787	0.804	0.871	0.692	0.577
MP+EPFP_6	448	161	299	65	0.736	0.821	0.873	0.650	0.540	133	55	82	30	0.707	0.732	0.816	0.599	0.427
MP+FCFP_6	484	125	315	49	0.795	0.865	0.908	0.716	0.642	157	31	82	30	0.835	0.732	0.840	0.726	0.566
MP+FPFP_6	449	160	312	52	0.737	0.857	0.896	0.661	0.576	147	41	90	22	0.782	0.804	0.870	0.687	0.571
MP+LCFP_6	508	101	302	62	0.834	0.830	0.891	0.749	0.652	141	47	85	27	0.750	0.759	0.839	0.644	0.496
MP+LPFP_6	426	183	318	46	0.700	0.874	0.903	0.635	0.555	153	35	83	29	0.814	0.741	0.841	0.703	0.549
MP+FCFP_4[b]	481.9	127.1	313.5	50.5	0.791	0.861	0.905	0.712	0.634	152.6	35.4	91	21	0.812	0.813	0.879	0.720	0.611

[a] MP represents molecular properties. [b] The results based on 20 times of the training−testing process.

predictions for the 300 tested compounds, and the corresponding C values are shown in Figure 3b. It is interesting to find that all models based on molecular properties and fingerprints perform better than the model solely based on molecular properties. Among all these fingerprint sets, FPFP_4 performs best for the test set, indicated by the highest C value (0.511). When molecular properties and FPFP_4 fingerprint set were used as the descriptors, for the 300 tested compounds the RP model achieves a sensitivity of 83.5%, a specificity of 67.0%, a prediction accuracy of 80.9% for the inhibitor class, and a prediction accuracy of 70.8% for the noninhibitor class.

In RP analysis, the depth of decision tree is a very important parameter, and it dominates the complexity of a decision tree. Usually, large tree depth tends to increase the accuracy on the training data but at the risk of overfitting. Small value tends to increase the applicability of a tree to new data sets, but at the risk of decreased accuracy and of failing to identify important features in the training data. So the tree depth should be carefully calibrated according to the predictions on the tested compounds. Here, the tree depth was changed from 3 to 10 and the corresponding performance of the model to the training and test sets was evaluated (Table S2 in the Supporting Information). The change of C versus tree depth is shown in Figure 4. For the training set the value of C increases as the tree depth increases; however for the test set the values of C do not always increase. For the test set the value of C increases significantly from depth 3 to 6, and then increases slightly to reach the maximum at depth 7 (C = 0.511). When depth is larger than 7, the C values for the test do not increase anymore. So for our analysis, depth of 7 is the best choice.

The decision tree with tree depth 6 is shown in Figure 5 as an example. This decision tree has 20 inner nodes and 21 leaves. In this decision tree, the discriminant descriptors include seven molecular properties (log S, MW, AlogP, log D, n_{N+O}, MSA and n_{ring}) and eight structural fragments. In the seven molecular properties chosen by the decision tree, log S, AlogP and log D are properties to describe hydrophobicity, MW, MSA and n_{ring} are properties to describe bulkiness, and n_{N+O} describes the electrostatic or H-bonding features. Therefore, the hydrophobicity, size,

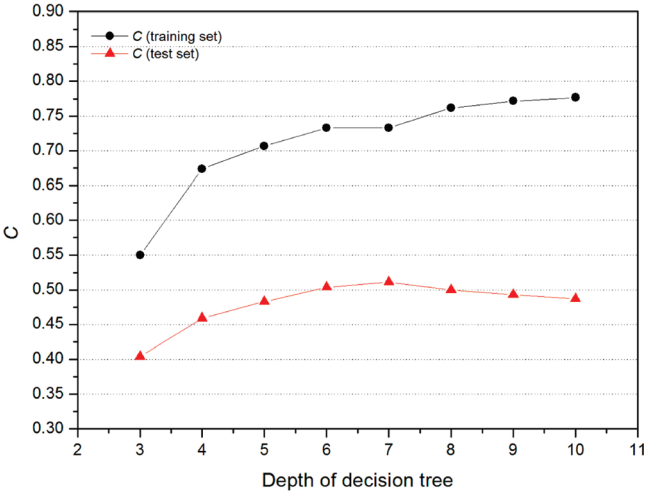


Figure 4. The change of C versus the tree depth for (a) the training set and (b) the test set. The RP model was constructed based on molecular properties and the FPFP_4 fingerprint set. The models with the highest C values are colored in red.

and electrostatic properties are important for P-gp transport. Moreover, the eight fingerprints shown in Figure 5 are also important to discriminate inhibitors from noninhibitors.

3. Naive Bayesian Classifiers. The decision tree given by RP is easily understandable. However, RP splits data at each node to "good" or "bad" class linearly. Moreover, RP is sensitive to some predefined parameters. Here, a more complicated statistical technique, naive Bayesian classification, was used to develop classification models. Naive Bayesian classification is an unsupervised learner, without fitting process and tuning parameters. The process of Bayesian learning is to search through each feature in an unbiased way for those with separation power.

Similar to the RP analysis, the performance of the naive Bayesian classifiers with different sets of fingerprints was evaluated and compared. The statistical results for these Bayesian classifiers are listed in Table 1. In Table 1, the statistical results for

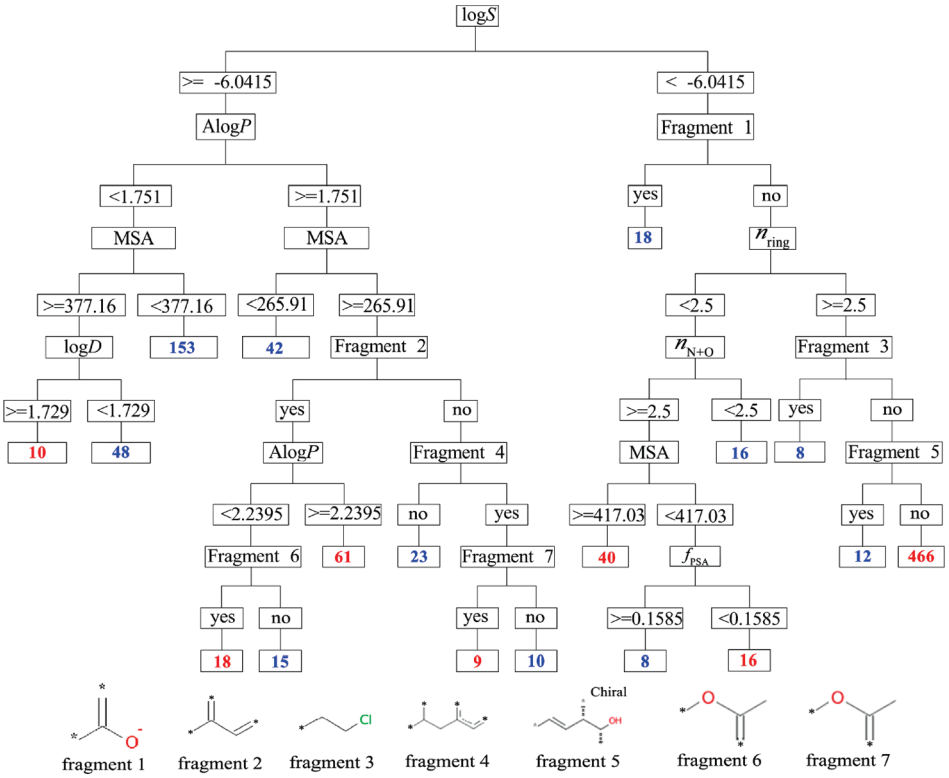


Figure 5. Decision tree to classify compounds into inhibitor and noninhibitor classes by using RP (tree depth is set to 6).

the training set were obtained using the leave-one-out (LOO) cross-validation process. It is obvious that the performance of the various fingerprint sets is quite different. According to the C values by the LOO cross-validation, ECFP_4, FCFP_4 and LCFP_6 are associated with better classifiers. The best classifier based on molecular properties and FCFP_4 fingerprint set has a sensitivity of 84.7, a specificity of 82.1, a prediction accuracy of inhibitors of 88.9%, and a prediction accuracy of noninhibitors of 76.1 for the training set. Compared with the classifier based on molecular properties, the addition of fingerprints can significantly improve the classification. All the Bayesian classifiers were then validated by the predictions on the external test set, and the validation results are shown in Table 1. According to the validation, four classifiers (MP+FCFP_4, MP+ECFP_6, MP+FCFP_6 and MP+FPFP_6) can give C values larger than 0.566 for the 300 tested compounds. Compared with the prediction accuracy of the PR models, the best Bayesian classifier performs much better. For the test set, the classifier based on molecular properties and FCFP_4 fingerprints gives a sensitivity of 83.5% and a specificity of 73.2%, and an overall prediction accuracy of ~80% ([TP + TN]/[TP + TN + FP + FN]). Then, we designed a more rigorous training-testing procedure to assess the statistical significance of the Bayesian classifiers. The whole data set was randomly split into a training set of 973 compounds and a test set of 300 molecules. The training−testing process was iteratively executed for 20 times to evaluate the average performance of the Bayesian classifier. As shown in Table 1, for the test set the Bayesian classifier based on molecular properties and FCFP_4 fingerprints gives an average sensitivity of 79.2% and an average specificity of 83.8%, and an average prediction accuracy of ~81%.

In our study, two threshold values of 4.0 and 5.0 were used to determine whether a compound is an inhibitor or not. To estimate the influence of these values on the performance of the classification models, the other four sets of threshold values (5.0 and 6.0; 6.0 and 7.0; 7.0 and 8.0; 8.0 and 9.0) were used to split the whole data into the inhibitor and noninhibitor classes and the classification models were reconstructed based on molecular properties and FCFP_4 fingerprints (Table S3 in the Supporting Information). As shown in Table S3 in the Supporting Information, the best performance for the training set can be achieved when the threshold values of 5.0 and 6.0 are used. However, the model with the threshold values of 7.0 and 8.0 has the best predictive capability for the tested molecules. It should be noted that the threshold values are really arbitrarily defined and we cannot determine which values are the best, but at least the data in Table S3 in the Supporting Information show that the classification models with reliable prediction for the tested molecules can be constructed even when a different set of threshold values were used.

The prediction accuracy of the Bayesian classifier based on molecular properties and FCFP_4 to discriminate inhibitors from noninhibitors was evaluated with two bimodal histograms of the training and test data sets (Figure 6). The histograms show that the inhibitors tend to have more positive values, while the noninhibitors tend to have more negative values. For the training set both classes of compounds have strong overlaps between −5 and 0. So the region between −5 and 0 can be defined as the "uncertain zone". When the Bayesian score of a molecule is in the uncertain zone, the prediction for this molecule is not reliable. For the 300 molecules in the test set, 100 of them are predicted in the uncertain zone. If these 100 molecules without accurate predictions were eliminated from the test set, the sensitivity and specificity were increased to 91.7% and 85.7%, respectively.

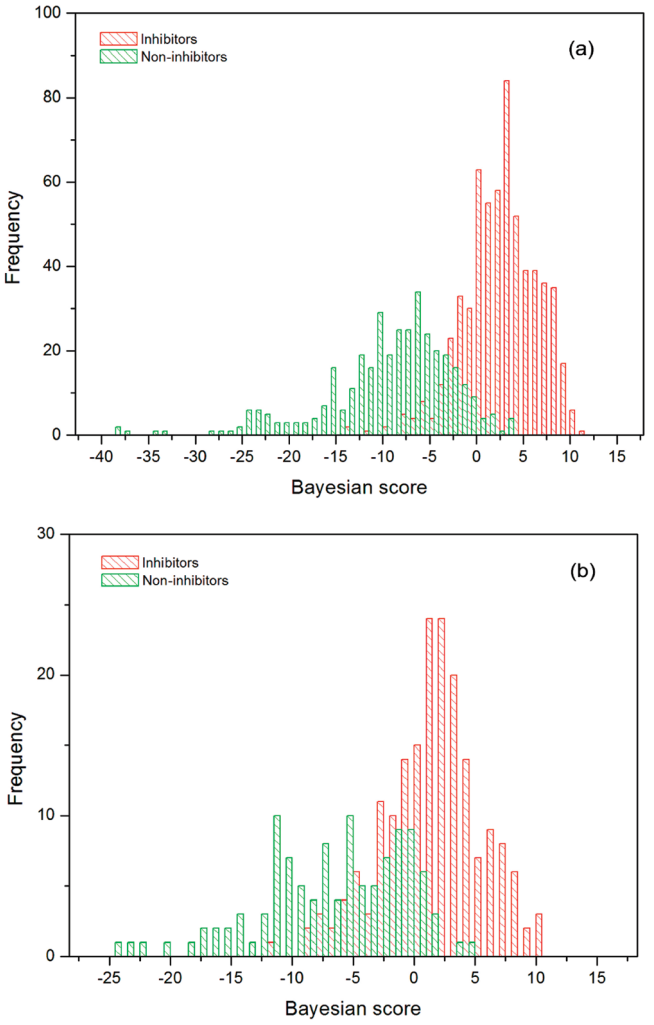


Figure 6. The distributions of Bayesian score predicted by the Bayesian classifier based on molecular properties and the FCFP_4 fingerprint set for the inhibitor and noninhibitor classes for (a) the training set and (b) the test set. The Bayesian scores for the training set were obtained by using the leave-one-out (LOO) cross-validation process.

4. Analysis of the Important Fragments Given by Naive Bayesian Classifier. An advantage of fingerprints is that they can be easily translated into 2-D fragments. The important fragments given by the Bayesian classifier may be useful for experimental scientists when designing molecules with better P-gp inhibition. The 15 good and 15 bad fragments ranked by the Bayesian scores are summarized in Figure 7.

By analyzing the fingerprints with positive contributions to P-gp inhibition shown in Figure 7a, we can observe that many fragments have nitrogen atoms, and these fragments include the top five fragments and fragments 7 and 8. According to previous studies, it is believed that a nitrogen atom attached to saturated carbons is a potential "binding center" with P-gp.[51,52] By analyzing 12 molecules with nitrogen atom, Ecker concluded that the interactions between nitrogen atom and P-gp were nonionic and were determined by the capability of the hydrogen acceptors.[51] Klopman and co-workers also reported that presence of a nitrogen atom linked to aliphatic atoms is relevant to MDR reversal activity.[52] Obviously, the nitrogen atoms in these important fragments can serve as strong hydrogen acceptors and form stable H-bonding interactions with P-gp. Another important fragment (fragment 6) includes an ester group linked to aromatic environment. The oxygen atom in this ester group is an electron donor and therefore acts as a H-bonding acceptor. Klopman and co-worker have observed a similar biophore for P-gp inhibition: a carbonyl group attached to a meta substituted aromatic ring.[52]

The 15 fingerprints shown in Figure 7b indicate that the existence of these fragments is unfavorable for P-gp inhibition. It is quite interesting to find that all the top five fragments have a negatively charged oxygen atom in an acid group, indicating that introducing an acid group would decrease the P-gp inhibition of a compound significantly. Sun analyzed 424 compounds as a training set by Bayesian classification and found that an acidic group would cause a compound to lose MDR activity.[17] In the 15 fingerprints shown in Figure 7b, we found that two of them have a nitrogen atom in a quaternary ammonium cation (fragments 9 and 13) and three of them have tertiary amine group (fragments 10, 14 and 15). Therefore, not all nitrogen atoms are favorable to increase the MDR activity of a molecule, and nitrogen in primary and secondary amines might be more favorable for P-gp inhibition.

5. Why Some Inhibitors Cannot Be Predicted Correctly. Finally, we checked the error sources of the Bayesian classifiers. The predictions using the best four Bayesian classifiers (classifiers based on MP+FCFP_4, MP+ECFP_6, MP+FCFP_6 and MP+FPFP_6) in Table 1 for the test set are listed in Table S4 in the Supporting Information. We found that in total 17 molecules out of 112 noninhibitors and 24 molecules out of 188 inhibitors cannot be correctly predicted by any of the four Bayesian classifiers. We believe that the following reasons account for the misclassification. First, the quality of the data set may be a primary source of prediction errors. In our analysis, if MDR-reversal ratio was less than 4.0, the compound was defined as a noninhibitor; if MDR-reversal ratio was larger than 5.0, the compound was defined as an inhibitor. MDR-reversal ratio is an important indicator but not a direct measure of P-gp inhibition. The cutoff of MDR-reversal ratio to define the noninhibitor class is also arbitrary. This cutoff used by Sun is 2.[17] For these 17 misclassified noninhibitors, seven of them have MDR-reversal ratio larger than 2.0, and these molecules are opipramol, boxidine, SPS 1841, methixene, keoxifene, chlorcyclizine and tixadil. It is understandable that these molecules are easily misclassified as inhibitors. Similarly, for these 24 misclassified inhibitors, some of them have relatively low MDR-reversal ratio (lower than 6.0), such as terodiline, halopemide and Ro-04-2287, and these compounds are easily to be predicted as noninhibitors. Moreover, the quality of the compounds collected from the literature without MDR-reversal activity cannot be completely guaranteed because of possible experimental errors. Finally, the misclassification may be caused by the intrinsic drawback of the models. The fingerprints can characterize the important structural fragments favorable or unfavorable for P-gp inhibition, but they cannot characterize the spatial arrangement of these important fingerprints if multiple important fingerprints are found simultaneously in one molecule.

CONCLUSIONS

In this work, we report an extensive P-gp inhibition database, which consists of 1273 molecules. On the basis of the large set of P-gp inhibition data, the relationships between eight important

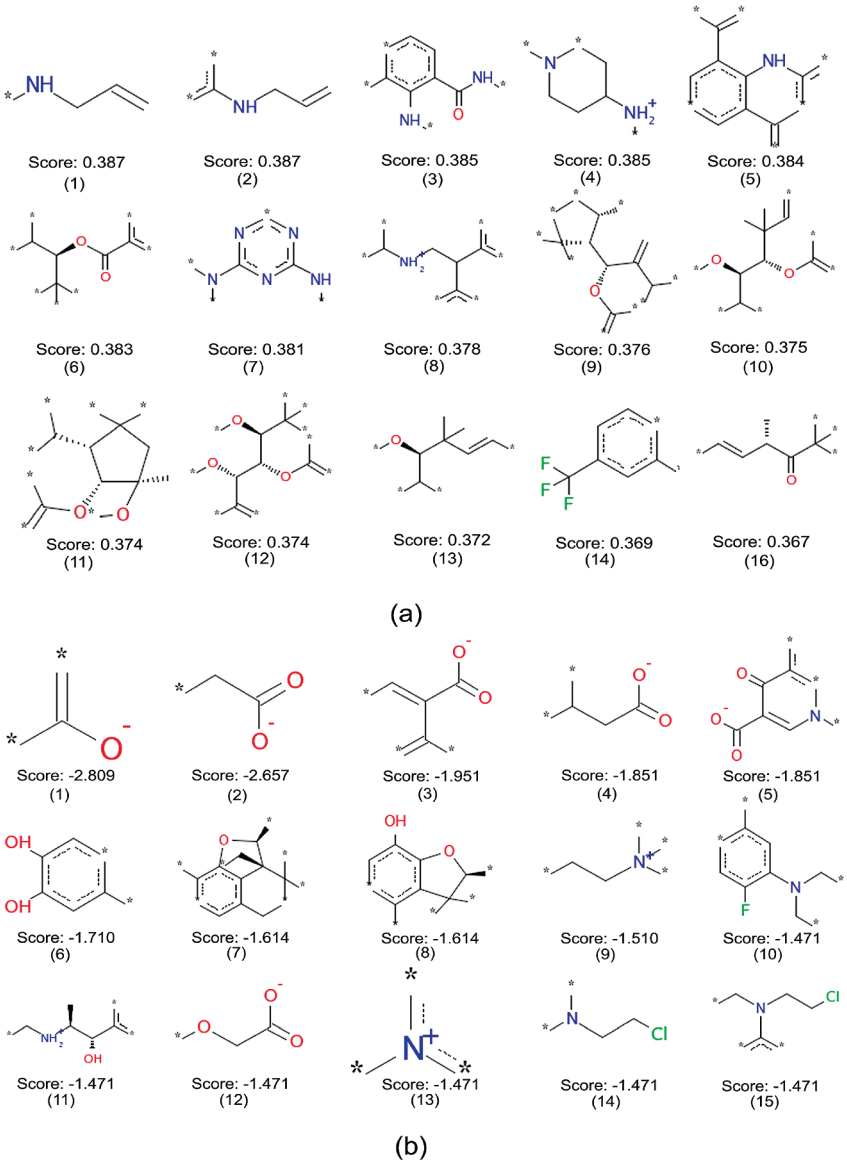


Figure 7. (a) The 15 good and (b) 15 bad fragments for P-gp inhibition identified by the Bayesian classifier based on molecular properties and the FCFP_4 fingerprint set.

molecular properties and P-gp inhibition have been examined systematically. We observed that some of them, especially solubility, log *D* and molecular weight, are important contributors of P-gp transport but a single molecular property is not enough to distinguish inhibitors from noninhibitors. Then, RP technique was applied to construct the decision trees to classify the whole data set into the inhibitor and noninhibitor classes. In order to characterize the structural features important for P-gp transport, structural fingerprints were introduced into our analysis. We found that the introduction of fingerprints improves the prediction accuracy significantly. Finally, naive Bayesian categorization modeling was applied to establish classifiers for P-gp inhibitors. The Bayesian classifier, based on molecular properties and fingerprints, demonstrates good predictivity, indicated by the high prediction accuracies for the training set and the test set (average correct prediction for 81.7% of 973 compounds in the training set using a leave-one-out cross-validation procedure and 81.2% of 300 compounds in the test set). The comparison analysis shows that naive Bayesian categorization modeling outperforms RP, indicating that the naive Bayesian classifiers show great potential for the predictions of P-gp inhibitors. The important molecular fragments favorable or unfavorable for P-gp transport given by Bayesian classifiers will be very helpful for the design of new inhibitors with better P-gp inhibitory activity.

The P-gp database of 1273 compounds with classification data and the P-gp database of 557 compounds with quantitative MDR-reversal activities are available online: http://modem.ucsd.edu/adme.

■ ASSOCIATED CONTENT

Supporting Information. The classification performance of thirteen RP models (Table S1). The classification performance of eight RP models using different tree depth (Table S2). The classification performance of five Bayesian classifiers based

on five sets of threshold values (Table S3). The predictions for the 300 tested compounds using the top four Bayesian classifiers (Table S3). The decision tree based on RP with a tree depth of 7 (Figure S1). This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*T.H.: e-mail, tjhou@suda.edu.cn or tingjunhou@hotmail.com; phone, +86-512-65882039. H.P. e-mail, p_h2002@hotmail.com; phone, +86-10-66932339.

ACKNOWLEDGMENT

The project is supported by the National Science Foundation of China (Grant No. 20973121 and 30873083), the National Basic Research Program of China (973 Program, Grant No. 2010CB934500) and the Priority Academic Program Development of Jiangsu Higher Education Institutions.

REFERENCES

(1) Sharom, F. J. ABC multidrug transporters: structure, function and role in chemoresistance. *Pharmacogenomics* **2008**, *9*, 105–127.

(2) Zhou, S. F. Structure, function and regulation of P-glycoprotein and its clinical relevance in drug disposition. *Xenobiotica* **2008**, *38*, 802–832.

(3) Ambudkar, S. V.; Kimchi-Sarfaty, C.; Sauna, Z. E.; Gottesman, M. M. P-glycoprotein: from genomics to mechanism. *Oncogene* **2003**, *22*, 7468–7485.

(4) Higgins, C. F.; Linton, K. J. The ATP switch model for ABC transporters. *Nat. Struct. Mol. Biol.* **2004**, *11*, 918–926.

(5) Bodo, A.; Bakos, E.; Szeri, F.; Varadi, A.; Sarkadi, B. The role of multidrug, transporters in drug availability, metabolism and toxicity. *Toxicol. Lett.* **2003**, *140*, 133–143.

(6) Lin, J. H. Role of P-glycoprotein in drug metabolism and disposition. *Drug Metab. Rev.* **2003**, *35*, 13–13.

(7) Lin, J. H. Drug-drug interaction mediated by inhibition and induction of P-glycoprotein. *Adv. Drug Delivery Rev.* **2003**, *55*, 53–81.

(8) Tanigawara, Y. Role of P-glycoprotein in drug disposition. *Ther. Drug Monit.* **2000**, *22*, 137–140.

(9) Varma, M. V. S.; Sateesh, K.; Panchagnula, R. Functional role of P-glycoprotein in limiting intestinal absorption of drugs: Contribution of passive permeability to P-glycoprotein mediated efflux transport. *Mol. Pharmaceutics* **2005**, *2*, 12–21.

(10) Mellor, H. R.; Callaghan, R. Resistance to chemotherapy in cancer: A complex and integrated cellular response. *Pharmacology* **2008**, *81*, 275–300.

(11) Borst, P.; Evers, R.; Kool, M.; Wijnholds, J. A family of drug transporters: The multidrug resistance-associated proteins. *J. Natl. Cancer Inst.* **2000**, *92*, 1295–1302.

(12) Gottesman, M. M.; Ling, V. The molecular basis of multidrug resistance in cancer: The early years of P-glycoprotein research. *FEBS Lett.* **2006**, *580*, 998–1009.

(13) Li, W. X.; Li, L. P.; Eksterowicz, J.; Ling, X. F. B.; Cardozo, M. Significance analysis and multiple pharmacophore models for differentiating P-glycoprotein substrates. *J. Chem. Inf. Model.* **2007**, *47*, 2429–2438.

(14) Ekins, S.; Kim, R. B.; Leake, B. F.; Dantzig, A. H.; Schuetz, E. G.; Lan, L. B.; Yasuda, K.; Shepard, R. L.; Winter, M. A.; Schuetz, J. D.; Wikel, J. H.; Wrighton, S. A. Three-dimensional quantitative structure-activity relationships of inhibitors of P-glycoprotein. *Mol. Pharmacol.* **2002**, *61*, 964–973.

(15) Ekins, S.; Kim, R. B.; Leake, B. F.; Dantzig, A. H.; Schuetz, E. G.; Lan, L. B.; Yasuda, K.; Shepard, R. L.; Winter, M. A.; Schuetz, J. D.; Wikel, J. H.; Wrighton, S. A. Application of three-dimensional quantitative structure-activity relationships of P-glycoprotein inhibitors and substrates. *Mol. Pharmacol.* **2002**, *61*, 974–981.

(16) Wang, Y. H.; Li, Y.; Yang, S. L.; Yang, L. Classification of substrates and inhibitors of P-glycoprotein using unsupervised machine learning approach. *J. Chem. Inf. Model.* **2005**, *45*, 750–757.

(17) Sun, H. M. A naive Bayes classifier for prediction of multidrug resistance reversal activity on the basis of atom typing. *J. Med. Chem.* **2005**, *48*, 4031–4039.

(18) Chang, C.; Bahadduri, P. M.; Polli, J. E.; Swaan, P. W.; Ekins, S. Rapid identification of P-glycoprotein substrates and inhibitors. *Drug Metab. Dispos.* **2006**, *34*, 1976–1984.

(19) Bakken, G. A.; Jurs, P. C. Classification of multidrug-resistance reversal agents using structure-based descriptors and linear discriminant analysis. *J. Med. Chem.* **2000**, *43*, 4534–4541.

(20) Ramu, A.; Ramu, N. Reversal of multidrug resistance by phenothiazines and structurally related compounds. *Cancer Chemother. Pharmacol.* **1992**, *30*, 165–173.

(21) Ramu, A.; Ramu, N. Reversal of multidrug resistance by bis(phenylalkyl)amines and structurally related compounds. *Cancer Chemother. Pharmacol.* **1994**, *34*, 423–430.

(22) Ford, J. M.; Hait, W. N. Pharmacology of Drugs That Alter Multidrug Resistance in Cancer. *Pharmacol. Rev.* **1990**, *42*, 155–199.

(23) Polli, J. W.; Wring, S. A.; Humphreys, J. E.; Huang, L. Y.; Morgan, J. B.; Webster, L. O.; Serabjit-Singh, C. S. Rational use of in vitro P-glycoprotein assays in drug discovery. *J. Pharmacol. Exp. Ther.* **2001**, *299*, 620–628.

(24) SYBYL molecular simulation package, http://www.sybyl.com. 2009.

(25) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: An analysis of ALOGP and CLOGP methods. *J. Phys. Chem. A* **1998**, *102*, 3762–3772.

(26) Tetko, I. V.; Tanchuk, V. Y.; Kasheva, T. N.; Villa, A. E. P. Estimation of aqueous solubility of chemical compounds using E-state indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1488–1493.

(27) Discovery Studio 2.5 Guide, Accelrys Inc., San Diego, http://www.accelrys.com, 2009.

(28) Rogers, D.; Brown, R. D.; Hahn, M. Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up. *J. Biomol. Screening* **2005**, *10*, 682–686.

(29) Pipeline Pilot 7.5, Accelrys Inc., 2009.

(30) Wang, J. M.; Krudy, G.; Xie, X. Q.; Wu, C. D.; Holland, G. Genetic algorithm-optimized QSPR models for bioavailability, protein binding, and urinary excretion. *J. Chem. Inf. Model.* **2006**, *46*, 2674–2683.

(31) Yoshida, F.; Topliss, J. G. QSAR model for drug human oral bioavailability. *J. Med. Chem.* **2000**, *43*, 2575–2585.

(32) Xia, X. Y.; Maliski, E. G.; Gallant, P.; Rogers, D. Classification of kinase inhibitors using a Bayesian model. *J. Med. Chem.* **2004**, *47*, 4463–4470.

(33) Hou, T. J.; Wang, J. M.; Li, Y. Y. ADME evaluation in drug discovery. 8. The prediction of human intestinal absorption by a support vector machine. *J. Chem. Inf. Model.* **2007**, *47*, 2408–2415.

(34) Beresford, A. P.; Selick, H. E.; Tarbit, M. H. The emerging importance of predictive ADME simulation in drug discovery. *Drug Discovery Today* **2002**, *7*, 109–116.

(35) Cruciani, C.; Crivori, P.; Carrupt, P. A.; Testa, B. Molecular fields in quantitative structure-permeation relationships: the VolSurf approach. *J. Mol. Struct.: THEOCHEM* **2000**, *503*, 17–30.

(36) Gola, J.; Obrezanova, O.; Champness, E.; Segall, M. ADMET property prediction: The state of the art and current challenges. *QSAR Comb. Sci.* **2006**, *25*, 1172–1180.

(37) Hou, T. J.; Li, Y. Y.; Zhang, W.; Wang, J. M. Recent Developments of In Silico Predictions of Intestinal Absorption and Oral Bioavailability. *Comb. Chem. High Throughput Screening* **2009**, *12*, 497–506.

(38) Hou, T. J.; Wang, J. M.; Zhang, W.; Wang, W.; Xu, X. Recent advances in computational prediction of drug absorption and permeability in drug discovery. *Curr. Med. Chem.* **2006**, *13*, 2653–2667.

(39) Hou, T. J.; Wang, J. M.; Zhang, W.; Xu, X. J. ADME evaluation in drug discovery. 7. Prediction of oral absorption by correlation and classification. *J. Chem. Inf. Model.* **2007**, *47*, 208–218.

(40) Hou, T. J.; Xia, K.; Zhang, W.; Xu, X. J. ADME evaluation in drug discovery. 4. Prediction of aqueous solubility based on atom contribution approach. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 266–275.

(41) Huuskonen, J. Estimation of water solubility from atom-type electrotopological state indices. *Environ. Toxicol. Chem.* **2001**, *20*, 491–497.

(42) Johnson, S. R.; Zheng, W. F. Recent progress in the computational prediction of aqueous solubility and absorption. *AAPS J.* **2006**, *8*, E27–E40.

(43) Jolivette, L. J.; Ekins, S. Methods for predicting human drug metabolism. *Adv. Clin. Chem.* **2007**, *43*, 131–176.

(44) Liu, R. F.; Sun, H. M.; So, S. S. Development of quantitative structure-property relationship models for early ADME evaluation in drug discovery. 2. Blood-brain barrier penetration. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1623–1632.

(45) Norinder, U.; Bergstrom, C. A. S. Prediction of ADMET properties. *ChemMedChem* **2006**, *1*, 920–937.

(46) Norinder, U.; Haeberlein, M. Computational approaches to the prediction of the blood-brain distribution. *Adv. Drug Delivery Rev.* **2002**, *54*, 291–313.

(47) van de Waterbeemd, H.; Gifford, E. ADMET in silico modelling: Towards prediction paradise? *Nat. Rev. Drug Discovery* **2003**, *2*, 192–204.

(48) Wang, J. M.; Krudy, G.; Hou, T. J.; Zhang, W.; Holland, G.; Xu, X. J. Development of reliable aqueous solubility models and their application in druglike analysis. *J. Chem. Inf. Model.* **2007**, *47*, 1395–1404.

(49) Zhao, Y. H.; Le, J.; Abraham, M. H.; Hersey, A.; Eddershaw, P. J.; Luscombe, C. N.; Boutina, D.; Beck, G.; Sherborne, B.; Cooper, I.; Platts, J. A. Evaluation of human intestinal absorption data and subsequent derivation of a quantitative structure-activity relationship (QSAR) with the Abraham descriptors. *J. Pharm. Sci.* **2001**, *90*, 749–784.

(50) Gleeson, M. P. Generation of a set of simple, interpretable ADMET rules of thumb. *J. Med. Chem.* **2008**, *51*, 817–834.

(51) Ecker, G.; Huber, M.; Schmid, D.; Chiba, P. The importance of a nitrogen atom in modulators of multidrug resistance. *Mol. Pharmacol.* **1999**, *56*, 791–796.

(52) Klopman, G.; Zhu, H.; Ecker, G.; Chiba, P. MCASE study of the multidrug resistance reversal activity of propafenone analogs. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 291–297.