Journal of Proteomics & Bioinformatics - Open Access

Research Article JPB/Vol.2/April 2009

Support Vector Machine Classifier for Predicting Drug Binding to P-glycoprotein

Karthikeyan Ramaswamy[1], Mohamed Sadiq[1], Sridhar V[1], Nagasuma Chandra[2*]

[1]Applied Research Group, Satyam Computer Services Limited,
SID Block, IISc Campus, Bangalore-560012, INDIA
[2]Bioinformatics Centre & Supercomputer Education and Research Centre,
Indian Institute of Science, Bangalore 560012, INDIA

*Corresponding author: Dr. Nagasuma Chandra, Bioinformatics Centre & SERC,
Raman Building, Indian Institute of Science, Bangalore 560 012, INDIA
Tel: +91-80-23601409, 22932892; Fax: +91-80-23600551;
E-mail: nchandra@serc.iisc.ernet.in

Received August 20, 2008; Accepted October 17, 2008; Published April 01, 2009

Citation: Karthikeyan R, Mohamed S, Sridhar V, Nagasuma C (2009) Support Vector Machine Classifier for Predicting Drug Binding to P-glycoprotein. J Proteomics Bioinform 2: 193-201. doi:10.4172/jpb.1000077

Abstract

Unforeseen reduction in bio-availability of drugs contribute heavily to late phase failure in drug discovery processes. P-glycoprotein, an efflux pump, that evicts a wide range of drugs is a major cause for reduction in bioavailability. Classification of potential drugs into binders and non-binders of this protein will aid greatly in weeding out the failures early in the discovery processes. The need to tap the power of computational approaches for such prediction is increasingly becoming evident, given the speed and ease with which predictions can be integrated into the discovery programs.

In this paper, we report development of a prediction method to identify substrates and nonsubstrates of P-glycoprotein, based on a support vector machine algorithm. The method uses a combination of descriptors, encoding substructure types and their relative positions in the drug molecule, thus considering both the chemical nature as well as the three dimensional shape information. A novel pattern recognition method, recently reported by us has been implemented for delineating substructures. The results obtained using the hybrid approach has been compared with those available in the literature for the same data set. An improvement in prediction accuracy with most methods is seen, with an accuracy reaching over 93%.

Keywords: Support vector machines; Predicting drug resistance; Efflux pump binding; Machine learning; QSAR

Introduction

Permeability glycoprotein (P-gp), a membrane transporter protein, is one of the major causes for the high attrition rate toward the end of the drug discovery pipeline, owing to its action as an efflux pump (McDevitt and Callaghan, 2007). This protein acts by evicting a range of chemically and pharmacologically diverse molecules out of the cells thereby making them unavailable to the intended receptors, causing multi drug resistance (MDR) in chemotherapy. P-gp is a product of the *MDR1* gene and belongs to the family of ATP Binding Cassette (ABC) transporters (Gottesman et al., 2002). It plays a major role in maintaining normal health by protecting the body against harmful cytotoxic and xenobiotic compounds (Hennessy and Spiers, 2007). During chemotherapy however, it can prove to be a major hindrance if the given drug is perceived by this protein as a threat to the cell and transported by this protein. The exact mechanism of action of P-gp in removal of these compounds is debated and various methods and models are proposed (Higgins and Linton, 2001; Schmitt and Tampe, 2002; Hennessy and Spiers, 2007). However, an undisputed fact is that, recognition of the compound by P-gp is a pre-requisite for it to be subsequently transported. If a useful drug compound has to be available at the site of action in the intended quantities, it would then be required to ensure that it would not be recognized and transported out by P-gp. This requirement has led to labeling the protein as an antitarget. Proteins such as human ether-a-go-go-related gene (hERG) protein channel and Cytochorome-P450 (CYP3A4 and CYP2B) also belong to this category (Recanatini et al., 2004).

Methods to determine P-gp activity utilize different experimental approaches, such as (a) the transporter is isolated followed by purification using a combination of anion exchange and affinity chromatography (Shapiro and Ling, 1994; Sharom, 1995), or (b) by using the transepithelial flux of digoxin across Caco-2 cells (Wandel et al., 1999) and (c) by measuring the displacement of ^{3}H-vinblastine and ^{3}H-verapamil from human intestinal Caco-2 cells over expressed with P-gp (Doppenschmitt et al., 1999) or (d) alternately, by the determination of binding affinities using the concepts of affinity chromatography, where a liquid chromatographic stationary phase containing immobilized P-glycoprotein was synthesized using cell membranes obtained from P-gp-expressing cells and the resulting P-gp stationary phase used in frontal and zonal chromatographic studies to investigate the binding of various drugs (Lu et al., 2001). Understandably these methods are not only time-consuming but also require significant resources.

Given the need for large scale screening of compounds for P-gp binding in multiple stages of almost every drug discovery process, it is no surprise that there is a lot of interest currently to develop appropriate computational methods for initial screening. Availability of screening data including those in organized databases renders exploration of newer computational approaches feasible. Several machine learning and other computational classification systems have been reported in the literature with differing prediction accuracies and differing mechanisms of classification for predicting the substrates and nonsubstrates of P-gp. Some examples are the pharmacophore models based on whole molecule descriptors (Penzotti et al., 2002), a Support Vector machines (SVM) method using various atom type counts and connectivity indices as descriptors (Xue et al., 2004), a topological substructural approach (Cabrera et al., 2006) and a SVM optimized by a particle swarm using topological and functional group based indices as descriptors (Huang et al., 2007). By classifying the compounds as substrates and nonsubstrates of P-gp, compounds which are substrates to P-gp can be avoided in the rational drug design. While each of the methods has their own merits and hence successes, it is clear that newer concepts in recognizing patterns are required to gain a more comprehensive understanding of the binding preferences and achieve higher prediction accuracies.

Support vector machines are popularly used classification technique for differentiating drugs for their classes using sets of molecular descriptors (Xu and Hagler, 2002; Burbidge et al., 2001; Zhao et al., 2006) due to their high performance in generalization, computational efficiency and robustness in high dimensions (Burges, 1998). 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 use of different features, also commonly referred to as descriptors in correlating molecular properties to pharmacological or toxicological activities. Recently we developed a toxicophore based SVM approach to predict torsadogenicity, which resulted in a classifier superior to that reported in the literature, for predicting Torsade de pointes (TdP), with prediction accuracies of around 90% (Bhavani et al., 2006). Encouraged by the success of the approach, we have applied it here for predicting binding potential to P-gp, with an improvement in feature representation. The method adopted by us uses a combination of statistical mining of important substructures representing possible toxicophores, their occurrence patterns as well as a three dimensional measure of the distribution of the substructures in the molecule. In addition, pair-wise Euclidian distances among the substructures in a compound are computed and used in combination.

Methods

Data Set

P-gp substrates and nonsubstrates were obtained from the literature by the report of Xue and co-workers (Penzotti et al., 2002; Xue et al., 2004). Using their data set, a total of 116 substrates and 85 nonsubstrates of P-gp were collected.

The three-dimensional coordinates of the substrates and nonsubstrates of P-gp are obtained in structure-data format (SDF) from publicly available databases such as Ligand.Info (http://ligand.info/) (von Grotthuss et al., 2004), Pubchem (http://pubchem.ncbi.nlm.nih.gov/) and Enhanced NCI database (http://129.43.27.140/ncidb2/). The training and test set from Xue and co-workers is used with available compounds from the above mentioned databases. So the training set is made of 138 compounds (73 substrates and 65 nonsubstrates) and the test set is made of 32 compounds (23 substrates and 9 nonsubstrates). The true positives and true negatives are in the ratio of 1.12: 1. The compounds which are in the independent validation set are used by random selection, to comprise the training and test sets to verify the results of the classification model. These compounds were converted to MOL2 format using openbabel-2.1.1, (http://sourceforge.net/project/showfiles.php?group_id=40728&packageid=32894&release_id=521581), using which, all further processing was carried out. The number and ratio of the number of positive to the negative compounds were maintained both in the training and the test sets.

Developing the support vector machine classifier

The development of the classifier involves extraction of substructures from the training set, followed by pruning the substructures for obtaining the optimal number of substructures that represents the entire training set. This is followed by extraction of different features such as (i) the number of instances of substructures present within each drug and (ii) the Euclidian distance of the centroid of the substructure(s) from the centroid of the drug molecule and (iii) Euclidian distances between centroids of each pair of substructures occurring in a molecule. Next, a SVM based classifier was trained, based on the extracted features. The test set compounds were predicted for their class as either substrates or nonsubstrates of P-gp.

Extraction of optimal discriminating substructures

The problem of identifying a set of substructures from a given set of molecules in their 3D representation can be mapped to that of determining a set of common sub graphs from a given set of graphs. Accordingly, the training set compounds were converted to their corresponding set of topological graphs and frequently occurring substructures were extracted using the Frequent Sub Graph (FSG) algorithm (Kuramochi and Karypis, 2004). The FSG algorithm was able to extract substructures efficiently and found to be computationally less expensive compared to other methods, hence leading us to choosing the FSG approach for the work. Substructures present in a database of graphs were extracted with a minimum support percentile or threshold (ó), which signifies that substructures are present in at least ó% of the input database of graphs. The frequent substructures were extracted from the training set compounds using the FSG implementation available in the PAFI toolkit (http://glaros.dtc.umn.edu/gkhome/pafi/overview) (Ghoting et al., 2005). Various thresholds ranging from 7% to 30% were tried out for extracting the substructures.

A very large number of frequent substructures were generated with the threshold value, of which most of them are not helpful in generating the classification model and hence the prediction. So the parent-child relationship among the substructures was exploited to obtain the optimal number of substructures that are representative of the entire training set compounds. The algorithm to find the optimal number of substructures and for finding discriminating substructures is as described by us previously (Bhavani et al., 2006).

Feature Generation

Two types of features were derived from the each of the parent substructures generated in the previous step. The first feature corresponds to the number of occurrences of a given substructure, while the second pertained to the geometric description of the substructure, with respect to each of the compounds that contained them. Feature extraction involved four pre-processing steps: (a) determining number of instances of substructures in the compounds of the training set, (b) determining 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 and (d) determining the pair-wise centroid distance among the substructures in a compound.

Discriminating substructures can be used either as binary features (the presence or absence of a substructure) or by determining the number of instances of substructures in the compounds in the dataset. We propose the use of number of instances of substructures since the properties of a drug are dependent on both the presence as well as the number of instances of substructures. The substructures obtained during the discriminating substructure selection were represented as SMILES patterns (http://www.daylight.com/dayhtml_tutorials/languages/smiles/index.html) 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 (http://sourceforge.net/project/showfiles.php?group_id=40728&package_id=100796&release_id=197201).

The probability of occurrence of the discriminating substructure in the training set compound was used to assign the weights for the various features that are generated for the classification. Also the weights were based on the ratio of positive compounds to negative compounds in the training set. The performance of the classifier can be enhanced by using the default from deviation probability as well as the ratio of number of positive compounds to negative compounds as described by us previously (Bhavani et al., 2006).

The spatial arrangement of substructures can be considered by calculating the Euclidian distances between the various substructures. In order to keep the number of features at an optimum level and to incorporate information about the geometry of the compound, we propose that the Euclidian distances between the centroid of the compound and the centroid of the substructures contained in the compound can be considered. Among the maximum/minimum/both Euclidian distances, maximum Euclidian distance is considered for the generation of the feature when more then one instance of the substructure is present in a compound. Also the pair-wise Euclidian distance among the substructures present in a compound is calculated using the maximum Euclidian distance of a substructure for the generation of an additional feature when there is more than one substructure is present in a compound. If the compound contains only one substructure then pair-wise Euclidian distance is 0. The weighted centroid was calculated by taking the atomic weights and the 3D coordinates that indicates the position of the atom in the 3D space as described by us previously (Bhavani et al., 2006).

The feature set generated for each of the compound consists of a combination of features i.e., the number of instances of discriminating substructures present in the compound, the number of instances of discriminating substructures present in the compound with weights, the Euclidian distances between the centroid of the substructure to the centroid of the compound and the pair-wise Euclidian distances among the substructures in a compound.

The classification and prediction model

The generated features were used to construct the classification model using support vector machines. The SVMLight package was used in training and classification of the compounds using the generated features (Joachims, 1999). The parameters in SVMLight package are used with their default value. The linear kernel function is used for the classification as it gives a good performance for the linearly separable classes. The radial basis kernel function was also tried but it did not offer any advantage as compared to the linear kernel. The threshold for extraction of frequent substructures was varied from 7% to 30% and parent child relationships among substructures were used to extract the most discriminating substructures. The features were generated for the different thresholds and the prediction accuracy was evaluated for the test set in terms of the correctly predicted substrates and nonsubstrates of P-gp.

Results and Discussion

Using the classifier developed, the test set compounds were tested for the P-gp binding potential. The results were obtained using 7 different types of features along with 8 different threshold values for each feature type, resulting in testing the algorithm with 56 parameter sets. The various features include (i) the number of instances of substructures (unweighted), (ii) the number of instances of substructures (weighted), (iii) the Euclidian distance between the centroid of the substructure to the centroid of the compound, (iv) all pair wise centroid distances between the substructures in a compound and (v) combinations of these features. The number of discriminating substructures was found to be 499 at 7% threshold, 435 at 8%, 330 at 9%, 237 at 10%, 92 at 15%, 38 at 20% and 28 at 25% and 21 at 30%. Of these, 9% threshold was found to be the least constrained but most informative in terms of the prediction accuracies obtained for the best set of features, as described below. Prediction accuracies for the seven different feature sets at different thresholds are shown in Table 1, where both the prediction accuracies as well as the number of support vectors are indicated for each combination. It is clear from the table that use of occurrences of substructures lead to a prediction accuracy in the range of about 59 - 81%, which on the whole improves marginally, when the number of instances of substructures are weighted by the probability of occurrence as described earlier.

The prediction rates however increased significantly when the Euclidian distances of the substructures to the centers of the respective molecules were considered. This is not surprising, because, by considering the distances between the substructure centroids, we are obtaining a reasonable idea of the shape of the molecule and the relative position of the two or more substructures that might be present in a given molecule. The highest prediction accuracy of 93.75% was obtained at 9% threshold (value of parameter C is 0.0006) for the feature set containing the combination of weighted instances of substructures and distances between the centroid of the substructure to the centroid of the whole molecule. Also the same prediction accuracy is obtained with all pair-wise centroid distances combined with weighted instance and Euclidian distance at 20% threshold (value of parameter C is 0.0004).

σ%	Prediction accuracy (%)						
	No of instances of substructure (Un weighted)	No of instances of substructures (Weighted) (a)	Euclidian distances (b)	(a+b)	All pair-wise centroid distance between substructures (c)	(b+c)	(a+b+c)
7%	59.38 (129)	65.62 (129)	84.38 (104)	87.50 (113)	68.75 (117)	68.75 (117)	71.88 (117)
8%	59.38 (130)	65.62 (130)	90.62 (105)	87.50 (115)	71.88 (119)	71.88 (118)	71.88 (116)
9%	62.50 (131)	65.62 (129)	90.62 (100)	93.75 (114)	78.12 (112)	78.12 (113)	78.12 (112)
10%	65.62 (131)	65.62 (129)	90.62 (105)	87.50 (113)	75.00 (111)	71.88 (111)	75.00 (111)
15%	71.88 (129)	68.75 (126)	84.38 (101)	81.25 (108)	78.12 (103)	78.12 (104)	81.25 (104)
20%	81.25 (123)	75.00 (123)	87.50 (93)	84.38 (103)	90.62 (99)	90.62 (99)	93.75 (97)
25%	71.88 (124)	75.00 (124)	81.25 (96)	81.25 (103)	90.62 (103)	87.50 (98)	87.50 (102)
30%	68.75 (121)	84.38 (123)	78.12 (96)	84.38 (100)	84.38 (96)	84.38 (97)	84.38 (98)

Table 1: Overall Prediction Accuracy (in %) using (i) Number of instances of substructures (Unweighted), (ii) Number of instances of substructures (Weighted), (iii) Euclidian distances, (iv) Number of instances of substructures (Weighted) and Euclidian distances, (v) All pair-wise centroid distance between substructures, (vi) Euclidian distances and all pair-wise centroid distance between substructures, and (vii) Number of instances of substructures (Weighted) and Euclidian distances and all pair-wise centroid distance between substructures at different values of ó.

Receiver Operating Characteristic (ROC) curve is a plot of true positive rate vs. false positive rate of a binary classification system as the score of decision threshold is varied (Fawcett, 2004; Provost and Fawcett, 2001). The ROC curve for the classification using Euclidian distances and weighted instances of substructures is given in Figure 1. The area under the ROC curve is 0.9758. The results are validated by random selection of training and test sets and the prediction accuracies for the original data as well as for the randomized sets are illustrated in Figure 2. Also 5 fold cross validation results for the training and test set correlate for the results of different descriptors. As an additional veri-

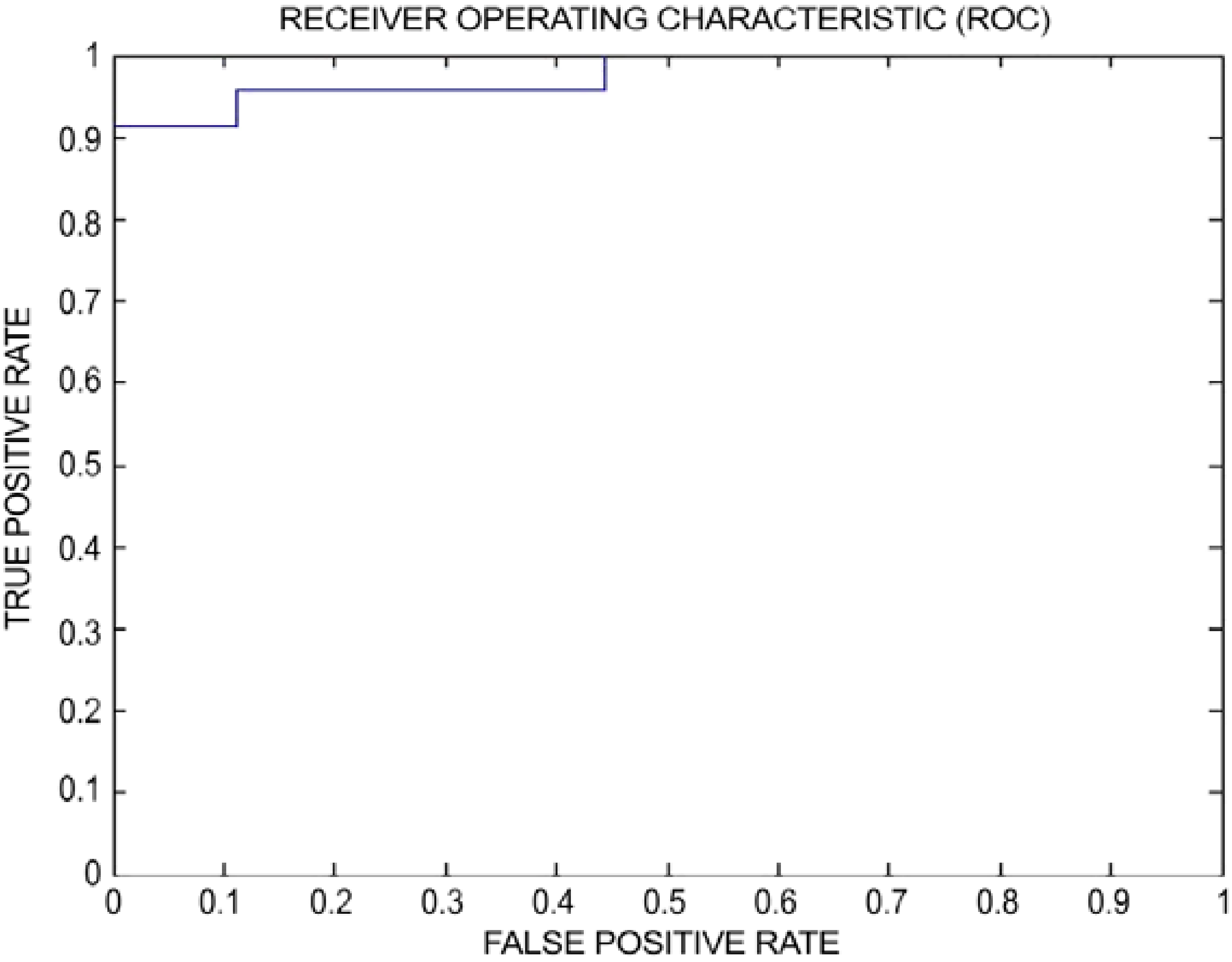

Figure 1: ROC curve for the classification using Euclidian distances and weighted instances of the substructures.

fication, the DrugBank (Wishart et al., 2006) compounds were obtained and converted to MOL2 format and tested for classification using them as unknown class. The Euclidian distance and weighted instances of the substructure feature is used in classification of the DrugBank Compounds. In the 1048 compounds of DrugBank, 437 compounds were found to have binding potential to P-gp. Of these we could validate 10 out of 12 binder and a single non binder compound common with our training data. Some of the DrugBank compounds classified as substrates and nonsubstrates are shown in Table 2. As the discriminating substructures vary with the training and test sets, the results are also varying in random sets and 5 fold cross validation. Most of the cases the Euclidian distance descriptor found to be giving good results and in some cases the combination of descriptors. The better results can be attributed to the presence of certain discriminating substructures that are specific to substrates and nonsubstrates of P-gp.

Many computational methods are available for predicting different pharmacological activities. Structure based approaches overcome some of the difficulties faced in classical QSAR methods based on various electronic, steric and hydrophobic properties, since they operate directly on the molecular structures and on their constituent substructures (Oprea and Matter, 2004). Inductive Logic Programming, Neural networks and SVM based on structure are available (Xu and Hagler, 2002; Srinivasan and King, 1999) to

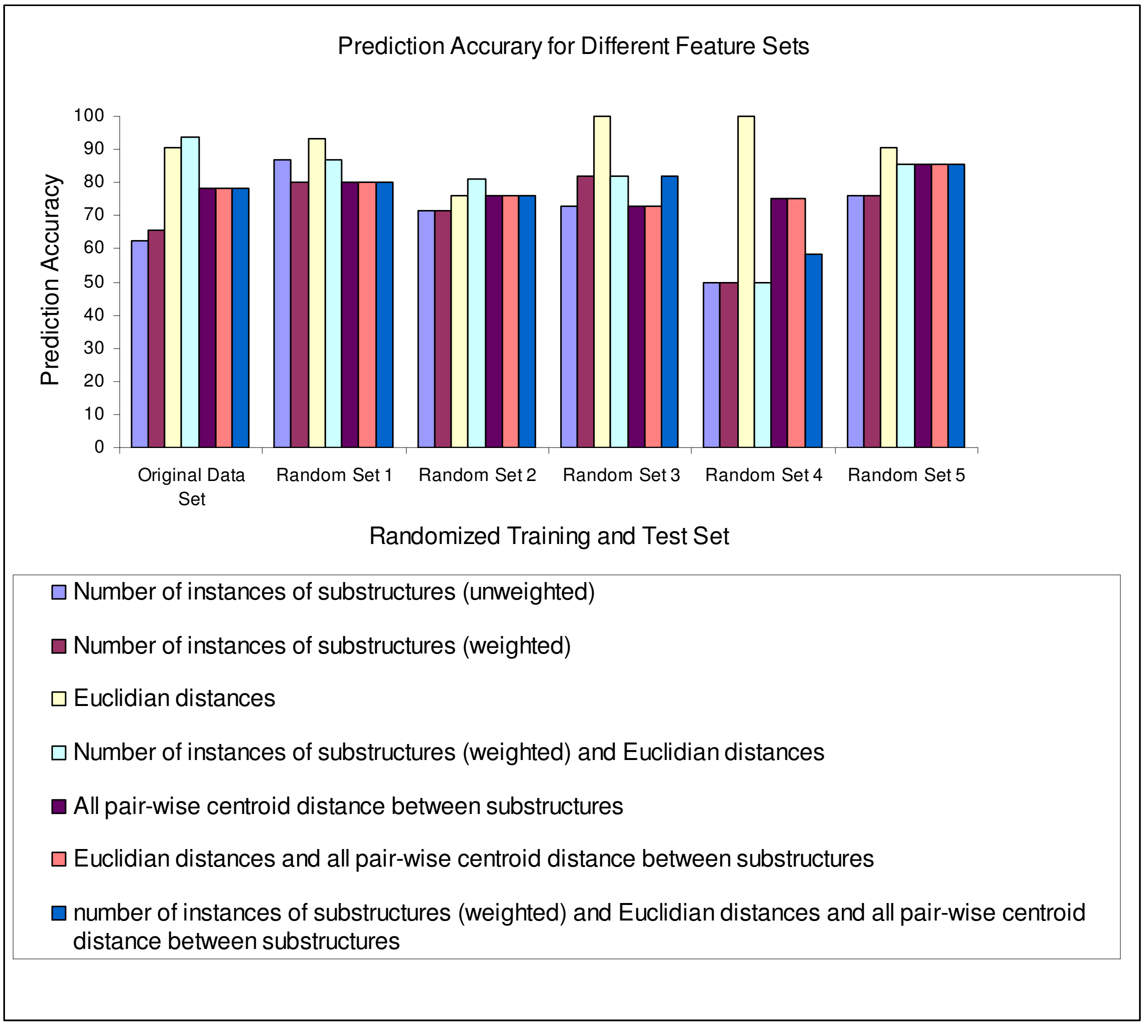

Figure 2: Comparison of the overall prediction accuracy (in %) using the following sets of features: (a) number of instances of substructures (unweighted), (b) number of instances of substructures (weighted), (c) Euclidian distances, (d) number of instances of substructures (weighted) and Euclidian distances (e) all pair-wise centroid distance between substructures (f) Euclidian distances and all pair-wise centroid distance between substructures, and (g) number of instances of substructures (weighted) and Euclidian distances and all pair-wise centroid distance between substructures for predicting binding potential to P-gp.

predict several adverse drug reactions such as torsadogenicity, carcinogenicity and mutagenicity based on molecular descriptors which can be physicochemical properties derived from structure or substructures extracted from the set of compounds (Burbidge et al., 2001; Deshpande et al., 2005; Yap et al., 2004). Support vector machine based on substructures have been shown to give better results compared to other descriptors based on physiochemical properties (Bhavani et al., 2006; Deshpande et al., 2005). In this study, it is interesting to note that we obtain an improvement of 13% over the SVM method of Xue et al which uses various atom type counts and connectivity indices as descriptors, a 15% increase in accuracy over the method of Cabrera et al that uses topological descriptors and 3% increase in accuracy over the method of Huang et al., that uses a SVM coupled with particle swarm optimization. Also combining our features with some of those already reported is likely to result in more useful classification. It is important

DrugBank Results	
Compounds classified as Substrates	Compounds classified as Nonsubstrates
Quinacrine	Cycloserine
Acebutolol	Phenytoin
Acetohexamide	Anastrozole
Cilostazol	Ethambutol
Dipyridamole	Triazolam
Toremifene	Methoxamine
Amiodarone	Methdilazine
Clotrimazole	Timolol
Erythromycin	Tamsulosin
Loperamide	Quinine

Table 2: DrugBank Compounds with generic names. Classification using Euclidian distances and weighted instances of the substructures

to recognize the different features that give rise to a high classification model such as that reported here, so that it will set the stage to explore in the future if combinations of such features can yield higher understanding of the correlations between the feature set, the classifier and the model and hence result in even higher prediction accuracies. Ability to detect those compounds that are likely to bind to P-gp, will help in weeding them out early in the drug discovery process, leading to significant reductions in the attrition rate. Given the speed and ease with which computational methods can be utilized for such purposes, they have the potential to be integrated as an essential component of the discovery processes.

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

1. Bhavani S, Nagargadde A, Thawani A, Sridhar V, Chandra N (2006) Substructure Based Support Vector Machine Classifiers for Prediction of Adverse Effects in Diverse Classes of Drugs. J Chem Inf Model 46: 2478-2486. » CrossRef » Pubmed » Google Scholar

2. Burbidge R, Trotter M, Buxton B, Holden S (2001) Drug design by machine learning: support vector machines for pharmaceutical data analysis. Comput Chem 26:5-14. » CrossRef » Pubmed » Google Scholar

3. Burges C (1998) A Tutorial on Support Vector Machines for Pattern Recognition. Data Min Knowl Discov 2: 121-167. » CrossRef » Google Scholar

4. Cabrera M, Gonzalez I, Fernandez C, Navarro C, Bermejo M (2006) A topological substructural approach for the prediction of P-glycoprotein substrates. J Pharm Sci 95: 589-606. » CrossRef » Pubmed » Google Scholar

5. Deshpande M, Kuramochi M, Wale N, Karypis G (2005) Frequent Substructure-Based Approaches for Classifying Chemical Compounds. IEEE Trans Knowl Data Eng 1036-1050. » CrossRef » Google Scholar

6. D ppenschmitt S, Langguth P, Regrdh CG, Andersson TB, Hilgendorf C, et al. (1999) Characterization of Binding Properties to Human P-Glycoprotein: Development of a [3H] Verapamil Radioligand Binding Assay. J Pharmacol Exp Ther 288: 348-357. » CrossRef » Pubmed » Google Scholar

7. Fawcett T (2004) ROC graphs: Notes and practical considerations for researchers. Machine Learning 38. » CrossRef » Google Scholar

8. Ghoting A, Buehrer G, Parthasarathy S, Kim D,

Nguyen A, et al. (2005) A characterization of data mining algorithms on a modern processor. Proceedings of the 1st international workshop on Data management on new hardwar. ACM New York NY USA p1. » CrossRef

9. Gottesman M, Fojo T, Bates S (2002) Multidrug resistance in cancer: role of ATP-dependent transporters. Nat Rev Cancer 2: 48-58. » CrossRef » Pubmed » Google Scholar

10. Hennessy M, Spiers J (2007) A primer on the mechanics of P-glycoprotein the multidrug transporter. Pharmacological Research 55: 1-15.» CrossRef » Pubmed » Google Scholar

11. Higgins C, Linton K (2001) Structural biology. The xyz of ABC transporters. Science 293: 1793-800. » CrossRef » Pubmed » Google Scholar

12. Huang J, Ma G, Muhammad I, Cheng Y (2007) Identifying P-Glycoprotein Substrates Using a Support Vector Machine Optimized by a Particle Swarm. J Chem Inf Model 47: 1638-1647. » CrossRef » Pubmed » Google Scholar

13. Joachims T (1999) Making large-scale support vector machine learning practical. Advances in kernel methods: support vector learning. MIT Press Cambridge, MA, USA 169 - 184. » CrossRef » Google Scholar

14. Kuramochi M, Karypis G (2004) An Efficient Algorithm for Discovering Frequent Subgraphs. IEEE Trans Knowl Data Eng 16: 1038-1051. » CrossRef » Pubmed » Google Scholar

15. Lu L, Leonessa F, Clarke R, Wainer I (2001) Competitive and allosteric interactions in ligand binding to P-glycoprotein as observed on an immobilized P-glycoprotein liquid chromatographic stationary phase. Mol Pharmacol 59: 62-68. » CrossRef » Pubmed » Google Scholar

16. McDevitt C, Callaghan R (2007) How can we best use structural information on P-glycoprotein to design inhibitors? Pharmacology and Therapeutics 113: 429-441. » CrossRef » Pubmed » Google Scholar

17. Oprea T, Matter H (2004) Integrating virtual screening in lead discovery. Current Opinion in Chemical Biology 8: 349-358.» CrossRef » Pubmed » Google Scholar

18. Penzotti J, Lamb M, Evensen E, Grootenhuis P (2002) A computational ensemble pharmacophore model for identifying substrates of P-glycoprotein. J Med Chem 45: 1737-1740. » CrossRef » Pubmed » Google Scholar

19. Provost F, Fawcett T (2001) Robust Classification for Imprecise Environments. Machine Learning 42: 203-231. » CrossRef » Google Scholar

20. Recanatini M, Bottegoni G, Cavalli A (2004) In silico antitarget screening. Drug Discovery Today: Technologies 1: 209-215.» CrossRef » Google Scholar

21. Schmitt L, Tampe R (2002) Structure and mechanism of ABC transporters. Curr Opin Struct Biol 12: 754-760. » CrossRef » Pubmed » Google Scholar

22. Shapiro A, Ling V (1994) ATPase activity of purified and reconstituted P-glycoprotein from Chinese hamster ovary cells. J Biol Chem 269: 3745-54. » CrossRef » Pubmed » Google Scholar

23. Sharom F (1995) Characterization and functional reconstitution of the multidrug transporter. J Bioenerg Biomembr 27: 15-22. » CrossRef » Pubmed » Google Scholar

24. Srinivasan A, King R (1999) Using Inductive Logic Programming to construct Structure-Activity Relationships. 64-73. » CrossRef » Google Scholar

25. von Grotthuss M, Koczyk G, Pas J, Wyrwicz L, Rychlewski L (2004) Ligand. Info small-molecule Meta-Database. Comb Chem High Throughput Screen 7 : 757-61. » CrossRef » Pubmed » Google Scholar

26. Wandel C, Kim R, Kajiji S, Guengerich P, Wilkinson G, et al., (1999) P-glycoprotein and cytochrome P-450 3A inhibition: dissociation of inhibitory potencies. Cancer Res 59: 3944-8. » CrossRef » Pubmed » Google Scholar

27. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M et al., (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res 34: D668-672. » CrossRef » Pubmed » Google Scholar

28. Xu J, Hagler A (2002) Chemoinformatics and Drug Discovery. Molecules 7: 566-600. » CrossRef » Google Scholar

29. Xue Y, Yap C, Sun L, Cao Z, Wang J, Chen Y et al., (2004) Prediction of P-glycoprotein substrates by a support vector machine approach. J Chem Inf Comput Sci 44: 1497-505. » CrossRef » Pubmed » Google Scholar

30. Yap C, Cai C, Xue Y, Chen Y (2004) Prediction of Torsade-Causing Potential of Drugs by Support Vector Machine Approach. Toxicol Sci 79: 170-177. » CrossRef » Pubmed » Google Scholar

31. Zhao C, Zhang H, Zhang X, Liu M, Hu Z, Fan B et al., (2006) Application of support vector machine (SVM) for prediction toxic activity of different data sets. Toxicology 217: 105-119.» CrossRef » Pubmed » Google Scholar