

Substructure-Based Support Vector Machine Classifiers for Prediction of Adverse Effects in Diverse Classes of Drugs

S. Bhavani,[†] A. Nagargadde,[†] A. Thawani,[†] V. Sridhar,[†] and N. Chandra^{*,‡}

Applied Research Group, Satyam Computer Services Limited, SID Block, IISc Campus, Bangalore, India, and
Bioinformatics Centre & Supercomputer Education and Research Centre, Indian Institute of Science,
Bangalore 560012, India

Received April 7, 2006

Unforeseen adverse effects exhibited by drugs contribute heavily to late-phase failure and even withdrawal of marketed drugs. Torsade de pointes (TdP) is one such important adverse effect, which causes cardiac arrhythmia and, in some cases, sudden death, making it crucial for potential drugs to be screened for torsadogenicity. The need to tap the power of computational approaches for the prediction of adverse effects such as TdP is increasingly becoming evident. The availability of screening data including those in organized databases greatly facilitates exploration of newer computational approaches. In this paper, we report the development of a prediction method based on a support machine vector algorithm. The method uses a combination of descriptors, encoding both the type of toxicophore as well as the position of the toxicophore in the drug molecule, thus considering both the pharmacophore and the three-dimensional shape information of the molecule. For delineating toxicophores, a novel pattern-recognition method that utilizes substructures within a molecule has been developed. The results obtained using the hybrid approach have been compared with those available in the literature for the same data set. An improvement in prediction accuracy is clearly seen, with the accuracy reaching up to 97% in predicting compounds that can cause TdP and 90% for predicting compounds that do not cause TdP. The generic nature of the method has been demonstrated with four data sets available for carcinogenicity, where prediction accuracies were significantly higher, with a best receiver operating characteristics (ROC) value of 0.81 as against a best ROC value of 0.7 reported in the literature for the same data set. Thus, the method holds promise for wide applicability in toxicity prediction.

INTRODUCTION

A high attrition rate toward the end of the drug discovery pipeline, mainly caused by adverse drug reactions exhibited by many seemingly good drug candidates,^{1,2} is a major concern in drug discovery. Worse still is the withdrawal of marketed drugs because of the detection of adverse drug reactions during the post-launch phase. One such adverse reaction is torsade de pointes (TdP), which is commonly described as an atypical ventricular tachyarrhythmia typified by undesired drug-induced QT prolongation, leading to an increased incidence of sudden death.³ Some examples of drugs withdrawn because of their torsadogenicity (or the TdP-causing potential) are terfenadine, ketoconazole, and cisapride.³ The importance of this adverse effect has been well-recognized now, leading to the emergence of suggested guidelines of screening of any new potential drug for QT liability in the early stages of drug discovery itself.⁴ The current common practices for such screening involve rigorous electrocardiographic evaluation of the QT-prolongation potential.⁵ The model behind this strategy is not fool-proof and can be problematic because it relies on QT prolongation as the measure of risk, a parameter which by itself is difficult to measure and shows high variance even within the same individual.³ The development of newer methods for measur-

ing torsadogenicity is therefore necessary. In silico approaches for detecting adverse drug reactions in general are increasingly being recognized as excellent options, not only because the computational models are quick, easy to use, and extremely cost-effective but also because the models themselves can be easily refined to better models as the knowledge about the mechanisms of pathogenesis increases.

Several computational methods have been developed over the years for predicting a given pharmacological activity of a set of molecules, which recently have been utilized for predicting adverse drug reactions as well.^{6,7} Structure–activity relationship and quantitative structure–activity relationship studies (SAR/QSAR) studies have been undertaken in order to understand the aspects of the molecule, directly in terms of its structure or more implicitly in terms of its physicochemical attributes, which in turn influence the activity of a drug as well as its side effects.^{8–11} Although conceptually very elegant and highly amenable to traditional classification techniques, QSAR suffers from the drawback of identifying the most relevant set of descriptors for a given combination of a molecular set and its possible activity. Structure-based approaches overcome this difficulty because they operate directly on the molecular structures and on their constituent substructures.¹¹ However, efficient and sensitive identification of the relevant substructures and subsequent discrimination among them has posed several challenges, necessitating exploration of different structural features as descriptors.

* Corresponding author tel.: +91-80-23601409, 22932892; fax: +91-80-2300551; e-mail: nchandra@serc.iisc.ernet.in.

[†] Satyam Computer Services Limited.

[‡] Indian Institute of Science.

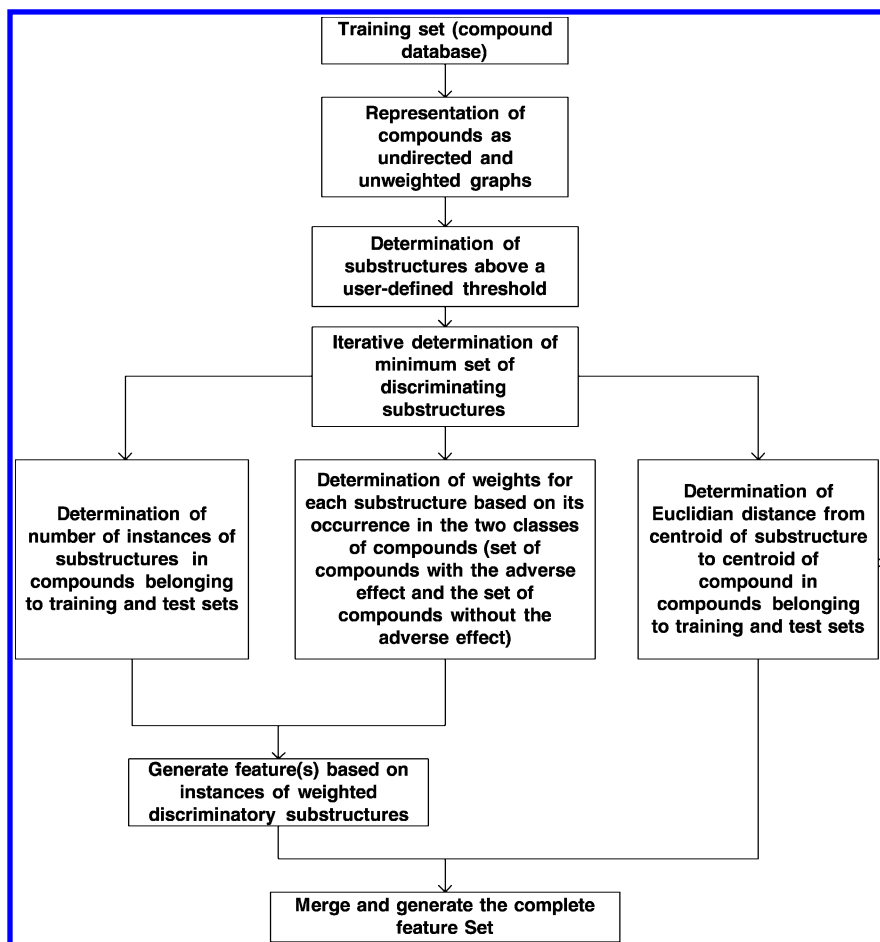


Figure 1. Schematic diagram illustrating feature generation and selection. An overview of the different stages in the algorithm, such as substructure extraction, identification of discriminating substructures, analysis of the instances of occurrence of substructures, and measuring the appropriate Euclidean distances, is shown in the diagram.

Structure-based models have been developed using various techniques such as inductive logic programming, neural networks, and support vector machines (SVM)^{12,13} to predict several adverse drug reactions such as carcinogenicity and mutagenicity on the basis of molecular descriptors which can be properties derived from the structure or substructures extracted from the set of compounds.^{14–17} Substructure-based classification models for drugs have been demonstrated to perform better than classification models based on physico-chemical properties.¹⁵ Support vector machines are by far the most popular technique for distinguishing various classes of drugs using sets of molecular descriptors^{13,14,18} owing to their good generalization performance, computational efficiency, and robustness in high dimensions.¹⁹ Support vector machines, motivated by Vapnik Chervonenkis theory,²⁰ when used for classification, create a hyperplane that separates the data into two classes. Given training examples labeled either “yes” or “no”, a maximum-margin hyperplane is identified which splits the “yes” from the “no” training examples, such that the distance between the hyperplane and the closest examples is maximized.¹⁹ The success in using machine learning algorithms for classification, in general, depends critically on the choice of features used for training them. Hence, it is important to explore the use of different features, also commonly referred to as descriptors, in correlating molecular properties to pharmacological or toxicological activities.

In this paper, we present a SVM-based technique to efficiently classify drug molecules and predict torsadogenicity on the basis of structural descriptors. The classifier is trained to recognize the frequent substructures of occurrence and their three-dimensional disposition in the respective molecules. The classification model based on SVM has been built using the following features, which were computed using the most discriminating substructures: (a) the number of instances of substructures, (b) the weighted values of the number of instances of substructures, and (c) the Euclidean distance between the centroid(s) of the discriminating substructure(s) to the centroid of drug molecules (see Figure 1). We have compared the results obtained using our approach with the results reported by Deshpande and co-workers¹⁵ and Yap and co-workers.¹⁷ The work by Yap and co-workers describes a SVM classifier to predict the torsadecausation potential of a diverse class of drugs using linear solvation energy relationship (LSER) descriptors, while Deshpande and co-workers report a frequent-subgraph-(FSG)-based algorithm to identify substructures which has been used to predict carcinogenicity. We demonstrate that the hybrid classifier we have used gives better results in predicting both torsadogenicity and carcinogenicity. The improvement in prediction performance can be attributed to a combination of the efficient extraction of frequent substructures and the identification of most discriminating substructures, on the basis of both the number of instances

and the overall geometry of the compound important for defining the drug–receptor interactions.

METHODS

Data Sets. A data set of compounds known to cause TdP has been compiled by the Arizona Center for Education and Research on Therapeutics (Arizona CERT) and made available over the Internet.²¹ Yap and co-workers¹⁷ utilized this data set along with an additional compilation of 39 compounds that were known to cause TdP, along with a further control set of 243 compounds known to be TdP-negative. A combination of the two sets, further grouped into training and test sets comprising 271 and 78 compounds, respectively, was also made available by the same authors,²² which was used by us. A second set of data that was provided as a part of the predictive toxicology challenge to evaluate the carcinogenicity of compounds²³ was also used here, to test the potential for the wide application of our approach.

A novel classification method, described below, has been applied to these data sets and validated on the basis of the experimentally derived annotation regarding their torsadogenicity. The three-dimensional structures of the drugs present in the training and test sets were obtained from a structural database.^{24,25} These drugs were represented in Mol2 format, using Openbabel-1.100.2.²⁶

The Algorithm. The algorithm developed here can be described to perform the following tasks. The first step in building the classifier is to extract substructures from the training set, followed by pruning the substructures to obtain an optimal number of substructures that are representative of the training set. The next step pertains to the extraction of different features such as (i) the number of instances of substructures present within each drug and (ii) the Euclidean distance of the centroid of the substructure(s) from the centroid of the drug molecule. The next step is to train a SVM-based classifier, on the basis of the extracted features, which can subsequently classify drugs on the basis of the property being modeled. For each drug in the test set, the values of the features that were identified using the training set are computed. The duly processed test set can then be classified.

Substructure Extraction and Pruning. The compounds in the training set were converted to a corresponding set of topological graphs and the frequently occurring substructures extracted using the FSG algorithm.²⁷ Substructures that were present above a predefined minimum threshold (σ) from a database of graphs were extracted, which corresponded to the substructures being contained in at least $\sigma\%$ of the compounds present in the input database. The extraction of frequent substructures from the compounds in the training set was performed using the PAFI toolkit^{28,29} that includes the implementation of the FSG algorithm. The threshold σ for the extraction of the discriminating substructures was chosen such that it was either equal to or less than the percentage of the positive compounds in the training set due to which substructures specific to a particular class of compounds were extracted.

The determination of frequent substructures above a predefined threshold leads to high-dimensional data sets containing a large number of substructures, most of which are not helpful for predicting the property being modeled.

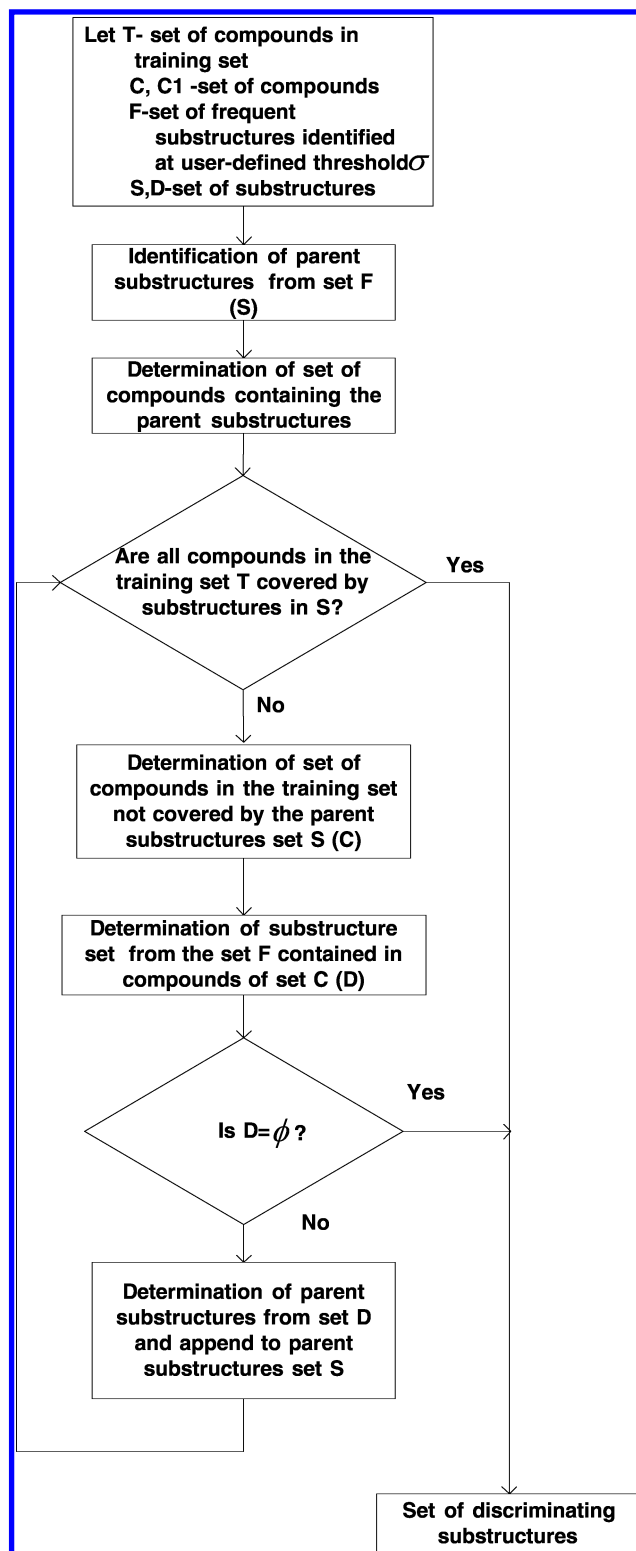


Figure 2. Algorithm for the selection of discriminating substructures. A flowchart indicating the pruning of the substructure set to include only the parent substructures. Φ is used to check whether the set D which contains the substructures present in the compounds which were not covered in the previous iteration of the discriminating substructure selection is a null set.

Here, we propose that the parent–child relationships between substructures can be used to eliminate a large number of unnecessary substructures. Parent–child relationships define substructure containment. If a substructure “a” (for example, C–C) is contained in a substructure “b” (for example,

C–C–C), and substructure “b” occurs in some of the compounds in the training set, it can be concluded that substructure “a” also occurs in the same set of compounds. This automatically implies that the substructure “b” can be considered as the parent substructure and the substructure “a” as the child substructure, thus helping us in safely eliminating the child substructure “a” from the feature set. An algorithm to iteratively determine the minimum number of discriminating substructures that cover all compounds in the training set was developed, as illustrated in Figure 2. This selection was applied until all compounds in the training set were covered by one or more of the substructures. Using parent–child relationships greatly reduces the time complexity in identifying the optimum set of discriminating substructures.

Feature Extraction. Feature extraction involved three preprocessing steps: (a) determining the number of instances of substructures in the compounds of the training set, (b) determining the normalized weights for the presence of each substructure in the compounds, and (c) determining the position(s) of each of the discriminating substructures in the compounds (if present) in the training set.

Because the properties of a drug are generally dependent on both the presence as well as the number of instances of substructures, we have used the number of instances of substructures as the first set of descriptors. The substructures obtained during the discriminating substructure selection were represented as SMARTS patterns,³⁰ and the structure information of compounds in mol2 format were used both to determine the number of instances of a particular substructure in a compound and their positions by employing the various functions present in the OELib package.³¹

Weights are assigned to various features on the basis of their probability of occurrence in the positive and negative compounds in the training set. The probability of occurrences of substructures plays an important role in the predictive accuracy of the classification model. It is used to determine the ratio of positive compounds to the negative compounds in which the substructure is present with respect to the ratio of the positive to the negative compounds in the training set. Thus, the probability of occurrence of a substructure takes the frequency of a substructure in the training set into account. Weights were assigned to various features in order to improve the performance of the classifier on the basis of the deviation from the default probability of occurrence of the feature in all compounds as well as on the ratio of positive compounds to negative compounds.

The deviation from the default probability³² is given by

$$p_a = \frac{f_a}{f_{\text{all}}} - \frac{n_a}{n_{\text{all}}}$$

where p_a = the deviation from the default probability, f_a = the number of positive compounds which contain the substructure, f_{all} = the number of compounds containing the substructure, n_a = the number of positive compounds, and n_{all} = the total number of compounds.

If $p_a > 0$, the substructure indicates activity. If $p_a < 0$, the substructure indicates an absence of activity.

The deviation from the default probability was then normalized using the equations given below:

$$w_i = \begin{cases} p_a \times \frac{R^-}{R^+ + R^-}; p_a > 0 \\ p_a \times \frac{R^+}{R^+ + R^-}; p_a < 0 \end{cases}$$

where, w_i = the weight associated with the presence of instances of discriminating substructures, p_a = the deviation from the default probability, R^+ = the number of positive instances in the data set, and R^- = the number of negative instances in the data set.

To incorporate information about the geometry of the compound, we propose that the Euclidean distances between the centroid of the compound and the centroid of the substructures contained in the compound can be considered as additional features. When multiple instances of a substructure were present in a particular compound, the maximum/minimum/both Euclidean distances can be taken into consideration. The weighted centroid was calculated by taking the atomic weights and the 3D coordinates that indicate the position of the atom in 3D space. The equations for calculating the weighted centroid and the Euclidean distance (dis) are given below:

$$(x, y, z) = \begin{cases} \frac{[(x_1 \times w_1) + (x_2 \times w_2) + \dots + (x_n \times w_n)]}{w_1 + w_2 + \dots + w_n}, \\ \frac{[(y_1 \times w_1) + (y_2 \times w_2) + \dots + (y_n \times w_n)]}{w_1 + w_2 + \dots + w_n}, \\ \frac{[(z_1 \times w_1) + (z_2 \times w_2) + \dots + (z_n \times w_n)]}{w_1 + w_2 + \dots + w_n} \end{cases}$$

where (x_i, y_i, z_i) denote the coordinates of the i th atom in the substructure of the compound/compound, w_i is the atomic weight of i th atom in the substructure of the compound/compound, and (x, y, z) are the coordinates of the centroid of the substructure of the compound/compound. The Euclidean distance (dis) is given by

$$\text{dis} = \sqrt{|x_1 - x_2|^2 + |y_1 - y_2|^2 + |z_1 - z_2|^2}$$

where (x_1, y_1, z_1) are the coordinates of the centroid of the substructure and (x_2, y_2, z_2) are the coordinates of the centroid of the compound.

The feature set for each compound thus consists of two sets of features, that is, the number of instances of discriminating substructures present in the compound with weights assigned (Fw) and the Euclidean distances between the centroids of the substructures to the centroid of the compound (Fe). Fw = {fi: fi = the number of instances of discriminating substructures with weights assigned; fi = si \times w_i , where si is the number of instances of a discriminating substructure in a compound}. Fe = {fi: fi = the maximum Euclidean distances between centroids of substructures to centroids of compounds}.

Construction of the Classification Model. The features computed for the compounds in the training set were in turn used to construct the classification model using support vector machines. The SVMlight package³³ was used to build

the SVM classifier. The different kernel functions, that is, linear, polynomial, and radial basis functions, which were available as a part of the SVMlight package were examined in order to build the classification model. The linear kernel was found to give a good performance because the type and number of features that were used presumably conform to linearly separable data in the feature space. The threshold for the extraction of frequent substructures was varied at regular intervals, and parent–child relationships among substructures were used to extract the most discriminating substructures. A grid search technique described by Rausch et al. was employed in order to choose an optimum value for the C parameter.³⁴ The prediction accuracy was evaluated for the test set in terms of the correctly predicted positive and negative instances. The equations for evaluating the overall accuracy and the accuracies in positive instances and negative instances have been given below:

$$\text{overall accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \times 100$$

$$\text{accuracy in positive set} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100$$

$$\text{accuracy in negative set} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100$$

where TP = the number of positive instances predicted correctly, TN = the number of negative instances predicted correctly, FP = the number of negative instances predicted as a positive instance, and FN = the number of positive instances predicted as a negative instance.

RESULTS AND DISCUSSION

Prediction of Torsade-Causing Potential Using the Developed Algorithm. Drugs that prolong the QT interval or induce torsade de pointes ventricular arrhythmia collated by the Arizona CERT were tagged as drugs that were generally accepted to have a risk of causing TdP.²⁸ These were annotated at four levels: (i) those with well-established TdP (31), (ii) those with a possible risk of TdP (35), (iii) those to be avoided by congenital long QT patients (35) and hence to be avoided for use in patients with diagnosed or suspected congenital long QT syndrome, and (iv) those that clearly do not have TdP (30). From these data, 67 drugs exhibiting TdP (referred to as positive drugs hereafter in the manuscript) were selected and used as the training set. An additional set of 39 drugs (again, positive for TdP) annotated for their torsadogenicity were collected from various other sources¹⁷ and used as part of the test set. A set of 243 drugs which were reported not to show torsadogenicity (referred to as negative drugs, hereafter) were collected from the various sources listed.¹⁷ This collation of data was undertaken by Yap and co-workers.³⁵ We not only used the same drugs in the training and test data sets as provided by Yap and co-workers (hereafter referred to as the original data set) but we also carried out randomization by interchanging the compounds in the training and test sets. The number and ratio of the number of positive to the number of negative compounds were maintained both in the training and test sets. Different thresholds, ranging from 3 to 20%, were tried out for extracting the substructures from the training set. A

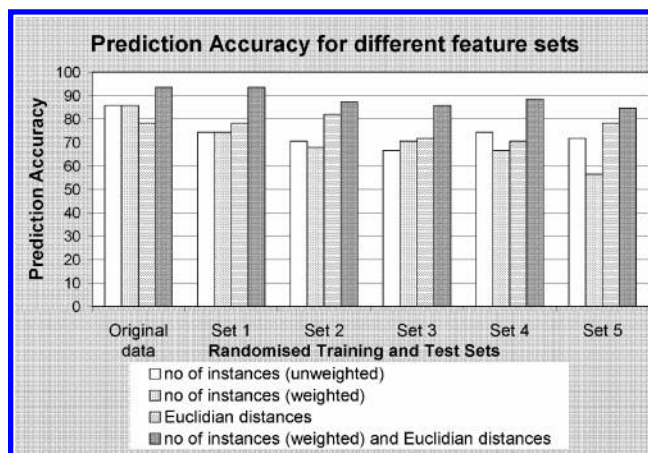




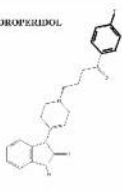
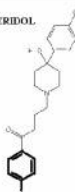
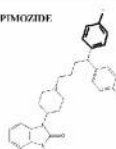
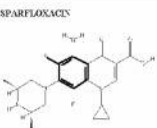


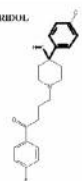
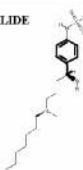
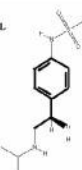

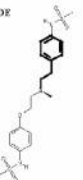
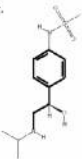

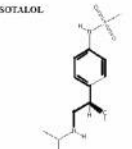
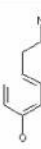


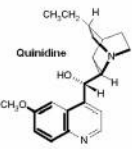

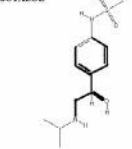

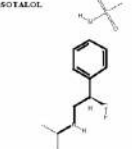

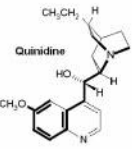

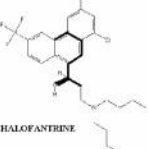
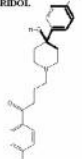
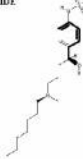

Figure 3. Comparison of the overall prediction accuracy (in %) using the following sets of features: (a) number of instances of substructures (unweighted), (b) number of instances of substructures (weighted), (c) Euclidean distances only (see text), and (d) number of instances of substructures (weighted) and Euclidean distances (see text) for predicting torsadogenicity.

5% threshold appeared to be the most informative because with higher values the number of substructures obtained reduced dramatically, while with lesser than five, too many substructures were being considered. This value was therefore used for further analysis. The total number of substructures used to build the TdP model was 1256 after the substructure extraction and selection.

When the feature vectors to represent the compounds consisted of the number of instances of substructures present, 36 of 39 were correctly predicted to be TdP-positive (true positives), while 31 of 39 were correctly predicted to be TdP-negative (true negatives). When the number of instances of substructures was weighted, 36 of 39 were correctly predicted to be TdP-positive (true positives), while 31 of 39 were correctly predicted to be TdP-negative (true negatives).

The prediction performance increased significantly when the instances with weights assigned were combined with the Euclidean distance measure, as illustrated in Figure 3. The overall accuracies went up to 93.59% (for the entire test set), of which the prediction accuracy for compounds with torsadogenicity was 97.4% (positive drugs in the test set) and the prediction accuracy for compounds without torsadogenicity was obtained as 89.74% (negative drugs in the test set). An improvement of 2.57% in overall accuracy was achieved over the classifier constructed by Yap and co-workers where LSER descriptors were used to construct the classifier. The addition of the distance information brings in an approximate measure of the shape (and hence conformation) of the drug molecule. It is therefore not surprising that its addition increases prediction performance.

A Structural Basis for Torsadogenicity. The mechanism through which several drugs exhibit TdP has now been unraveled, which reveals that the adverse effect emerges from an unwanted blockade of the human ether-a-go-go-related gene (hERG) protein channel.³⁶ Each subunit of the hERG protein is a 1159 amino acid polypeptide chain, the tetramer of which functions as a voltage-gated potassium ion channel. The domain architecture of the subunit is similar to that of other ion channels in that it has a six-trans-membrane (TM) helix (helices S1–S6) TM domain, flanked by large cytoplasmic domains on either side. The conduction pore of the

(a) e1cece c1F		<div>CISAPRIDE </div> <div>DROPERIDOL </div> <div>HALOPERIDOL </div> <div>PIMOZIDE </div> <div>SPARFLOXACIN </div>
e1cececl C(OH)C		<div>HALOFANTRINE </div> <div>HALOPERIDOL </div> <div>IBUTILIDE </div> <div>SOTALOL </div>
e1cececl CCN		<div>DOPETILIDE </div> <div>SOTALOL </div>
e1ccc.c (C(O)C N)c1		<div>SOTALOL </div>
e1cc(CC N)c.cc1O		<div>Quinidine </div>
NCCe1cc c(O).cc1		<div>Quinidine </div>
e1cc(C (O)CN) cc.c1		<div>SOTALOL </div>
e1cececl CC NCC		<div>SOTALOL </div>
Oe1c.cc (CCN) cc1		<div>Quinidine </div>
OC(C) cccc		<div>HALOFANTRINE </div> <div>HALOPERIDOL </div> <div>IBUTILIDE </div> <div>SOTALOL </div>

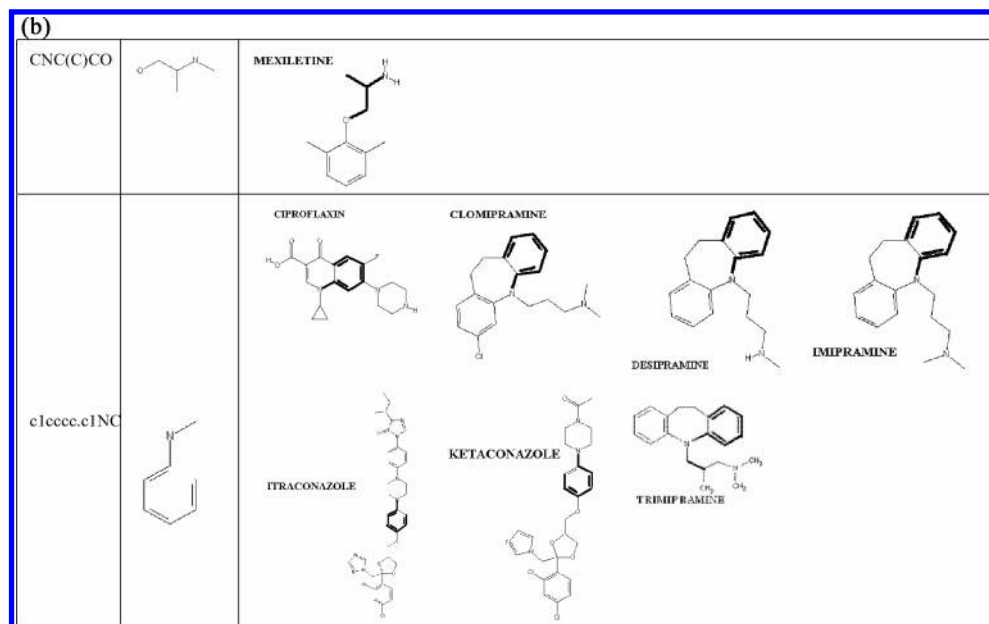


Figure 4. (a) List of the top 10 most discriminating substructures for drugs with positive TdP potentials. (b) List of the top two most discriminating substructures, present in drugs with negative TdP potentials. The SMARTS codes, the substructures, and the drugs that contain them are illustrated. The relevant substructures contained in the drug molecules are highlighted in bold.

channel, formed primarily by TM helices S5 and S6, conforms to the signature motifs of K⁺ channels, exemplified by the KcSA crystal structure.³⁷ Residues in this region contribute significantly to the selective conduction of K⁺ ions. The binding site for interaction with several drugs that block hERG is within the central cavity of the channel pore. Homology models of the pore domain of hERG both in the closed and open states on the basis of the crystal structures of homologous proteins KcSA and KvAP have been reported.³⁷ The models indicate that residues Thr623, Ser624, Val625, Gly648, Tyr652, Phe656, and Val659 are present at the binding site. A number of studies have indicated that, by binding to hERG and plugging the conduction pathway, these drugs interfere with the normal flow of K⁺ ions through the channel.³⁷

Many of these residues are highly conserved and have been identified by alanine scanning mutagenesis to be important for drug binding.³⁶ In particular, Tyr652 and Phe656 have been shown to be crucial for binding.³⁸ In addition, Thr623, Ser624, Val625, and Val629 have all been implicated in drug binding. The two critical aromatic residues are thought to confer the key difference between hERG and other non-ether-a-go-go voltage-sensitive channels. Further characterizations have suggested that Phe656 and Tyr652 may be involved in aromatic stacking interactions with the drug.³⁹ Consistent with this hypothesis, several independently carried out ligand-based statistical and structural computational studies have identified a protonated/charged nitrogen linked to a phenyl ring and a further group that may be either aromatic or hydrophobic as an important 2D pharmacophore to bind to hERG.

The top 10 substructures identified by our approach that are most discriminatory for positive binding potential are illustrated in Figure 4a. Drugs that contain these substructures are also shown. The most common substructures that top the list are seen to contain an aromatic nucleus followed by a -CH₂-CH₂- linker segment connecting to a protonated nitrogen (SMARTS: c1ccccc1CCN). It is interesting to note

that the same substructure has been identified by independent studies using QSAR approaches to be highly correlating with positive TdP potential.³⁷ In addition, even simpler substructures containing an aromatic nucleus with a fluoro or a 1-ethyl hydroxy substituent were also seen to correlate strongly with positive potential while an amino-propyl-ether or even an amino alkyl/aryl moiety was seen to correlate with negative potential (Figure 4b). These observations suggest that, while the c1ccccc1CCN substructure is important for binding to hERG and hence causes TdP, other substructures too can bind to the same binding site and exhibit similar effects. Some examples of drugs that contain the positive substructures as identified by our approach are cisapride, quinidine, dofetilide, sparfloxacin, ibutilide, sotalol, and haloperidol. These drugs are known to cause significant TdP. While drugs such as cisapride, quinidine, dofetilide, and sotalol contain the classical substructure identified by several approaches, the reason for the torsadogenicity of ibutilide had remained hazy. The identification of other substructures that correlate with TdP, as in the case of ibutilide, highlights the usefulness of our approach. In summary, these results indicate (a) the presence of a clear structural pattern in these molecules that will enable their recognition by the hERG channel, despite belonging to diverse classes, and (b) our ability to recognize these patterns or toxicophores through computational approaches. For reasons of restricting unwanted conformational freedom, so as to achieve predictable pharmacophore presentation to their respective receptors, aromatic nuclei have often been used as desirable scaffolds over which various classes of drugs are built. Understanding substructures that are recognized by hERG becomes even more important, in this context.

Proof of Concept of Generality of the Approach: Prediction of Carcinogenicity. To establish that the classifier reported here is superior to that previously reported in a more general context and not exclusive in any way for the TdP predictions, we applied the algorithm to a second set

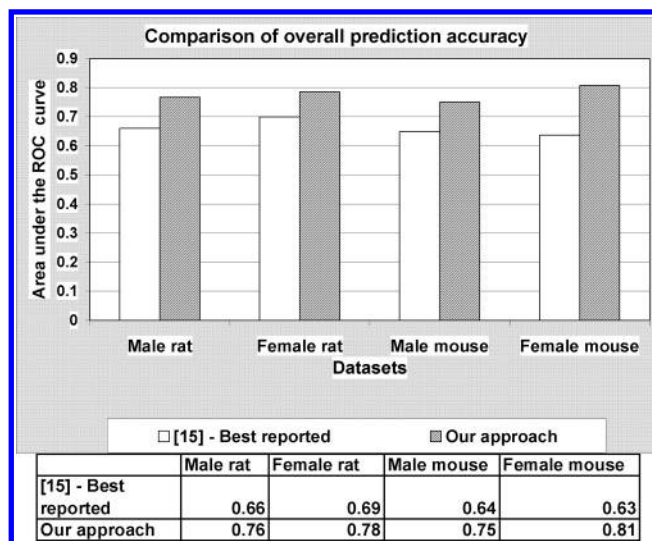


Figure 5. Area under the ROC curve obtained using our approach and the best results reported by Deshpande and co-workers¹⁵ for the predictive toxicology challenge data sets. The improvement in performance in predicting carcinogenicity using our approach can be clearly seen in all four data sets.

Table 1. Overall Prediction Accuracy (in %) Using (a) Number of Instances of Substructures (Unweighted), (b) Number of Instances of Substructures (Weighted), and (c) Number of Instances of Substructures (Weighted) and Euclidean Distances

	number of instances of substructures (unweighted)	number of instances of substructures (weighted)	Euclidean distances	number of instances of substructures (weighted) and Euclidean distances
male rat	65.95	89.19	67.03	92.43
female rat	65.41	72.97	52.97	82.7
male mouse	62.7	54.59	51.35	78.38
female mouse	61.08	58.92	46.49	71.35

of data, for predicting carcinogenicity. A database containing four data sets for such a purpose was collected and analyzed as part of the Predictive Toxicology Challenge 2001.^{23,40} The carcinogenicity potential of different compounds on four laboratory animals (male rat, female rat, male mouse, and female mouse) was provided. Two different classifiers were employed by Deshpande and co-workers on the basis of a C4.5 decision tree and support vector machines. Of these, the SVM-based classification involving substructures was reported to be the best by these authors. Here, we used the same data set and applied our SVM classifier, which involves an additional component of distance measures besides the substructures. The threshold for extracting the discriminating substructures was varied from 5% to 10% and the resulting accuracy measured. Four classification models were constructed, corresponding to each of the rodents (see Figure 5 and Table 1).

CONCLUSIONS

In parallel with the classical paradigm in drug discovery research of identifying primary targets, attention is now being paid to a new paradigm for simultaneously avoiding recognition by certain molecules, the latter in fact are termed as "antitargets".⁷ hERG channel protein, whose inhibition leads to TdP, is one such antitarget. It has become necessary to

either screen or predict the ability of a potential new drug to bind and inhibit the hERG protein, so that such molecules can be weeded out early in the discovery phase. The tools that exist for in silico prediction are directed toward recognizing the toxicophore patterns responsible for binding to hERG protein and hence causing TdP. Both statistical tools and 3D modeling methods have been used for this purpose. While each of them have their own merits and hence successes, it is clear that newer concepts in recognizing patterns are required to achieve higher prediction accuracies. The method adopted by us uses a combination of the statistical mining of important substructures representing possible toxicophores and their occurrence patterns as well as a three-dimensional measure of the distribution of the substructures in the molecule. This hybrid approach appears to result in a classifier superior to that reported in the literature so far, for predicting TdP. Improvement in prediction accuracies for an entirely different set of data as in the case of compounds exhibiting carcinogenicity in fact serves as a proof of concept of the generality of this approach for predicting any adverse drug reaction. Given that TdP is such an important adverse effect, algorithms such as this can be used for screening compounds at a very early stage in discovery, even before the actual molecules can be synthesized or purified. Such screening saves a lot of time, effort, and money and helps in better planning and focusing of the available resources for drug discovery. Such tools may soon become indispensable to the pharmaceutical industry, in the same way that computer-aided design and testing (commonly called CAD) tools are in the automobile or circuit manufacturing industries.

ACKNOWLEDGMENT

Financial support from the Department of Biotechnology (DBT), Government of India, is gratefully acknowledged. The use of facilities at the Super Computer Education & Research Centre, Bioinformatics Centre and Interactive Graphics facility supported by DBT is also acknowledged.

REFERENCES AND NOTES

- (1) Johnson, D. E.; Wolfgang, G. H. I. Predicting Human Safety: Screening and Computational Approaches. *Drug Discovery Today* **2000**, 5 (10), 445–454.
- (2) Whitebread, S.; Hamonb, J.; Bojanica, D.; Urban, L. Keynote Review: In Vitro Safety Pharmacology Profiling: An Essential Tool for Successful Drug Development. *Drug Discovery Today* **2005**, 10, 1421–1433.
- (3) Fitzgerald, P. T.; Ackerman, M. J. Drug-Induced Torsades de Pointes: The Evolving Role of Pharmacogenetics. *Heart Rhythm* **2005**, 2 (Suppl. 2), 30–37.
- (4) Fermini, B.; Fossa, A. A. The Impact of Drug-Induced QT Interval Prolongation on Drug Discovery and Development. *Nat. Rev. Drug Discovery* **2003**, 2, 439–447.
- (5) De Ponti, F.; Poluzzi, E.; Cavalli, A.; Recanatini, M.; Montanaro, N. Safety of Non-antiarrhythmic Drugs That Prolong the QT Interval or Induce Torsades de Pointes: An Overview. *Drug Saf.* **2002**, 25, 263–286.
- (6) Clark, D. E.; Grootenhuys, P. D. Progress in Computational Methods for the Prediction of ADMET Properties. *Curr. Opin. Drug Discovery Dev.* **2002**, 5 (3), 389–390.
- (7) Recanatini, M.; Bottegoni, G.; Cavalli, A. In Silico Antitarget Screening. *Drug Discovery Today Technol.* **2004**, 1, 209–215.
- (8) Hansch, C.; Muir, R. M.; Fujita, T.; Maloney, C. F.; Streich, M. The Correlation of Biological Activity of Plant Growth-Regulators and Chloromycetin Derivatives with Hammett Constants and Partition Coefficients. *J. Am. Chem. Soc.* **1963**, 85, 2817–1824.

- (9) Guner, F. History and Evolution of the Pharmacophore Concept in Computer-Aided Drug Design. *Curr. Top. Med. Chem.* **2002**, 2, 1321–1332.
- (10) Kubinyi, H. *3D QSAR in Drug Design. Theory Methods and Applications*; ESCOM Science Publishers: Leiden, The Netherlands, 1993.
- (11) Oprea, T. I.; Matter, H. Integrating virtual screening in lead discovery. *Curr. Opin. Chem. Biol.* **2004**, 8(4), 349–58.
- (12) King, R. D.; Srinivasan, A.; Dehaspe, L. Warmr: A Data Mining Tool for Chemical Data. *J. Comput.-Aided Mol. Des.* **2001**, 15, 173–181.
- (13) Xu, J.; Hagler, A. Chemoinformatics and Drug Discovery. *Molecules* **2002**, 7, 566–600.
- (14) Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. Drug Design by Machine Learning: Support Vector Machines for Pharmaceutical Data Analysis. *Comput. Chem.* **2002**, 26, 5–14.
- (15) Deshpande, M.; Kuramochi, M.; Karypis, G. Frequent Sub-Structure-Based Approaches for Classifying Chemical Compounds. Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining, Edmonton, Canada, 2002, 35–42.
- (16) Kriegl, J. M.; Arnhold, T.; Beck, B.; Fox, T. A Support Vector Machine Approach To Classify Human Cytochrome P450 3A4 Inhibitors. *J. Comput.-Aided Mol. Des.* **2005**, 19, 189–201.
- (17) Yap, C. W.; Cai, C. Z.; Xue, Y.; Chen, Y. Z. Prediction of Torsade Causing Potential of Drugs by Support Vector Machine Approach. *Toxicol. Sci.* **2004**, 79, 170–177.
- (18) Zhao, C. Y.; Zhang, H. X.; Zhang, X. Y.; Liu, M. C.; Hu, Z. D.; Fan, B. T. Application of Support Vector Machine (SVM) for Prediction Toxic Activity of Different Data Sets. *Toxicology* **2006**, 217, 105–119.
- (19) Burges, C. J. C. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Min. Knowl. Discuss.* **1998**, 2, 127–167.
- (20) Cortes, C.; Vapnik, V. Support Vector Networks. *Mach. Learn.* **1995**, 20, 273–297.
- (21) Arizona CERT. <http://www.arizonacert.org/medical-pros/drug-lists/drug-lists.htm> (accessed Sept 2006).
- (22) TdP Data Set. <http://toxsci.oxfordjournals.org/cgi/content/full/kfh082/DC1> (accessed Sept 2006).
- (23) Predictive Toxicology Challenge Data Set. <http://www.predictive-toxicology.org/ptc/> (accessed Sept 2006).
- (24) von Grotthuss, M.; Pas, J.; Rychlewski, L. Ligand-Info, Searching for Similar Small Compounds Using Index Profiles. *BMC Bioinf.* **2003**, 19, 1041–1042.
- (25) Ligand.Info. <http://ligand.info/> (accessed Sept 2006).
- (26) Openbabel-1.100.2. http://sourceforge.net/project/showfiles.php?group_id=40728 (accessed Sept 2006).
- (27) Kuramochi, M.; Karypis, G. An Efficient Algorithm for Discovering Frequent Subgraphs. *IEEE Trans. Knowl. Data Eng.* **2004**, 16 (9), 1038–1051.
- (28) Ghoting, A.; Buehrer, G.; Parthasarathy, S. A Characterization of Data Mining Algorithms on a Modern Processor. Proceedings of the 1st International Workshop on Data Management on New Hardware DAMON '05, 1–6.
- (29) PAFI Toolkit. <http://glaros.dtc.umn.edu/gkhome/pafi/overview> (accessed Sept 2006).
- (30) SMARTS Tutorial. http://www.daylight.com/dayhtml_tutorials/languages/smarts/index.html (accessed Sept 2006).
- (31) OELib. http://sourceforge.net/project/showfiles.php?group_id=40728&package_id=100796&release_id=197201 (accessed Sept 2006).
- (32) Helma, C. Data Mining and Knowledge Discovery in Predictive Toxicology. *SAR QSAR Environ. Res.* **2004**, 15 (5–6), 367–383.
- (33) Joachims, T. Making Large-Scale SVM Learning Practical. In *Advances in Kernel Methods – Support Vector Learning*; MIT-Press: Cambridge, MA, 1999.
- (34) Rausch C.; Weber T.; Kohlbacher O.; Wohlleben W.; Huson D. H.; Specificity Prediction of Adenylation Domains in Nonribosomal Peptide Synthetases (NRPS) using Transductive Support Vector Machines (TSVMs). *Nucleic Acids Res.* **2005**, 33 (18), 5799–808.
- (35) Arizona CERT. <http://www.arizonacert.org/medical-pros/drug-lists/browse-drug-list.cfm?alpha=Z> (accessed Sept 2006).
- (36) Mitcheson, J. S.; Chen, J.; Lin, M.; Culberson, C.; Sanguinetti, M. C. A Structural Basis for Drug-Induced Long QT Syndrome. *Proc. Natl. Acad. Sci.* **2000**, 97 (22), 12329–12333.
- (37) Stansfeld, P. J.; Sutcliffe, M. J.; Mitchenson, J. S. Molecular Mechanisms for Drug Interactions with hERG that Cause Long QT Syndrome. *Expert Opin. Drug Metab. Toxicol.* **2006**, 2 (1), 81–94.
- (38) Lees-Miller, J. P.; Duan, Y.; Teng, G. Q.; Duff, H. J. Molecular Determinant of High-Affinity Dofetilide Binding to HERG1 Expressed in Xenopus Oocytes: Involvement of S6 sites. *Mol. Pharmacol.* **2000**, 57 (2), 367–374.
- (39) Fernandez, D.; Ghanta, A.; Kauffman, G. W.; Sanguinetti, M. C. Physicochemical Features of the HERG Channel Drug Binding Site. *J. Biol. Chem.* **2004**, 279 (11), 10120–10127; Epub Dec 29, 2003.
- (40) Srinivasan, A.; King, R. D.; Muggleton, S. H.; Sternberg, M. The Predictive Toxicology Evaluation Challenge. Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence (IJCAI)-97, 1–6.

CI060128L