Identifying P-Glycoprotein Substrates Using a Support Vector Machine Optimized by a Particle Swarm

Jianping Huang, Guangli Ma, Ishtiaq Muhammad, and Yiyu Cheng*

Pharmaceutical Informatics Institute, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310027, China

Received March 4, 2007

P-Glycoprotein (P-gp) contributes to extruding a structurally, chemically, and pharmacologically diverse range of substrates out of cells. This function may result in the failure of chemotherapy in cancer and influence pharmacokinetic properties of many drugs. Although a great deal of research has been devoted to the investigation of P-gp and its substrate specificity, still we do not have a clear understanding of the resolution of the three-dimensional structure of P-gp and its working role as a drug efflux pump at a molecular level. Hence to identify whether a compound is a P-gp substrate or not, computational methods are promising both in cancer treatment and the drug discovery processes. We have established more effective models for prediction of P-gp substrates with an average accuracy of >90% using a Particle Swarm (PS) algorithm and a Support Vector Machine (SVM) approach. The applied models yielded higher accuracies and contained fewer variables in comparison with previous studies. An analysis of P-gp substrate specificity based on the data set is also presented by a PS and a SVM. The aim of this study is 3-fold: (i) presentation of a modified PS algorithm that is applicable for selection of molecular descriptors in quantitative structure−activity relationship (QSAR) model construction, (ii) application of this modified PS algorithm as a wrapper to undertake feature selection in construction of a QSAR model to predict P-gp substrates with a multiple linear (ML) and SVM approach, and (iii) also finding factors (molecular descriptors) that most likely are associated with P-gp substrate specificity by using a PS and a SVM from the data set.

INTRODUCTION

P-Glycoprotein, the product of the MDR-1 gene, belongs to a superfamily of ATP-binding cassette (ABC) transporters, and it functions as an energy-dependent efflux pump that exports substrates out of cells.[1] It plays a key physiological role in the protection of the body from harmful xenobiotics by extruding them out of cells, and it also plays a significant role in many important pharmacological barriers.[2] The expression of P-gp is detected not only in various human tumors (i.e., colon, renal, adrenal carcinomas, lung and gastric carcinomas, and certain germ cell tumors) but also in certain normal human tissues (i.e., epithelia of the liver, kidney, small, large intestine, capillary endothelial cells in brain, ovary, and testis).[1,3] P-Glycoprotein recognizes and transports a structurally, chemically, and pharmacologically diverse range of hydrophobic compounds, a phenomenon known as multidrug resistance (MDR) that is responsible for one of the major reasons of the failure of chemotherapy in cancer.[1] Increasing attention has also been given to exploring the important modulating role of P-gp in pharmacokinetic properties of many clinically important therapeutic agents because it has a significant influence on drug absorption,[4−6] distribution,[4,7] metabolism,[8] excretion,[4] toxicity,[8] and drug−drug interactions.[9] Recently, polymorphism has been reported to affect the pharmacokinetics of many commonly used drugs, including anticancer drugs.[1,3] Thus, the knowledge of factors associated with the P-gp substrate specificity and the identification of whether a compound is a P-gp substrate or not is promising both in cancer treatment and the drug discovery process.

Many efforts have been carried out to investigate the structure and the mechanism of P-gp. A number of models and hypotheses proposed in previous studies also provide us with a deep insight not only into how the mechanism of P-gp transports its substrates but also about its biochemical and pharmacological characterization.[1,5] Preliminary studies show that P-gp is a 170 kDa membrane-bound protein, composed of two homologous halves joined by a flexible linker region. Each halve consists of six transmembrane (TM) domains and an ATP-binding/utilization domain.[1] The proper interaction of the two halves of human P-gp is important to form a single transporter. Several of the TMs contribute to the substrate binding, and different substrates have different and perhaps overlapping binding sites on P-gp.[1] Multiple binding sites and a binding site with different regions of interaction are essential to the diversity of substrates and the complex interactions between different kinds of drugs. Nevertheless, detailed knowledge of these binding sites and their interaction mechanisms are still uncertain. Owing to the deficiency of the resolution of the three-dimensional structure of P-gp and its working role as a drug efflux pump at a molecular level, it is still a complicated task to identify a P-gp substrate and its specificity.

However, many investigations have been conducted to characterize P-gp substrate specificity, and useful inferences have also been presented based on various experimental results. For instance, Ford and Hait argued that P-gp drug

* Corresponding author phone: +86-571-87952509; fax: +86-571-87951138; e-mail: chengyy@zju.edu.cn.

10.1021/ci700083n CCC: $37.00
Published on Web 07/03/2007

substrates appear to be hydrophobic, with a molecular mass of 300−2000 Da.[10] Seelig and co-workers suggested that molecules can be predicted to P-gp substrates if they contain recognition elements that are formed by two or three electron donor groups with a fixed spatial separation.[11] Many researchers also demonstrated that the number and strength of hydrogen bonds is crucial for the P-gp and drug substrates interaction.[12] Didziapetris and co-workers suggested that P-gp substrate specificity can be roughly estimated by a "rule of fours". They found that compounds with (N+O) ≥ 8, MW > 400, and acid pK_a > 4 are likely to be Pgp substrates, whereas compounds with (N+O) ≤ 4, MW < 400, and base pK_a < 8 are likely to be nonsubstrates.[13] Other researchers have also suggested that P-gp substrate specificity depends on multiple factors, such as logP, molecular weight (MW), surface area (SA), aromaticity, amphiphilicity, proton basicity, hydrogen bonding ability, and so on.[13]

Although various approaches have been developed for prediction and physicochemical properties characterization of P-gp substrates, computational technologies based on identification of structure−property relationships (SPR) and pharmacophore appear to attract more attention. Numerous computational QSAR models have been presented in recent literature to predict the P-gp substrates.[14−20] For example, Xue and co-workers constructed a QSAR model to predict P-gp substrates using a support vector machine approach.[15] More recently, De Cerqueira Lima and co-workers explored combinatorial QSAR models by the combination of several computational methods.[18] Cabrera also presented a TOPS-MODE approach for the prediction of P-glycoprotein substrates by QSAR modeling.[20] The pharmacophore model was another widely concerned technology in recent research.[21−24] For example, Penzotti and co-workers proposed a computational ensemble pharmacophore model with a prediction accuracy of 63% to differentiate substrates from nonsubstrates of P-pg.[21] By virtue of continuous development of statistical and data mining technologies, computational methods have shown more prosperous results and provide more choices and efficiency in identification of P-gp substrates.

QSAR techniques have long been used to develop quantitative correlations between biological activity and structural or physicochemical properties of molecules.[25] Therefore, to construct a good QSAR model we need to consider both the molecular descriptors (physicochemical properties) and the mathematical algorithm (relationship). In other words, two steps and their corresponding algorithms should be taken into consideration while using machine learning approaches to construct QSAR models: (1) feature selection and the corresponding algorithm and (2) data training and the corresponding machine learning algorithm. The most popular machine learning algorithms used include Artifical Neural Network, Decision Tree, SVM, and so on. Feature selection algorithms can be divided into three main categories:[26] (1) those where the selection is embedded within the basic induction algorithm, (2) those which use feature selection as a filter prior to induction, and (3) those which use feature selection as a wrapper around the induction process. The embedded approaches apply the feature selection process as an integral part of the learning algorithm. The filtering techniques attempt to identify features that are related to or predictive of the outcome interest: they operate independently of the learning algorithm. The wrapper approach, which will be employed in this paper, differs in that it evaluates subsets based upon accuracy estimation provided by a classifier built with a feature subset.[27] Comparing with a filter, it is obvious that a wrapper can produce better results because it takes the bias of the classifier into account and evaluates features in context. The filter technique had been used as the approach in the previous QSAR models studies to predict the P-gp substrates and to select molecular descriptors.[15,18] However, in our study, we will focus on the wrapper approach and introduce a relatively new PS algorithm to undertake the molecular descriptors selection task.

The PS algorithm has been a powerful competitor to other evolutionary algorithms such as the genetic algorithm (GA)[28,29] as well as industry standard algorithms such as the J48 algorithm.[30] Analysis shows that PS outperforms GA in solving many kinds of optimization problems, especially continuous design problems.[28] Although the PS and the GA on average yield the same effectiveness (solution quality), the PS is simpler and more computationally efficient than the GA.[31] A number of binary PS algorithms have also been proposed in feature selection[32,33] and QSAR model construction[34,35] recently.

In this paper, a modified PS algorithm that is applicable for selection of molecular descriptors in the QSAR model construction was presented. For P-gp substrates prediction purpose, this modified PS algorithm was employed as a wrapper to undertake feature selection in the construction of ML and SVM models. In addition, an attempt was made to find out which factors (molecular descriptors) may have the most contributions to the P-gp substrate specificity based on the data set. A PS and SVM prediction model was constructed, and an overview of correlations between P-gp substrate specificity and molecular descriptors based on this prediction model was also presented.

METHODS

Data Set. A total of 203 compounds classified as P-gp substrates or nonsubstrates used in this work were taken from ref 20, in which all these compounds have been carefully collected and identified. These compounds can also be obtained from the original literature.[11,15,16,21,36] To warrant a molecular diversity of the data set is of great importance to qualify the model's prediction ability. Therefore, Cabrera et al. had performed a k-means cluster analysis (k-MCA) process on the selection of a training and validation set.[20] To make a comparison, we used the same training and validation set. These compounds were divided into two parts: a training set with 163 compounds, containing 81 substrates and 82 nonsubstrates, and a validation set with 40 compounds, containing 22 substrates and 18 nonsubstrates. The training set was used to train the learning machine. The validation set was not used in the training but to validate the model and keep it from overfitting.

An external prediction data set used in ref 20 with 35 compounds containing 7 substrates and 28 nonsubstrates was also employed in this work to access the actual prediction ability of applied models.

Descriptors. Molecular descriptors used to characterize the molecular structure are indispensable to QSAR. The molecular descriptors involved in our study were calculated

by the Dragon software,[37] including all 13 categories of the 2D molecular descriptors (total 929 descriptors) and one category of the 3D molecular descriptors (74 geometrical descriptors). But only 11 categories and 79 numbers of them were selected after several preprocessing steps in order to reduce their redundancy and to improve their quality. That is, (i) those with too many zero values (>%80), (ii) those that have very small standard deviation values (<0.5), and (iii) those that have high correlation coefficients (>%90) with others were eliminated from the data set. Before being applied to the next feature selection step, these molecular descriptors values were scaled to the range (0,1) using a min-max normalization method[38]

$$\upsilon_{\text{new}} = (\upsilon - \min)/(\max - \min) + \min \qquad (1)$$

where min and max are the minimum and maximum values of a feature, and υ represents the value to be scaled.

As a result, the final data included 7 constitutional descriptors, 20 topological descriptors, 6 walk and path counts, 5 information indices, 7 edge adjacency indices, 1 topological charge indices, 3 eigenvalue-based indices, 4 functional group counts, 4 atom-centered fragments, 7 molecular properties, and 15 geometrical descriptors (Table 1).

Implementation. The PS algorithm and relative programs were implemented in the Java programming language, running on the Java (TM) 2 Runtime Environment, Standard Edition (build 1.5.0_02-b09). The Java package of libsvm (version 2.8)[39] used in this work is a free support vector machine tool available online. Molecular descriptors were generated using the Dragon professional version 5.4. All programs were run on Windows XP platforms on a Dell personal computer equipped with an Intel (R) Pentium (R) 4 CPU 2.80 GHz and 512M memory.

Particle Swarm Algorithm.[40−43] The PS algorithm was originally introduced in terms of social and cognitive behavior by Kennedy and Eberhart in 1995, motivated by the social behavior of movement organisms such as bird flocking and fish schooling. It is a stochastic and population based search algorithm with each particle randomly initialized with an original position and velocity. An individual particle's search direction (behavior) consists of two elements: the cognitive (individual) part and the social (swarm) part. When particles are flying through the search space, their velocities are dynamically adjusted according to their historical behaviors, and they will have a tendency to fly toward the better and better search area over the course of the search process. The performance of each particle is measured according to a predefined fitness function, which is related to the problem to be solved. The original PS algorithm is described as below

$$\upsilon_{id} = \upsilon_{id} + c_1 * \text{rand}() * (p_{id} - x_{id}) + c_2 * \text{Rand}() * (p_{gd} - x_{id}) \qquad (2a)$$

$$x_{id} = x_{id} + \upsilon_{id} \qquad (2b)$$

where c_1 and c_2 are positive constants, and rand() and Rand() are two random functions in the range [0,1]; $X_i = (x_{i1}, x_{i2}, ..., x_{iD})$ represents the ith particle; $P_i = (p_{i1}, p_{i2}, ..., p_{iD})$ represents the best previous position (the position giving the best fitness value) of the ith particle; the symbol g represents the index of the best particle among all particles in the population; and $V_i = (\upsilon_{i1}, \upsilon_{i2}, ..., \upsilon_{iD})$ represents the rate of the position change (velocity) for particle i.

The second part of eq 2a is the "cognitive" part, which represents the private thinking of the particle itself. The third part is the "social" part, which represents the collaboration among the particles. Equation 2a is used to calculate the particle's new velocity according to its previous velocity and the distances of its current position from its own best experience (position) and the group's best experience. Then the particle flies toward a new position according to eq 2b.

The performance of different problems can be improved by making us of the tradeoff between exploration and exploitation through adjusting the cognitive (c_1) and social learning rates (c_2) constants. Shi and Eberhart proposed another improved model:

$$\upsilon_{id} = w * \upsilon_{id} + c_1 * \text{rand}() * (p_{id} - x_{id}) + c_2 * \text{Rand}() * (p_{gd} - x_{id}) \qquad (3a)$$

$$x_{id} = x_{id} + \upsilon_{id} \qquad (3b)$$

An inertia weight which is also a positive constant is brought into the equation, collaborating with the c_1 and c_2 constants and contributing to the balance of the global search and local search coop. Equation 3, already proven to be more efficient by many researchers, has become a PS prototype of most of the variations presented, and it is the basic PS employed used in this work.

The PS algorithm is simple in concept, easy to implement, and less computational in cost. However, it has been found to be robust and possess a fast convergence in solving problems featuring nonlinearity and nondifferentiability, multiple optima, and high dimensionality.

A Modified Particle Swarm Algorithm Applicable for Feature Selection for QSAR. The PS algorithm was originally proposed for continuous problems, and attempts have been made to extend it to discrete optimization problems. Kennedy and Eberhart proposed the first discrete version with particle position composed of a set of bits that contain either '1' or '0', which denote being selected or not and are able to alternate according to the comparison result between a random number and a logistic transformation $S(\upsilon_{id})$. Although many other binary particle swarm algorithms were also proposed for feature selection and QSAR model construction recently,[32,34] most of them lose the consistency forms or ways of evolution in comparing with the continuous particle swarm algorithm. For instance, the position of a particle can only be composed of a set of '1' or '0' in the discrete version, which may result in the efficient loss compared with continuous particle swarm on the consideration of convergence velocity. Furthermore, we have to design two distinct particle swarm algorithm systems to deal with these two types of problems.

The basic PS algorithm presented here uses the same equation as the continuous PS algorithm, and the position and velocity of a particle are computed in the continuous space. The conversion of the new candidate position from continuous space to discrete space processes only when passed to the fitness function. Suppose that the particle

Table 1. Molecular Descriptors Used in This Work

abbrev	constitutional descriptors	abbrev	constitutional descriptors
MW	molecular weight	nCIR	number of circuits
AMW	average molecular weight	nH	number of hydrogen atoms
nBM	number of multiple bonds	nO	number of oxygen atoms
nCIC	number of rings		

abbrev	topological descriptors	abbrev	topological descriptors
VDA	average vertex distance degree	DECC	eccentric
TI1	first Mohar index	Lop	Lopping centric index
TI2	second Mohar index	D/Dr05	distance/detour ring index of order 5
Rww	reciprocal hyper-detour index	D/Dr06	distance/detour ring index of order 6
Wap	all-path Wiener index	D/Dr09	distance/detour ring index of order 9
J	Balaban distance connectivity index	D/Dr10	distance/detour ring index of order 10
Jhetp	Balaban-type index from polarizability weighted distance matrix	T(N..N)	sum of topological distances between N..N
		T(N..O)	sum of topological distances between N..O
MAXDN	maximal electrotopological negative variation	T(N..S)	sum of topological distances between N..S
MAXDP	maximal electrotopological positive variation	T(O..O)	sum of topological distances between N..O
TIE	E-state topological parameter		

abbrev	walk and path counts	abbrev	walk and path counts
SRW10	self-returning walk count of order 10 (ten the number of non-H bonds)	piPC10	molecular multiple path count of order 10
		piID	conventional bond-order ID number
MPC08	molecular path count of order 08	PCD	difference between multiple path count and path count
piPC03	molecular multiple path count of order 03		

abbrev	information indices	abbrev	information indices
IDDE	mean information content on the distance degree equality	IC2	information content index (neighborhood symmetry of 2-order)
CIC0	complementary information content (neighborhood symmetry of 0-order)	CIC2	complementary information content (neighborhood symmetry of 2-order)
CIC1	complementary information content (neighborhood symmetry of 1-order)		

abbrev	edge adjacency indices	abbrev	edge adjacency indices
EEig10x	eigenvalue 10 from edge adj. matrix weighted by edge degrees	EEig04r	eigenvalue 04 from edge adj. matrix weighted by resonance integrals
EEig03d	eigenvalue 03 from edge adj. matrix weighted by dipole moments	ESpm07u	spectral moment 07 from edge adj. matrix
		ESpm01d	spectral moment 01 from edge adj. matrix weighted by dipole moments
EEig05d	eigenvalue 05 from edge adj. matrix weighted by dipole moments		
EEig14d	eigenvalue 14 from edge adj. matrix weighted by dipole moments		

abbrev	topological charge indices	abbrev	topological charge indices
GGI4	topological charge index of order 4		

abbrev	eigenvalue-based indices	abbrev	eigenvalue-based indices
SEigZ	eigenvalue sum from Z weighted distance matrix (Barysz matrix)	VEA1	eigenvector coefficient sum from adjacency matrix
SEigv	eigenvalue sum from van der Waals weighted distance matrix		

abbrev	functional group counts	abbrev	functional group counts
nCs	number of total secondary C(sp3)	nCrt	number of ring tertiary C(sp3)
nCt	number of total tertiary C(sp3)	nCb-	number of substituted benzene C(sp2)

abbrev	atom-centered fragments	abbrev	atom-centered fragments
H-046	H^a attached to C^0(sp3) no X attached to next C. X represents any electronegative atom (O, N, S, P, Se, halogens)	H-047	H^a attached to C^1(sp3)/C^0(sp2)
		H-048	H^a attached to C^2(sp3)/C^1(sp2)/C^0(sp)
		H-051	H attached to alpha-C^b

abbrev	molecular properties	abbrev	molecular properties
Hy	hydrophilic factor	MLOGP2	squared Moriguchi octanol-water partition coeff (logP∧2)
AMR	molar refractivity	ALOGP	Ghose-Crippen octanol-water partition coeff. (logP)
TPSA(Tot)	topological polar surface area using N, O, S, P polar contributions	BLTF96	Verhaar model of Fish baseline toxicity for Fish (96h) from MLOGP (mmol/L)
MLOGP	Moriguchi octanol-water partition coefficient		

abbrev	geometrical descriptors	abbrev	geometrical descriptors
G1	gravitational index G1	HOMT	HOMA total (trial)
SPAN	span R	DISPm	d COMMA2 value/weighted by atomic masses
SPH	sphericity	QXXm	Qxx COMMA2 value/weighted by atomic masses
ASP	asphericity	QYYm	Qyy COMMA2 value/weighted by atomic masses
FDI	folding degree index	DISPv	d COMMA2 value/weighted by atomic van der Waals volumes
PJI3	3D Petijean shape index		
L/Bw	length-to-breadth ratio by WHIM	DISPe	d COMMA2 value/weighted by atomic Sanderson electronegativities
HOMA	harmonic oscillator model of aromaticity index		
RCI	Jug RC index		

position values is limited to interval [0, 1), conversion can be accomplished by mapping each value hits in the interval [0, 0.5) to 1 and every other value to 0.

The number of molecular descriptors is considered to be one of the important factors reflecting the quality of a QSAR model. In general, the fewer is the better, and 3−8 numbers of molecular descriptors are preferable. Considering this factor, a punish factor is often used in the fitness evaluation expression. This is usually the case when stochastic search algorithms such as PS and GA have been reported in much of the literature. One of these examples is referred to in ref 44. However, the number of the candidate molecular descriptors (features) can be very large, ranging from several tens to a thousand. It is inefficient to use such traditional PS algorithms directly for feature selection under this condition. Because the number of features selected in the evolution process obey normal distribution, the probability that only 3−8 features are selected in a generation is very small. To reduce this inefficiency, the mapping area can be adjusted from a continuous space to a discrete space in the conversion step of the PS algorithm we presented by shrinking the interval that represents the feature being selected. For example, the conversion can be adjusted by mapping each value hits in the interval [0, 0.1) to 1 and other values to 0. Thus, each feature has a probability of 1/10 to be selected and only about 1/10 of all features being selected in each generation. It can be ensured that only 3−8 numbers of features are being selected in more than 90% of the generations after adjustment and the computing efficiency can be improved dramatically.

Support Vector Machine (SVM).[45−47] A SVM has been widely used as a powerful machine learning tool presented by Vladimir Vapnik and co-workers in 1992 although the groundwork for a SVM was started in the 1960s. To date, a great deal of attention in diverse areas has been attracted due to its high accuracy and its less proneness to overfitting than other methods in machine learning. A SVM is based on the Statistical Learning Theory and aims to find the maximum marginal hyperplane (the hyperplane with the largest margin) that best separates two classes according to the training data. Depending on the problem, SVM methodology is usually divided into four categories: linearly separable, linearly inseparable, nonlinearly separable, and nonlinearly inseparable.

Let the training patterns be given as a tuple $(\mathbf{x}_i, y_i)$ $(I = 1,2, ..., N)$, where vector $\mathbf{x}_i$ corresponds to the molecular descriptors set for the ith pattern with class labels y_i. Each y_i can take the value of either $+1$ or -1.

In the simple case of the data that are linearly separable, the following inequality is satisfied

$$y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 \tag{4}$$

which is the combination of the following two inequalities

$$\mathbf{H}_1 : \; y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 \quad \text{for } y_i = +1 \tag{5a}$$

and

$$\mathbf{H}_2 : \; y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \leq 1 \quad \text{for } y_i = -1 \tag{5b}$$

where $\mathbf{H}_1$ and $\mathbf{H}_2$ are two parallel hyperplanes defining the sides of the maximum margin; and $\mathbf{w}$ is a weight vector normal to the hyperplane. Any tuple that falls on or above $\mathbf{H}_1$ belongs to class $+1$, and any tuple that falls on or below $\mathbf{H}_2$ belongs to class -1. Any training tuples that fall on hyperplane $\mathbf{H}_1$ or $\mathbf{H}_2$ are called support vectors. It is obvious that the support vectors are of essential importance and give the most information regarding classification. The distance from the hyperplane to any point on $\mathbf{H}_1$ is $1/|\mathbf{w}||$, where $||\mathbf{w}||$ is the Euclidean norm of $\mathbf{w}$, and the maximal margin is $2/|\mathbf{w}||$. The learning task, hence, can be accomplished by solving the constrained optimization problem $\min_w ||\mathbf{w}||^2/2$ subject to inequality (4).

Having the concepts given above, we turn to focus on the most general case of the data which are nonlinearly inseparable. To accommodate the inseparable data, the inequality constraints can be relaxed by introducing a positive slack variable (ξ). With regard to the nonlinear problem, a nonlinear mapping is used to transform the original training data into a higher dimension. Within this new dimension, it searches for the linear optimal separating hyperplane (decision boundary). The transform can be done by using a kernel function $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i)\cdot\Phi(\mathbf{x}_j)$. For instance, $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = e^{-||x_i - x_j||^2/2\sigma^2}$ is one of the commonly used kernel functions known as the Gaussian radial basis function (RBF) kernel. Hence, the objective function is modified to the following equation

$$f(\mathbf{w}) = \frac{1}{2}||\mathbf{w}||^2 + C\left(\sum_{i=1}^{N}\xi_i\right)^k \tag{6}$$

where C and k are user-specified parameters of penalty. To simplify the problem, we assume $k = 1$ in the remainder of this section. Then, the Lagrangian for the modified constrained optimization problem comes out to be as follows

$$L_P = \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i\{y_i[\mathbf{w}\cdot\Phi(\mathbf{x}_i) + b] - 1 + \xi_i\} - \sum_{i=1}^{N}\mu_i\xi_i \tag{7}$$

where the parameters α_i and μ_i are Lagrange multipliers. This equation is difficult for solving, so the following dual Lagrangian needed to maximize is presented

$$L_D = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j \cdot \mathbf{K}(\mathbf{x}_i,\mathbf{x}_j) \tag{8}$$

where $0 \leq \alpha_i \leq C \; \forall \; i$ and $\sum_{i=1}^{N}\alpha_i y_i = 0$. This dual equation can be solved numerically using quadratic programming techniques under Karush-Kuhn-Tucker (KKT) conditions to obtain the Lagrange multipliers α_i and then also $\mathbf{w}$ and b.

Therefore, the ultimate equation used to predict a test instance $\mathbf{x}$ can be given as follows:

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{N} y_i\alpha_i\mathbf{K}(\mathbf{x}_i, \mathbf{x}) + b\right) \tag{9}$$

RESULTS AND DISCUSSION

Multiple Linear Model. To show the effectiveness of a PS algorithm in the QSAR models construction, the simplest

Table 2. Correlation Coefficients between Descriptors in the ML Model for the Training Set

descriptor	nCIC	nH	nO	TPSA(Tot)
AMW	−0.017	−0.438	−0.102	−0.046
nCIC		0.393	0.470	0.301
nH			0.624	0.519
nO				0.824

Table 3. SVM Model and Its Prediction Accuracy

molecular descriptors	Tr[a] (%)	Va[b] (%)	Pr[c] (%)	CV[d] (%)
AMW, nH, nO, nCrt, nCb-, TPSA(Tot),MLOGP	95.7	90	88.6	80.8

[a] Tr represents the training set. [b] Va represents the validation set. [c] Pr represents the prediction set, respectively. [d] CV represents 5-fold full cross-validation on the couple of training set and validation set.

case constructing an ML model using a PS was considered. The number of descriptors used in the model construction were limited between 3 and 8. The fitness function of a PS algorithm was given below

$$\text{fitness} = 0.33 \times \text{Tr}_{\text{correct}}/\text{Tr}_{\text{total}} + 0.67 \times \text{Va}_{\text{correct}}/\text{Va}_{\text{total}} \quad (10)$$

where $\text{Tr}_{\text{correct}}$ and Tr_{total} represent the numbers of the training set that correctly predicted and the total numbers of the training set, respectively; and $\text{Va}_{\text{correct}}$ and Va_{total} represent the numbers of the validation set that correctly predicted and the total numbers of the training set, respectively.

The result model was given as follows

$$F = \text{sign}(-1.16 \times \text{AMW} + 2.93 \times \text{nCIC} + 8.8 \times \text{nH} - 5.70 \times \text{nO} + 2.03 \times \text{TPSA(Tot)} + 2.2)$$

where the positive or negative sign of F represents whether the molecular is a P-gp substrate or not.

Although this linear model was very simple and contained only 5 descriptors, it yielded a relatively high prediction accuracy that was above 80%. The prediction accuracies of the training set, validation set, and external prediction set were 80.3%, 80%, and 82.9%, respectively. Correlation coefficients between descriptors in the ML model for the training set were shown in Table 2. There were no highly correlated descriptors found in this model.

Support Vector Machine Model. A SVM is an excellent machine learning approach and also plays well in our models construction. The commonly used RBF kernel function was employed in the SVM learning process, and two other parameters need to be defined have been indicated in eq 6. So a particle position vector contained 81 components in SVM training models. The fitness function was also given below:

$$\text{fitness} = 0.5 \times \text{Tr}_{\text{correct}}/\text{Tr}_{\text{total}} + \text{Va}_{\text{correct}}/\text{Va}_{\text{total}} \quad (11)$$

Taking into account both accuracy and interpretability, the top SVM model and molecular descriptor set selected were given in Table 3. It included 7 molecular descriptors, and the prediction accuracies were 95.7%, 90%, and 88.6% on the training set, validation set, and external prediction set, respectively, with an average accuracy of 91.4%. The result of the 5-fold full cross-validation of 80.8% was a guarantee of its quality. Correlation coefficients between descriptors in the SVM model for the training set were also shown in Table 4, and there were no highly correlated descriptors. It is clear that this SVM model generated a much higher accuracy than the ML model.

Table 4. Correlation Coefficients between Descriptors in the SVM Model for the Training Set

descriptor	nH	nO	nCrt	nCb-	TPSA(Tot)	MLOGP
AMW	−0.438	−0.102	0.022	−0.004	−0.046	0.096
nH		0.624	0.365	0.044	0.519	−0.053
nO			0.335	0.251	0.824	−0.459
nCrt				−0.133	0.049	0.054
nCb-					0.116	0.137
TPSA(Tot)						−0.528

Discussion and Comparison of Our Models with the Previous Studies. Both of the result models contained only 2D molecular descriptors, and the geometrical descriptors used in this work did not appear to have a great effect upon the identification of substrates of P-gp. Only three categories of molecular descriptors were presented. The ML model consisted of 4 constitutional descriptors (AMW, nCIC, nH, nO) and 1 molecular property (TPSA(Tot)), while the SVM model consisted of 3 constitutional descriptors (AMW, nH, nO), 2 functional group counts (nCrt, nCb-), and 2 molecular properties (TPSA(Tot), MLOGP). The co-occurrence of 4 molecular descriptors in these two models, i.e., AMW, nH, nO, and TPSA(Tot), may suggest that they play a key role in the prediction of substrates of P-gp. The meanings of all these descriptors in the result models can be found in Table 1. Most of them seemed to be intelligible and facile, and a more detailed analysis on these descriptors will be given in the next section.

In order to access the prediction accuracy for the substrates and nonsubstrates of P-gp with our models, we introduced true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) here to achieve this aim. The prediction accuracy for the substrates of P-gp then can be given by the sensitivity (SE) = TP/(TP + FN), and the prediction accuracy for the nonsubstrates of P-gp can be given by specificity (SP) = TN/(TN + FP). The prediction accuracy for the substrates and nonsubstrates of P-gp produced by the ML and SVM models were given in Table 5.

As shown in Table 5, both of the ML and SVM models yielded higher prediction accuracies for the substrates (sensitivity) than for the nonsubstrates of P-gp (specificity) on all of the training sets, validation sets, and prediction sets. With an average accuracy greater than 91%, the SVM model exhibited much better prediction ability.

A direct comparison between our models and the one presented by Cabrera et al.[20] which used the same data set was given in Table 6. In their work, Cabrera et al. employed a TOPS-MODE approach in the generation of a discriminant function by a Linear Discriminant Analysis (LDA) to classify P-gp substrates, resulting in a model consisting of 5 variables called spectral moments. As can be seen from Table 6, the prediction accuracy obtained from our ML model had a slight improvement on the validation set when compared with that obtained by the TOPS-MODE approach. But with the prediction set, prediction accuracy of this ML model improved significantly. Besides, descriptors used in our ML

Table 5. Prediction Accuracy for the Substrates and Nonsubstrates of P-gp of the ML and SVM Models[a]

	training set						validation set						prediction set					
n	TP	FN	TN	FP	Se (%)	Sp (%)	TP	FN	TN	FP	Se (%)	Sp (%)	TP	FN	TN	FP	Se (%)	Sp (%)
1[b]	78	13	53	19	86	74	18	4	14	4	82	78	6	1	23	5	86	82
2[c]	89	2	67	5	98	93	20	2	16	2	91	89	7	0	24	4	100	86

[a] TP (true positive), FN (false negative), TN (true negative), FP (false positive), Se (sensitivity) is the prediction accuracy for substrates, and SP (specificity) is the prediction accuracy for nonsubstrates. [b] Represents the ML model. [c] Represents the SVM model.

Table 6. Comparison between Our Models and TOPS-MODE Approach[a]

	training set			validation set			prediction set		
references	Se (%)	Sp (%)	A (%)	Se (%)	Sp (%)	A (%)	Se (%)	Sp (%)	A (%)
Cabrera et al.[20]	82	79	80.5	82	72	78	71	71	71
ML model	86	74	80	82	78	80	86	82	84
SVM model	98	93	95.5	91	89	90	100	86	93

[a] Se, Sp, and A represent sensitivity, specificity, and accuracy, respectively.

Table 7. Overview of Several Computational Models Proposed by Previous Studies

	data set		test set		
references	N[a]	V[b]	Se (%)[c]	Sp (%)[d]	A (%)[e]
Penzotti et al.[21]	195		53	79	63
Gombar et al.[16]	140	27	94	78	86
Xue et al.[15]	201	22	84	67	80
De Cerqueira Lima et al.[18]	195		78	84	81
Cabrera et al.[20]	203	5	82	72	78
this work (the ML model)	203	5	82	78	80
this work (the SVM model)	203	7	91	89	90

[a] N is the total number of compounds in the coupling of training set and test set. [b] V is the number of variables used in the model. [c] Se represents sensitivity. [d] Sp represents specificity. [e] A represents accuracy.

model seemed to be more intuitive. Therefore, it may lead to the conclusion that our ML model was more practicable and provided more prediction power than Cabrera's. Moreover, the SVM model showed further effectiveness in that it gave much higher prediction ability.

Many other computational models also have been developed to identify P-gp substrates. Taking into account the different data set and data set size used, a direct comparison among all these computational models is inappropriate. However, several computational models results obtained by using a relatively large data set are listed in Table 7, which can provide us with an overview of these studies.

The results showed that our SVM model provided the highest accuracy with fewer variables. In the simplest case, the ML model also yielded relatively high accuracy compared with the previous studies. From the results given in Table 7, it can be concluded that the PS algorithm is an efficacious and powerful tool to undertake the feature selection task and is promising in QSAR model construction. As a search optimization algorithm, PS needs to search in a large features space, which can incur more computational cost. But its dramatic improvement in results may compensate for this cost. Moreover, descriptors contained in the result models are more intelligible. As is well-known that predictive power and interpretability are two main factors needed to compose a good QSAR model, the result models seem to satisfy this criterion better than the previous studies.

Correlations between P-gp Substrate Specificity and Descriptors. Albeit the mechanism of P-gp substrate specificity has not been understood completely, yet it will be useful to predict which molecular descriptors have the most contributions to declare a compound to be a P-gp substrate. Although the Pearson correlations can give us an instinctive understanding of the linear associations between P-gp substrate specificity and its molecular descriptors (Figure 1), it is difficult for us to get enough information about the P-gp substrate specificity from the present data set. Moreover, the specificity of a P-gp substrate does not appear to consist of one factor alone. We tried our best to extract as much information as possible from the available data set and to see if other means can offer us more insight into their relationships. The currently emerging data mining technologies make it possible to achieve this goal. As a stochastic algorithm, PS has played a more important role in data mining.[30] It is capable of searching in a very large space which can provide us an outlook of this space and without exhausting it.

In this model, we used the same fitness function as given in the previous SVM models section. The PS was set with 30 particles and 50 iterations and covered 30 runs. In each evaluation, a descriptor got a hit if this descriptor was selected and the fitness function was greater than or equal to a given threshold (0.8 was used in this study). This was based on the assumption that if the fitness function was greater than or equal to a threshold, then the descriptors selected were probably having the contribution to the specificity of a P-gp substrate. Hence, a descriptor selected more often will get more hits.

Correlations between P-gp substrate specificity and descriptors based on our PS and SVM prediction model are given in Figure 2. A descriptor with more hits indicates it may have more contribution to the P-gp substrate specificity.

It is obvious that AMW, nCIC, nH, nCrt, and nCb-descriptors in this figure contain significantly high value. Many studies have stressed the importance of molecular weight (MW) in the prediction of P-gp substrate specificity.[13,48−51] However, AMW seems to play a more important role in our ML and SVM models, and the greater value of AMW than MW shown in Figure 2 suggests that AMW may be more efficient than MW in the prediction of P-gp substrate specificity based on the data set used. The aromatic rings were taken as factors on the consideration of P-gp substrate specificity in the early studies.[48,49] The high value of nCIC

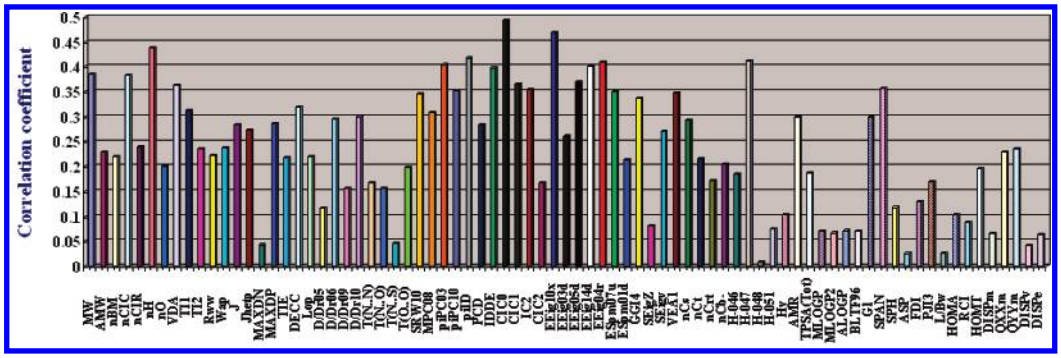

Figure 1. Pearson correlations (+) between P-gp substrate specificity and their molecular descriptors. A pattern is labeled 0 or 1 according to whether it is a P-gp substrate or not.

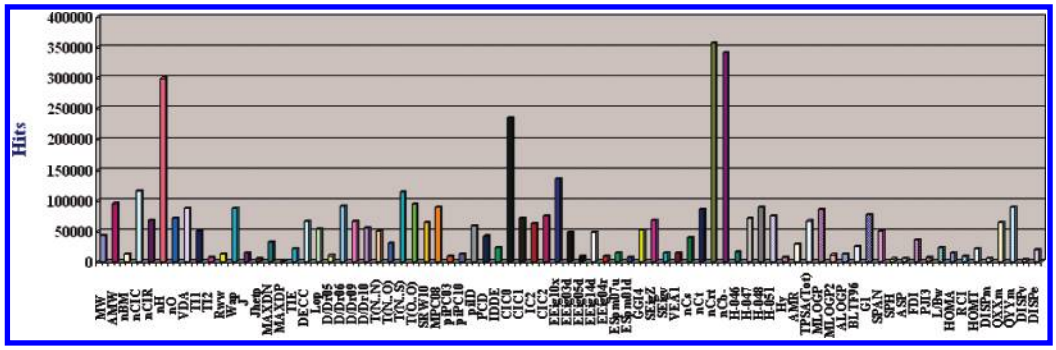

Figure 2. Correlations between P-gp substrate specificity and descriptors based on the PS and the SVM prediction models. A descriptor with more hits indicates it may have more contributions to the P-gp substrate specificity.

(number of rings) obtained from this model seems to agree with these studies well. Moreover, as reported in the previous studies, the number of hydrogen bonds is crucial for the P-gp substrate specificity.[12,13] Although we cannot draw the conclusion that the number of hydrogen atoms is a good displacement of the number of hydrogen bonds in the identification of a substrate of P-gp, it seems that nH (number of hydrogen atoms) can be taken as an effective factor.

The two top values contained by nCrt and nCb-descriptors indicated their essential roles in the prediction of P-gp substrate specificity based on the data set, and nCrt and nCb- represent the number of ring tertiary C(sp3) and the number of substituted benzene C(sp2), respectively. We believe that the information provided here will be helpful in further research for the prediction of P-gp substrates.

Surface area (SA) is another popularly concerned factor in the prediction of P-gp substrate specificity.[48,52,53] The TPSA(Tot) descriptor shown in Figure 2 is a topological polar surface area using N, O, S, and P polar contributions. Results show that it is also a prominent factor to be considered. More studies have also been concerned about the LogP[48,49,54−58] as an important factor. As can be seen from Figure 2, although the ALOGP and MLOGP2 values are not noticeable, MLOGP (Moriguchi octanol−water partition coefficient) may be a valuable descriptor to be concerned when predicting P-gp substrate specificity.

It is also worthwhile to mention that some other descriptors may also be useful in the prediction of P-gp substrate specificity, for instance, topological descriptors such as T(N..S) and T(O..O), information indices such as CIC0, and atom-centered fragments such as H-047, H-048, and H-051. Every of these descriptors alone may not be very conspicuous, but it could be valuable in the prediction of P-gp substrate specificity while coupling with other factors.

As a stochastic algorithm and robust data mining technology, the PS algorithm has shown its powerful ability to retrieve valuable information from a large data set. The results could be helpful in the identification of P-gp substrate and its specificity. However, it should be noted that, suffering from the limitation of the data set used in this work, further research should be undertaken to improve the results. Furthermore, accurate and effective results can be obtained with working on massive data incorporating a more diverse range of P-gp substrates and nonsubstrates.

CONCLUSION

To identify whether a compound is a P-gp substrate or not is very important both in cancer treatment and drug discovery. Without a clear understanding of the resolution of the three-dimensional structure of P-gp and its working role as a drug efflux pump at a molecular level, computational methods are still a desirable means to achieve this goal. As widely accepted out of powerful data mining technologies, a PS has shown its efficiency in solving optimization problems. In this paper, we presented a modified PS algorithm which is suitable for the selection of molecular descriptors in the QSAR model construction. Then we employed it as our wrapper to undertake a feature selection in the construction of QSAR model for predicting P-gp substrates. Finally, we derived factors mostly associated with the P-gp substrate specificity by using a PS and a SVM based on the data set. Results showed that a PS is an efficacious and powerful tool to undertake a feature selection task in the construction of a QSAR model for predicting P-gp substrates. Moreover, as a searching algorithm, a PS can give us an overview and provide us with an extraordinary amount of information in a large space without exhausting it.

We anticipate that our models can speed up the virtual screening in the early stages of drug discovery while considering whether a compound contains a P-gp substrate specificity or not. This study is promising in the identification of a P-gp substrate and may facilitate further research on the identification of P-gp substrate specificity, their mechanism of action, and the unraveling of drug interaction on a specific tissue or organ.

ACKNOWLEDGMENT

This work was financially supported by the National Basic Research Program of China (No. 2005CB523402).

REFERENCES AND NOTES

(1) Ambudkar, S. V.; Kimchi-Sarfaty, C.; Sauna, Z. E.; Gottesman, M. M. P-glycoprotein: from genomics to mechanism. *Oncogene* **2003**, *22*, 7468−7485.

(2) 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* (4), 361−398.

(3) Tandon, T. L. R.; Kapoor, K. B.; Bano, B. G.; Gupta, G. S.; Gillani, G. Z.; Gupta, G. S.; Kour, K. D. P-glycoprotein: Pharmacological relevance. *Indian J. Pharmacol.* **2006**, *38* (1), 13−24.

(4) Welwyn, U. K. Role of transport proteins in drug absorption, distribution and excretion. *Xenobiotica* **2001**, *31* (8), 469−497.

(5) Seelig, A.; Landwojtowicz, E.; Fischer, H.; Blatter, X. L. Towards P-glycoprotein structure−activity relationships. *Drug bioavailability/ Estimation of solubility, permeability, absorption and bioavailability*; 2003.

(6) Varma, M. V.; Sateesh, K.; Panchagnula, R. Functional role of P-glycoprotein in limiting intestinal absorption of drugs: contribution of passive permeability to P-glycoprotein mediated efflux transport. *Mol. Pharmaceutics* **2005**, *2* (1), 12−21.

(7) Tanigawara, Y. Role of P-glycoprotein in drug disposition. *Ther. Drug Monit.* **2000**, *22* (1), 137−40.

(8) Bodo, A.; Bakos, E.; Szeri, F.; Varadi, A.; Sarkadi, B. The role of multidrug transporters in drug availability, metabolism and toxicity. *Toxicol. Lett.* **2003**, *140* (141), 133−43.

(9) Lin, J. H. Drug-drug interaction mediated by inhibition and induction of P-glycoprotein. *Adv. Drug Delivery Rev.* **2003**, *55* (1), 53−81.

(10) Ford, J. M.; Hait, W. N. Pharmacology of drugs that alter multidrug resistance in cancer. *Pharm. Rev.* **1990**, *42* (3), 155−199.

(11) Seelig, A. A general pattern for substrate recognition by P-glycoprotein. *FEBS J.* **1998**, *251* (1), 252−261.

(12) Ecker, G.; Huber, M.; Schmid, D.; Chiba, P. The Importance of a Nitrogen Atom in Modulators of Multidrug Resistance. *Mol. Pharmacol.* **1999**, *56* (4), 791−796.

(13) Didziapetris, R.; Japertas, P.; Avdeef, A.; Petrauskas, A. Classification Analysis of P-Glycoprotein Substrate Specificity. *J. Drug Targeting* **2003**, *11* (7), 391−406.

(14) Ekins, S.; Kim, R. B.; Leake, B. F.; Dantzig, A. H.; Schuetz, E. G.; Lan, L. B.; Yasuda, K.; Shepard, R. L.; Winter, M. A.; Schuetz, J. D. Three-Dimensional Quantitative Structure-Activity Relationships of Inhibitors of P-Glycoprotein. *Mol. Pharmacol.* **2002**, *61* (5), 964−973.

(15) Xue, Y.; Yap, C. W.; Sun, L. Z.; Cao, Z. W.; Wang, J. F.; Chen, Y. Z. Prediction of P-glycoprotein substrates by a support vector machine approach. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (4), 1497−505.

(16) Gombar, V. K.; Polli, J. W.; Humphreys, J. E.; Wring, S. A.; Serabjit-Singh, C. S. Predicting P-glycoprotein substrates by a quantitative structure-activity relationship model. *J. Pharm. Sci.* **2004**, *93* (4), 957−968.

(17) Wang, Y. H.; Li, Y.; Yang, S. L.; Yang, L. Classification of substrates and inhibitors of P-glycoprotein using unsupervised machine learning approach. *J. Chem. Inf. Model.* **2005**, *45* (3), 750−757.

(18) de Cerqueira Lima, P.; Golbraikh, A.; Oloff, S.; Xiao, Y.; Tropsha, A. Combinatorial QSAR Modeling of P-Glycoprotein Substrates. *J. Chem. Inf. Model.* **2006**, *46* (3), 1245−1254.

(19) Crivori, P.; Reinach, B.; Pezzetta, D.; Poggesi, I. Computational Models for Identifying Potential P-Glycoprotein Substrates and Inhibitors. *Mol. Pharmaceutics* **2006**, *3* (1), 33−44.

(20) Cabrera, M. A.; González, I.; Fernández, C.; Navarro, C.; Bermejo, M. A topological substructural approach for the prediction of P-glycoprotein substrates. *J. Pharm. Sci.* **2006**, *95* (3), 589−606.

(21) Penzotti, J. E.; Lamb, M. L.; Evensen, E.; Grootenhuis, P. D. A computational ensemble pharmacophore model for identifying substrates of P-glycoprotein. *J. Med. Chem.* **2002**, *45* (9), 1737−1740.

(22) Ekins, S.; Kim, R. B.; Leake, B. F.; Dantzig, A. H.; Schuetz, E. G.; Lan, L. B.; Yasuda, K.; Shepard, R. L.; Winter, M.; Schuetz, J. D. Application of Three-Dimensional Quantitative Structure-Activity Relationships of P-Glycoprotein Inhibitors and Substrates. *Mol. Pharmacol.* **2002**, *61* (5), 974−981.

(23) Pajeva, I. K.; Wiese, M. Pharmacophore model of drugs involved in P-glycoprotein multidrug resistance: explanation of structural variety (hypothesis). *J. Med. Chem.* **2002**, *45* (26), 5671−5684.

(24) Cianchetta, G.; Singleton, R. W.; Zhang, M.; Wildgoose, M.; Giesing, D.; Fravolini, A.; Cruciani, G.; Vaz, R. J. A pharmacophore hypothesis for P-glycoprotein substrate recognition using GRIND-based 3D-QSAR. *J. Med. Chem.* **2005**, *48* (8), 2927−2935.

(25) Cedefto, W.; Agraflotis, D. Particle swarms for drug design. *The 2005 IEEE Congress on Evolutionary Computation*, 2005; p 2.

(26) Blum, A.; Langley, P. Selection of Relevant Features and Examples in Machine Learning. *Artif. Intelligence* **1997**, *97* (1−2), 245−271.

(27) Kohavi, R.; John, G. H. Wrappers for Feature Subset Selection. *Artif. Intelligence* **1997**, *97* (1−2), 273−324.

(28) Al-kazemi, B.; Mohan, C. K. Multi-phase discrete particle swarm optimization. *Fourth International Workshop on Frontiers in Evolutionary Algorithms*, 2000.

(29) Hassan, R.; Cohanim, B.; De Weck, O.; Venter, G. A Comparison Of Particle Swarm Optimization And The Genetic Algorithm. *46th AIAA/ ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference*, 2005; pp 1−13.

(30) Sousa, T.; Silva, A.; Neves, A. Particle swarm based Data Mining Algorithms for classification tasks. *Parallel Computing* **2004**, *30* (5−6), 767−783.

(31) Fourie, P. C.; Groenwold, A. A. In *Particle Swarms in Size and Shape Optimization*, Proceedings of the International Workshop on Multidisciplinary Design Optimization, Pretoria, South Africa, 2000; pp 97−106.

(32) Tang, E. K.; Suganthan, P. N.; Yao, X. In *Feature Selection for Microarray Data Using Least Squares SVM and Particle Swarm Optimization*, Proceedings of the 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, 2005, CIBCB'05; 2005; pp 1−8.

(33) Firpi, H. A.; Goodman, E. Swarmed feature selection. Proceedings of the 33rd Applied Imagery Pattern Recognition Workshop, 2004; pp 112−118.

(34) Agrafiotis, D. K.; Cedeno, W. Feature selection for structure-activity correlation using binary particle swarms. *J. Med. Chem.* **2002**, *45* (5), 1098−1107.

(35) Wang, Z.; Durst, G. L.; Eberhart, R. C.; Boyd, D. B.; Miled, Z. B. In *Particle swarm optimization and neural network application for QSAR*. Proceedings of the 18th International Parallel and Distributed Processing Symposium, 2004.

(36) Polli, J. W.; Wring, S. A.; Humphreys, J. E.; Huang, L.; Morgan, J. B.; Webster, L. O.; Serabjit-Singh, C. S. Rational Use of in Vitro P-glycoprotein Assays in Drug Discovery. *J. Pharmacol. Exp. Ther.* **2001**, *299* (2), 620−628.

(37) Todeschini, R.; Consonni, V.; Mauri, A.; Pavan, M. M. *Dragon 5.4*; Chemometrics and QSAR Research Group, University of Milano-Bicocca: Milan, Italy, 2006.

(38) Han, J.; Kamber, M. *Data Mining: Concepts and Techniques*, 2nd ed.; Morgan Kaufmann: 2006.

(39) Chang, C. C.; Lin, C. J. LIBSVM: a library for support vector machines. **2001**, *80*, 604−611. Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm.

(40) Kennedy, J.; Eberhart, R. In *Particle swarm optimization*, Proceedings of the IEEE International Conference on Neural Networks, 1995; p 4.

(41) Eberhart, R. C.; Shi, Y. In *Comparing inertia weights and constriction factors in particles warm optimization*. Proceedings of the 2000 Congress on Evolutionary Computation, 2000; p 1.

(42) Shi, Y.; Eberhart, R. C.; Center, E.; Carmel, I. N. In *Empirical study of particle swarm optimization*. Proceedings of the 1999 Congress on Evolutionary Computation, CEC 99, 1999; p 3.

(43) Shi, Y.; Eberhart, R. C. Parameter selection in particle swarm optimization. *Evol. Programming* **1998**, *7*, 611−616.

(44) Liu, Y.; Qin, Z.; Xu, Z.; He, X. *Feature Selection with Particle Swarms*; Springer: Berlin/Heidelberg, 2004; Vol. 3314.

(45) Vapnik, V. N.; Vapnik, V. *The Nature of Statistical Learning Theory*; Springer: 2000.

(46) Cristianini, N.; Shawe-Taylor, J. *An introduction to Support Vector Machines*; 2000.

(47) Burges, C. J. C. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Min. Knowledge Discovery* **1998**, *2* (2), 121−167.

(48) Bain, L. J.; Leblanc, G. A. Interaction of Structurally Diverse Pesticides with the Human MDR1 Gene Product P-Glycoprotein. *Toxicol. Appl. Pharmacol.* **1996**, *141* (1), 288−298.

(49) Klopman, G.; Shi, L. M.; Ramu, A. Quantitative Structure-Activity Relationship of Multidrug Resistance Reversal Agents. *Mol. Pharmacol.* **1997**, *52* (2), 323−334.
(50) Hansch, C.; Kurup, A.; Garg, R.; Gao, H. Chem-bioinformatics and QSAR: a review of QSAR lacking positive hydrophobic terms. *Chem. Rev.* **2001**, *101* (3), 619−72.
(51) Lee, J. S.; Paull, K.; Alvarez, M.; Hose, C.; Monks, A.; Grever, M.; Fojo, A. T.; Bates, S. E. Rhodamine efflux patterns predict P-glycoprotein substrates in the National Cancer Institute drug screen. *Mol. Pharmacol.* **1994**, *46* (4), 627−638.
(52) 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* (2), 159−168.
(53) Oesterberg, T.; Norinder, U. Prediction of Polar Surface Area and Drug Transport Processes Using Simple Parameters and PLS Statistics. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (6), 1408−1411.
(54) Ecker, G.; Chiba, P.; Hitzler, M.; Schmid, D.; Visser, K.; Cordes, H. P.; Cs?llei, J.; Seydel, J. K.; Schaper, K. J. Structure-Activity Relationship Studies on Benzofuran Analogs of Propafenone-Type Modulators of Tumor Cell Multidrug Resistance. *J. Med. Chem.* **1996**, *39*, 4767−4774.
(55) Schmitt, L.; Tampe, R. Structure and mechanism of ABC transporters. *Curr. Opin. Struct. Biol.* **2002**, *12*, 754−760.
(56) Schmid, D.; Ecker, G.; Kopp, S.; Hitzler, M.; Chiba, P. Structure-activity relationship studies of propafenone analogs based on P-glycoprotein ATPase activity measurements. *Biochem. Pharmacol.* **1999**, *58* (9), 1447−56.
(57) Chiba, P.; Ecker, G.; Schmid, D.; Drach, J.; Tell, B.; Goldenberg, S.; Gekeler, V. Structural requirements for activity of propafenone-type modulators in P-glycoprotein-mediated multidrug resistance. *Mol. Pharmacol.* **1996**, *49* (6), 1122−1130.
(58) Yasuda, K.; Lan, L.; Sanglard, D.; Furuya, K.; Schuetz, J. D.; Schuetz, E. G. Interaction of Cytochrome P450 3A Inhibitors with P-Glycoprotein. *J. Pharmacol. Exp. Ther.* **2002**, *303* (1), 323−332.

CI700083N