TOXICOLOGICAL SCIENCES **79,** 170–177 (2004)
DOI: 10.1093/toxsci/kfh082
Advance Access publication February 19, 2004

Prediction of Torsade-Causing Potential of Drugs by Support Vector Machine Approach

C. W. Yap,* C. Z. Cai,*,† Y. Xue,*,‡ and Y. Z. Chen*,[1]

*Department of Computational Science, National University of Singapore, Blk SOC1, Level 7, 3 Science Drive 2, Singapore 117543; †Department of Applied Physics, Chongqing University, Chongqing 400044, P. R. China; and ‡Department of Chemistry, Sichuan University, Chengdu 610064, P. R. China

Received December 17, 2003; accepted January 16, 2004

In an effort to facilitate drug discovery, computational methods for facilitating the prediction of various adverse drug reactions (ADRs) have been developed. So far, attention has not been sufficiently paid to the development of methods for the prediction of serious ADRs that occur less frequently. Some of these ADRs, such as torsade de pointes (TdP), are important issues in the approval of drugs for certain diseases. Thus there is a need to develop tools for facilitating the prediction of these ADRs. This work explores the use of a statistical learning method, support vector machine (SVM), for TdP prediction. TdP involves multiple mechanisms and SVM is a method suitable for such a problem. Our SVM classification system used a set of linear solvation energy relationship (LSER) descriptors and was optimized by leave-one-out cross validation procedure. Its prediction accuracy was evaluated by using an independent set of agents and by comparison with results obtained from other commonly used classification methods using the same dataset and optimization procedure. The accuracies for the SVM prediction of TdP-causing agents and non-TdP-causing agents are 97.4 and 84.6% respectively; one is substantially improved against and the other is comparable to the results obtained by other classification methods useful for multiple-mechanism prediction problems. This indicates the potential of SVM in facilitating the prediction of TdP-causing risk of small molecules and perhaps other ADRs that involve multiple mechanisms.

Key Words: **support vector machine; torsade de pointes; linear solvation energy relationship; prediction.**

Adverse drug reaction (ADR) is one of the main reasons for the failure of investigational drugs and the withdrawal of marketed drugs (Johnson and Wolfgang, 2000; van de Waterbeemd and Gifford, 2003). It accounts for up to one-third of all drug failures during drug development (Kennedy, 1997). In an effort to improve the efficiency of drug discovery, computational tools for ADR prediction have been developed, aimed at facilitating the elimination of ADR causing agents in early stages of drug development (Kennedy, 1997; van de Waterbeemd and Gifford, 2003). Mechanism-based knowledge systems (Sanderson and Earnshaw, 1991; Smithing and Darvas, 1992) and statistical models describing the correlation between specific ADR and structure-derived physicochemical features (Klopman, 1992; Prival, 2001) have been developed. Moreover, ligand-protein docking methods have also been explored for the prediction of ADR by screening ADR-inducing drug-protein interactions (Chen and Ung, 2001; Rockey and Elcock, 2002). These methods have shown promising potential in the prediction of such ADRs as carcinogenicity, mutagenicity, teratogenicity, irritation, sensitization, immunotoxicity, and neurotoxicity (Benigni *et al.*, 2000 Cronin and Basketter, 1994; Devillers, 2000; Kulkarni and Hopfinger, 1999).

So far, attention has not been sufficiently paid to the development of methods for prediction of serious ADRs that occur less frequently. While these ADRs are tolerated to a certain extent for the approval of drugs used in serious diseases urgently needing effective or more treatment options such as AIDS and cancer (Somers *et al.*, 1990), they are nonetheless important safety issues for the approval of drugs intended for minor illnesses with availability of alternative treatment options. Examples of these illnesses are rhinitis, cough, pain, inflammation, and hypertension. Therefore, there is a need to develop computational methods for facilitating the prediction of these ADRs.

One such ADR is torsade de pointes (TdP), which is an atypical rapid ventricular tachycardia with periodic waxing and waning of amplitude of the QRS complexes on the electrocardiogram as well as rotation of the complexes about the isoelectric line (*Dorland's Illustrated Medical Dictionary*, 2000). TdP may be self-limited or may progress to ventricular fibrillation (*Dorland's Illustrated Medical Dictionary*, 2000). This ADR is uncommon (Darpo, 2001) and thus difficult to detect during clinical trials. There are cases of TdP-causing drugs

No funding was used to assist in conducting the study and the authors do not have any conflicts of interest directly relevant to the contents of the manuscript.

[1] To whom correspondence should be addressed. Fax: 65-6774-6756. E-mail: yzchen@cz3.nus.edu.sg.

which were initially approved and later withdrawn after post-marketing surveillance revealed their TdP-causing potential (De Ponti *et al.*, 2002 Layton *et al.*, 2003).

Not all mechanisms of TdP are completely understood (Moss, 1999). TdP is frequently associated with QT prolongation, which is the lengthening of the time between the start of ventricular depolarization and the end of ventricular repolarization. This arises from the disruption of the balance between inward and outward currents during the cardiac action potential repolarization phase (Malik and Camm, 2001). Drugs that induce QT prolongation usually cause disruption of the outward potassium currents by blocking potassium ion channels, particularly HERG K^+ channel (Vandenberg *et al.*, 2001). This correlation between QT prolongation and blockade of relevant channels had been exploited in the development of computational methods for the prediction of the QT prolongation risk of drugs using artificial neural network (Roche *et al.*, 2002) and pharmacophore models (Cavalli and Poluzzi, 2002).

There is no definitive correlation between QT prolongation and TdP (Malik and Camm, 2001; Muzikant and Penland, 2002). For instance, verapamil causes QT prolongation but does not induce TdP, whereas procainamide and disopyramide cause TdP but are not potent inhibitors of the HERG K^+ channel (Muzikant and Penland, 2002). Thus, it is desirable to develop a method capable of prediction of TdP of multiple mechanisms without complete knowledge of these mechanisms.

A useful method for classification of systems with multiple mechanisms without requiring their knowledge is the support vector machine (SVM), a relatively new and promising statistical learning algorithm for binary classification by means of supervised learning. SVM was originally developed by Vapnik and his coworkers (Burges, 1998; Vapnik, 1995) and has been applied to a wide range of problems including drug blood-brain barrier penetration prediction (Doniger *et al.*, 2002 Trotter *et al.*, 2001), cancer diagnosis (Guyon *et al.*, 2002 Scridhar *et al.*, 2001 Terrence *et al.*, 2000), microarray gene expression data analysis (Brown *et al.*, 2000), and protein function prediction (Cai *et al.*, 2003a). This work explores the use of SVM as a potential tool for TdP prediction.

MATERIALS AND METHODS

Selection of TdP and non-TdP causing agents. TdP^+ agents were collected from ArizonaCERT (2003). These agents were identified from human studies and can be divided into four classes: Class 1 contains agents with risk of TdP; class 2 includes agents with possible risk of TdP; class 3 is composed of agents to be avoided by congenital long QT patients; and class 4 contains agents which have been weakly associated with TdP. Only agents from class 1, 2, and 3 were used for training the SVM system. Agents in class 4 were not considered because it is unclear which of the agents definitely induces TdP. Thus 67 TdP^+ agents (shown in Table 1 of Supplementary Data) were selected and used as the training set.

To objectively assess the prediction accuracy of our SVM system, an additional set of TdP^+ agents, also identified from human studies, were collected from Micromedex (MICROMEDEX Edition expires 12/2003), *Drug Information Handbook* (Lacy *et al.*, 2002), *Meyler's Side Effects of Drugs* (Dukes, 1996), and a list of agents compiled by De Ponti *et al.* (2001), The selection criteria for the agents are (1) agents with known TdP side effects and (2) agents from De Ponti's list satisfying either criterion Ia or IIIa. Criterion Ia is the existence of clinical studies and/or case reports associating the compound with the occurrence of TdP/ventricular tachyarrhythmias. Criterion IIIa is the presence of official warnings in the labeling on QT prolongation or occurrence of TdP. The exclusion criteria are (1) agents known to be involved in QT prolongation without information about their effect on TdP, (2) agents in class 1, 2, 3, or 4 of the ArizonaCERT list. This gives an independent validation set of 39 TdP^+ agents, which are listed in Table 2 of Supplementary Data.

Like in the case of other classification systems, training of a SVM system requires information about TdP^- agents. In this work, 243 TdP^- agents were obtained from the search of Micromedex, *Drug Information Handbook,* and American Hospital Formulary Service (AHFS) for agents with no reported case of TdP in humans. Thirty-nine of these agents were randomly selected and used as part of the independent validation set (Table 2 of Supplementary Data) to assess the prediction accuracy of the SVM system on TdP^- agents, while the rest were used in the training set (Table 1 of Supplementary Data).

Chemical descriptors. In this work, linear solvation energy relationships (LSER) descriptors (Abraham, 1993; Kamlet *et al.*, 1981, 1987) were used for the modeling of TdP-causing potential of compounds. LSER descriptors describe solvent-solute interactions and contain three main terms: a cavity term, a polar term, and hydrogen-bond term. The cavity term is a measure of the endoergic cavity-forming process, which is the free energy necessary to separate the solvent molecules, overcoming solvent-solvent cohesive interactions, and provides a suitably sized cavity for the solute. The polar term measures the exoergic balance of solute-solvent and solute-solute dipolarity/polarizability interactions and the hydrogen-bond term measures the exoergic effects of the complexation between solutes and solvents.

LSER was initially developed for the estimation of the effects of different solvents on properties of specific solutes or the solubilities, lipophilicities, or

TABLE 1
Results of Various Classification Methods on Independent Validation Set

		TdP^+			TdP^-			
Method	Optimum parameter	TP	FN	Accuracy (%)	TN	FP	Accuracy (%)	Overall accuracy (%)
C4.5 decision tree	—	15	24	38.5	36	3	92.3	65.4
KNN	3	35	4	89.7	34	5	87.2	88.5
PNN	0.1	28	11	71.8	33	6	84.6	78.2
SVM	0.3	38	1	97.4	33	6	84.6	91.0

other properties of a set of different solutes in a specific solvent. It has since been extended for analysis of biological properties including toxicological properties of compounds (Dai *et al.*, 2001 He *et al.*, 1995 Liu *et al.*, 2003 Sixt *et al.*, 1995 Wilson and Famini, 1991; Yu *et al.*, 2002), cell permeation (Platts *et al.*, 2000), intestinal absorption (Zhao *et al.*, 2001), and blood-brain barrier penetration (Platts *et al.*, 2001). LSER descriptors encode the size, polarity, and hydrogen bonding capability of a chemical that has been found to be important for the passive transport of a chemical through biological membranes (Gratton *et al.*, 1997 Kramer and Wunderli-Allenspach, 2001). In addition, it has been shown that complex systems, such as receptor sites, can be approximately described as a solvent system and LSER methods provide useful insights into important binding features (Cramer and Truhlar, 1992). Thus, the polar term may represent the binding action via dispersion forces of a chemical in the polar regions of a receptor molecule and the hydrogen bond term represents the hydrogen-bonding effect between the chemical and the receptor molecule (Liu *et al.*, 2003 Lowrey *et al.*, 1997). Since toxicity of a compound involves the transport of the compound to a site and its interaction with a molecular target, LSER descriptors are thus likely to be useful for TdP modeling.

The LSER descriptors used in this study was calculated using our own developed software based on the method developed by Platts (1999) and are given in Tables 1 and 2 of Supplementary Data. The accuracy of these calculated descriptors for some of the compounds has been verified using the demo version of the software Absolv (Sirius, 2000). These descriptors are excess molar refraction, combined dipolarity/polarizability, overall solute hydrogen bond acidity, overall solute hydrogen bond basicity, and McGowan's characteristic volume.

SVM algorithm. The theory of SVM has been extensively described in literatures (Burges, 1998; Evgeniou and Pontil, 2001; Vapnik, 1995). Thus only a brief description is given here. SVM is based on the structural risk minimization (SRM) principle from statistical learning theory (Vapnik, 1995). In linearly separable cases, SVM constructs a hyperplane that separates the two classes of vectors (TdP$^+$ class and TdP$^-$ class) with a maximum margin. Each TdP$^+$ or TdP$^-$ agent is represented by a vector $\mathbf{x}_i$, which is its LSER descriptors. This is accomplished by finding another vector w and a parameter b that minimizes $\|\mathbf{w}\|^2$ and satisfies the following conditions:

$$\mathbf{w} \cdot \mathbf{x}_i + b \geq +1, \quad \text{for } y_i = +1 \quad \text{Class 1 (positive)} \tag{1}$$

$$\mathbf{w} \cdot \mathbf{x}_i + b \leq -1, \quad \text{for } y_i = -1 \quad \text{Class 2 (negative)} \tag{2}$$

where y_i is the class index, $\mathbf{w}$ is a vector normal to the hyperplane, $|b|/\|\mathbf{w}\|$ is the perpendicular distance from the hyperplane to the origin, and $\|\mathbf{w}\|^2$ is the Euclidean norm of $\mathbf{w}$. After the determination of $\mathbf{w}$ and b, a given vector $\mathbf{x}_i$ can be classified by

$$sign[(\mathbf{w} \cdot \mathbf{x}) + b] \tag{3}$$

In nonlinearly separable cases, SVM maps the vectors into a high dimensional feature space using a kernel function K($\mathbf{x}_i$, $\mathbf{x}_j$). An example of a kernel function is the Gaussian kernel, which has been extensively used in different studies with good results (Burbidge *et al.*, 2001 Czerminski *et al.*, 2001 Trotter *et al.*, 2001).

$$K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\|\mathbf{x}_j - \mathbf{x}_i\|^2/2\sigma^2} \tag{4}$$

Linear support vector machine is applied to this feature space and then the decision function is given by

$$f(\mathbf{x}) = sign(\sum_{i=1}^{l} \alpha_i^0 y_i K(\mathbf{x}, \mathbf{x}_i) + b) \tag{5}$$

where the coefficients α_i^0 and b are determined by maximizing the following Langrangian expression:

$$\sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \tag{6}$$

under the following conditions:

$$a_i \geq 0 \quad \text{and} \quad \sum_{i=1}^{l} \alpha_i y_i = 0 \tag{7}$$

A positive or negative value from Equation 3 or Equation 5 indicates that the vector $\mathbf{x}$ belongs to the positive (TdP$^+$) or negative (TdP$^-$) class, respectively.

Validation of SVM classification system. In this work, the SVM classification system was optimized and validated using leave-one-out (LOO) cross-validation. In LOO cross-validation, a compound is left out of the training set and the remaining compounds are used to derive a SVM classification system. The classification system is then used to classify the left-out compound. This process is repeated until every compound in the training set has been left out once. The TdP$^+$, TdP$^-$ and overall accuracies are calculated using the following equations:

$$\text{TdP}^+ \text{ accuracy} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\% \tag{8}$$

$$\text{TdP}^- \text{ accuracy} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100\% \tag{9}$$

$$\text{Overall accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}} \times 100\% \tag{10}$$

where TP is number of the true positives, TN is the number of true negatives, FP is number of the false positives, and FN is the number of false negatives.

Y-randomization was also used to validate the trained SVM classification system. A portion of TdP$^+$ agents in the training set is randomly exchanged with TdP$^-$ agents in the training set, creating new training sets with false TdP$^+$ and TdP$^-$ agents. A SVM classification system is trained using this scrambled training set. The randomization is repeated 10 times and LOO accuracies of the new classification system from each run are compared to that of the original classification system. If the scrambled training set gives significantly lower LOO accuracies than the original training set, the original classification system is considered as not resulting from chance correlation.

The final SVM classification system was then tested by using the independent validation set to objectively assess its predictive capability. Prediction accuracy of the final SVM classification system using this independent validation was compared with those derived from three other classification methods useful for the prediction of multiple mechanisms. These methods are probabilistic neural network (PNN; Specht, 1990), k nearest neighbor (KNN; Fix and Hodges, 1951), and C4.5 decision tree (Quinlan, 1993). PNN is a form of neural network that is designed for classification through the use of Bayes optimal decision rule. Unlike traditional neural networks like feed-forward back-propagation neural network where there are multiple parameters and network architectures to be optimized, PNN only has a single adjustable parameter, a smoothing factor σ for the radial basis function in the Parzen's nonparameteric estimator (Parzen, 1962). Thus PNN usually trains a system orders of magnitude faster than the traditional neural networks.

In KNN, the Euclidean distance between an unclassified point and each individual datum in the training data is measured (Fix and Hodges, 1951). A total of k number of data points which are nearest to the unclassified point are then used to determine the data class of the unclassified point. The data class making up the majority of the k nearest neighbors will be predicted data class of the unclassified point.

C4.5 decision tree is a classifier in the form of a decision tree where a leaf indicates a data class and a decision node specifies a test to be carried out on a single attribute value, with one branch and subtree for each possible outcome of the test (Quinlan, 1993). C4.5 decision tree uses recursive partitioning where each attribute of the data is examined in turn and ranked according to its ability to partition the remaining data to construct the decision tree. A case is classified by starting at the root of the tree and moving through it until a leaf is encountered. At each nonleaf decision node, the case's outcome for the test at the node is determined and attention shifts to the root of the subtree corresponding to this outcome. When this process finally leads to a leaf, the class of the case is predicted to be that recorded at the leaf.

The three classification systems were trained using the same training set, descriptors, and procedure as those used in SVM. They were tested using the same independent validation set. SVM was performed using SVM★, which has recently been developed and tested for the classification of DNA-binding proteins (Cai *et al.*, 2003b). Gaussian kernel shown in Equation 4 was used by SVM★. PNN and KNN were conducted by using our own software and C4.5 decision tree was performed by using the code from Quinlan (1993).

RESULTS

A principal component analysis (PCA; Wold *et al.*, 1987) on all of the five LSER descriptors was performed using the training set. PCA resulted in two principal components that explained 84.6% of the total variance in the five LSER descriptors. Component one and two explained 70.2% and 14.4% of the variance, respectively. Figure 1 shows a score plot of the compounds in the training set using the first two principal components. Octreotide, a TdP^+ compound, and desmopressin, a TdP^- compound, were found to be far out to the right of the score space. Both of these compounds are large in size, with molecular weights of approximately 1019 and 1069, respectively. There is also a cluster of TdP^- compounds at the top of the score plot. This cluster mainly contains the aminoglycoside antibiotics like amikacin and gentamicin together with two other compounds, acarbose and zanamivir. Other than the aminoglycoside's cluster, the score plot showed that TdP^+ and TdP^- compounds cannot be easily separated using their principal components.

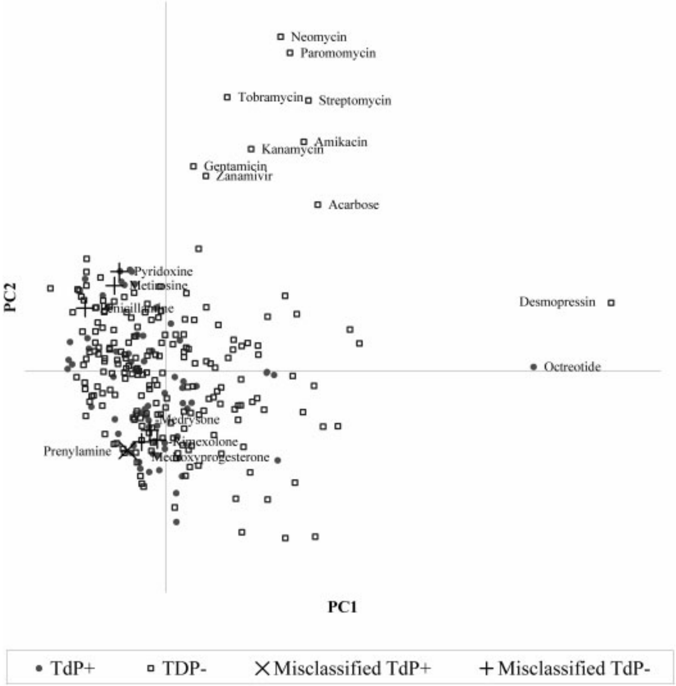

FIG. 1. Score plot of first two principal components for training set. ● represents the TdP^+ agents, ■ represents the TdP^- agents, × represents the TdP^+ agents in the independent validation set that are misclassified and + represents the TdP^- agents in the independent validation set that are misclassified.

LOO cross-validation was used to derive the optimum sigma parameter for the Gaussian kernel (Equation 4) used by SVM and the optimum SVM classification system was found to have a LOO TdP^+ accuracy of 71.6% and LOO TdP^- accuracy of 86.3%. The coefficients for the decision function of the optimum SVM classification system (Equation 5) are given in Table 3 of Supplementary Data. Both of these accuracies are significantly greater than 50%, indicating that the trained SVM classification system is significantly better than a random classifier.

To determine whether it results from chance correlation, the SVM classification system was further tested by repeating y randomization 10 times. The average LOO TdP^+ accuracy from these 10 scrambled classification systems is 21.2% and the average LOO TdP^- accuracy is 77.3%. Both of these accuracies are worse than that of the original SVM classification system, indicating that the SVM classification system is produced as a result of actual correlation between LSER descriptors and TdP-causing potential of the chemicals and not due to chance.

There has been no reported computational study of the TdP-causing potential of a compound. Thus to objectively assess the usefulness of SVM for TdP prediction, its prediction accuracy is compared with those obtained from three other classification methods, C4.5 decision tree, KNN, and PNN, using the same independent validation set. The optimum parameters, k for KNN and σ for PNN, were found by using LOO cross-validation. The optimum parameters for SVM, PNN, and KNN and the accuracy results are given in Table 1. SVM has the highest overall accuracy among the four classification methods. Its TdP^+ accuracy of 97.4% is substantially higher than the other three classification methods that have TdP^+ accuracies of 38.5–89.7%. Its TdP^- accuracy of 84.6% is comparable to the other three methods that have TdP^- accuracies of 84.6–92.3%. These results suggest that SVM is potentially useful for facilitating the prediction of TdP causing risk of investigative agents and likely other ADRs with multiple mechanisms.

In the training set, there are several aminoglycoside antibiotics grouped together in a cluster that does not overlap significantly with the main cluster of compounds. To examine whether this cluster of aminoglycoside antibiotics contributes in some way to the high TdP^+ accuracy, a new SVM classifi-

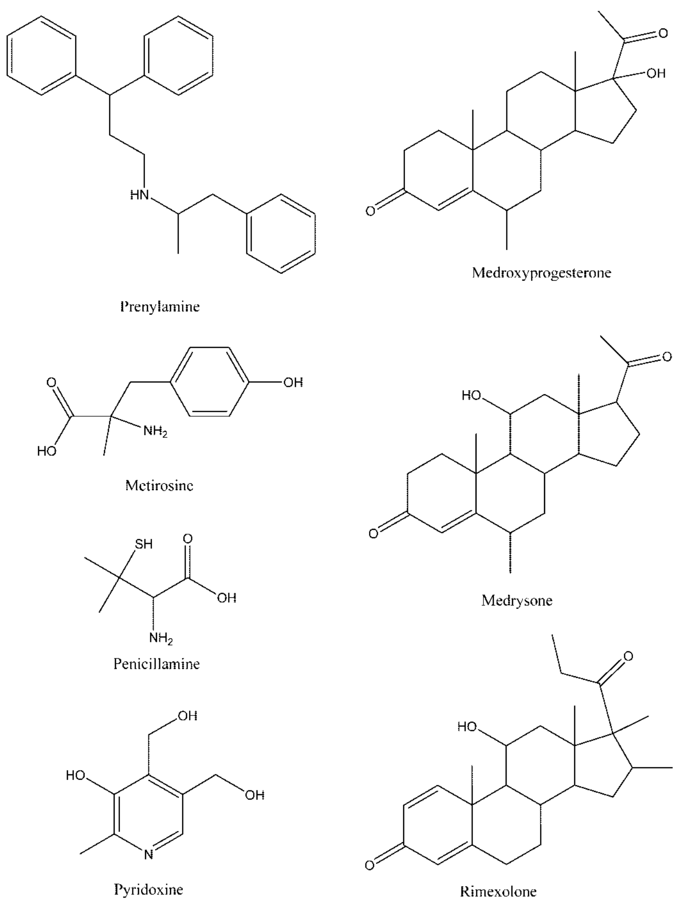

FIG. 2. Incorrectly classified compounds in the independent validation set.

cation system was trained with all of the aminoglycoside antibiotics removed from the training set. The new SVM classification system gives the same TdP$^+$ and TdP$^-$ accuracies as the original system. This suggests that the aminoglycoside antibiotics are not responsible for the high TdP$^+$ accuracy of the SVM classification system.

There are seven agents incorrectly classified by our SVM system, which are shown in Figure 2. These include one TdP causing agent (prenylamine) and six non-TdP causing agents (medroxyprogesterone, medrysone, metirosine, penicillamine, pyridoxine, rimexolone). Their location on the score plot of the training set is shown in Figure 1. Prenylamine is incorrectly classified by SVM, PNN, and C4.5 decision tree. Metirosine and pyridoxine are incorrectly classified by SVM, KNN, and PNN, while penicillamine is incorrectly classified by both SVM and PNN. Medroxyprogesterone, medrysone, and rimexolone have a common steroidal structure and are consistently misclassified by all the four classification methods. This may indicate that the LSER descriptors are unable to fully describe the properties of steroidal compounds thus resulting in their misclassifications by all the four classification methods.

To determine whether the LSER descriptors are sufficient for TdP prediction, we analyzed 490 commonly used descriptors for their relevance in TdP classification and used those essential descriptors to construct a separate SVM classification system. Results using that system are compared with the results using LSER descriptors. These descriptors can be broadly classified into four classes. The first class includes descriptors for global properties of a molecule such as molecular weight, count of atoms, rings, and rotatable bonds. The second class contains topological descriptors such as molecular connectivity indices (Kier and Hall, 1986), electrotopological indices (Kier and Hall, 1999), shape indices (Kier, 1985), and flexibility indices (Kier, 1990). The third class is composed of geometric descriptors including molecular volume, surface area, and polar surface area. The fourth class contains chemical descriptors such as dipole moment, polarizability, and some of the VolSurf descriptors (Cruciani *et al.*, 2000). A preliminary screening was done to reduce the pool of descriptors by eliminating those descriptors that contained little information. Descriptors that have the same value for more than 50% of the compounds were also removed. Backward elimination was then used to produce an optimum subset of descriptors. During backward elimination, LOO cross-validation was used to assess the performance of each subset of descriptors. In the end, the best subset of descriptors consists of 108 descriptors that are not highly correlated with one another. These 108 descriptors were used to train the SVM classification system and the resultant system has TdP$^+$ and TdP$^-$ accuracies of 92.3% and 84.6% on the independent validation set. These results are comparable to that of the current study. This suggests that LSER descriptors are equally useful for prediction of TdP as those using a more diverse set of descriptors.

DISCUSSION

In this study, SVM classification system is compared with three other classification methods and the results suggest that SVM classification system has the best predictive ability among the four methods. All of these classification methods were developed primarily in the machine learning literature and use different algorithms than standard statistical methods. Thus to fully evaluate the performance of SVM classification system, a standard statistical method, logistic regression, was applied to the classification of the same TdP$^+$ and TdP$^-$ datasets. The TdP$^+$ prediction accuracy using the independent validation set using logistic regression is only 20.5%. In addition, y randomization validation tests showed that the LOO TdP$^-$ accuracy of the logistic regression model is less than the mean LOO TdP$^-$ accuracies of the scrambled models. Thus the logistic regression model, as a method for systems with unique mechanism, is not suitable for TdP classification that is intrinsically a multi-mechanism problem.

The possible reason for the usefulness of LSER descriptors for TdP prediction is that they roughly encode most of the essential characteristics related to the TdP causing capability of a compound. Excess molar refraction represents the tendency of a compound to interact with a receptor through n- and π-electron pairs and thus is a measure of the hydrophobic interaction between the compound and receptor. The combined dipolarity/polarizability, on the other hand, represents the ability of electrons to move and be delocalized in the chemical and is a measure of the polar interaction between the compound and receptor.

The overall solute hydrogen bond acidity, overall solute hydrogen bond basicity represents the ability of the compound to form hydrogen bonds with the receptor. This, together with the hydrophobic and polar interactions encoded by the excess molar refraction and combined dipolarity/polarizability, determines the binding affinity of the chemical for the receptor.

The McGowan's characteristic volume influences the passage of a chemical through biological membranes. A compound with a large volume may have difficulty passing through biological membranes and thus may not exhibit toxicity as it is unable to reach its toxicity receptor. In addition, the binding site of a receptor is usually a cavity that can accommodate compounds of a specific range of sizes and shapes.

Currently, with the exception of C4.5 decision tree, which is able to generate decision rules, the other three classification methods are unable to determine the relative importance of individual LSER descriptor. This limits the scope of the application of SVM classification systems in drug design to tasks such as high-throughput screening. With further improvement of SVM algorithm such as the introduction of weighting function to the descriptors (Chapelle *et al.*, 2002), specific rules of the descriptors may be derived which in turn extend the application range of SVM classification systems.

As with all other *in silico* predictions of toxicological properties of chemical compounds, prediction of TdP-causing potential by SVM should be assessed together with pharmacokinetic and pharmacodynamic properties of the chemical compounds in order to determine their clinical significance. This is because a potential TdP-causing drug is not the sole factor in precipitating TdP in a patient. Variability in drug concentrations, drug/drug interactions, and individual patient's susceptibility are some of the numerous factors that affect the occurrence of TdP in patients. Thus a positive TdP-causing risk of a drug-like molecule may not preclude its use in the clinical setting (Malik and Camm, 2001). For example, both halofantrine and terfenadine can potentially cause TdP. However, halofantrine is still in use whereas terfenadine has been withdrawn from the U.S. market as halofantrine is useful for resistant malaria treatment but for terfenadine, there are other safer alternatives, like fexofenadine, available (Malik and Camm, 2001). Despite the limitations of *in silico* prediction of TdP, it may be used as part of the overall risk-benefit analysis of investigative drugs to evaluate their usefulness in the clinical setting.

As a statistical learning method for the prediction of systems with multiple mechanisms, SVM is potentially useful for facilitating the prediction of TdP causing risk of investigative agents. The availability of more extensive information about various ADR-causing agents and associated mechanisms and more comprehensive descriptors for toxicity prediction will enable the development of SVM and other computational methods into useful tools for facilitating the prediction of different types of ADRs in the early stage of drug development.

REFERENCES

Abraham, M. H. (1993). Scales of solute hydrogen-bonding: Their construction and appplication to physicochemical and biochemical processes. *Chem. Soc. Rev.* **22,** 73–83.

AHFS Drug Information (2001). American Society of Health-System Pharmacists, Bethesda, MD.

ArizonaCERT. Drugs that prolong the QT interval and/or induce torsades de pointes ventricular arrhythmia, Vol. 2003. University of Arizona CERT.

Benigni, R., Guiliani, A., Franke, R., and Gruska, A. (2000). Quantitative structure-activity relationships of mutagenic and carcinogenic aromatic amines. *Chem. Rev.* **100,** 3697–3714.

Brown, M. P. S., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C., Ares, J. M., and Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. U.S.A.* **97,** 262–267.

Burbidge, R., Trotter, M., Buxton, B., and Holden, S. (2001). Drug design by machine learning: Support vector machines for pharmaceutical data analysis. *Comput. Chem.* **26,** 5–14.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Disc.* **2,** 127–167.

Cai, C. Z., Han, L. Y., Ji, Z. L., Chen, X., and Chen, Y. Z. (2003a). SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res.* **31,** 3692–3697.

Cai, C. Z., Wang, W. L., and Zong, C. Y. (2003b). Support vector machine classification of physical and biological datasets. *Inter. J. Mod. Phys. C.* **14,** 575–585.

Cavalli, A., and Poluzzi, E. (2002). Toward a pharmacophore for drugs inducing the long QT syndrome: Insights from a CoMFA study of HERG K+ channel blockers. *J. Med. Chem.* **45,** 3844–3853.

Chapelle, O., Vapnik, V., Bousquet, O., and Mukherjee, S. (2002). Choosing multiple parameters for support vector machines. *Mach. Learn.* **46,** 131–159.

Chen, Y. Z., and Ung, C. Y. (2001). Prediction of potential toxicity and side effect protein targets of a small molecule by a ligand-protein inverse docking approach. *J. Mol. Graph. Mod.* **20,** 199–218.

Cramer, C. J., and Truhlar, D. G. (1992). An SCF solvation model for the hydrophobic effect and absolute free energies of aqueous solvation. *Science* **256,** 213–217.

Cronin, M. T. D., and Basketter, D. A. (1994). Multivariate QSAR analysis of a skin sensitization database. *SAR QSAR Environ. Res.* **2,** 159–179.

Cruciani, G., Crivori, P., Carrupt, P. A., and Testa, B. (2000). Molecular fields in quantitative structure-permeation relationships: The VolSurf approach. *J. Mol. Struc.-Theochem.* **503,** 17–30.

Czerminski, R., Yasri, A., and Hartsough, D. (2001). Use of support vector machine in pattern classification: Application to QSAR studies. *Quant. Struct.-Act. Relat.* **20,** 227–240.

Dai, J., Jin, L., Yao, S., and Wang, L. (2001). Prediction of partition coefficient and toxicity for benzaldehyde compounds by their capacity factors and various molecular descriptors. *Chemosphere* **42,** 899–907.

Darpo, B. (2001). Spectrum of drugs prolonging QT interval and the incidence of torsade de pointes. *Eur. Heart J.* **2001,** K70–K80.

De Ponti, F., Poluzzi, E., Cavalli, A., Recanatini, M., and Montanaro, N. (2002). Safety of non-antiarrhythmic drugs that prolong the QT interval or induce torsade de pointes: An overview. *Drug Saf.* **25,** 263–286.

De Ponti, F., Poluzzi, E., and Montanaro, N. (2001). Organising evidence on QT prolongation and occurrence of torsades de pointes with non-antiarrhythmic drugs: A call for consensus. *Eur. J. Clin. Pharmacol.* **57,** 185–209.

Devillers, J. (2000). A neural network SAR model for allergic contact dermatitis. *Toxicol. Methods* **10,** 181–193.

Doniger, S., Hofmann, T., and Yeh, J. (2002). Predicting CNS permeability of drug molecules: Comparison of neural network and support vector machine algorithms. *J. Comp. Biol.* **9,** 849–864.

Dorland's Illustrated Medical Dictionary (2000). W. B. Saunders, London.

Dukes, M. N. G. (1996). *Meyler's Side Effects of Drugs*. Excerpta Medica, Amsterdam.

Evgeniou, T., and Pontil, M. (2001). Support vector machines: Theory and applications. In *Machine Learning and Its Applications. Advanced Lectures* (G. Paliouras, V. Karkaletsis, and C. D. Spyropoulos, Eds.), pp. 249–257. Springer, New York.

Fix, E., and Hodges, J. L. (1951). Discriminatory analysis: Non-parametric discrimination: Consistency properties, pp. 261–279. USAF School of Aviation Medicine, Randolph Field, TX.

Furey, T. S., Cristiani, N., Duffy, N., Bednarski, D. W., Schummer, M., and Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* **16,** 906–914.

Gratton, J. A., Abraham, M. H., Bradbury, M. W., and Chadha, H. S. (1997). Molecular factors influencing drug transfer across the blood-brain barrier. *J. Pharm. Pharmacol.* **49,** 1211–1216.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Mach. Learn.* **46,** 389–422.

He, Y. B., Wang, L. S., Liu, Z. T., and Zhang, Z. (1995). Acute toxicity of alkyl (1-phenylsulfonyl)cycloalkane-carboxylates to Daphnia magna and quantitative structure–activity relationships. *Chemosphere* **31,** 2739–2746.

Johnson, D. E., and Wolfgang, G. H. (2000). Predicting human safety: Screening and computational approaches. *Drug Discov. Today* **5,** 445–454.

Kamlet, M. J., Abbound, J.-L. M., and Taft, R. W. (1981). An examination of linear solvation energy relationships. In *Progress in Physical Organic Chemistry* (R. W. Taft, Ed.), Vol. 13, pp. 485–630. Wiley, New York.

Kamlet, M. J., Doherty, P. J., Taft, R. W., Abraham, M. H., Veith, G. D., and Abraham, D. J. (1987). Solubility properties in polymers and biological media. 8. An analysis of the factors that influence toxicities of organic nonelectrolytes to the golden orfe fish (*Leuciscus idus melanotus*). *Environ. Sci. Technol.* **21,** 149–155.

Kennedy, T. (1997). Managing the drug discovery/development interface. *Drug Discov. Today* **2,** 436–444.

Kier, L. B. (1985). A shape index from molecular graphs. *Quant. Struct.-Act. Relat.* **4,** 109–116.

Kier, L. B. (1990). Indexes of molecular shape from chemical graphs. In *Computational Chemical Graph Theory* (D. H. Rouvray, Ed.), pp. 151–174. Nova Science Publishers, New York.

Kier, L. B., and Hall, L. H. (1986). *Molecular Connectivity in Structure-Activity Analysis*. Research Studies Press, Wiley, Letchworth, Hertfordshire, England.

Kier, L. B., and Hall, L. H. (1999). *Molecular Structure Description: The Electrotopological State*. Academic Press, San Diego.

Klopman, G. (1992). MULTI-CASE: 1. A hierarchical computer automated structure evaluation program. *Quant. Struct.-Act. Relat.* **11,** 176–184.

Kramer, S. D., and Wunderli-Allenspach, H. (2001). Physicochemical properties in pharmacokinetic lead optimization. *Farmaco* **56,** 145–148.

Kulkarni, A. S., and Hopfinger, A. J. (1999). Membrane-interaction QSAR analysis: Application to the estimation of eye irritation by organic compounds. *Pharm. Res.* **16,** 1244–1252.

Lacy, C. F., Armstrong, L. L., Goldman, M. P., and Lance, L. L. (2002). *Drug information handbook*. Lexi-Comp, Hudson, OH.

Layton, D., Key, C., and Shakir, S. A. (2003). Prolongation of the QT interval and cardiac arrhythmias associated with cisapride: Limitations of the pharmacoepidemiological studies conducted and proposals for the future. *Pharmacoepidemiol. Drug Saf.* **12,** 31–40.

Liu, X. H., Wang, B., Huang, Z., Han, S. K., and Wang, L. S. (2003). Acute toxicity and quantitative structure-activity relationships of alpha-branched phenylsulfonyl acetates to Daphnia magna. *Chemosphere* **50,** 403–408.

Lowrey, A. H., Famini, G. R., Loumbev, V., Wilson, L. Y., and Tosk, J. M. (1997). Modeling drug-melanin interaction with theoretical linear solvation energy relationships. *Pigm. Cell Res.* **10,** 251–256.

Malik, M., and Camm, A. J. (2001). Evaluation of drug-induced QT interval prolongation: Implications for drug approval and labelling. *Drug Saf.* **24,** 323–351.

MICROMEDEX (Edition expires 12/2003). MICROMEDEX. MICROMEDEX, Greenwood Village, CO.

Moss, A. J. (1999). The QT interval and torsade de pointes. *Drug Saf.* **21,** 5–10.

Muzikant, A. L., and Penland, R. C. (2002). Models for profiling the potential QT prolongation risk of drugs. *Curr. Opin. Drug Discov. Devel.* **5,** 127–135.

Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Stat.* **33**, 1065–1076.

Platts, J. A., Abraham, M. H., Hersey, A., and Butina, D. (2000). Estimation of molecular linear free energy relationship descriptors. 4. Correlation and prediction of cell permeation. *Pharm. Res.* **17,** 1013–1018.

Platts, J. A., Abraham, M. H., Zhao, Y. H., Hersey, A., Ijaz, L., and Butina, D. (2001). Correlation and prediction of a large blood-brain distribution data set—an LFER study. *Eur. J. Med. Chem.* **36,** 719–730.

Platts, J. A., Butina, D., Abraham, M. H., and Hersey, A. (1999). Estimation of molecular free energy relation descriptors using a group contribution approach. *J. Chem. Inf. Comput. Sci.* **39**, 835–845.

Prival, M. J. (2001). Evaluation of the TOPKAT system for predicting the carcinogenicity of chemicals. *Environ. Mol. Mutagen.* **37,** 55–69.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA.

Roche, O., Trube, G., Zuegge, J., Pflimlin, P., Alanine, A., and Schneider, G. (2002). A virtual screening method for prediction of the hERG potassium channel liability of compound libraries. *Chembiochem.* **3,** 455–459.

Rockey, W. M., and Elcock, A. H. (2002). Progress toward virtual screening for drug side effects. *Proteins* **48,** 664–671.

Sanderson, D. M., and Earnshaw, C. G. (1991). Computer prediction of possible toxic action from chemical structure: The DEREK system. *Hum. Exp. Toxicol.* **10,** 261–273.

Scridhar, R., Pablo, T., Rifkin, R., and *et al.* (2001). Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 15149–15154.

Sirius (2000). Absolv. Sirius Analytical Instruments.

Sixt, S., Altschuh, J., and Brueggemann, R. (1995). Quantitative structure-toxicity relationships for 80 chlorinated compounds using quantum chemical descriptors. *Chemosphere* **30,** 2397–2414.

Smithing, M. P., and Darvas, F. (1992). HazardExpert: An expert system for predicting chemical toxicity. In *Food Safety Assessment* (J. W. Finlay, S. F. Robinson, and D. J. Armstrong, Eds.), pp. 191–200. American Chemical Society, Washington, DC.

Somers, E., Kasparek, M. C., and Pound, J. (1990). Drug regulation–the Canadian approach. *Regul. Toxicol. Pharmacol.* **12,** 214–223.

Specht, D. F. (1990). Probabilistic neural networks. *Neural Netw.* **3,** 109–118.

Trotter, M. W. B., Buxton, B. F., and Holden, S. B. (2001). Support vector machines in combinatorial chemistry. *Measurement and Control* **34,** 235–239.

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

Vandenberg, J. I., Walker, B. D., and Campbell, T. J. (2001). HERG K+ channels: Friend and foe. *Trends Pharmacol. Sci.* **22,** 240–246.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer, New York.

Wilson, L. Y., and Famini, G. R. (1991). Using theoretical descriptors in quantitative structure-activity relationships: Some toxicological indices. *J. Med. Chem.* **34,** 1668–1674.

Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemom. Intell. Lab. Sys.* **2,** 37–52.

Yu, R. L., Hu, G. R., and Zhao, Y. H. (2002). Comparative study of four QSAR models of aromatic compounds to aquatic organisms. *J Environ Sci (China)* **14,** 552–557.

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