Available online at www.sciencedirect.com

ScienceDirect

Bioorganic & Medicinal Chemistry 16 (2008) 6252–6260

Bioorganic & Medicinal Chemistry

Support vector machines classification of hERG liabilities based on atom types

Lei Jia[†] and Hongmao Sun*

Department of Discovery Chemistry, Hoffmann-La Roche, Nutley, NJ 07110, USA

Received 4 March 2008; revised 8 April 2008; accepted 14 April 2008
Available online 16 April 2008

Abstract—Drug-induced long QT syndrome (LQTS) can cause critical cardiovascular side effects and has accounted for the withdrawal of several drugs from the market. Blockade of the potassium ion channel encoded by the human *ether-a-go-go*-related gene (hERG) has been identified as a major contributor to drug-induced LQTS. Experimental measurement of hERG activity for each compound in development is costly and time-consuming, thus it is beneficial to develop a predictive hERG model. Here, we present a hERG classification model formulated using support vector machines (SVM) as machine learning method and using atom types as molecular descriptors. The training set used in this study was composed of 977 corporate compounds with hERG activities measured under the same conditions. The impact of soft margin and kernel function on the performance of the SVM models was examined. The robustness of SVM was evaluated by comparing the predictive power of the models built with 90%, 50%, and 10% of the training set data. The final SVM model was able to correctly classify 94% of an external testing set containing 66 drug molecules. The most important atom types with respect to discriminative power were extracted and analyzed.

1. Introduction

Drug safety is a critical issue in pharmaceutical research and development. The well-known high attrition rate in the drug development phase—approximately 90% for drugs entering clinical trials between 1991 and 2000[1]—consumes a significant amount of R&D investment and in turn contributes to the cost of marketed prescription drugs.[2] Some failures are hard to prevent at an early stage of drug discovery, such as the lack of expected biological effects in humans, but some failures might be avoidable even in the preclinical stage, such as those due to drug toxicity.[1,3] Removing compounds with liabilities from the drug discovery pipeline as early as possible should save the overall cost and shorten the time of the R&D process.

The human acquired long QT syndrome (LQTS) is a cardiac repolarization disorder that may cause arrhythmia, or rapid irregular heart beats that can lead to fainting and sudden death. Drug-induced LQTS has received much recent attention.[4] In the last decade, a number of drugs were reclassified (see www.qtdrugs.org) or withdrawn from the market due to their potential to cause LQTS.[5–7] It is widely accepted that drug-induced LQTS is due to the blocking of potassium ion channels [8,9] encoded by the human *ether-a-go-go*-related gene (*hERG*).[10] This gene was originally discovered in *Drosophila ERG* on the basis of its leg-shaking mutant phenotype.[11] hERG potassium ion channels are voltage-dependent K^+ permeable transmembrane proteins consisting of 6 subunits. Although a three-dimensional experimental structure of the hERG channel is not yet available, other available structures of potassium ion channels such as bacteria MthK[12] and KvAP,[13] and mammalian Kv1.2,[14] which are homologous to hERG, shed light onto the likely structural features of hERG. Four of the six transmembrane subunits serve as the probes for potential sensor, while the other two are involved in forming a pathway for K^+, also known as the channel pore, in a tetramer. As the cross membrane potential changes, the conformation of the two subunits switches in a concerted fashion between open (channel being activated) and closed (channel being inactivated) conformations by depolarization and repolarization.[4,15,16] Along the ion transition path, there are several possible

Keywords: HERG prediction; SVM; Atom typing.

* Corresponding author. Tel.: +1 973 562 3870; fax: +1 973 235 6084; e-mail: Hongmao.sun@roche.com

[†] Present address: Department of Chemistry, New York University, New York, NY 10003, USA.

0968-0896/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2008.04.028

binding sites to accommodate small molecules.[15,17] Thus, hERG can be blocked by diverse range of compound structures including a wide spectrum of drugs.[4] Drug blockage of hERG may cause side effects of cardiac arrhythmia, including LQTS which could lead to sudden death and other serious disorders.

Recently, determining hERG blockage activity of relevant compounds has become an important step in pharmaceutical research and development. In vitro and in vivo methods have been developed to measure hERG activity more accurately and efficiently.[18,19] However, even high-throughput experimental assays take longer time than in silico methods. In addition, scaling up compounds for hERG assay might significantly slow down the pace of a drug discovery project. Furthermore, in silico methods can predict hERG activities of compounds that have not yet been synthesized. Therefore, an accurate in silico hERG predictor would be of high value as a complementary alternative to experimental assays. In the past several years, a number of in silico models have been published to predict hERG activity by QSAR approaches[20–24] and classification methods[25,26] including naive Bayesian,[27] decision tree,[28] random forest,[29] partial least squares (PLS),[29,30]and support vector machines (SVM).[29,31,32]

SVM is a supervised machine learning method, capable of detecting subtle patterns in noisy or complex datasets.[33,34] SVM is one of the most popular kernel methods for performing discriminative classification. The algorithm is based on a strong theoretical foundation, and has proven to be robust and accurate in a wide range of applications, including handwriting recognition, face detection, speaker identification, macroarray expression data analysis, and virtual screening. What makes SVM different from other classification methods is that SVM searches for a hyper-plane to not only separate negatives from positives in a training set, but also to minimize generalization errors by maximizing the separating margin, which is defined as the distance from the separating hyper-plane to the nearest expression vectors. SVM assumes improved predictive power to be contained within those previously unseen data.[35] In addition, the introduction of a soft margin, which allows a few anomalous observations to fall into the wrong side of the hyper-plane, effectively prevents SVM from over-training, and offers an extra handle to balance hyper-plane violations against the maximal width of the margin. Previous studies showed that a SVM-based learner can offer remarkably robust performances in solving biological and drug discovery problems, which typically involve high-dimensional and noisy data.[36–38] In this study, we introduced a novel SVM model in conjunction with atom typing descriptors[39] to predict hERG activities. The SVM model was trained with an in-house hERG dataset containing nearly a thousand diverse compounds, and it was then applied to predict the hERG activity of 66 drug molecules available in the public domain.[24]

2. Methods

2.1. Datasets

The training set in this study was obtained by extracting corporate compounds for which hERG activities had been measured under identical conditions. The compounds in the training set are mostly drug-like, indicated by the fact that over 90% of the compounds violate none or only one of Lipinsky's rule-of-five.[40] The standard procedure for measuring hERG activity of a compound was as follows: Chinese hamster ovary (CHO) cells were used for stable expression, and standard patch clamp (MultiClamp 700A-2) and automated patch clamp (PatchXpress 7000A) were used to measure the hERG channel currents at a temperature of 37 °C. The solution conditions were set as follows: NaCl 150 mM, KCl 4 mM, $MgCl_2$ 1 mM, $CaCl_2$ 1.2 mM, HEPES 10 mM, with the solution buffered to pH 7.4 with NaOH. For the standard patch clamp procedure, currents were measured using a whole cell patch clamp technique with the voltage step pattern at 0.1 Hz. Compounds were introduced at a flow rate of 2 ml/min. The measurement of hERG current response with respect to compound concentration was performed from the end of the incubation period until a steady state was attained. For the automated patch clamp procedure, once a stable whole-cell configuration was achieved, 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 based on at least three-point measures.[41]

For the purpose of validation, the original 977-compound training set was randomly split into 9:1, 1:1, and 1:9 ratios to serve as training sets and testing sets. An external testing set was obtained consisting of 66 drugs with reported hERG activities,[24] resulting from the removal of two duplicate compounds, A-56268 and Hismanal, from the original 68-compound dataset. Their hERG activities had been measured under experimental conditions similar to those of the training set.[27]

Compounds in the training set were binned into positives and negatives according to their IC_{50} values. 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.[7,42] Assuming the required plasma concentration of a moderately potent drug to be 1 μM, the threshold for hERG safety should be set at around 30 μM. Although there are compounds with IC_{50} over 30 μM that are cardiotoxic, most compounds that reach this value have been found to be safe.[24,43] The setting of the 30 μM threshold resulted in 322 negative and 655 positive compounds in the training set, and 13 negative and 53 positive compounds in the external testing set.

2.2. Descriptors—atom type classification

The descriptors used in this study were interpretable atom types, which were assigned according to each atom's own chemical properties and its neighboring atoms and bonds.[39] A classification tree was designed to assign a particular atom type to each atom in an input molecule. The classification tree was trained by optimizing the log *P* predictions of the compounds in Starlist, a high-quality dataset containing nearly 11,000 structurally diverse compounds, to make sure each atom type can reasonably reflect its chemical environment. The optimized atom types have been proven applicable to predicting various molecular properties.[27,39,44]

2.3. Support vector machines (SVM)

SVM is a kernel function based, supervised machine learning technique. The SVM package *Gist*[45] was employed in this study. A typical protocol for SVM is to first train the learner with a set of compounds with known classification—'to learn', and then to use the trained model to the classify previously unseen compounds—'to predict'. There are four key elements in SVM:

(1) The separating hyper-plane: Given a training set $T = \{(\mathbf{x}_i,\ y_i)_\mathrm{N}\}$, where $y_i \in \{+1,\ -1\}$, the most binary classification problems can be converted to identifying a hyper-plane, such that this hyper-plane, an affine subspace of dimension $N-1$, can divide the space into two half spaces with respect to the inputs of the two distinct classes (Fig. 1A). For a linear classifier, a hyper-plane is a linear function of **x**, $f(\mathbf{x}) = \langle \mathbf{w} \cdot \mathbf{x} \rangle + b$, such that

$$y_i(f(\mathbf{x})) = y_i(\langle \mathbf{w} \cdot \mathbf{x} \rangle + b) > 0$$

Here, **w** is the weight vector and *b* is the bias, a scalar value.

Therefore, the separating hyper-plane can be expressed as the following equation:

$$f(\mathbf{x}) = \langle \mathbf{w} \cdot \mathbf{x} \rangle + b = 0$$

(2) The maximum-margin hyper-plane: A fundamental question for the classification problems is how well a derived model performs on classifying unseen data. It is not trivial to estimate the generalization error solely based on a training dataset. According to Novikoff's Theorem, minimizing the generalization error is equivalent to maximizing the separating margin. Therefore, the binary classification with minimized generalization error problem is transformed to a constrained optimization problem (Fig. 1B):

$$\text{Maximize the margin } \frac{2}{\|\mathbf{w}\|},\ \text{or minimize}$$

$$\frac{1}{2}\|\mathbf{w}\|^2,\ \text{subject to } y_i(\langle \mathbf{w} \cdot \mathbf{x} \rangle + b) \geqslant 1.$$

This constrained optimization problem can be tackled with the method of Lagrange multipliers. By introducing an unknown scalar α, then finding derivatives of the Lagrangian *L*, we obtain

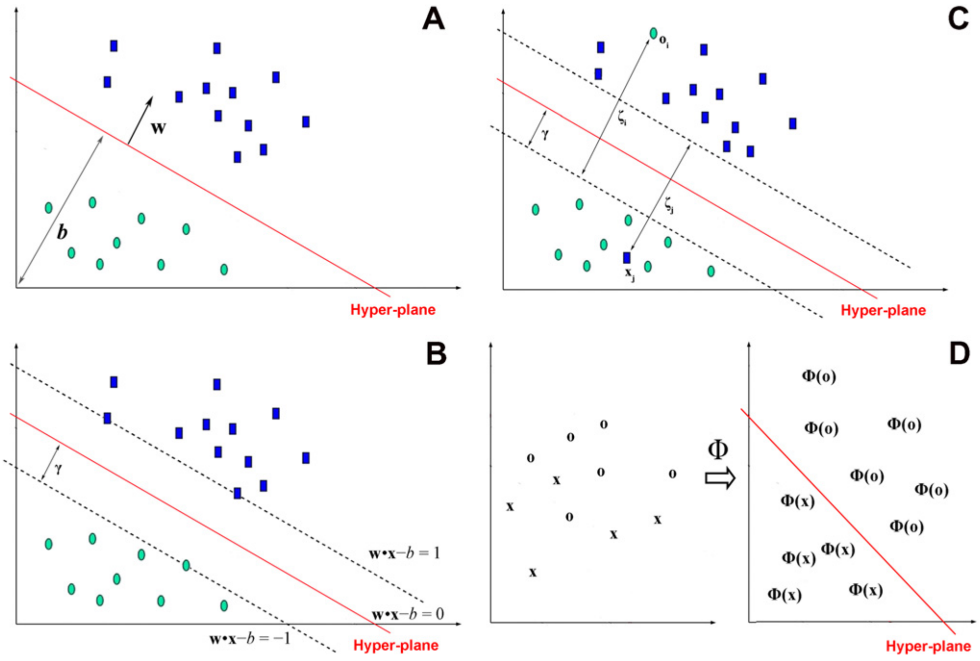

Figure 1. (A) A separating hyper-plane. (B) A maximal-margin hyper-plane. The support vectors are those data points adjacent to the margin hyper-planes and misclassified. (C) Soft margin tolerating a number of data points being misclassified with penalty associate with the 'distance' between the misclassified data point to the margin hyper-plane. (D) Kernel trick allowing application of a linear algorithm to solve a non-linear problem.

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i$$

and

$$L = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j$$

the coefficients of α_i are obtained by maximizing the Lagrangian L, subject to the constraints:

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \quad \text{and} \quad \alpha_i > 0,$$

Once the coefficients of α_i are determined, the final hypothesis is a linear combination of the training points. The decision function can be expressed as:

$$\begin{aligned} h(\mathbf{x}) &= \operatorname{sgn}(\langle \mathbf{w} \cdot \mathbf{x} \rangle + b) \\ &= \operatorname{sgn}\left(\left\langle \sum_{j=1}^{N} \alpha_j y_j \mathbf{x}_j \cdot \mathbf{x} \right\rangle + b \right) \\ &= \operatorname{sgn}\left(\sum_{j=1}^{N} \alpha_j y_j \langle \mathbf{x}_j \cdot \mathbf{x} \rangle + b \right) \qquad (1) \end{aligned}$$

It becomes clear that SVM learning depends only on dot products of input pairs, while prediction of an unseen sample depends only on dot product of the sample with the training data.

(3) The soft-margin: The maximum margin classifier can work properly on linearly separable problems, but it may encounter technical difficulties when applied to noisy and complex datasets from the real world—the optimization problems might be unsolvable. Introduction of the concept of a soft margin can turn these unsolvable problems into solvable ones by tolerating noise and misclassified data points (Fig. 1C). The program *Gist* implements both 1-norm and 2-norm soft margins. For simplicity, 1-norm is to box the constraint, while 2-norm is to weight the diagonal. Therefore, soft margin machines enhance the applicability of maximal margin classifiers, and guarantee a solution to real-world optimization problems even without utilizing powerful kernels. In this study, 1-norm soft margin constraints were applied.

(4) *The kernel function*: The maximal margin classifier we have discussed is a linear classifier, but non-linear classifiers can be easily implemented in SVM by applying the so-called 'kernel trick' to the algorithm. The kernel trick is a method to use a linear algorithm to solve a non-linear classification problem by mapping the original non-linear observations into a higher dimensional feature space. As illustrated earlier, both learning and prediction of a linear classifier depend solely on dot products of input pairs or testing sample and training data. Thus, assuming $\Phi(\mathbf{x_i})$ is the function to project $\mathbf{x_i}$ to a high-dimensional space, what is needed for learning and prediction is $\Phi(\mathbf{x_i}) \cdot \Phi(\mathbf{x_j})$, instead of projection function Φitself. Therefore, the kernel trick allows a simple replacement of the dot products in Eq. 1 with a selected kernel function to realize a non-linear transformation, assuming that the kernel function meets the requirements of Mercer's conditions, such that $K(\mathbf{x_1}, \mathbf{x_2}) = \sum_{i=1}^{N} \lambda_i \phi_i(\mathbf{x_1}) \cdot \phi_i(\mathbf{x_2})$. (Mercer's Theorem).[46] Introduction of kernel functions substantially increases the expressive power of the learning machines, making it possible to formulate a linear classifier in higher dimensional feature space (Fig. 1D).

3. Results and discussions

3.1. Robustness of SVM

In the program *Gist*, learning and prediction are two separate steps, carried out by the modules gist-train-svm and gist-classify. Prediction results were evaluated by the gist-score-svm module. To examine the robustness of SVM modeling, 977 compounds in the original training set were divided into sets with ratios of 9:1 and 1:1, and SVM models were constructed based on 90%, 50%, and 10% of the dataset to predict the remaining 10%, 50%, and 90%, data respectively. As shown in Table 1, models built on the basis of 90% and 50% of data yielded equally high accuracies of 85% and 86% in prediction. A reduced accuracy was observed for the model based on 10% data, yet the value of 74% was still satisfactory considering the possibility that a

Table 1. Validation and prediction results

	Validation 1	Validation 2	Validation 3	Prediction
No. of compounds in training set	879	489	98	977
No. of compounds in testing set	98	488	879	66
No. of active/inactive compounds in training set	587/292	322/167	68/30	655/322
No. of active/inactive compounds in testing set	68/30	333/155	587/292	53/13
Positive constraints	200	100	10	200
Negative constraints	10	100	10	50
FP	11	58	141	3
FN	4	10	87	1
Positive prediction accuracy	84%	83%	76%	94%
Negative prediction accuracy	87%	94%	70%	92%
Average accuracy	85%	86%	74%	94%
Training_ROC	0.90881	0.94765	0.9299	0.91604
Testing_ROC	0.82353	0.90495	0.76772	0.90711

Table 2. Results of testing soft-margins with kernel power = 1

Positive constraint	Negative constraint	Number of support vectors	False positive	False negative	Total	Testing ROC	Training ROC
0.1	0.1	686	13	0	13	0.76488	0.82575
0.2	0.2	671	13	0	13	0.76778	0.82727
0.5	0.5	651	12	0	12	0.76923	0.84031
1	1	618	11	3	14	0.77213	0.85344
2	2	583	11	2	13	0.791	0.87096
5	5	528	8	2	10	0.82729	0.89357
10	10	485	6	1	7	0.87518	0.90889
20	20	454	5	2	7	0.88679	0.92128
50	50	402	2	6	8	0.90276	0.93179
100	100	376	3	4	7	0.91582	0.93764
200	200	357	4	1	5	0.91727	0.94349
500	500	327	2	9	11	0.91001	0.95005
1000	1000	310	2	8	10	0.90711	0.9543
2	200	726	2	30	32	0.76488	0.85122
4	200	715	5	15	20	0.76633	0.85539
10	200	664	5	14	19	0.80842	0.89135
20	200	593	6	2	8	0.86647	0.92543
40	200	493	1	21	22	0.88389	0.94375
50	200	463	0	30	30	0.88679	0.94601
66	200	438	8	1	9	0.88389	0.94994
100	200	417	2	19	21	0.8984	0.94856
200	100	352	2	21	23	0.90856	0.93316
200	66	349	1	32	33	0.90566	0.9176
200	**50**	**355**	**3**	**1**	**4**	**0.90711**	**0.91604**
200	40	353	12	0	12	0.91292	0.91598
200	20	367	1	27	28	0.89985	0.91059
200	10	388	5	1	6	0.90276	0.90339
200	4	397	13	0	13	0.91292	0.90094

large fraction of the compounds in the testing set may be dissimilar to the 10% compounds in the training set. The area under curve (AUC) values of the receiver operating characteristic (ROC) curves,[47] an assessment of the goodness of a binary classifier, were all above 0.9 for the training sets, but the ROC accuracies for testing sets dropped to 0.82 and 0.77 for the models using 90% and 10% training data. The results indicated that SVM classification in combination with atom types as molecular descriptors could achieve highly accurate models when a proper training set was employed. On the other hand, in the case where a training set was relatively small, a reasonably predictive model was still obtained. This robust nature is a significant strength of SVM, since there are always cases where the size of a uniform training set is limited by various reasons, such as high expense or difficulties in obtaining in vivo data.

3.2. Effects of soft margins and kernel functions

Soft margins offer a handle to improve separating margins by forgiving a number of misclassified data points. A light penalty against misclassification may forgive more misclassified data points and gain benefits from a larger separating margin, while a heavy penalty will produce less misclassification at a cost of sacrificing the margin. By using 977 corporate compounds as a training set and 66 drugs as a testing set, two different sets of soft margins were tested in this study, namely balanced and imbalanced soft margins. In a balanced setting, both positive and negative constraints were set equal and were increased from 1 to 1000, while in the imbalanced case by setting the maximum constraint to 200, the positive/negative constraint ratio varied in a stepwise manner, as shown in Table 2. There was a clear trend in balanced soft margins that the number of support vectors decreased as the penalty increased. In addition, a decreasing trend for the number of false positives (FP) and an increasing trend for the number of false negatives (FN) were also observed as the constraints increased. The total misclassified drugs (FP + FN) reached a minimum of 5 when both positive and negative constraints equaled 200. When the ratio of the two constraints varied, the number of support vectors decreased to a minimum value of 349 and then slightly increased along with the increment of the positive/negative constraint ratio. The minimum number of support vectors was reached at the ratio of 3:1 of soft margins, which is close to the ratio of the positive/negative data in the training set. The minimal misclassified drugs were achieved with 3 FP and 1 FN when positive constraint was set to 200 and negative constraint to 50 (see Table 2, marked as bold). The results support the assumption that an optimal classifier can be reached when the ratio of the soft margin constraints matches that of the positive/negative data in the training set.[45]

A non-linear kernel function, a polynomial function with an order of 2, was also utilized to derive a non-linear classifier. The non-linear classifier allowed satisfactory predictions with smaller soft-margin constraints (Table 3). The number of support vectors was lower in the polynomial learner than the corresponding linear machine. The non-linear model also exhibited unsmooth

Table 3. Results of testing soft-margins with kernel power = 2

Positive constraint	Negative constraint	Number of support vectors	False positive	False negative	Total	Testing ROC	Training ROC
0.1	0.1	588	11	2	13	0.79971	0.87506
0.2	0.2	558	9	2	11	0.81858	0.89164
0.5	0.5	493	7	1	8	0.88679	0.91241
1	1	457	5	2	7	0.8984	0.92594
2	2	421	5	2	7	0.89115	0.93518
5	5	388	3	2	5	0.93179	0.94918
10	10	371	3	2	5	0.91727	0.95802
20	20	356	5	0	5	0.91582	0.96518
50	50	342	5	2	7	0.91727	0.97569
100	100	317	3	8	11	0.90711	0.98134
200	200	324	4	8	12	0.88824	0.98498
500	500	319	9	0	9	0.8984	0.99379
1000	1000	309	0	53	53	0.88244	0.9949
2	200	570	9	1	10	0.87228	0.9357
4	200	527	0	53	53	0.8926	0.94374
10	200	466	0	53	53	0.91727	0.95434
20	200	415	13	0	13	0.92453	0.97448
40	200	373	0	53	53	0.92453	0.9811
50	200	364	13	0	13	0.93179	0.9842
66	200	347	5	4	9	0.93033	0.98609
100	200	334	0	53	53	0.92453	0.9878
200	200	324	0	53	53	0.87373	0.98489
200	100	311	0	45	45	0.88824	0.97874
200	66	313	13	0	13	0.87808	0.97509
200	50	306	0	46	46	0.87228	0.97065
200	40	316	3	4	7	0.87083	0.96847
200	20	319	5	1	6	0.88389	0.96481
200	10	325	3	2	5	0.8955	0.94966

variation of FP and FN numbers. For example, changing the soft margin constraints from 500 to 1000 caused FP to drop from 9 to 0, while FN increased from 0 to 53. This kind of non-linear behavior of FP and FN numbers was less severe in the linear models. Judging from the total misclassified drugs, the introduction of non-linear learner did not improve the prediction power in this study. This result might indicate that hERG liability of a compound could be classified by linear rules based on atom types. If a linear classifier is as good as non-linear algorithms, the simpler linear model is suggested to be used unless new data prove non-linear rules.[48] Higher order polynomial and Gaussian kernel functions were not examined in this study, out of concerns over the increased possibility of overfitting associated with non-linearity.

3.3. Prediction results

The best linear SVM classifier with optimal soft margins was able to predict the testing set of 66 drug molecules with an accuracy of 94% (Table 4). The ROC curve of the SVM model using the 977-compound Roche dataset as the training set is shown in Figure 2. There were only 3 FP and 1 FN compounds misclassified in this model. These four misclassified drugs are a subset of the eight misclassified drugs obtained using naive Bayesian method in our previous publication.[27] The pIC_{50} values of the four misclassified drugs, 4.92, 4.52, 4.30, and 3.61, fall into a narrow range around the cut-off value of 4.52. Compounds with their target properties close to the threshold would be expected to have a higher likelihood of being misclassified. Another possible reason for misclassification lies in the fact that the drugs are dissimilar to the compounds in the training set (Fig. 3). As shown in Eq. 1, the decision function relies heavily on the dot products of the atom types vector of testing compound and those of the compounds in training set. The dot product of atom types of two compounds is, to some degree, a measure of the similarity of the compounds. Thus, the SVM learning method can be thought of projecting a test compound to a multiple dimensional space consisting of compounds selected as support vectors, and classifying via linear combinations of weighted ‘similarities’. Since it is not the individual atom types that are analyzed across the training compounds, correlation between certain atom types will not affect the performance in SVM. For example, the drug nicotine is small in size, thus it might not be similar to any of the ‘drug-like’ compounds in the training set. As a result, many dot products of nicotine and training set compounds might have a value close to zero, and its calculated decision function value would tend to be noisy and less discriminative, compared to those compounds which are similar to a number of compounds in the training set. Considering the great structural diversity of the drugs in the testing set, our model shows very high robustness in predicting hERG liabilities.

3.4. Interpretation of the model

An interpretable model is more appreciated by users than a ‘black-box’ model for it supplies practical guidelines for synthetic chemists to improve the properties of next generation of a molecule. Since atom types are employed in this study as molecular descriptors, the use of

Table 4. The predicted classifications of 66 drug molecules in the testing set, together with their experimental hERG activities, and their corresponding classifications

ID	Name	pIC_{50}	Class	Prediction	ID	Name	pIC_{50}	Class	Prediction
1	Alosetron	5.49	P	1.457	34	Mesoridazine	6.49	P	2.638
2	Amiodarone	5.00	P	3.593	35	Mibefradil	5.84	P	0.646
3	Amitriptyline	5.00	P	2.020	36	Mizolastine	6.36	P	1.176
4	Astem izole	8.00	P	2.488	37	Moxifloxacin	3.89	N	**−0.785**
5	Azimilide	5.85	P	2.144	38	Nicotine	3.61	N	1.715
6	Bepridil	6.26	P	1.680	39	Nifedipine	4.30	N	0.307
7	Carvedilol	4.98	P	1.271	40	Nitrendipine	5.00	P	0.730
8	Cetirizine	4.52	N	1.279	41	Norclozapine	5.35	P	1.398
9	Chlorpheniramine	4.68	P	1.185	42	Ofloxacin	2.85	N	**−0.800**
10	Chlorpromazine	5.83	P	1.903	43	Ondansetron	6.09	P	0.699
11	Ciprofloxacin	3.02	N	**−1.286**	44	Perhexiline	5.11	P	0.758
12	Cisapride	7.40	P	2.503	45	Pimozide	7.30	P	2.598
13	Clarithromycin	4.23	N	**−0.185**	46	Quinidine	6.49	P	1.543
14	Clozapine	6.49	P	2.167	47	Risperidone	6.82	P	0.786
15	Cocaine	5.14	P	1.306	48	Sertindole	8.00	P	2.301
16	Ziprasidone	6.92	P	2.054	49	Sildenafil	5.48	P	0.652
17	Desipramine	5.86	P	1.856	50	Sparfloxacin	4.74	P	0.088
18	Diltiazem	4.76	P	1.387	51	Terfenadine	6.70	P	1.957
19	Diphenhydramine	4.57	P	0.502	52	Terikalant	6.60	P	0.616
20	Disopyramide	4.04	N	**−0.078**	53	Thioridazine	6.44	P	2.680
21	Dofetilide	8.00	P	1.047	54	Verapamil	6.85	P	0.684
22	Dolasetron	4.92	P	**−0.592**	55	Vesnarinone	5.96	P	0.168
23	E4031	7.70	P	1.656	57	Citalopram	5.40	P	1.157
24	Epinastine	4.00	N	**−0.874**	58	Desmethylastemizole	9.00	P	2.692
25	Gatifloxacin	3.89	N	**−0.970**	59	Droperidol	7.49	P	2.772
26	Grepafloxacin	4.30	N	**−0.825**	60	Flecainide	5.41	P	0.801
27	Halofantrine	6.70	P	2.885	61	Fluoxetine	5.82	P	1.489
28	Haloperidol	7.52	P	2.093	63	MDL-74156	5.23	P	0.082
29	Ibutilide	8.00	P	0.371	64	Mefloquine	5.25	P	1.183
30	Imipramine	5.47	P	1.377	65	Norastemizole	7.55	P	2.372
31	Ketoconazole	5.72	P	0.478	66	Olanzapine	6.74	P	1.414
32	Levofloxacin	3.04	N	**−0.800**	67	RP-58866	6.70	P	0.616
33	Loratadine	6.77	P	1.233	68	Trimethoprim	3.62	N	**−1.461**

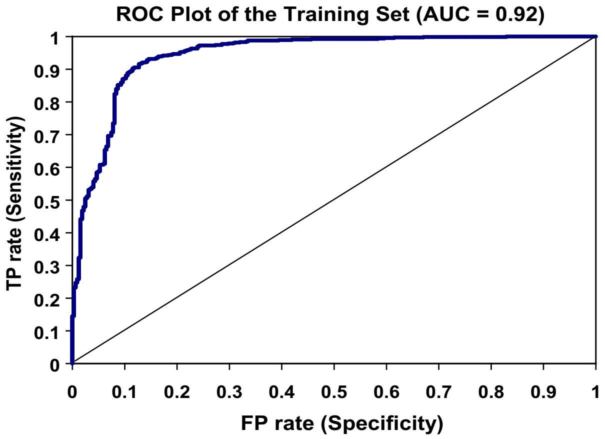

Figure 2. The receiver operating characteristic (ROC) curve of the SVM model built on the basis of the whole training set. Area under curve (AUC) of 0.92 indicates that the model is excellent.

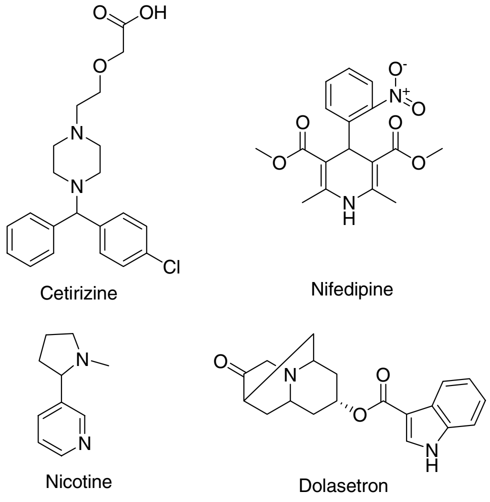

Figure 3. Chemical structures of the four misclassified drugs in the external test set.

the feature selection (Fselect) module from *Gist* can extract a list of atoms or functional groups exhibiting the most positive or negative impact on the hERG activity of a compound. Table 5 shows the ten most influential atom types or correction factors contributing to the hERG activity of a compound. The most important atom type was N16,[39] which represented a nitrogen atom in an aliphatic ring connected to another aliphatic atom. Atom type N16 exists mostly in piperidine and piperizine rings, and usually carries a positive charge. Indeed, 308 of the 332 total occurrence of atom type N16 are found in hERG positive compounds. Atom type C17[39] is structurally correlated to N16—an unsub-

Table 5. Top 10 important atom types and correction factors

Rank	Atom type	Fisher score	Count of the atom type in positive dataset	Count of the atom type in negative dataset	Total number of the atom type
1	N16	0.432985	308	24	332
2	N4	0.185248	201	24	225
3	C17	0.183688	1318	317	1635
4	C14	0.176259	538	112	650
5	H4	0.158155	5	55	60
6	M12	0.132254	1814	767	2581
7	C21	0.132134	223	37	260
8	C3	0.129025	4153	1620	5773
9	H1	0.115019	12,764	4807	17,571
10	C6	0.101007	10	44	54

stituted carbon atom next to the N16 in a ring belongs to the atom type C17. Atom type C17 occurs 1635 times in the training compounds, and 1318 of these occurrences are in hERG positive compounds. Conversely, atom type H4[39] is an acidic hydrogen, and it occurs 55 times in hERG negatives but only 5 times in hERG positives. These results are in agreement with a generally accepted observation that basic compounds tend to be hERG active, while acidic groups eliminate hERG activity.[27] The only correction factor in the top 10 list was M12,[39] the total number of aromatic rings in a molecule. In the training set, there were 535 compounds containing more than 3 aromatic rings, and only 94 of these were hERG negative. On the contrary, 14 out of 19 compounds with only one aromatic ring were hERG negative. Thus, reducing the number of aromatic rings and lipophilicity of a molecule is another general rule to dial away hERG liability.

4. Conclusion

Drug-induced LQTS causes critical cardiovascular side effects, and should be addressed in the early stage of drug discovery. This study aimed to establish a reliable in silico model to predict hERG liability of a molecule based on a large and uniform dataset. By using atom types as molecular descriptors, SVM classification models based on 977 corporate compounds were built to examine the robustness of the method. Models constructed from 90% and 50% of the training set resulted in a high accuracy rate of over 85% in predicting hERG liabilities of the remaining compounds, while the model established on 10% of training data still yielded a satisfactory accuracy of 74%. These results demonstrate that maximal margin SVM classifiers are robust and capable of producing reasonably predictive models even based on a small training set. The SVM model built on the basis of the whole training set correctly classified 94% of 66 drugs of diverse structural classes in a test set. Extraction of the important atom types with high discerning power supplied practical chemical guidelines for eliminating hERG liabilities. The optimized model presented in this work can serve as a filter to remove compounds from drug discovery pipeline with potential hERG liabilities, or to prioritize compounds to be synthesized. The predicted hERG activity can also be used as a flag in comparing libraries to be synthesized or purchased. A detailed analysis of the atom types in hERG active compounds may offer advice toward designing compounds without this undesirable activity.

Acknowledgments

The authors thank Dr. David Fry, for critical reading of this manuscript. L.J. appreciates the financial support for his summer internship at Hoffmann-La Roche Inc., in Nutley.

References and notes

1. Kola, I.; Landis, J. *Nat. Rev. Drug. Discov.* **2004**, *3*, 711–715.
2. Lang, L. *Gastroenterology* **2004**, *127*, 1026.
3. Whitebread, S.; Hamon, J.; Bojanic, D.; Urban, L. *Drug Discovery Today* **2005**, *10*, 1421–1433.
4. Maurizio Recanatini, E. P.; Matteo Masetti; Andrea Cavalli; Fabrizio De Ponti *Med. Res. Rev.* **2005**, *25*, 133–166.
5. Redfern, W. S.; Carlsson, L.; Davis, A. S.; Lynch, W. G.; MacKenzie, I.; Palethorpe, S.; Siegl, P. K.; Strang, I.; Sullivan, A. T.; Wallis, R.; Camm, A. J.; Hammond, T. G. *Cardiovasc. Res.* **2003**, *58*, 32–45.
6. Vincent, G. M. *J. Cardiovasc. Electrophysiol.* **2001**, *12*, 546–547.
7. Aronov, A. M. *Drug Discovery Today* **2005**, *10*, 149–155.
8. Curran, M. E.; Splawski, I.; Timothy, K. W.; Vincent, G. M.; Green, E. D.; Keating, M. T. *Cell* **1995**, *80*, 795–803.
9. Trudeau, M. C.; Warmke, J. W.; Ganetzky, B.; Robertson, G. A. *Science* **1995**, *269*, 92–95.
10. Sanguinetti, M. C.; Jiang, C.; Curran, M. E.; Keating, M. T. *Cell* **1995**, *81*, 299–307.
11. Ganetzky, B.; Wu, C. F. *J. Neurogenet.* **1983**, *1*, 17–28.
12. Jiang, Y.; Lee, A.; Chen, J.; Cadene, M.; Chait, B. T.; MacKinnon, R. *Nature* **2002**, *417*, 515–522.
13. Jiang, Y.; Lee, A.; Chen, J.; Ruta, V.; Cadene, M.; Chait, B. T.; MacKinnon, R. *Nature* **2003**, *423*, 33–41.
14. Long, S. B.; Campbell, E. B.; Mackinnon, R. *Science* **2005**, *309*, 897–903.
15. Sanguinetti, M. C.; Tristani-Firouzi, M. *Nature* **2006**, *440*, 463–469.
16. Witchel, H. J. *Expert Opin. Ther. Targets* **2007**, *11*, 321–336.
17. Thomas, D.; Karle, C. A.; Kiehn, J. *Curr. Pharm. Des.* **2006**, *12*, 2271–2283.
18. Wood, C.; Williams, C.; Waldron, G. J. *Drug Discovery Today* **2004**, *9*, 434–441.

19. Netzer, R.; Ebneth, A.; Bischoff, U.; Pongs, O. *Drug Discovery Today* **2001**, *6*, 78–84.
20. Obrezanova, O.; Csanyi, G.; Gola, J. M.; Segall, M. D. *J. Chem. Inf. Model.* **2007**.
21. Coi, A.; Massarelli, I.; Murgia, L.; Saraceno, M.; Calderone, V.; Bianucci, A. M. *Bioorg. Med. Chem.* **2006**, *14*, 3153–3159.
22. Cianchetta, G.; Li, Y.; Kang, J.; Rampe, D.; Fravolini, A.; Cruciani, G.; Vaz, R. J. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 3637–3642.
23. Aptula, A. O.; Cronin, M. T. *SAR QSAR Environ. Res.* **2004**, *15*, 399–411.
24. Keseru, G. M. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 2773–2775.
25. Roche, O.; Trube, G.; Zuegge, J.; Pflimlin, P.; Alanine, A.; Schneider, G. *ChemBiochem* **2002**, *3*, 455–459.
26. Chen, X.; Li, H.; Yap, C. W.; Ung, C. Y.; Jiang, L.; Cao, Z. W.; Li, Y. X.; Chen, Y. Z. *Cardiovasc. Hematol. Agents Med. Chem.* **2007**, *5*, 11–19.
27. Sun, H. *ChemMedChem* **2006**, *1*, 315–322.
28. Gepp, M. M.; Hutter, M. C. *Bioorg. Med. Chem.* **2006**, *14*, 5325–5332.
29. Song, M.; Clark, M. *J. Chem. Inf. Model.* **2006**, *46*, 392–400.
30. Gavaghan, C. L.; Arnby, C. H.; Blomberg, N.; Strandlund, G.; Boyer, S. *J. Comput. Aided Mol. Des.* **2007**, *21*, 189–206.
31. Tobita, M.; Nishikawa, T.; Nagashima, R. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 2886–2890.
32. Leong, M. K. *Chem. Res. Toxicol.* **2007**, *20*, 217–226.
33. Vapnik, V. *Statistical Learning Theory*; John Wiley and Sons, Inc.: New York, 1998.
34. Vapnik, V. *The Nature of Statistical Learning Theory*; Springer-Verlag: New York, 1995.
35. Noble, W. S. *Nat. Biotechnol.* **2006**, *24*, 1565–1567.
36. Noble, W. S. In *Kernel Methods in Computational Biology B*; Schoelkopf, K. T., Vert, J.-P., Eds.; MIT Press, 2004.
37. Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. *Comput. Chem.* **2001**, *26*, 5–14.
38. Brown, M. P.; Grundy, W. N.; Lin, D.; Cristianini, N.; Sugnet, C. W.; Furey, T. S.; Ares, M., Jr.; Haussler, D. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 262–267.
39. Sun, H. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 748–757.
40. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. *Adv. Drug Delivery Rev.* **2001**, *46*, 3–26.
41. Guo, L.; Guthrie, H. *J. Pharmacol. Toxicol. Methods* **2005**, *52*, 123–135.
42. Crumb, W.; Cavero, I. I. *Pharm. Sci. Technol. Today* **1999**, *2*, 270–280.
43. De Ponti, F.; Poluzzi, E.; Cavalli, A.; Recanatini, M.; Montanaro, N. *Drug Saf.* **2002**, *25*, 263–286.
44. Sun, H. *J. Med. Chem.* **2005**, *48*, 4031–4039.
45. Pavlidis, P.; Wapinski, I.; Noble, W. S. *Bioinformatics* **2004**, *20*, 586–587.
46. Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines*; Cambridge University Press: United Kindom, 2000.
47. Fawcett, T. *Pattern Recognit. Lett.* **2006**, *27*, 861–874.
48. Rognvaldsson, T.; You, L. *Bioinformatics* **2004**, *20*, 1702–1709.