DOI: 10.1002/cmdc.200500047

An Accurate and Interpretable Bayesian Classification Model for Prediction of hERG Liability

Hongmao Sun*[a]

Drug-induced QT interval prolongation has been identified as a critical side-effect of non-cardiovascular therapeutic agents and has resulted in the withdrawal of many drugs from the market. As almost all cases of drug-induced QT prolongation can be traced to the blockade of a voltage-dependent potassium ion channel encoded by the hERG (the human ether-à-go-go-related gene), early identification of potential hERG channel blockers will decrease the risk of cardiotoxicity-induced attritions in the later and more expensive development stage. Presented herein is a naive Bayes classifier to categorize hERG blockers into active and inactive classes, by using a universal, generic molecular descriptor system.[1] The naive Bayes classifier was built from a training set containing 1979 corporate compounds, and exhibited an ROC accuracy of 0.87. The model was validated on an external test set of 66 drugs, of which 58 were correctly classified. The cumulative probabilities reflected the confidence of prediction and were proven useful for the identification of hERG blockers. Relative performance was compared for two classifiers constructed from either an atom-type-based molecular descriptor or the long range functional class fingerprint descriptor FCFP_6. The combination of an atom-typing descriptor and the naive Bayes classification technique enables the interpretation of the resulting model, which offers extra information for the design of compounds free of undesirable hERG activity.

Introduction

The QT interval of the electrocardiogram (ECG) is a widely used measure of the ventricular repolarization process. The prolongation of the QT interval is associated with an increased risk of torsades de pointes (TdP). Once considered a trivial physiological finding, drug-induced long QT syndrome has been identified as a critical side effect of non-cardiovascular drugs and has caused the withdrawal of many drugs from the market.[2] Further investigation has placed focus upon a voltage-dependent potassium ion channel, encoded by the hERG (the human ether-à-go-go-related gene),[3] as almost all cases of drug-induced QT prolongation can be traced to the blockade of hERG. As drug-induced QT prolongation increases the likelihood of TdP, which may cause sudden death, the current regulatory guidelines scrutinize all hERG blockers to determine whether they cause QT prolongation.[4] Furthermore, unlike other ion channels that interact only with structurally specific ligands, the hERG potassium ion channel can be blocked by a broad spectrum of structurally diverse drugs. Therefore, removal of potential hERG blockers from the drug-discovery pipeline is an important issue for projects across all therapeutic areas.[5] To include hERG information on a compound early in the decision-making process, many in vitro and in silico methods have been developed to estimate hERG activity.[6] If hERG blockers can be identified at the early stages of drug development, the focus of research resources may be directed to druglike compounds without potential hERG liability.

Structural approaches to hERG blockade include the efforts to understand molecular recognition from both the protein side, by solving crystal structures of potassium ion channels or constructing homology models, and the ligand side, through pharmacophore models and QSAR (quantitative structure–activity relationship) models. Several crystal structures of potassium ion channels have been solved recently, covering voltage-gated, calcium-gated, and families of inwardly rectifying potassium channels.[7–10] These structures brought us closer to fully understanding potassium ion channel function, but did not shed much light upon the puzzle of why the hERG channel can accept molecules of all different chemotypes. Some key residues involved in the binding of hERG blockers have been identified through mutagenesis approaches and homology modeling.[11–13] Homology models constructed from the closed-state crystal structure of a K^+ channel from *Streptomyces lividans* (KcsA) clearly revealed a ligand-binding site that is too small to accommodate some of the known hERG blockers. More recently, Reynolds and co-workers developed multiple homology models of the hERG potassium ion channel on the basis of both the closed-state structure of KcsA and the opened-state structure of a K^+ channel from *Methanobacterium thermoautotrophicum* (MthK), attempting to address the flexibility of the hERG channel.[14] The opened state represented a bigger cavity through a bending of the S6 helix at the GLY hinge.

[a] *Dr. H. Sun*
Discovery Chemistry, Hoffmann–La Roche, Inc.
340 Kingsland Street
Nutley, NJ 07110 (USA)
Fax: (+1) 973-235-6084
E-mail: hongmao.sun@roche.com

Pharmacophore models proposed to date are quite similar. Cavalli et al. derived a four-point pharmacophore model, with one basic nitrogen atom surrounded by three aromatic moieties, based on the analysis of 31 QT-prolonging drugs.[15] Although the model gave a reasonable prediction for six drugs in a validation set, pharmacophore models, in general, performed well only on analogues structurally similar to those in the training set, which limits their application to the problems involving structurally diverse hERG blockers. The "general hERG pharmacophore model" developed by Ekins and co-workers was generated by using 15 compounds selected from the literature that consisted of four hydrophobes and one positive charge.[16] Again, the model was not general enough to be applicable to different compound series. Alternatively, a comparative molecular similarity analysis (CoMSiA) was performed by Pearlstein and co-workers for a data set containing 22 sertindole analogues and 10 structurally diverse hERG inhibitors.[17] Their CoMSiA and homology models revealed some detailed intermolecular interactions between hERG blockers and the ion channel. Roche et al. applied various techniques, including substructure analysis, self-organizing maps (SOM), principal component analysis (PCA), partial least squares regression (PLS), and supervised artificial neural network (ANN) to model hERG activities of 472 druglike compounds.[18] The ANN model correctly classified 93% of the hERG-inactive agents and 71% of the hERG channel blockers in a 95-compound test set. Keserü and co-workers collected 68 druglike compounds from Roche's data set and Fenichel's list,[19] and both 5-parameter traditional QSAR and 6-component hologram QSAR (HQSAR) models gave reasonable predictions of the whole data set.[20]

Judging from the fact that the hERG channel can accommodate a wide spectrum of structurally diverse compounds, hERG blockade might be a multiple mechanism problem. Even though it has been widely accepted that hERG blockers interact directly with the transmembrane domain of the channel, as supported by mutagenesis evidence, there is yet no observation which can exclude the involvement of the PAS domain at the N terminus of the channel in interactions with hERG blockers. Multiple mechanisms make it challenging to construct a global model for the prediction of hERG activity, but data-driven QSAR modeling simplifies the problem by projecting from the problem domain to the chemical domain, avoiding the need to invoke specific mechanisms.[21] For an attempt at this kind of global model, the availability of a large and structurally diverse training set is a preliminary requirement. In the study reported herein, a large corporate data set was employed for construction of the hERG classification model.

Methods

Data sets

The training set consisted of 1979 compounds from Roche corporate compound library with measured hERG activities. The compounds were selected from more than 30 projects to maximize the structural diversity. The average Tanimoto distance[22] of the training set was computed to be 0.233, with the pairwise Tanimoto distances ranging from 0.013 to 0.989. As shown in Figure 1, the distributions of molecular weight, calculated log *P*, and the counts of hydrogen-bond donors and acceptors illustrate that the compounds in the training data set aligned well with the chemical space of druglike compounds. To measure hERG activity, Chinese hamster ovary (CHO) cells stably expressing the hERG cardiac potassium channel were used. Standard whole-cell patch-clamp electrophysiology was performed to record hERG channel currents from these cells at 35–37 °C. Cells were held at a resting voltage of −80 mV and then stimulated by a voltage pattern to activate hERG channels and conduct IK_{hERG} current (both inward and outward). After the fibers stabilized for a few minutes, the amplitude and kinetics of IK_{hERG} were recorded at a stimulation frequency of 0.1 Hz. The test compounds were added to the cells in ascending concentrations, and apparent IC_{50} values for hERG channel inhibition were calculated.

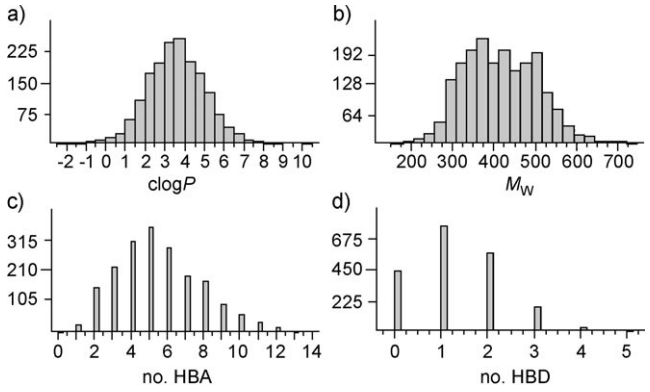

Figure 1. Histograms of the molecular properties of the 1979 compounds in the training set: a) calculated log *P*, b) molecular weight, c) hydrogen-bond acceptors (HBA), and d) hydrogen-bond donors (HBD).

The test set used in this study comprised 66 drugs reported in the paper by Keserü after removing the two duplicate compounds A-56268 and Hismanal.[20] All IC_{50} values for the inhibition of hERG potassium ion channels were measured under experimental conditions similar to those of the training set.

Like many other toxicity endpoints, a qualitative description of the hERG activity of a compound is often sufficient for decision-making. In the laboratory, more accurate measurement generally requires more time and greater compound supply, which significantly slows down the process. On the other hand, in silico models based on qualitative data are more tolerant of uncertainty in the training data. In this study, compounds were classified into hERG positives and negatives according to their hERG activities or IC_{50} values. The overriding question is at which IC_{50} value a compound is considered free of hERG liability. There exists no universal answer to this question. Considering the possibility of accidental overdose, a compound is considered safe if its hERG activity, as measured by IC_{50}, is at least 10- to 30-fold higher than the anticipated plasma or tissue concentration necessary for its therapeutic activity.[5] Assuming the required plasma concentration of a moderately potent drug to be 1 μM, the threshold for hERG safety

should be set at 30 μM. Although there are compounds with IC_{50} values of over 30 μM that are cardiotoxic, most compounds that reach this threshold have been found to be safe.[20,23] With 30 μM as the cutoff activity, 307 compounds in the training set were set as hERG negative, the rest as positive.

Atom-type classification

The typing of an atom, assigned according to its own chemical properties and the neighboring atoms and bonds, is not only straightforward but interpretable. As described in a previous paper,[1] atom types were assigned to each atom through a classification tree. The key steps towards an accurate classification tree are to determine where to split and where to stop splitting the tree. To make sure each atom type can reasonably reflect its chemical environment, the classification tree was trained by optimizing the log*P* predictions of the compounds in Starlist. Starlist is a high-quality data set, which contains nearly 11 000 structurally diverse compounds. The optimized atom types have been proven applicable to deriving models for the prediction of different molecular properties.[1,24] Details on the method of atom-type classification were described in reference [1].

Naive Bayes classifier

The theory of naive Bayes classification has been described in detail in a previous paper,[24] thus, only some key points are summarized herein. Using atom types as molecular descriptors, each molecule can be represented as a vector $a=\langle a_1, a_2, \ldots, a_n\rangle$, for which $a_1, a_2, \ldots, a_n$ are the occurrences of the atom types $A_1, A_2, \ldots, A_n$. The probability of a new compound belonging to a certain class *C*, say being hERG positive, can be expressed as $P(C=+|A_1=0, A_2=0, \ldots, A_i=1, \ldots, A_j=1, \ldots, A_n=0)$, in which A_i-A_j corresponds to the atom types of the compound.

According to Bayes' theorem,[25]

$$P(+|A_1, A_2, \ldots, A_n.) = \frac{P(A_1, A_2, \ldots, A_n|+)P(+)}{P(A_1, A_2, \ldots, A_n)} \quad (1)$$

The three probabilities on the right side of Equation [1] can be learned from a training set that contains a number of compounds with known hERG activities. The prior probability $P(+)$ is simply the percentage of positive compounds in the training set, whereas the marginal probability $P(A_1, A_2, \ldots, A_n)$ can be ignored, as it is the same to all classes. Therefore, the problem is simplified to estimating $P(A_1, A_2, \ldots, A_n|+)$.

By using Bayes' theorem recursively and assuming that each atom type is conditionally independent of every other atom type, we get Equation [2].

$$P(A_1, A_2, \ldots, A_n|+) = \prod_{i=1}^{n} P(A_i = a_i|+) \quad (2)$$

This is a key decoupling step in which the molecule expressed as a vector of atom types in the conditional probability is broken down to individual atoms on the right side of the equation. Now each factor in the product can be easily estimated from a training set:

$$P(A_i = a_i|+) = \frac{\text{count}(A_i = a_i \cap C = +)}{\text{count}(C = +)} \quad (3)$$

The marginal probability can be cancelled out by simple mathematical operation. Let $p_- = P(C=-)$ and $p_+ = P(C=+)$, let $p_{i-} = P(A_i=a_i|C=-)$ and $p_{i+} = P(A_i=a_i|C=+)$, then

$$p = P(C = +|A_1 = a_1, A_2 = a_2, \ldots, A_n = a_n) = \left(\prod_{i=1}^{n} p_{i+}\right)\frac{p_+}{z} \quad (4)$$

and

$$q = P(C = -|A_1 = a_1, A_2 = a_2, \ldots, A_n = a_n) = \left(\prod_{i=1}^{n} p_{i-}\right)\frac{p_-}{z} \quad (5)$$

for which *z* is marginal probability, a constant. As $p+q=1$, then

$$\log\frac{p}{q} = \log\frac{p}{1-p} = \sum_{i=1}^{n}(\log p_{i+} - \log p_{i-}) + (\log p_+ - \log p_-) \quad (6)$$

Here, marginal probability is cancelled, and *p* can be evaluated by exponentiating both sides and rearranging the terms.

There are only a couple of practical issues remaining to be addressed: zero counts and missing values. Zero counts were generally overcome by using Laplace correction. For a 2-class problem, the Laplace corrected $P(A_i=a_i|C=+)$ could be expressed $(\text{count}(A_i=a_i|C=+) + 0.5)/(\text{count}(C=+) + 1.0)$, as adopted by the program Pipeline Pilot.[26] Missing atom types are ignored to avoid introducing unproven information.

SciTegic's Pipeline Pilot (version 3.0) was used to perform the naive Bayes classification.[26]

Results and Discussion

Atom-type classification

After being trained by log*P*,[1] 218 atom types were identified by the atom-type classification tree. Twenty seven correction factors introduced in this study were the same as those used in reference [1]. Finally, a 246×1979 matrix in which the last column is hERG activity membership (positive/negative) was prepared as the input for naive Bayes analysis.

Naive Bayes classifier

The activities of hERG were classified into two classes: hERG-positive compounds with pIC_{50} values greater than 4.52, and

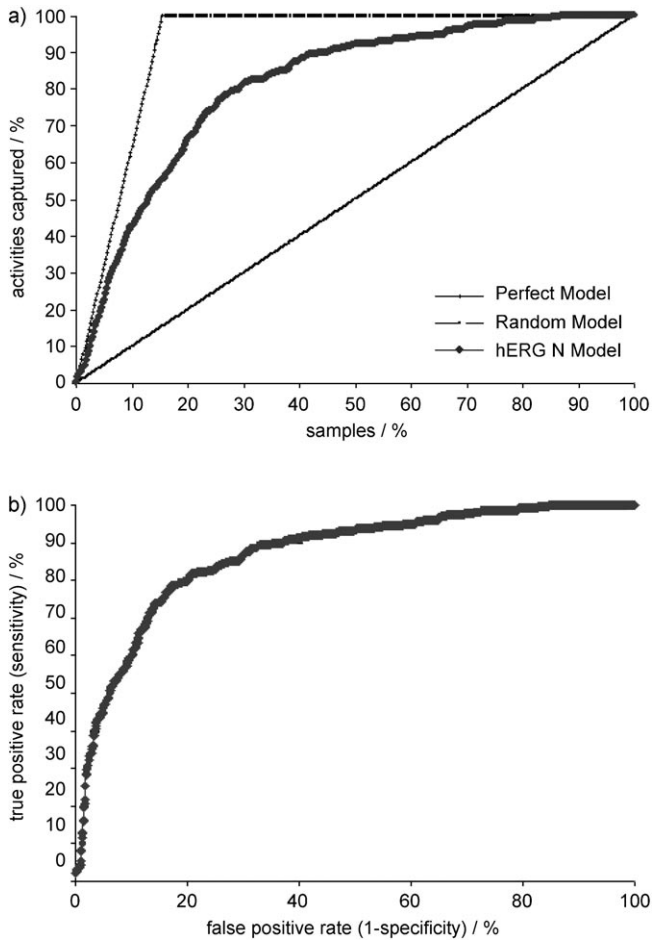

Figure 2. a) Enrichment and b) ROC plots of the atom-typing model for the 1979-compound training set.

hERG-negative compounds with pIC_{50} values less than or equal to 4.52. By setting the less populated hERG-negative compounds (15.5%) as "good" samples and the highly populated hERG-positive compounds as "background", the classifier derived from the 1979-compound training set gave a model with receiver operating characteristic (ROC) accuracy of 0.87. Figure 2 depicts the enrichment plot and ROC plot of the model. The enrichment plot illustrates how fast all the hERG-negative compounds could be identified if the compounds were resorted according to the model. An enrichment curve close to the perfect model is a good indication of the high prioritization power of the model. In this model, 50% of hERG-negative compounds would be found if only 12% of the compounds were tested, compared with 7.75% of a perfect model and 50% of a random model.

Although the naive Bayes classifier is known to be optimal when attributes are independent given the class, it has been illustrated from the analysis of a large amount of both artificial and real-world data that naive Bayes classifiers performed surprisingly well, even when the independence assumption was seriously violated.[27] It has also been proven that the probability estimates of the naive Bayes classifier were only optimal under quadratic loss if the independence assumption held, whereas the classifier itself could be optimal under zero–one loss even when this assumption was violated by a wide margin.[24,27–30] In practice, the naive Bayes classifier has been demonstrated to outperform other sophisticated classifiers in text characterization and anti-spam filtering, and has recently started to be applied in drug discovery.[31–33] In this study, a traditional naive Bayes classification was carried out, in which each atom type contributed equally and objectively in determining the final class membership of a compound. The high ROC accuracy and enrichment performance indicated that highly correlated atom-type descriptors did not have a clear negative impact on the performance of the classifier.

The way that the naive Bayes classifier treated the occurrence of an atom type in a molecule was qualitative. In other words, an atom type occurring three times in a molecule was a subfeature, and the same atom type occurring six times was another independent sub-feature; therefore, three and six lost their numerical meanings. This qualitative treatment of atom types reflected the real situation implied by atom typing, but resulted in more missing values. As missing value means missing information in prediction, a good model is always based on a large and structurally diverse training set with minimal negative effect of missing values. Not surprisingly, a naive Bayes classifier tends to give better prediction to the compounds similar to those in the training set, owing to the minimized missing value effect.

Validation of the model

To validate its predictive performance, the model was applied to predict an independent external data set. Table 1 lists the calculated cumulative possibilities of the 66 compounds in the test set, together with their experimentally determined hERG activities and class membership. It turned out that 58 out of 66 drugs, or 87.9%, in the test set were correctly classified according to their cumulative probabilities. Encouragingly, the 41 most hERG-active drugs with $IC_{50} < 5$ μM were all correctly predicted by the model. Table 1 was sorted according to the cumulative probabilities, which could be considered as the confidence level of predictions. The first 43 drugs, most likely to be hERG positive based on prediction, were all shown to be hERG positive by the assay. Similarly, if only the last one-third of the drugs were to be selected for further development, the set would include all the hERG-negative drugs, except cetirizine, with an IC_{50} value of 30 μM, which is exactly the cutoff concentration. Further inspection of the eight misclassified drugs indicated that their measured activities, expressed as the pIC_{50} values, were mostly between 4.0 and 5.0, except Nicotine (3.61) and MDL-74156 (5.23). In other words, the hERG activities of these six misclassified drugs were within a threefold difference from the cutoff activity of 30 μM. Even the most sophisticated hERG assay system cannot guarantee to control the experimental error within a threefold scale; thus, in this sense, the model was very predictive. Misclassification of nicotine might be a result of its low molecular weight of 162.26, as there were very few compounds in the training set that had a molecular weight less than 200 (Figure 1). It became clear by investigation of the ligand-binding site of the hERG potassium

Table 1. Cumulative possibilities of the 66 compounds in the test set computed from the atom-typing model together with their experimentally determined hERG activities and class membership (sorted according to prediction).

ID	Name	pIC_{50}	Class	Prediction	ID	Name	pIC_{50}	Class	Prediction
46	quinidine	6.49	P	−11.4969	19	diphenhydramine	4.57	P	−4.1074
53	thioridazine	6.44	P	−10.8210	43	ondansetron	6.09	P	−4.0533
34	mesoridazine	6.49	P	−10.6060	33	loratadine	6.77	P	−4.0385
14	clozapine	6.49	P	−10.1042	5	azimilide	5.85	P	−3.9157
3	amitriptyline	5.00	P	−9.5267	60	flecainide	5.41	P	−3.7638
58	desmethylastemizole	9.00	P	−9.1441	49	sildenafil	5.48	P	−3.2451
51	terfenadine	6.70	P	−9.0785	2	amiodarone	5.00	P	−3.2033
10	chlorpromazine	5.83	P	−8.3113	7	carvedilol	4.98	P	−3.0620
27	halofantrine	6.70	P	−8.3030	9	chlorpheniramine	4.68	P	−2.8490
28	haloperidol	7.52	P	−8.2536	31	ketoconazole	5.72	P	−2.8310
52	terikalant	6.60	P	−7.9100	8	cetirizine	4.52	N	−2.7321
67	RP-58866	6.70	P	−7.9100	24	epinastine	4.00	N	−2.5712
41	norclozapine	5.35	P	−7.7977	55	vesnarinone	5.96	P	−2.3463
66	olanzapine	6.74	P	−7.7080	29	ibutilide	8.00	P	−2.2196
59	droperidol	7.49	P	−7.5467	38	nicotine	3.61	N	−1.9022
61	fluoxetine	5.82	P	−7.5254	1	alosetron	5.49	P	−1.3344
57	citalopram	5.40	P	−7.5125	44	perhexiline	5.11	P	−1.3179
16	ziprasidone	6.92	P	−7.3375	15	cocaine	5.14	P	−0.5110
4	astemizole	8.00	P	−7.0462	20	disopyramide	4.04	N	−0.3536
30	imipramine	5.47	P	−7.0289	18	diltiazem	4.76	P	−0.0886
64	mefloquine	5.25	P	−6.9943	63	MDL-74156	5.23	P	2.7902
17	desipramine	5.86	P	−6.8341	11	ciprofloxacin	3.02	N	3.0836
6	bepridil	6.26	P	−6.2496	40	nitrendipine	5.00	P	3.1676
54	verapamil	6.85	P	−6.1638	42	ofloxacin	2.85	N	3.4216
35	mibefradil	5.84	P	−6.0297	32	levofloxacin	3.04	N	3.4216
65	norastemizole	7.55	P	−5.7983	68	trimethoprim	3.62	N	3.5334
45	pimozide	7.30	P	−5.7970	22	dolasetron	4.92	P	4.5953
23	E4031	7.70	P	−5.6031	26	grepafloxacin	4.30	N	4.7111
21	dofetilide	8.00	P	−5.4991	37	moxifloxacin	3.89	N	4.9413
48	sertindole	8.00	P	−4.9358	39	nifedipine	4.30	N	5.1602
36	mizolastine	6.36	P	−4.8675	25	gatifloxacin	3.89	N	5.3897
47	risperidone	6.82	P	−4.3696	50	sparfloxacin	4.74	P	5.8530
12	cisapride	7.40	P	−4.2547	13	clarithromycin	4.23	N	15.6526

ion channel and the published pharmacophore models that small molecules like nicotine were too small to bind tightly to the channel. MDL-74156 was a close analogue of another misclassified drug dolasetron, and both contained a rarely occurring fused ring system, as shown in Figure 3, which was not well represented by the training set. Both drugs were weak hERG blockers, but misclassified as hERG negative.

Comparison with fingerprint descriptors

As an unsupervised learner, naive Bayes classifier has no fitting process and no tuning parameters, making it an appropriate algorithm for comparing the effectiveness of different molecular descriptor systems. As FCFP_6 is a well-known fingerprint type molecular descriptor system supplied by Pipeline Pilot, it is interesting to compare the performance of relatively low-dimensional atom-type-based descriptors to that of fingerprint-based descriptors

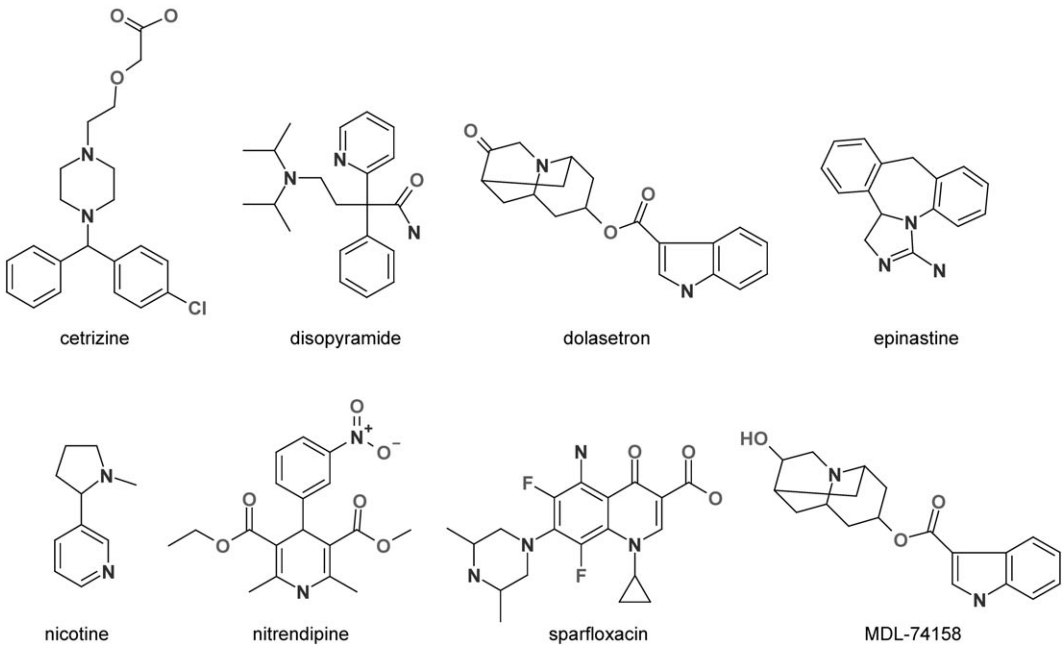

Figure 3. Structures of the eight misclassified drugs in the test set.

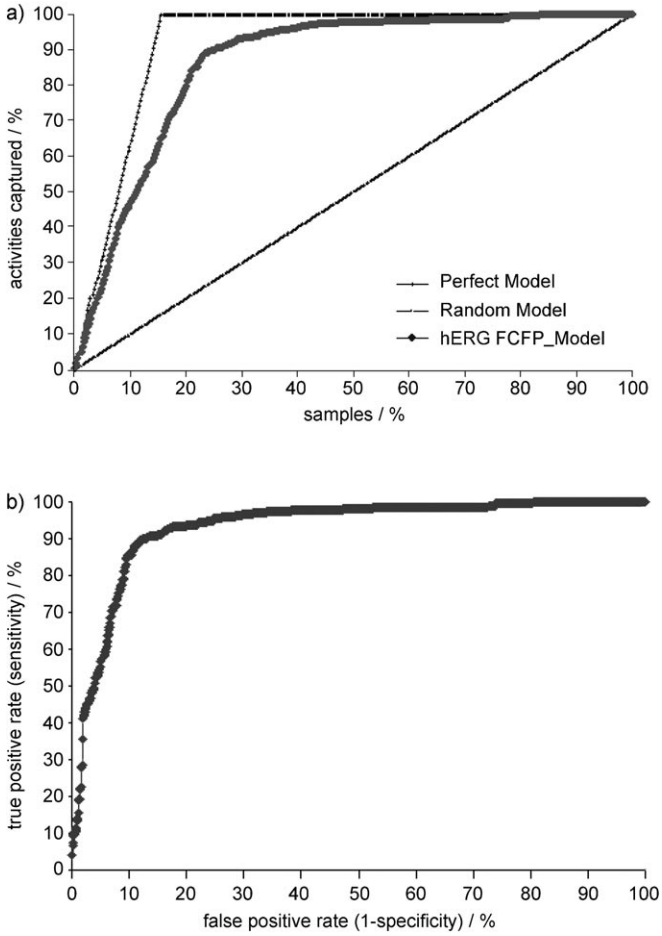

Figure 4. a) Enrichment and b) ROC plots of the FCFP_6 model for the 1979-compound training set.

of high dimension. The classifier built from FCFP_6, together with other physicochemical properties, including *Alog P*, molecular weight, number of hydrogen-bond donors and acceptors, and number of rotatable bonds, gave an excellent ROC accuracy of 0.93, which was significantly better than the atom-typing model (Figure 4). However, the predictive power of both models was similar, as evaluated from the success rate of predicting the hERG activities of the 66-compound testing set. Interestingly, six drugs were misclassified by both classifiers, namely cetirizine, nicotine, disopyramide, epinastine, nitrendipine, and sparfloxacin (Table 2). FCFP_6 model correctly predicted the pair of the similar analogues of dolasetron and MDL-74156, but failed to categorize sildenafil.

Model interpretation

A major difference between the two descriptor systems mentioned above is the model interpretability. Although they carry more specific structural information, high-dimensional fingerprint-based molecular descriptors tend to lose the interpretation of chemical features. In contrast, atom-type-based descriptors are capable of extracting fragmental information understandable to chemists. The assignment of a particular atom type depends on not only its own properties, but also its neighborhood environment; therefore, atom types often implicitly carry fragmental information. On the other hand, atom typing does not explicitly predefine any fragments, and this decreases the possibility of introducing bias at the beginning of the model construction and enables learners to identify structural features that are not well-defined by usual fragments.

The discriminating power of an atom type is determined by at least two factors, namely the uneven occurrence of the atom type across different classes and its total occurrence in a training set. Table 3 lists the most important atom types and correction factors in terms of discerning power. Accordingly, acidic groups abolish hERG activity, whereas basic groups such as piperidines and piperazines are warning signals for hERG liability. All 11 amino acids (M7), which were not close analogues, were hERG negative. Similarly, 25 out of 28 acids containing an acidic oxygen atom O6 were hERG negative. The observation that acidic groups abolish hERG activity was successfully applied to remove hERG affinity from the drug terfenadine and resulted in a close analogue fexofenadine which was hERG-free.[34] Many pharmacophore models suggested that a positive-charge center was essential for hERG affinity of a compound,[6] which might be true for hERG blockers effective in the nM range, but hardly held for μM-range hERG inhibitors. Piperidine and piperizine are the most frequently used functional groups to introduce a positive-charge center to a molecule. Indeed, only 24 out of 522 compounds containing one N16 atom, a charged aliphatic cyclic nitrogen, were hERG negative; within 38 compounds carrying two or three N16 atoms, only one was hERG negative. However, neutral nitrogen atoms in an aliphatic ring also played an important role in affecting the hERG activity of a compound. N15, an aliphatic cyclic nitrogen atom next to a carbonyl group, appeared in 61 compounds in the training set, but only one of these was hERG negative.

The number of aromatic rings seemed to be a good predictor of hERG activity. Ninety-four compounds had less than three unsubstituted aromatic carbon atoms (C3), 49 of which were hERG negative, whereas only two out of 32 compounds with more than 13 C3 atoms fell into the same category. All 43 compounds with four C4 atoms, implying two biphenyl groups in a molecule, were hERG positive. Another atom type with strong discerning power was C68, a carboxylic carbon atom attached to any aromatic atom. Fifty out of 77 compounds containing C68 were hERG negative, a ratio that was significantly higher than 15.5% hERG negatives in the whole training set. C7, an aromatic carbonyl carbon atom, appeared in seven compounds, six of which were hERG negative. There were 200 compounds in the training set carrying one C55 atom, a tertiary carbon atom bonded to an aromatic ring; only 13 of these compounds were hERG negative. Thirty-one compounds containing two C55 atoms were all hERG positive. The observation implied that compounds which branched immediately after an aromatic ring tended to be hERG blockers.

Like other machine learning methods, the naive Bayes classifier cannot learn anything beyond the training data set. As

Table 2. Cumulative possibilities of the 66 compounds in the test set computed from the FCFP_6 model together with their experimentally determined hERG activities.

ID	pIC_{50}	Name	Prediction	ID	pIC_{50}	Name	Prediction
1	5.49	alosetron	−8.1128	34	6.49	mesoridazine	−20.2439
2	5.00	amiodarone	−17.2389	35	5.84	mibefradil	−15.0664
3	5.00	amitriptyline	−10.4950	36	6.36	mizolastine	−13.3043
4	8.00	astemizole	−27.0853	37	3.89	moxifloxacin	34.0105
5	5.85	azimilide	−16.6013	38	3.61	nicotine	−5.1061
6	6.26	bepridil	−17.4115	39	4.30	nifedipine	13.2423
7	4.98	carvedilol	−8.2639	40	5.00	nitrendipine	11.9937
8	4.52	cetirizine	−11.3888	41	5.35	norclozapine	−11.2082
9	4.68	chlorpheniramine	−9.7790	42	2.85	ofloxacin	24.8823
10	5.83	chlorpromazine	−14.4107	43	6.09	ondansetron	−15.1415
11	3.02	ciprofloxacin	18.2765	44	5.11	perhexiline	−2.8562
12	7.40	cisapride	−19.3406	45	7.30	pimozide	−31.0103
13	4.23	clarithromycin	60.0559	46	6.49	quinidine	−18.1679
14	6.49	clozapine	−18.4460	47	6.82	risperidone	−14.4783
15	5.14	cocaine	−5.9731	48	8.00	sertindole	−30.5150
16	6.92	ziprasidone	−18.5950	49	5.48	sildenafil	6.5217
17	5.86	desipramine	−5.3075	50	4.74	sparfloxacin	14.9438
18	4.76	diltiazem	−6.8601	51	6.70	terfenadine	−16.0668
19	4.57	diphenhydramine	−10.3002	52	6.60	terikalant	−12.3389
20	4.04	disopyramide	−3.1246	53	6.44	thioridazine	−19.0038
21	8.00	dofetilide	−11.9990	54	6.85	verapamil	−11.4427
22	4.92	dolasetron	−11.5793	55	5.96	vesnarinone	−21.0527
23	7.70	E4031	−14.2971	57	5.40	citalopram	−15.1844
24	4.00	epinastine	−2.0248	58	9.00	desmethylastemizole	−34.5987
25	3.89	gatifloxacin	32.2819	59	7.49	droperidol	−28.4837
26	4.30	grepafloxacin	24.8195	60	5.41	flecainide	−8.7831
27	6.70	halofantrine	−21.9482	61	5.82	fluoxetine	−19.1338
28	7.52	haloperidol	−29.1959	63	5.23	MDL-74156	−11.8898
29	8.00	ibutilide	−11.2523	64	5.25	mefloquine	−7.1699
30	5.47	imipramine	−12.5002	65	7.55	norastemizole	−10.9182
31	5.72	ketoconazole	−8.4830	66	6.74	olanzapine	−17.3489
32	3.04	levofloxacin	24.8823	67	6.70	RP-58866	−12.3389
33	6.77	loratadine	−6.6196	68	3.62	trimethoprim	10.7152

Table 3. The atom types and correction factors showing strong discerning power.

Atom Type[a]	Occurrence in Molecule	Normalized Probability	Feature Count[b]	Subset Count[c]	Atom Type	Occurrence in Molecule	Normalized Probability	Feature Count[b]	Subset Count[c]
O6	1	1.582	28	25	C2	4	−1.449	21	0
M7	1	1.489	11	11	N30	1	−1.422	47	1
M8	1	1.424	82	56	N9	2	−1.373	19	0
M2	2	1.379	65	43	C46	5	−1.333	18	0
C68	1	1.371	77	50	F5	1	−1.307	65	2
C7	1	1.211	7	6	M3	6	−1.293	64	2
C3	0, 1	1.210	32	19	F2	6	−1.293	64	2
O3	2	1.206	9	7	M4	2	−1.291	17	0
C4	4	−2.037	43	0	C18	2	−1.249	61	2
N3	3	−1.836	34	0	C70	1	−1.248	16	0
C55	2	−1.759	31	0	N4	2	−1.238	38	1
N15	1	−1.625	59	1	N16	2	−1.215	37	1
C71	1	−1.485	22	0	C58	1	−1.202	15	0

[a] The definition of atom types and correction factors is the same as in reference [1]. [b] Feature count represents the number of molecules in the training set with the indicated feature. [c] Subset count is the number of hERG-negative compounds with the same feature.

training sets are not perfect, the resulting classifiers can only be as good as the training sets. As a result, one needs to be very cautious in drawing any conclusion from analysis of the statistics table to avoid over-interpretation. For example, all 34 compounds in the training set with exactly three N3, an aromatic nitrogen atom next to another aromatic nitrogen atom, were hERG positive, but there was still no evidence showing that triazole-containing compounds were all hERG blockers.

Traditional naive Bayes classification, as applied in this study, is very objective, in which each atom type contributed equally to the determination of the membership of a new compound and missing values were ignored. There are at least two modifications which might result in an improved model. 1. If the contributions of different atom types were weighted to optimize the prediction of the training set, a better model would be reached. However, introduction of extra variables will inevitably decrease the predictive power of the model. 2. Instead of ignoring missing values, using the adjusted probability of its closest neighbors could minimize information loss.

Conclusion

The application of a universal molecular descriptor system has been extended to the field of hERG activity prediction. The naive Bayes classifier built from a training set containing 1979 corporate compounds was predictive with an ROC accuracy of 0.87. In the external validation test, the classifier correctly categorized 87.9% of a 66-drug testing set, whereas the cumulative probabilities were useful in selecting the compounds with high hERG activities. Model interpretation through a statistics table identified the specific atom types and fragments that contribute most significantly to hERG affinities. This information can be used as a guideline for the rational design of compounds free of hERG liability.

Acknowledgements

The author is grateful to the MDO team at Hoffmann–La Roche for measuring the hERG activities of the training compounds. The author is also greatly indebted to Dr. David Fry and Dr. Sung-Sau So for their critical reading of the manuscript and insightful suggestions.

Keywords: atom-typing molecular descriptors · hERG · molecular modeling · naive Bayes classification · structure–activity relationships

[1] H. Sun, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 748 – 757.
[2] R. Pearlstein, R. Vaz, D. Rampe, *J. Med. Chem.* **2003**, *46*, 2017 – 2022.
[3] W. S. Redfern, L. Carlsson, A. S. Davis, W. G. Lynch, I. MacKenzie, S. Palethorpe, P. K. Siegl, I. Strang, A. T. Sullivan, R. Wallis, A. J. Camm, T. G. Hammond, *Cardiovasc. Res.* **2003**, *58*, 32 – 45.
[4] B. Fermini, A. A. Fossa, *Nat. Rev. Drug Discov.* **2003**, *2*, 439 – 447.
[5] W. Crumb, I. I. Cavero, *Pharm. Sci. Technol. Today* **1999**, *2*, 270 – 280.
[6] M. Recanatini, E. Poluzzi, M. Masetti, A. Cavalli, F. De Ponti, *Med. Res. Rev.* **2004**, *25*, 133 – 166.
[7] Y. Jiang, A. Lee, J. Chen, V. Ruta, M. Cadene, B. T. Chait, R. MacKinnon, *Nature* **2003**, *423*, 33 – 41.
[8] Y. Jiang, A. Pico, M. Cadene, B. T. Chait, R. MacKinnon, *Neuron* **2001**, *29*, 593 – 601.
[9] A. Kuo, J. M. Gulbis, J. F. Antcliff, T. Rahman, E. D. Lowe, J. Zimmer, J. Cuthbertson, F. M. Ashcroft, T. Ezaki, D. A. Doyle, *Science* **2003**, *300*, 1922 – 1926.
[10] J. M. Gulbis, M. Zhou, S. Mann, R. MacKinnon, *Science* **2000**, *289*, 123 – 127.
[11] P. L. Smith, T. Baukrowitz, G. Yellen, *Nature* **1996**, *379*, 833 – 836.
[12] J. A. Sanchez-Chapula, T. Ferrer, R. A. Navarro-Polanco, M. C. Sanguinetti, *Mol. Pharmacol.* **2003**, *63*, 1051 – 1058.
[13] J. S. Mitcheson, J. Chen, M. Lin, C. Culberson, M. C. Sanguinetti, *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 12329 – 12333.
[14] R. Rajamani, B. A. Tounge, J. Li, C. H. Reynolds, *Bioorg. Med. Chem. Lett.* **2005**, *15*, 1737 – 1741.
[15] A. Cavalli, E. Poluzzi, F. De Ponti, M. Recanatini, *J. Med. Chem.* **2002**, *45*, 3844 – 3853.
[16] S. Ekins, W. J. Crumb, R. D. Sarazan, J. H. Wikel, S. A. Wrighton, *J. Pharmacol. Exp. Ther.* **2002**, *301*, 427 – 434.
[17] R. A. Pearlstein, R. J. Vaz, J. Kang, X. L. Chen, M. Preobrazhenskaya, A. E. Shchekotikhin, A. M. Korolev, L. N. Lysenkova, O. V. Miroshnikova, J. Hendrix, D. Rampe, *Bioorg. Med. Chem. Lett.* **2003**, *13*, 1829 – 1835.
[18] O. Roche, G. Trube, J. Zuegge, P. Pflimlin, A. Alanine, G. Schneider, *ChemBioChem* **2002**, *3*, 455 – 459.
[19] R. R. Fenichel, http://www.fenichel.net/pages/Professional/subpages/QT/Tables/pbydrug.htm.
[20] G. M. Keserü, *Bioorg. Med. Chem. Lett.* **2003**, *13*, 2773 – 2775.
[21] H. Sun, *Curr. Comput.-Aided Drug Des.* **2005**, *1*, 179 – 193.
[22] J. W. Godden, L. Xue, J. Bajorath, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 163 – 166.
[23] F. De Ponti, E. Poluzzi, A. Cavalli, M. Recanatini, N. Montanaro, *Drug Safety* **2002**, *25*, 263 – 286.
[24] H. Sun, *J. Med. Chem.* **2005**, *48*, 4031 – 4039.
[25] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*; 2nd ed., Springer, New York, **1993**.
[26] *Pipeline Pilot* http://www.scitegic.com.
[27] P. Domingos, M. J. Pazzani in *13th International Conference on Machine Learning*, Bari, Italy, **1996**, pp. 105–112.
[28] I. Rish, *An Empirical Study of the Naive Bayes Classifier*, IBM T. J. Watson Research Center, **2001**.
[29] P. Domingos, M. J. Pazzani, *Mach. Learn.* **1997**, *29*, 103 – 130.
[30] H. Sun, *J. Med. Chem.* **2005**, *48*, 4031 – 4039.
[31] A. Bender, H. Y. Mussa, R. C. Glen, S. Reiling, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 170 – 178.
[32] X. Xia, E. G. Maliski, P. Gallant, D. Rogers, *J. Med. Chem.* **2004**, *47*, 4463 – 4470.
[33] A. E. Klon, M. Glick, M. Thoma, P. Acklin, J. W. Davies, *J. Med. Chem.* **2004**, *47*, 2743 – 2749.
[34] C. R. Scherer, C. Lerche, N. Decher, A. T. Dennis, P. Maier, E. Ficker, A. E. Busch, B. Wollnik, K. Steinmeyer, *Brit. J. Pharmacol.* **2002**, *137*, 892 – 900.

Received: September 14, 2005
Published online on January 20, 2006

