Prediction of P-Glycoprotein Substrates by a Support Vector Machine Approach

Y. Xue,[†,‡,§] C. W. Yap,[†] L. Z. Sun,[†] Z. W. Cao,[†] J. F. Wang,[†] and Y. Z. Chen*[,†]

Department of Computational Science, National University of Singapore, Blk SOC1, Level 7, 3 Science Drive 2, Singapore 117543, Singapore-MIT Alliance, E-04-10, 4 Engineering Drive 3, Singapore, 117576, and Department of Chemistry, Sichuan University, Chengdu, 610064, P. R. China

Received January 12, 2004

P-glycoproteins (P-gp) actively transport a wide variety of chemicals out of cells and function as drug efflux pumps that mediate multidrug resistance and limit the efficacy of many drugs. Methods for facilitating early elimination of potential P-gp substrates are useful for facilitating new drug discovery. A computational ensemble pharmacophore model has recently been used for the prediction of P-gp substrates with a promising accuracy of 63%. It is desirable to extend the prediction range beyond compounds covered by the known pharmacophore models. For such a purpose, a machine learning method, support vector machine (SVM), was explored for the prediction of P-gp substrates. A set of 201 chemical compounds, including 116 substrates and 85 nonsubstrates of P-gp, was used to train and test a SVM classification system. This SVM system gave a prediction accuracy of at least 81.2% for P-gp substrates based on two different evaluation methods, which is substantially improved against that obtained from the multiple-pharmacophore model. The prediction accuracy for nonsubstrates of P-gp is 79.2% using 5-fold cross-validation. These accuracies are slightly better than those obtained from other statistical classification methods, including *k*-nearest neighbor (*k*-NN), probabilistic neural networks (PNN), and C4.5 decision tree, that use the same sets of data and molecular descriptors. Our study indicates the potential of SVM in facilitating the prediction of P-gp substrates.

INTRODUCTION

P-glycoprotein (P-gp), encoded by the highly conserved multidrug (MDR) resistant genes, is an ATP-dependent drug efflux pump which can transport a diverse range of structurally and functionally unrelated substrates across the plasma membrane.[1,2] Overexpression of this protein may result in multidrug resistance and is a major cause of the failure of cancer chemotherapy[3,4] and diminished efficacy of antibiotics and antiviral agents.[5,6] Two approaches have been explored to circumvent MDR. One is the design of P-gp inhibitors[7,8] and another is to identify and eliminate drug candidates that are substrates of P-gp in the early stage of drug discovery.[9−12] Methods that facilitate the identification of P-gp substrates and inhibitors in a cost efficient and fast-speed manner are therefore useful for facilitating drug discovery.

Efforts have been directed at the development of computational methods for P-gp substrate prediction.[9−12] Molecular mechanism of P-gp mediated transport is not well understood, and the high-resolution structure of P-gp is unavailable.[1,2] Thus prediction methods are primarily based on statistical models derived from identification of structure−activity relationships,[9,10] structural recognition elements,[11] and multiple pharmacophores.[12] In particular, the multiple-pharmacophore model showed promising capability of P-gp substrate prediction for a large variety of compounds that conform to the known pharmacophores,[12] achieving a prediction accuracy of 63% for a set of 195 compounds.

Not all of the pharmaceutically interested substrates, agonists, and antagonists have available pharmacophore models. Therefore methods that extend the prediction range beyond those agents covered by known pharmacophore models are desired. One interesting method is the statistical learning method, support vector machine (SVM), which is useful for classification of systems with multiple mechanisms without requiring either the knowledge about their mechanisms or the intrinsic relationships between activities and molecular properties.[13−20] SVM was originally developed by Vapnik and co-workers[14,15] and have shown promising capability for solving a number of biological classification problems including prediction of blood-brain barrier penetration,[13] prediction of torsade-causing potential of drugs,[20] microarray gene expression data analysis,[16] protein fold recognition,[17] prediction of protein−protein interaction,[18] and protein function.[19] These studies have demonstrated that SVM is consistently superior to other supervised classification learning methods.[13,16,20,21]

This work explored the use of SVM as a potential tool for the prediction of P-gp substrates. Known P-gp substrates and nonsubstrates were used for training and testing a SVM classification system for recognition of physicochemical features of P-gp substrates. Through this learning-by-examples process, the trained SVM system can then be used for classifying a chemical compound as either a substrate or a nonsubstrate of P-gp. The classification accuracy of this system was evaluated by using two methods, evaluation by using an independent set of compounds and 5-fold cross-validation, and it is compared to the 5-fold cross-validation prediction accuracies derived from three other statistical classification methods using the same sets of data and

* Corresponding author phone: 65-6874-6877; fax: 65-6774-6756; e-mail: yzchen@cz3.nus.edu.sg.
† National University of Singapore.
‡ Singapore-MIT Alliance.
§ Sichuan University.

10.1021/ci049971e CCC: $27.50
Published on Web 05/08/2004

molecular descriptors, so as to objectively examine whether SVM is useful for P-gp substrate prediction.

METHODS

Selection of Substrates and Nonsubstrates of P-gp. P-gp substrates were collected from the literature such that each compound has been either described as being transported by P-gp or reported to induce overexpression of P-gp which directly contributes to MDR. Nonsubstrates of P-gp are those specifically described as not transportable by P-gp. A total of 116 substrates and 85 nonsubstrates of P-gp were collected. These compounds were further separated into training and testing sets by two different methods. The first method is an independent evaluation set to evaluate the classification accuracy. The second method is a 5-fold cross-validation.

In the first method, these compounds were further separated into three sets: training, testing, and independent validation set. The training set is used by SVM to develop a statistical model. The testing set is used by SVM to optimize the parameters of SVM classification algorithm, and the independent validation set is used for assessing the classification accuracy of the model. These compounds were divided into the three sets according to their distribution in the chemical space. Here, chemical space is defined by the commonly used structural and chemical descriptors.[22] Compounds of similar structural and chemical features are evenly assigned into separate sets. For those compounds without enough structurally and chemically similar counterparts, they were assigned, in order of priority, to the training and then the testing set, respectively. The training, testing, and independent validation sets are listed in Tables 1−3, respectively.

In the second method, the data set of 201 compounds was divided into five subsets of approximately equal size. After training the SVM with a collection of four subsets, the performance of the SVMs was tested against the fifth subset. This process is repeated five times so that every subset is once used as the test data.

Molecular Descriptors. Molecular descriptors have been routinely used for quantitative description of structural and physicochemical properties of molecules in statistical study of drugs and small molecules.[22−26] In this study, a set of 159 molecular descriptors was selected from the more than 1000 descriptors described in the literature by eliminating those descriptors that are obviously redundant or unrelated to the problem studied here. These descriptors, given in Table 4, include 18 descriptors in the class of simple molecular properties, 28 descriptors in the class of molecular connectivity and shape, 84 descriptors in the class of electrotopological state, 13 descriptors in the class of quantum chemical properties, and 16 descriptors in the class of geometrical properties. They were computed from the 3D structure of each compound using our own designed molecular descriptor computing program.[27] The remaining redundant and unrelated descriptors are further reduced by using feature selection methods.[28,29]

Examples of topological descriptors include number of rings and rotatable bonds, number of hydrogen bond acceptors and donors, molecular length vectors, molecular connectivity chi indices, molecular shape Kappa indices, electrotopological state indices, and atom type electrotopological state indices. Molecular connectivity chi indices and shape Kappa indices provide information about molecular size, shape, branching, unsaturation, heteroatom content, and cyclicity.[30,31] Electrotopological state indices encode information about both the topological environment of a particular atom and the electronic interactions from all the other atoms in the molecule.[32,33]

Quantum chemical descriptors are used for describing electrostatic and electronic properties of a molecule. These descriptors were computed by using molecular orbital energies and wave functions of electronic motion in a molecule, which were derived from the approximate solutions of the Schrödinger equation of electronic motion.[34] The computed quantum chemical descriptors include partial atomic charges, the highest occupied and lowest unoccupied molecular orbital energies, dipole moment, polarizability, and other descriptors derived from them.[35]

Geometric descriptors represent 3D-structural features of molecules. These include the van der Waals volume, solvent accessible surface area, molecular surface area, van der Waals surface area, and the related properties from combining them with partial atomic charges.[36,37]

Feature Selection Method. Feature selection methods have been introduced for the improvement of classification performance of statistical learning methods and for the selection of features meaningful for discriminating two data sets.[28,29,38−41] One approach, recursive feature elimination (RFE) method, has gained popularity due to its effectiveness for discovering informative features or attributes in cancer classification and drug activity analysis.[28,40] Thus, the RFE method was used in this work for selecting features relevant to P-gp substrate classification.

It has been suggested that the ranking criterion for feature selection can be formulated from the variation in an objective function upon removing each feature.[42] To improve the efficiency of SVM training, this objective function is represented by a cost function J for the ith feature, and it is computed by using the training set only. When the ith feature is removed or its weight w_i is reduced to zero, the variation of the cost function $DJ(i)$ is given by

$$DJ(i) = \frac{1}{2}\frac{\partial^2 J}{\partial w_i^2}(Dw_i)^2 \qquad (1)$$

The case of $Dw_i = w_i - 0$ corresponds to the removal of feature i.

Guyon et al. have used RFE to reduce the descriptors of a linear SVM classification system for cancer detection from gene selection data.[40] In the corresponding linear SVM classifier, the cost function is $J = (1/2)||\mathbf{w}||^2 - \alpha^T \mathbf{l}$, where $\mathbf{1}$ is an m dimensional identity vector (m is the number of compounds in the training set). Therefore $DJ(i) = (1/2)\, w_i^2$ and w_i^2 can be used as a feature ranking criterion. Yu et al. have used RFE to reduce the descriptors of a nonlinear SVM classification system of polynomial kernels for prediction of drug activity.[28] However, because of the diversity and complexity of the compounds to be classified, the use of linear and polynomial kernels may not always be sufficient for accurate prediction of various pharmaceutical and biological properties. Thus, in this work, SVM classification

Table 1. Substrates (Class 1) and Nonsubstrates (Class −1) of P-Glycoprotein in the Training Set

no.	compound	actual class	no.	compound	actual class
1	corticosterone	1	72	safingol	1
2	doxorubicin	1	73	phenoxazine	1
3	quinidine	1	74	vindoline	1
4	vinblastine	1	75	4 (adopted from ref 12)	−1
5	acetamido-deoxypodophyllotoxin	1	76	NSC667558	−1
6	fluphenazine	1	77	NSC676602	−1
7	hydrocortisone	1	78	NSC667532	−1
8	digoxin	1	79	prednisolone	−1
9	dexamethasone	1	80	aminodeoxy	−1
10	daunomycin	1	81	cortexolone	−1
11	HOE33342	1	82	methoxychlor	−1
12	GF120918−1	1	83	chlorambucil	−1
13	diltiazem	1	84	NSC674570	−1
14	colchicine	1	85	NSC49899	−1
15	cyclosporin-A	1	86	deoxypodophyllotoxin	−1
16	dibucaine	1	87	PSC833	−1
17	phodamine123	1	88	NSC630148	−1
18	digitoxigenin	1	89	NSC630721	−1
19	staurosporine	1	90	3 (adopted from ref 12)	−1
20	isosafrole	1	91	progesterone	−1
21	lovastatin	1	92	aldoxycarb	−1
22	fexofenadine	1	93	L767679	−1
23	nimodipine	1	94	BIBW22	−1
24	nelfinavir	1	95	NSC633528	−1
25	methadone	1	96	nigericin	−1
26	trifluoperazine	1	97	NSC653278	−1
27	monensin	1	98	NSC623083	−1
28	ondansetron	1	99	NSC668354	−1
29	indinavir	1	100	reserpic_acid	−1
30	dexniguldipine	1	101	fluazifop-butyl	−1
31	saquinavir	1	102	NSC664565	−1
32	S-farnesylcysteine-methylester	1	103	tamoxifen	−1
33	reserpine	1	104	NSC667560	−1
34	LY335979	1	105	cytarabine	−1
35	mitoxantrone	1	106	NSC615985	−1
36	topotecan	1	107	NSC678047	−1
37	dipyridamole	1	108	NSC676610	−1
38	haloperidol	1	109	carbaryl	−1
39	estradiol	1	110	aldicarb	−1
40	azidopine	1	111	carmustine	−1
41	toremifene	1	112	cyclophosphamide	−1
42	paclitaxel	1	113	epinephrine	−1
43	thioridazine	1	114	fluorouracil	−1
44	morphine-6-glucuronide	1	115	lindane	−1
45	nifedipine	1	116	NSC314622	−1
46	actinomycin_d	1	117	midazolam	−1
47	cefoperazone	1	118	NSC268251	−1
48	triflupromazine	1	119	NSC606532	−1
49	amiodarone	1	120	NSC617286	−1
50	cefazolin	1	121	NSC639677	−1
51	cefotetan	1	122	NSC648403	−1
52	clotrimazole	1	123	NSC666331	−1
53	erythromycin	1	124	NSC671400	−1
54	flunitrazepam	1	125	NSC686028	−1
55	loperamide	1	126	S_farnesyl_cysteine	−1
56	methotrexate	1	127	aminocarb	−1
57	phenobarbital	1	128	atrazine	−1
58	phenytoin	1	129	chaps	−1
59	prazosin	1	130	dialifos	−1
60	promazine	1	131	dieldrin	−1
61	ritonavir	1	132	leptophos	−1
62	tetraphenylphosphonium	1	133	mirex	−1
63	bisantrene	1	134	phosmet	−1
64	endosulfan	1	135	systeine_methylester	−1
65	estriol	1	136	triforine	−1
66	ivermectin	1	137	trypan_blue	−1
67	leupeptin	1	138	vinclozolin	−1
68	mithramycin	1	139	NSC667551	−1
69	pararosaniline	1	140	NSC676615	−1
70	rapamycin	1	141	epipodophyllotoxin	−1
71	S9788	1	142	deoxycorticosterone	−1

Table 2. Substrates (Class 1) and Nonsubstrates (Class −1) of P-Glycoprotein in the Testing Set

no.	compound	actual class	predicted class	no.	compound	actual class	predicted class
1	epirubicin	1	1	18	docetaxel	1	1
2	quinine	1	1	19	mitomycin-C	1	1
3	vincristine	1	1	20	morphine	1	1
4	cis-flupenthixol	1	1	21	valinomycin	1	1
5	digitoxin	1	1	22	teniposide	1	1
6	methylprednisolone	1	1	23	epothilone_a	1	1
7	idarubicin	1	1	24	1 (adopted from ref 12)	−1	−1
8	verapamil	1	1	25	2 (adopted from ref 12)	−1	−1
9	pafenolol	1	1	26	farnesol	−1	−1
10	digoxigenin	1	1	27	melphalan	−1	−1
11	terfenadine	1	1	28	mevinphos	−1	−1
12	spiperone	1	1	29	paraquat	−1	−1
13	cinchonidine	1	1	30	propiconazole	−1	−1
14	methylreserpate	1	1	31	NSC676593	−1	−1
15	celiprolol	1	1	32	NSC676618	−1	−1
16	cepharanthine	1	1	33	NSC674508	−1	−1
17	puromycin	1	1	34	NSC309132	−1	−1

Table 3. Substrates (Class 1) and Nonsubstrates (Class −1) of P-Glycoprotein in the Independent Validation Set

no.	compound	actual class	predicted class	no.	compound	actual class	predicted class
1	acebutolol	1	1	14	k02	1	1
2	adriamycin	1	1	15	losartan	1	1
3	aldosterone	1	1	16	nicardipine	1	1
4	calphostin_c	1	1	17	perphenazine	1	1
5	catharantine	1	−1	18	rifampicin	1	1
6	chlorpromazine	1	1	19	yohimbine	1	−1
7	CP100356	1	1	20	NSC364080	−1	1
8	depredil	1	−1	21	NSC630357	−1	1
9	domperidone	1	1	22	NSC667533	−1	−1
10	emetine	1	1	23	NSC676617	−1	−1
11	etoposide	1	1	24	NSC676616	−1	−1
12	gallopamil	1	1	25	podophyllotoxin	−1	−1
13	hydroxyrubicin	1	1				

Table 4. Molecular Descriptors Used in This Work

descriptor class	number of descriptors in class	descriptors
simple molecular properties	18	molecular weight, number of ring structures, number of rotatable bonds, number of H-bond donors, number of H-bond acceptors, element counts
molecular connectivity and shape	28	molecular connectivity indices, valence molecular connectivity indices, molecular shape Kappa indices, Kappa alpha indices, flexibility index
electro-topological state	84	electrotopological state indices and atom type electrotopological state indices
quantum chemical properties	13	atomic charge on the most positively charged H atom, largest negative charge on an non-H atom, polarizability index, hydrogen bond acceptor basicity (covalent HBAB), hydrogen bond donor acidity (covalent HBDA), molecular dipole moment, absolute hardness, softness, ionization potential, electron affinity, chemical potential, electronegativity index, electrophilicity index
geometrical properties	16	molecular size vectors (distance of the longest separated atom pairs, combined distance of the longest separated three atoms, combined distance of the longest separated four atoms), molecular van der Waals volume, solvent accessible surface area, molecular surface area, van der Waals surface area, polar molecular surface area, sum of solvent accessible surface areas of positively charged atoms, sum of solvent accessible surface areas of negatively charged atoms, sum of charge weighted solvent accessible surface areas of positively charged atoms, sum of charge weighted solvent accessible surface areas of negatively charged atoms, sum of van der Waals surface areas of positively charged atoms, sum of van der Waals surface areas of negatively charged atoms, sum of charge weighted van der Waals surface areas of positively charged atoms, sum of charge weighted van der Waals surface areas of negatively charged atoms

systems of Gaussian kernels were used. In this case, the cost function to be minimized, under the constraints $0 \le \alpha_k \le C$ and $\sum_k \alpha_k y_k = 0$, is

$$J = (1/2)\alpha^T \mathbf{H} \alpha - \alpha^T \mathbf{1} \qquad (2)$$

where $\mathbf{H}$ is the matrix with elements $y_i y_j \exp(-|x_i - x_j|^2/(2\sigma^2))$.

To compute the variation in the cost function upon removal of the input component i, the parameters α's were kept unchanged and the matrix $\mathbf{H}$ was recomputed. The resulting ranking coefficient is

$$DJ(i) = (1/2)\alpha^T \mathbf{H} \alpha - (1/2)\alpha^T \mathbf{H}(-i)\alpha \qquad (3)$$

where $\mathbf{H}(-i)$ is the matrix computed by using the same

method as that of matrix **H** but with its *i*th component removed. One or more of the features with the smallest *DJ*(*i*) can thus be eliminated.

The computation procedure in this work is outlined as the follows: The SVM classification system for this study was trained by using a Gaussian kernel function. The training was conducted by sequential variation of the parameter σ in the special region against the whole training data set. The prediction accuracy of this SVM system during the training process was evaluated by means of 5-fold cross-validation. In the first step, for a fixed σ, the SVM classifier is trained by using the complete set of features (molecular descriptors) described in the previous section. The second step is to compute the ranking criterion score *DJ*(*i*) for each feature in the current set by using eq 3. All of the computed *DJ*(*i*) is subsequently ranked in descending order. The third step is to remove the *m* features with the smallest criterion scores. In this work, *m* was chosen to be 5 as that used in earlier studies.[28] In the fourth step, the SVM classification system is retrained by using the remaining set of features, and the corresponding prediction accuracy is computed by means of 5-fold cross-validation. The first to fourth steps are then repeated for other values of σ. After the completion of these procedures, the set of features and parameter σ that give the best prediction accuracy are selected.

SVM Algorithm. The theory of SVM has been extensively described in the literature.[14,15,23] Thus only a brief description is given here. SVM is based on the structural risk minimization (SRM) principle from statistical learning theory.[14] In linearly separable cases, SVM constructs a hyperplane which separates two different classes of vectors with a maximum margin. A vector corresponds to the features of a drug in this work, and this vector is represented by $\mathbf{x}_i$, with structural and physicochemical descriptors of a molecule as its components. This is done by finding another vector **w** and a parameter *b* that minimizes $||\mathbf{w}||^2$ and satisfies the following conditions

$$\mathbf{w}\cdot\mathbf{x}_i + b \geq +1, \text{ for } y_i = +1 \text{ class 1 (positive)} \quad (4)$$

$$\mathbf{w}\cdot\mathbf{x}_i + b \leq -1, \text{ for } y_i = -1 \text{ class 2 (negative)} \quad (5)$$

where y_i is the class index, **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 **w**. After the determination of **w** and *b*, a given vector $\mathbf{x}_i$ can be classified by

$$sign[(\mathbf{w}\cdot\mathbf{x}) + b] \quad (6)$$

In nonlinearly separable cases, SVM maps the input variable into a higher 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.[13,43,44]

$$K(\mathbf{x}_i,\mathbf{x}_j) = e^{-||\mathbf{x}_j-\mathbf{x}_i||^2/2\sigma^2} \quad (7)$$

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

$$f(\mathbf{x}) = sign\left(\sum_{i=1}^{l}\alpha_i^0 y_i K(\mathbf{x},\mathbf{x}_i) + b\right) \quad (8)$$

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

$$\sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i,\mathbf{x}_j) \quad (9)$$

under the following conditions:

$$a_i \geq 0 \text{ and } \sum_{i=1}^{l}\alpha_i y_i = 0 \quad (10)$$

A positive or negative value from eq 6 or eq 8 indicates that the vector **x** belongs to the positive or negative class, respectively.

As in the case of all discriminative methods,[45,46] the performance of SVM classification can be measured by the quantity of true positives TP, true negatives TN, false positives FP, false negatives FN, sensitivity SE = TP/(TP+FN) which is the prediction accuracy for the substrates of P-gp in this work, and specificity SP = TN/(TN+FP) which is the prediction accuracy for the nonsubstrates of P-gp in this work. The overall prediction accuracy (*Q*) is also used to measure the prediction accuracies and can be given by

$$Q = \frac{TP + TN}{TP + TN + FP + FN} \quad (11)$$

Other Statistical Classification Systems. To objectively examine whether SVM is useful for P-gp substrate prediction, prediction accuracies of the trained SVM system were compared with those derived from three other classification methods by using 5-fold cross-validation. These methods are *k*-nearest neighbor (KNN),[47,48] probabilistic neural network (PNN),[49] and C4.5 decision tree.[50] In KNN, the Euclidean distance between an unclassified vector **x** and each individual vector $\mathbf{x}_i$ in the training set is measured.[47,48] A total of *k* number of vectors nearest to the unclassified vector **x** are used to determine the class of that unclassified vector. The class of the majority of the *k*-nearest neighbors is chosen as the predicted class of the unclassified vector **x**.

PNN is a form of neural network that uses Bayes optimal decision rule for classification.[49] Traditional neural networks such as feed-forward back-propagation neural network rely on multiple parameters and network architectures to be optimized. In contrast, PNN only has a single adjustable parameter, a smoothing factor σ for the radial basis function in the Parzen's nonparameteric estimator. Thus the training process of PNN is usually orders of magnitude faster than those of the traditional neural networks.

C4.5 decision tree is a branch-test-based classifier.[50] A branch in a decision tree corresponds to a group of classes and a leaf represents a specific class. A decision node specifies a test to be conducted on a single attribute value, with one branch and its subsequent classes as possible outcomes of the test. C4.5 decision tree uses recursive partitioning to examine every attribute of the data and rank them according to their ability to partition the remaining data, thereby constructing a decision tree. A vector **x** is classified

Table 5. SVM Prediction Accuracy for the Substrates and Nonsubstrates of P-gp by Using Independent Evaluation Sets[a]

		testing set				independent validation set					
		substrates		nonsubstrates		substrates			nonsubstrates		
training set		TP	FN	TN	FP	TP	FN	SE (%)	TN	FP	SP (%)
74	68	22	0	12	0	16	3	84.2	4	2	66.7

[a] Predicted results are given in TP (true positive), FN (false negative), TN (true negative), FP (false positive), SE (sensitivity) which is the prediction accuracy for substrates, and SP (specificity) which is the prediction accuracy for nonsubstrates. Number of substrates or nonsubstrates in testing and independent evaluation sets is TP+FN or TN+FP, respectively.

Table 6. SVM Prediction Accuracy of the Substrates and Nonsubstrates of P-Glycoprotein by Using 5-Fold Cross-Validation

cross-validation	substrates			nonsubstrates			
	TP	FN	SE (%)	TN	FP	SP(%)	*Q* (%)
1	17	7	70.8	12	4	75.0	72.5
2	15	2	88.2	11	5	68.8	78.8
3	30	8	78.9	13	1	92.9	82.7
4	15	4	78.9	15	3	83.3	81.1
5	16	2	88.9	16	5	76.2	82.1
av			81.2			79.2	79.4
SE			7.5			9.2	4.2

Table 7. Comparison of the Prediction Accuracy of the Substrates and Nonsubstrates of P-Glycoprotein from Different Classification Methods by Using 5-Fold Cross-Validation[a]

method	substrates SE (%)	nonsubstrates SP (%)	*Q* (%)
k-NN	79.2	61.6	70.8
PNN	77.3	71.4	74.4
C4.5 decision tree	74.6	69.9	71.5
SVM	81.2	79.2	79.4

[a] These methods include *k*-NN, PNN, and C4.5 decision tree as well as SVM.

by starting at the root of the tree and moving through the tree until a leaf is encountered. At each nonleaf decision node, a test is conducted, and the classification process proceeds to the branch selected by the test. Upon reaching the destination leaf, the class of the vector **x** is predicted to be that associated with the leaf.

RESULTS AND DISCUSSION

SVM prediction of both substrates and nonsubstrates of P-gp was evaluated by means of both the use of independent evaluation set and 5-fold cross-validation. The results of these two methods are given in Tables 5 and 6, respectively. The accuracy for the prediction of the P-gp substrate using 5-fold cross-validation is 81.2% and that by using the independent validation set is 84.2% respective. Thus both methods appear to give consistent assessment about the prediction accuracy.

A direct comparison with results from a previous study is inappropriate because of differences in the use of a data set, molecular descriptors, and classification methods. Although desirable, it is impossible to conduct a separate comparison with results from other studies without full information about the algorithms of molecular descriptors and classification methods used in each study. Nonetheless, a tentative comparison may provide some crude estimate regarding the approximate level of accuracy of our method with respect to those achieved by other studies. The prediction accuracy for P-gp substrates is substantially improved with respect to the value of 63% derived from the ensemble pharmacophore model.[12]

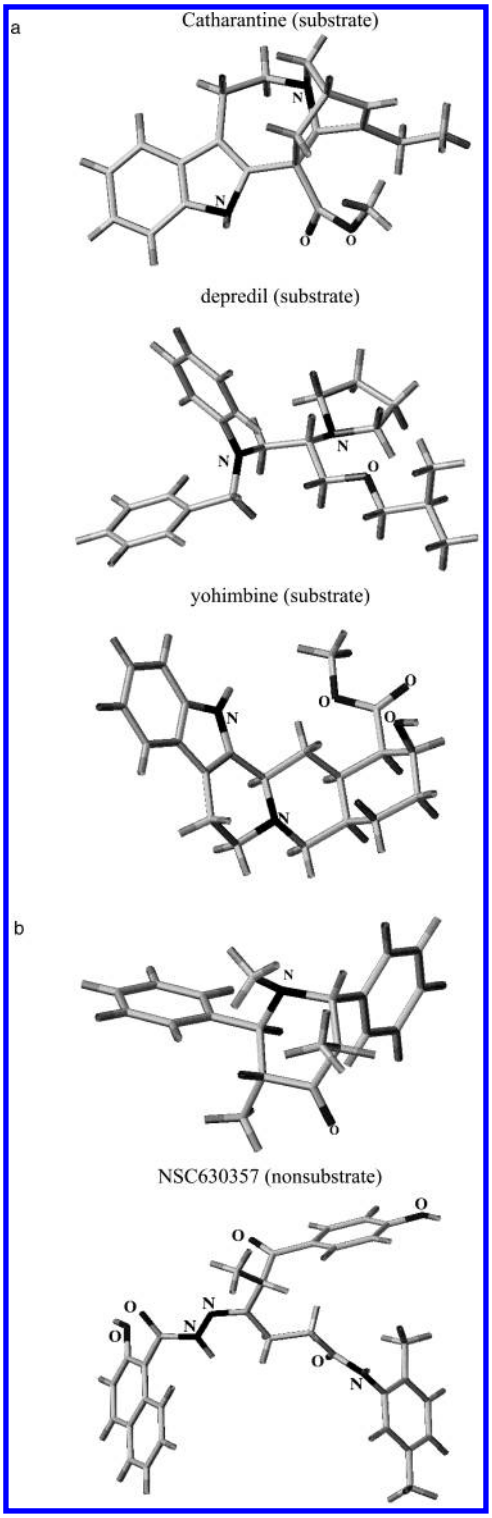

Figure 1. Unoptimized structures of misclassified compounds in the independent validation set.

The prediction accuracy for nonsubstrates of P-gp is 79.2% using 5-fold cross-validation and 66.7% using independent evaluation set. The substantially lower accuracy derived from the independent evaluation set likely arises because of the small number of P-gp nonsubstrates in the set. Another factor

Table 8. Molecular Descriptors Selected from the Feature Selection Method for Classification of P-gp Substrates and Nonsubstrates

descriptors	description	class
Ncocl	count of Cl atoms	simple molecular properties
Ndonr	number of H-bond donors	simple molecular properties
$^5\chi_{CH}$	simple molecular connectivity Chi indices for cycle of 5 atoms	connectivity and shape
$^3\chi^v_P$	valence molecular connectivity Chi indices for path order 3	connectivity and shape
$^5\chi^v_{CH}$	valence molecular connectivity Chi indices for cycle of 5 atoms	connectivity and shape
Scar	sum of Estate indices of carbon atoms	geometrical properties
dis2	length vector (longest third atom)	geometrical properties
Sapcw	sum of charge weighted solvent accessible surface areas of positively charged atoms	geometrical properties
S(1)	atom-type H Estate sum for −O**H**	electro-topological state
S(9)	atom-type H Estate sum for ═C**H**− (sp^2)	electro-topological state
S(12)	atom-type H Estate sum for C$\mathbf{H}_n$ (saturated)	electro-topological state
S(13)	atom-type H Estate sum for C$\mathbf{H}_n$ (unsaturated)	electro-topological state
S(16)	atom-type Estate sum for −CH_3	electro-topological state
S(18)	atom-type Estate sum for >CH_2	electro-topological state
S(20)	atom-type Estate sum for ═CH−	electro-topological state
S(21)	atom-type Estate sum for: CH: (aromatic)	electro-topological state
S(25)	atom-type Estate sum for ═C<	electro-topological state
S(36)	atom-type Estate sum for >N−	electro-topological state
π_i	polarizability index	quantum chemical properties
q^+	atomic charge on the most positively charged H atom	quantum chemical properties
μ	molecular dipole moment	quantum chemical properties
ω	electrophilicity index	quantum chemical properties

is the inadequate sampling of the chemical space covered by nonsubstrates of P-gp. It is likely that the 85 nonsubstrates collected in this work only represent a portion of all possible classes of nonsubstrates of P-gp. Protein nonsubstrates are rarely described in the literature, thus additional efforts are needed to enable the collection of this information.

SVM classification results were further compared to those from other statistical classification methods including *k*-nearest neighbor (*k*-NN), probabilistic neural networks (PNN), and C4.5 decision tree. The same sets of data and descriptors are used in these computations. The results are shown in Table 7, and it is found that the accuracy from SVM classification is slightly better than those from other classification methods. This suggests that SVM is capable of prediction of P-gp substrates and P-gp nonsubstrates at a comparable or perhaps better accuracy with respect to that from other classification methods without requiring either the knowledge of mechanism or the intrinsic structure−activity relationships.

SVM typically uses a portion of the training set as support vectors for classification. In contrast, *k*-NN and PNN use the whole training set for classification. Our own studies suggest that the number of support vectors of SVM is in the range of 40−70% of the training set. Thus the classification speed of SVM is usually 30−60% faster than that of *k*-NN and PNN. On the other hand, the classification speed of SVM is slower than that of decision tree methods which conduct tests on descriptors to reach a decision leaf.

In the independent evaluation set, there are 3 and 2 incorrectly classified substrates and nonsubstrates of P-gp, respectively, which are shown in Figure 1. The three P-gp substrates are catharantine, depredil, and yohimbine, and the two nonsubstrates of P-gp are NSC364080 and NSC630357. Each of the three misclassified P-gp substrates is composed of an inflexible multiring structure and a short flexible hydrophilic tail. The inflexible region of catharantine is composed of 6 tightly connected stereo-rings. Yohimbine contains a rigid 21-membered ring fraction. Depredil has three separate rings triangularly connected to each other by a few rotatable bonds. The two misclassified nonsubstrates of P-gp, NSC364080 and NSC630357, both contain the three ring structure, similar to depredil, but without a short flexible hydrophilic chain. This flexible hydrophilic chain appears to be a factor that distinguishes between a substrate and a nonsubstrate of P-gp.

While encoding molecular shape and flexibility features, topological descriptors may not adequately describe the detailed configuration of a large rigid structure combined with a short flexible hydrophilic tail in the molecule. Therefore our analysis seems to suggest that the incorrect classification of these five compounds arises from an inadequate description of the flexibility about the short hydrophilic tail attached to bulky rigid ring structures.

Table 8 gives the molecular descriptors selected from the feature selection method RFE. Those from the class of topological descriptors constitute the largest percentage of the descriptors selected. This is consistent with the findings from the classification of multidrug-resistant (MDR) agents, many of which are P-gp substrates, by using structure-based descriptors and linear discriminant analysis, which showed that 60% of the molecular descriptors important for MDR are topological in nature.[8] A study of quantitative structure−activity relationships (QSAR) of MDR agents also identified several biophores, e.g., a generic form of C−C−X−C−C with X = N, NH, or O (preferably a tertiary nitrogen), as a key structural element for MDR.[7] These biophores are primarily determined by electrotopological features and bond connectivity. In addition to the large percentage of electro-topological, the RFE method also selected three molecular connectivity descriptors, which seem to correlate with the features of the biophores identified from the QSAR study of MDR agents.

The rest of the RFE selected descriptors are from the quantum chemical class and simple molecular property class. The selected quantum chemical descriptors determine polarizability, molecular dipole moment, electrophilicity, and the atomic charge of the positively charged hydrogen atoms in a molecule. The selected simple molecular property descriptors give the number of hydrogen bond donors and that of Cl atoms. With the exception of the last descriptor,

the MolSurf counterparts of these quantum chemical and simple molecular property descriptors have been used for the prediction of P-gp-interacting drugs by means of the multivariate statistics method.[51] Based on structural comparison, it has been found that the number of electron donors and hydrogen bond acceptor groups are important elements for P-gp substrate recognition.[11] An analysis of multiple pharmacophores of P-gp substrates has identified hydrophobe, hydrogen bond donor and acceptor as important elements for P-gp substrates.[12] Thus these studies consistently suggested the importance of the selected quantum chemical features and hydrogen-bond property for prediction of P-gp substrates and nonsubstrates.

The other RFE selected descriptor, the count of Cl atoms, has not been specifically used in other P-gp substrate studies. One possible reason is that the molecules used in those studies do not contain a Cl atom, thus it is unnecessary to introduce this descriptor in those studies. In this work, the descriptor for hydrogen bond acceptor was not selected by RFE, which has been found to be an important element for P-gp substrates in other studies.[11,12] One likely reason for the exclusion of this descriptor is that it has a high level of redundancy with the relevant features covered by the quantum chemical descriptors such as electrophilicity, polarizability, and molecular dipole moment when they are combined with the hydrogen bond donor descriptor.

CONCLUSION

SVM is a potentially useful computational method for facilitating the prediction of P-gp substrates. Prediction accuracy may be further improved by consideration of factors such as hydrogen bonding, active transport, and relationship with pharmacodynamic properties. Moreover, recent works on the introduction of weighting function into SVM descriptors[52] may also be helpful in developing SVM into a practical tool for the prediction of P-gp substrates and thus facilitate new drug development.

REFERENCES AND NOTES

(1) Schmitt, L.; Tampe, R. Structure and mechanism of ABC transporters. *Curr. Opin. Struct. Biol.* **2002**, *12*, 754−760.

(2) van Veen, H. W.; Konings, W. N. Structure and function of multidrug transporters. *Adv. Exp. Med. Biol.* **1998**, *456*, 145−158.

(3) Gottesman, M. M.; Pastan, I.; Ambudkar, S. V. P-glycoprotein and multidrug resistance. *Curr. Opin. Genet. Dev.* **1996**, *6*, 610−617.

(4) Ambudkar, S. V.; Dey, S.; Hrycyna, C. A.; Ramachandra, M.; Pastan, I.; Gottesman, M. M. Biochemical, cellular, and pharmacological aspects of the multidrug transporter. *Annu. Rev. Pharmacol. Toxicol.* **1999**, *39*, 361−398.

(5) Delph, Y. P-glycoprotein: a tangled web waiting to be unraveled. http://www.aidsinfonyc.org/tag/science/pgp.html 2000.

(6) Kim, R. B.; Fromm, M. F.; Wandel, C.; Leake, B.; Wood, A. J.; Roden, D. M.; Wilkinson, G. R. The drug transporter P-glycoprotein limits oral absorption and brain entry of HIV-1 protease inhibitors. *J. Clin. Invest.* **1998**, *101*, 289−294.

(7) Klopman, G.; Shi, L. M.; Ramu, A. Quantitative structure−activity relationship of multidrug resistance reversal agents. *Mol. Pharmacol.* **1997**, *52*, 323−334.

(8) Bakken, G. A.; Jurs, P. C. Classification of multidrug-resistance reversal agents using structure-based descriptors and linear discriminant analysis. *J. Med. Chem.* **2000**, *43*, 4534−4541.

(9) Bain, L. J.; McLachlan, J. B.; LeBlanc, G. A. Structure−activity relationships for xenobiotic transport substrates and inhibitory ligands of P-glycoprotein. *Environ. Health Perspect.* **1997**, *105*, 812−818.

(10) Litman, T.; Zeuthen, T.; Skovsgaard, T.; Stein, W. D. Structure−activity relationships of P-glycoprotein interacting drugs: kinetic characterization of their effects on ATPase activity. *Biochim. Biophys. Acta* **1997**, *1361*, 159−168.

(11) Seelig, A. A general pattern for substrate recognition by P-glycoprotein. *Eur. J. Biochem.* **1998**, *251*, 252−261.

(12) Penzotti, J. E.; Lamb, M. L.; Evensen, E.; Grootenhuis, P. D. J. A computational ensemble pharmacophore model for identifying substrates of P-glycoprotein. *J. Med. Chem.* **2002**, *45*, 1737−1740.

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

(14) Vapnik, V. N. *The nature of statistical learning theory*; Springer: New York, 1995.

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

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

(17) Ding, C. H. Q.; Dubchak, I. Multi class protein fold recognition using support vector machines and neural networks. *Bioinformatics* **2001**, *17*, 349−358.

(18) Bock, J. R.; Gough, D. A. Predicting protein−protein interactions from primary structure. *Bioinformatics* **2001**, *17*, 455−460.

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

(20) 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.

(21) Cai, Y. D.; Liu, X. J.; Xu, X. B.; Chou, K. C. Support vector machines for predicting HIV protease cleavage sites in protein. *J. Comput. Chem.* **2002**, *23*, 267−274.

(22) Todeschini, R.; Consonni, V. *Handbook of molecular descriptors*; Wiley-VCH: Weinheim, 2000.

(23) Katritzky, A. R.; Gordeeva, E. V. Traditional topological indices vs electronic, geometrical and combined molecular descriptors in QSAR/QSPR research. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 835−857.

(24) Cruciani, G.; Pastor, M.; Guba, W. Volsurf: a new tool for the pharmacokinetic optimization of lead compounds. *Eur. J. Pharm. Sci.* **2000**, *11*, S29−S39.

(25) Kier, L. B.; Hall, L. H. *Molecular structure description: The electrotopological state*; Academic Press: San Diego, 1999.

(26) Karelson, M.; et al. Quantum-chemical descriptors in QSAR/QSPR studies. *Chem. Rev.* **1996**, *96*, 1027−1043.

(27) Xue, Y.; Yap, C. W.; Li, Z. R.; Chen, Y. Z. Evaluation of a method for improving the computation speed of molecular descriptors for drug property analysis. *Acta Pharmacol. Sin.* **2004**, Submitted for publication.

(28) Yu, H.; Yang, J.; Wang, W.; Han, J. Discovering compact and highly discriminative features or feature combinations of drug activities using support vector machines. *Proc. IEEE Comput. Soc. Bioinformatics Conf. (CSB)* **2003**, 220−228.

(29) Degroeve, S.; De Baets, B.; Van de Peer, Y.; Rouzé, P. Feature subset selection for splice site prediction. *Bioinformatics* **2002**, *18*, S75−S83.

(30) Kier, L. B.; Hall, L. H. *Molecular connectivity in structure−activity analysis*; Research Studies Press: Wiley: Letchworth, Hertfordshire, England, New York, 1986.

(31) Hall, L. H.; Kier, L. B. The molecular connectivity chi indices and kappa shape indices in structure−property modeling. In *Reviews of Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers: New York; 1991; Vol. 2, pp 367−412.

(32) Hall, L. H.; Mohney, B. K.; Kier, L. B. The electrotopological state: Structure information at the atomic level for molecular graphs. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 76−82.

(33) Hall, L. H.; Kier, L. B. Electrotopological state indices for atom types: A novel combination of electronic, topological, and valence state information. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039−1045.

(34) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. Development and use of quantum mechanical molecular models, 76. AM1: a new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.

(35) Thanikaivelan, P.; Subramanian, V.; Raghava, J.; Rao, J. R.; Nair, B. U. Application of quantum chemical descriptors in quantitative structure activity and structure property relationship. *Chem. Phys. Lett.* **2000**, *323*, 59−70.

(36) Hopfinger, A. J. A QSAR investigation of dihydrofolate reductase inhibition by Baker triazines based upon molecular shape analysis. *J. Am. Chem. Soc.* **1980**, *102*, 7196−7206.

(37) Tsodikov, O. V.; Record, M. T. J.; Sergeev, Y. V. Novel computer program for fast exact calculation of accessible and molecular surface areas and average surface curvature. *J. Comput. Chem.* **2002**, *23*, 600−609.

(38) Bayada, D. M.; Hamersma, H.; van Geerestein, V. J. Molecular diversity and representativity in chemical databases. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1−10.
(39) Furey, T. S.; Cristianini, N.; Duffy, N.; Bednarski, D. W.; Schummer, M.; Haussler, D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* **2000**, *16*, 906−914.
(40) Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **2002**, *46*, 389−422.
(41) Furlanello, C.; Serafini, M.; Merler, S.; Jurman, G. An accelerated procedure for recursive feature ranking on microarray data. *Neural Networks* **2003**, *16*, 641−648.
(42) Kohavi, R.; John, G. H. Wrappers for feature subset selection. *Artif. Intelligence* **1997**, *97*, 273−324.
(43) Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Comput. Chem.* **2001**, *26*, 5−14.
(44) Czerminski, R.; Yasri, A.; Hartsough, D. Use of support vector machine in pattern classification: Application to QSAR studies. *Quant. Struct.-Act. Relat.* **2001**, *20*, 227−240.
(45) Roulston, J. E. Screening with tumor markers. *Mol. Pharmacol.* **2002**, *20*, 153−162.
(46) Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C. A.; Nielsen, H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **2000**, *16*, 412−424.
(47) Huberty, C. J. *Applied discriminant analysis*; John Wiley & Sons: New York, 1994.
(48) Johnson, R. A.; Wichern, D. W. *Applied multivariate statistical analysis*; Prentice Hall: Englewood Cliffs, NJ, 1982.
(49) Specht, D. F. Probabilistic neural networks. *Neural Networks* **1990**, *3*, 109−118.
(50) Quinlan, J. R. *C4.5: programs for machine learning*; Morgan Kaufmann: San Mateo, CA, 1993.
(51) Osterberg, T.; Norinder, U. Theoretical calculation and prediction of P-glycoprotein-interacting drugs using MolSurf parametrization and PLS statistics. *Eur. J. Pharm. Sci.* **2000**, *10*, 295−303.
(52) Chapelle, O.; Vapnik, V.; Bousquet, O.; Mukherjee, S. Choosing multiple parameters for support vector machines. *Mach. Learn.* **2002**, *46*, 131−159.

CI049971E