QSAR Models for the Prediction of Binding Affinities to Human Serum Albumin Using the Heuristic Method and a Support Vector Machine

C. X. Xue,[†] R. S. Zhang,[†,‡] H. X. Liu,[†] X. J. Yao,[†,§] M. C. Liu,[†] Z. D. Hu,*[,†] and B. T. Fan[§]

Department of Chemistry, Lanzhou University, Lanzhou 730000, China, Department of Computer Science, Lanzhou University, Lanzhou 730000, China, and Université Paris 7-Denis Diderot, ITODYS 1, Rue Guy de la Brosse, 75005 Paris, France

Received May 30, 2004

The binding affinities to human serum albumin for 94 diverse drugs and drug-like compounds were modeled with the descriptors calculated from the molecular structure alone using a quantitative structure−activity relationship (QSAR) technique. The heuristic method (HM) and support vector machine (SVM) were utilized to construct the linear and nonlinear prediction models, leading to a good correlation coefficient (R^2) of 0.86 and 0.94 and root-mean-square errors (rms) of 0.212 and 0.134 albumin drug binding affinity units, respectively. Furthermore, the models were evaluated by a 10 compound external test set, yielding R^2 of 0.71 and 0.89 and rms error of 0.430 and 0.222. The specific information described by the heuristic linear model could give some insights into the factors that are likely to govern the binding affinity of the compounds and be used as an aid to the drug design process; however, the prediction results of the nonlinear SVM model seem to be better than that of the HM.

1. INTRODUCTION

Drugs in the human circulatory system often bind to the components of the plasma such as albumin, acidglycoprotein (AGP), or lipoproteins. Such binding can take place in association with single, or multiple, plasma elements. Albumin, which comprises more than half of all blood proteins, is the most significant plasma component involved in the binding of drugs.[1] This protein is extremely important from a biopharmacological point of view because it is the major transporter of nonesterified fatty acids as well as of different drugs and metabolites to different tissues.[2] Drug binding to Human Serum Albumin (HSA) is an area of intense research. The pharmacokinetics and pharmacodynamics of drugs are strongly affected by their binding to this protein. Oral bioavailability (%F) is directly affected by the extent to which a drug binds to plasma proteins because the bound drug is not available to the mechanisms that govern first pass metabolism. Drugs with high protein binding activity values tend to have a greater half-life compared to those with lower values.[1] The development of computational models for the prediction of drug pharmacokinetics is an area of current intense research in the pharmaceutical industry.[3,4] An undesirable proportion of compounds with good biological activity fails to progress to later stages of drug development because of inappropriate pharmacokinetic and pharmacodynamic properties.[5] Computational models of this type are useful because they rationalize a large number of experimental observations and therefore allow for saving time and money in the drug design process. In addition, they are useful in areas such as design of virtual compound collections, computational-chemical optimization of compounds, and design of combinatorial libraries with appropriate ADME (absorption, distribution, metabolism, and excretion) properties.

Quantitative structure−activity relationships (QSAR) have been successfully established to predict different important biopharmaceutical properties, such as metabolism,[6] toxicity,[7] oral bioavailability,[8] etc. Given the importance of drug binding to HSA, it should be extremely useful to develop QSAR models to predict the binding affinity to HSA. This would allow speeding up of the design of new compounds with appropriate HSA binding properties and therefore the optimization of the pharmacokinetics.

In the previous studies of drug albumin binding affinities, Colmenarejo et al. has experimentally determined through high-performance affinity chromatography the binding affinities to HSA of 95 diverse drugs and drug-like compounds and then developed 7 QSAR models for specific well-known families of drugs and for the whole database of 94 drugs based on genetic algorithms.[2] The best global model yields correlation coefficient (R^2) of 0.83. Hall et al. modeled the binding affinities to HSA for the same data set of 94 drugs using E-state topological descriptors, providing R^2 of 0.77, and a standard error (s) of 0.29.[9] However, some compounds have somewhat large residuals in both studies.

One of the important problems for the QSAR applications is the numerical representation (often called molecular descriptor) of the chemical structure. The built model performance and the accuracy of the results are strongly dependent on the structural representation. Various numerical representations of the compounds were proposed in the QSAR studies: constitutional and topological descriptors; numerical code; quantum chemistry descriptors, etc. The software CODESSA, developed by Katritzky group, enables the calculation of a large number of quantitative descriptors

* Corresponding author phone: +86-931-891-2578; fax: +86-931-891-2582; e-mail: snowmoun@21cn.com.
[†] Department of Chemistry, Lanzhou University.
[‡] Department of Computer Science, Lanzhou University.
[§] Université Paris 7-Denis Diderot.

10.1021/ci049820b CCC: $27.50
Published on Web 08/20/2004

based solely on the molecular structural information and combines diverse methods with advanced statistical analysis to establish molecular structure−property/activity relationships.[10,11] CODESSA has been applied successfully in a variety of QSAR analyses.[12,13]

After the calculation of the molecular descriptors, linear methods, such as MLR, principal component regression (PCR), and partial least squares (PLS) or nonlinear methods, e.g. neural networks, can be used in the development of a quantitative relationship between the structural descriptors and the property. Machine learning techniques such as neural networks, genetic algorithm, etc., have been applied to the QSAR analysis since the late 1980s, mainly in response to increased accuracy demands. The most popular neural networks model is the back-propagation (BP) neural network due to its simple architecture yet powerful problem-solving ability. However, the BP neural network suffers from a number of weaknesses which include the need for a large number of controlling parameters, difficulty in obtaining a stable solution, and the danger of overfitting. Other problems with the use of neural networks concern the reproducibility of results, due largely to random initialization of the networks and variation of stopping criteria.[14] Genetic algorithms can suffer in a similar manner. The stochastic nature of both population initialization and the genetic operators used during training can make results hard to reproduce.[15] Owing to the reasons outlined above, there is a continuing need for the application of more accurate and informative techniques in QSAR analysis.

The support vector machine (SVM) is a new algorithm developed from the machine learning community. Due to its remarkable generalization performance, the SVM has attracted attention and gained extensive application, such as pattern recognition problems,[16,17] drug design,[18] QSAR,[19] and quantitative structure−property relationship (QSPR) analysis.[20−23]

In the present work, the CODESSA program was used for the calculation of the descriptors and for the statistical analysis to obtain the multiparameter QSAR equations describing the binding affinities of drugs. The heuristic method (HM) in the CODESSA program and the SVM were utilized to establish a quantitative linear and nonlinear relationship between the binding affinity and the molecular structure, respectively. The principal objective of this investigation is the development of an accurate quantitative model that will establish a relation between the structural descriptors and the binding affinity and, at the same time, seek the important structural features related to the drug albumin binding affinity.

2. EXPERIMENTAL SECTION

2.1. Data Preparation. The study investigated QSAR model development for a diverse, heterogeneous group of commercially available drugs whose albumin binding affinity has been determined by a high-performance affinity chromatography by using an immobilized HSA column. The compounds and its albumin binding affinity values were taken from the paper published by Colmenarejo et al. and shown in Table 1.[2] The binding affinity was calculated in the logarithmic scale as $\log k(\mathrm{HSA}) = \log((t - t_0)/t_0)$, where t and t_0 are the retention times of the drug and nonretained material, respectively. The binding affinities in the data set fall in the range of −1.39 for acetylsalicylic acid to +1.34 for clotrimazole, respectively, with a mean value of −0.06. To compare the results with the literatures, the separation of the drugs in the training and test sets is identical with that in refs 2 and 9. The training set of 84 compounds was used to adjust the parameters of the models, and the test set of 10 compounds was used to evaluate its prediction ability.

2.2. Descriptor Calculation. The structures of the compounds were drawn with the HyperChem program and exported in a file format suitable for MOPAC.[24] The geometry optimization was performed with the semiempirical AM1 method in the MOPAC 6.0 program.[25,26] All the geometries had been fully optimized without symmetry restrictions. In all cases frequency calculations had been performed in order to ensure that all the calculated geometries correspond to true minima. The MOPAC output files were used by the CODESSA program to calculate five classes of descriptors: constitutional (number of various types of atoms and bonds, number of rings, molecular weight, etc.); topological (Wiener index, Randic indices, Kier-Hall shape indices, Balaban index, etc.); geometrical (moments of inertia, molecular volume, molecular surface area, etc.); electrostatic (minimum and maximum partial charges, polarity parameter, charged partial surface area descriptors, etc.); and quantum chemical (reactivity indices, dipole moment, HOMO and LUMO energies, etc.).[10]

3. METHODOLOGY

3.1. The Heuristic Method.[10] The heuristic multilinear regression procedures available in the framework of the CODESSA program were used to perform a complete search for the best multilinear correlations with a multitude of descriptors. These procedures provide collinearity control (i.e., any two descriptors intercorrelated above 0.8 are never involved in the same model) and implement heuristic algorithms for the rapid selection of the best correlation, without testing all the possible combinations of the available descriptors. The heuristic method of the descriptor selection proceeds with a preselection of the descriptors by eliminating (i) those descriptors that are not available for each structure, (ii) descriptors having a small variation in magnitude for all structures, (iii) descriptors that give a F-test's value below 1.0 in the one-parameter correlation, and (iv) descriptors whose t-values are less than the user-specified value, etc. This procedure orders the descriptors by decreasing correlation coefficient when used in one-parameter correlations. The next step involves correlation of the given property with (i) the top descriptor in the above list with each of the remaining descriptors and (ii) the next one with each of the remaining descriptors, etc. The best pairs, as evidenced by the highest F-values in the two-parameter correlations, are chosen and used for further inclusion of descriptors in a similar manner.

The goodness of the correlation is tested by the correlation coefficient (R^2), the F-test (F), and the squared standard error (s^2). The stability of the correlations was tested against the cross-validated coefficient, R^2_{cv}. The R^2_{cv} describes the stability of a regression model obtained by focusing on the sensitivity of the model to the elimination of any single data point. Briefly, for each data point, the regression is recalculated with the same descriptors but for the data set without this point. The obtained regression is used to predict the value

Table 1. Experimental and Calculated Binding Affinities of Drugs to HSA

no.	name	log*k* (HSA)	HM[b]	residue	SVM[c]	residue	no.	name	log*k* (HSA)	HM[b]	residue	SVM[c]	residue
1[a]	acetylsalicylic acid	−1.39	−0.48	0.91	−1.22	0.17	48	camptothecin	−0.08	0.49	0.57	−0.06	0.02
2	cefuroxime	−1.33	−1.26	0.07	−1.22	0.11	49	tetracycline	−0.08	−0.35	−0.27	−0.18	−0.10
3	amoxicillin	−1.21	−0.92	0.29	−1.04	0.17	50[a]	bupropion	−0.05	0.13	0.18	−0.03	0.02
4	cephalexin	−1.11	−0.73	0.38	−0.93	0.18	51	sumatriptan	−0.05	−0.29	−0.24	−0.12	−0.07
5	5-fluorocytosine	−1.11	−1.13	−0.02	−1.01	0.10	52	warfarin	−0.04	0.30	0.34	0.03	0.07
6	cromolyn	−1.07	−1.31	−0.24	−1.02	0.05	53	bumetanide	−0.03	0.05	0.08	−0.12	−0.09
7	ebselen	−1.04	−1.07	−0.03	−1.04	0.00	54	oxyphenbutazone	−0.02	0.06	0.08	0.07	0.09
8	zidovudine	−1.02	−1.26	−0.24	−1.15	−0.13	55	acrivastine	−0.02	0.44	0.46	−0.22	−0.20
9	caffeine	−0.92	−0.71	0.21	−0.89	0.03	56	phenytoin	0.00	0.03	0.03	0.05	0.05
10[a]	acetaminophen	−0.81	−0.92	−0.11	−0.74	0.07	57	doxiciclyne	0.01	−0.61	−0.62	0.04	0.03
11	l-tryptophan	−0.78	−0.56	0.22	−0.73	0.05	58	ketoprofen	0.03	−0.01	−0.04	0.11	0.08
12	methotrexate	−0.77	−0.52	0.25	−0.84	−0.07	59	alprenolol	0.04	−0.06	−0.10	−0.13	−0.17
13	propylthiouracil	−0.75	−0.77	−0.02	−0.58	0.17	60[a]	prazosin	0.06	−0.21	−0.27	−0.28	−0.34
14	antipyrine	−0.69	−0.24	0.45	−0.45	0.24	61	digitoxin	0.13	0.25	0.12	−0.03	−0.16
15	phenoxymethyl-penicillinic acid	−0.69	−0.55	0.14	−0.63	0.06	62	levofloxacin	0.14	−0.03	−0.17	−0.12	−0.26
							63	ciprofloxacin	0.14	0.02	−0.12	−0.15	−0.29
16	salicylic acid	−0.66	−0.77	−0.11	−0.78	−0.12	64	labetalol	0.14	0.08	−0.06	−0.31	−0.45
17	cefuroxime axetil	−0.56	−0.61	−0.05	−0.52	0.04	65	norfloxacin	0.14	−0.16	−0.30	−0.20	−0.34
18	etoposide	−0.49	−0.27	0.22	−0.24	0.25	66	phenylbutazone	0.19	0.38	0.19	0.14	−0.05
19	atenolol	−0.48	−0.32	0.16	−0.40	0.08	67	sancicline	0.21	−0.02	−0.23	0.12	−0.09
20[a]	chloramphenicol	−0.46	−0.81	−0.35	−0.77	−0.31	68	minocycline	0.21	0.11	−0.10	0.26	0.05
21	cimetidine	−0.44	−0.65	−0.21	−0.55	−0.11	69	naproxen	0.25	0.03	−0.22	0.33	0.08
22	chlorpropamide	−0.44	−0.50	−0.06	−0.57	−0.13	70[a]	clofibrate	0.27	−0.03	−0.30	−0.09	−0.36
23	sotalol	−0.44	−0.13	0.31	−0.34	0.10	71	propranolol	0.28	0.05	−0.23	0.14	−0.14
24	hydrochlorothiazide	−0.42	−0.43	−0.01	−0.27	0.15	72	tetracaine	0.32	0.31	−0.01	0.35	0.03
25	tolazamide	−0.42	−0.54	−0.12	−0.29	0.13	73	fusidic acid	0.33	0.60	0.27	0.48	0.15
26	hydrocortisone	−0.40	−0.22	0.18	−0.35	0.05	74	novobiocin	0.35	0.30	−0.05	0.40	0.05
27	nadolol	−0.40	−0.30	0.10	−0.35	0.05	75	ondansetron	0.37	0.33	−0.04	0.15	−0.22
28	prednisolone	−0.40	−0.29	0.11	−0.45	−0.05	76	droperidol	0.43	0.63	0.20	0.49	0.06
29	scopolamine	−0.34	−0.26	0.08	−0.29	0.05	77	quinidine	0.44	0.41	−0.03	0.54	0.10
30[a]	timolol	−0.33	−0.51	−0.18	−0.35	−0.02	78	indomethacin	0.47	0.31	−0.16	0.36	−0.11
31	metoprolol	−0.29	−0.03	0.26	−0.40	−0.11	79	quinine	0.49	0.40	−0.09	0.38	−0.11
32	trimethoprim	−0.26	−0.35	−0.09	−0.29	−0.03	80[a]	verapamyl	0.52	0.98	0.46	0.76	0.24
33	dansylglycine	−0.26	−0.30	−0.04	−0.21	0.05	81	sulfasalazine	0.56	0.21	−0.35	0.61	0.05
34	lidocaine	−0.23	−0.01	0.22	−0.24	−0.01	82	progesterone	0.59	0.49	−0.10	0.64	0.05
35	methylprednisolone	−0.22	−0.25	−0.03	−0.37	−0.15	83	desipramine	0.61	0.56	−0.05	0.70	0.09
36	tolbutamide	−0.22	−0.14	0.08	−0.18	0.04	84	estradiol	0.68	0.37	−0.31	0.79	0.11
37	sulfaphenazole	−0.21	−0.12	0.09	−0.11	0.10	85	glibenclamide	0.68	0.58	−0.10	0.57	−0.11
38	acebutolol	−0.21	−0.05	0.16	−0.22	−0.01	86	testosterone	0.74	0.30	−0.44	0.72	−0.02
39	procaine	−0.19	−0.19	0.00	−0.09	0.10	87	imipramine	0.75	0.77	0.02	0.81	0.06
40[a]	terazosin	−0.16	−0.08	0.08	−0.07	0.09	88	ketoconazole	0.84	0.86	0.02	0.79	−0.05
41	oxprenolol	−0.15	−0.04	0.11	−0.03	0.12	89	promazine	0.92	0.81	−0.11	0.97	0.05
42	lamotrigine	−0.13	−0.26	−0.13	−0.22	−0.09	90[a]	itraconazole	1.04	1.70	0.66	0.81	−0.23
43	clonidine	−0.13	−0.18	−0.05	−0.29	−0.16	91	triflupromazine	1.05	1.02	−0.03	1.15	0.10
44	pindolol	−0.13	−0.25	−0.12	−0.24	−0.11	92	chlorpromazine	1.10	0.89	−0.21	0.98	−0.12
45	frusemide	−0.13	−0.25	−0.12	−0.18	−0.05	93	terbinafine	1.17	0.82	−0.35	0.83	−0.34
46	carbamazepine	−0.10	0.34	0.44	0.09	0.19	94	clotrimazole	1.34	1.20	−0.14	1.19	−0.15
47	ranitidine	−0.10	−0.08	0.02	−0.06	0.04							

[a] Compounds in the test set. [b] Predicted binding affinity by HM. [c] Predicted binding affinity by SVM.

of this point, and the set of estimated values calculated in this way is correlated with the experimental values.

The heuristic method usually produces correlations 2−5 times faster than other methods, with comparable quality.[27] The rapidity of calculations from the heuristic method renders it the first method of choice in practical research. Thus, in this work, the heuristic method was used to build the linear model.

3.2. Support Vector Machine.[28,29] The foundation of support vector machines has been developed by Vapnik and is gaining popularity due to many attractive features and promising empirical performance.[30,31] The formulation embodies the Structural Risk Minimization (SRM) principle,[28,29] which has been shown to be superior to the traditional Empirical Risk Minimization (ERM) principle, employed by conventional neural networks. SRM minimizes an upper bound on VC dimension ("generalization error"), as opposed to ERM that minimizes the error on the training data. It is the difference that equips the SVM with good generalization performance, which is the goal in statistical learning. Originally, the SVM was developed for classification problems.[32] And now, with the introduction of ϵ-insensitive loss function, the SVM has been extended to solve nonlinear regression estimation.[33]

Compared to other neural network regressors, there are three distinct characteristics when a SVM is used to estimate the regression function. First of all, SVM estimates the regression using a set of linear functions that are defined in a high dimensional space. Second, SVM carries out the regression estimation by risk minimization where the risk is measured using Vapnik's ϵ-insensitive loss function. Third, SVM uses a risk function consisting of the empirical error and a regularization term which is derived from the SRM principle.

In support vector regression (SVR), the basic idea is to map the data x into a higher-dimensional feature space F via a nonlinear mapping Φ and then to do linear regression in this space. Therefore, regression approximation addresses the problem of estimating a function based on a given data set $G = \{(x_i,d_i)\}_i^n$ (x_i is the input vector, d_i is the desired value, and n is the total number of data patterns). SVM approximates the function using the following

$$y = f(x) = w\Phi(x) + b \tag{1}$$

where $\Phi(x)$ denotes the element wise mapping from x into feature space. The coefficients w and b are estimated by minimizing

$$R_{SVM}(C) = C\frac{1}{n}\sum_{i=1}^{n} L_\epsilon(d_i,y_i) + \frac{1}{2}||w||^2 \tag{2}$$

$$L_\epsilon(d,y) = \begin{cases} |d-y| - \epsilon & |d-y| \geq \epsilon \\ 0 & otherwise \end{cases} \tag{3}$$

In eq 2, R_{SVM} is the regularized risk function, and the first term $C(1/n)\sum_{i=1}^{n} L_\epsilon(d_i,y_i)$ is the empirical error (risk). They are measured by the ϵ-insensitive loss function (L_ϵ) given by eq 3. This loss function provides the advantage of enabling one to use sparse data points to represent the decision function given by eq 1. The second term $(1/2)||w||^2$, on the other hand, is the regularization term. C is referred to as the regularized constant, and it determines the tradeoff between the empirical risk and the regularization term. Increasing the value of C will result in the relative importance of the empirical risk with respect to the regularization term to grow. ϵ is called the tube size, and it is equivalent to the approximation accuracy placed on the training data points. Both C and ϵ are user-prescribed parameters.

Finally, by introducing Lagrange multipliers (a_i,a_i^*) and exploiting the optimality constraints, the decision function given by eq 4 has the following explicit form:

$$f(x,a_i,a_i^*) = \sum(a_i - a_i^*)K(x,x_i) + b \tag{4}$$

Based on the Karush-Kuhn-Tucker (KKT) conditions of quadratic programming, only a number of coefficients $(a_i - a_i^*)$ will assume nonzero values, and the data points associated with them could be referred to as support vectors. In eq 4, the kernel function K corresponds to $K(x,x_i) = \Phi(x)\cdot\Phi(x_i)$. One has several possibilities for the choice of this kernel function, including linear, polynomial, splines, and radial basis function. The elegance of using kernel function lies in the fact that one can deal with feature spaces of arbitrary dimensionality without having to compute the map $\Phi(x)$ explicitly. In SVR, a commonly used kernel function is the Gaussian Radial Basis Function.

The overall performances of HM and SVM were evaluated in terms of root-mean-square (rms) error which was defined as below

$$rms = \sqrt{\frac{\sum_{i=1}^{n_s}(y_k - \hat{y}_k)^2}{n_s}} \tag{5}$$

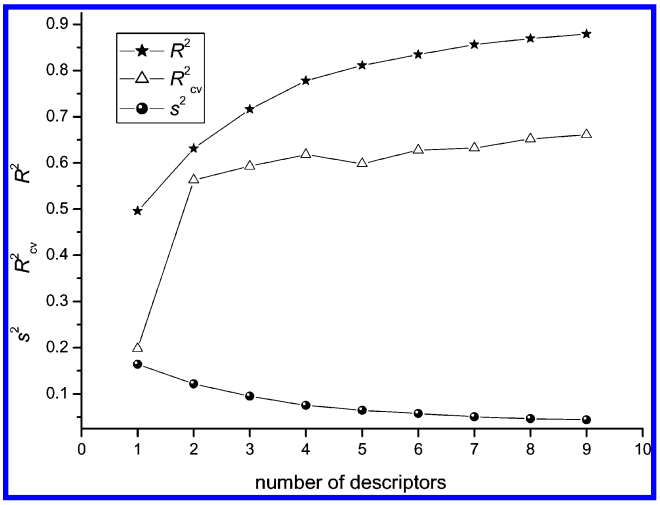

Figure 1. Influence of the number of descriptors on R^2, R^2_{cv}, and s^2 of the regression models.

Table 2. Seven-Descriptor Linear Model for the Binding Affinity[a]

descriptor	chemical meaning	coefficient	*t*-test
(constant)	intercept	−2.513 ± 0.388	−6.472
HDCA-2	HA dependent HDCA-2 [Zefirov's PC]	−0.401 ± 0.078	−5.136
MSA	molecular surface area	0.007 ± 0.001	12.801
NO	number of O atoms	−0.149 ± 0.017	−8.877
RNR	relative number of rings	9.210 ± 1.395	6.605
RNN	relative number of N atoms	−3.945 ± 0.663	−5.950
BI	Balaban index	0.403 ± 0.097	4.147
RNCS	relative negative charged SA (SAMNEG*RNCG) [quantum-chemical PC]	−0.045 ± 0.013	−3.392

[a] $R^2 = 0.86$; $s^2 = 0.050$; rms = 0.212; $n = 84$; $F = 63.89$; $R^2_{cv} = 0.63$.

where y_k is the desired output, $\hat{y}_k$ is the actual output of the model, and n_s is the number of compounds in analyzed set.

All calculation programs implementing SVM were written in R-file based on the R script for SVM.[34] The scripts were compiled using an R 1.7.1 compiler running on a Pentium IV PC with 256M RAM.

4. RESULTS AND DISCUSSION

4.1. Results of the Heuristic Method. About 600 descriptors were calculated by the CODESSA program for each of the compounds. After the heuristic reduction, the pool of the descriptors was reduced to 243. A variety of subset sizes was investigated to determine the optimum number of the descriptors in a model. When adding another descriptor did not improve significantly the statistics of a model, it was determined that the optimum subset size had been achieved. To avoid the "overparametrization" of the model, an increase of the R^2 value of less than 0.02 was chosen as the breakpoint criterion. The influences of the number of the descriptors on the correlation coefficient (R^2), the cross-validated coefficient (R^2_{cv}), and the squared standard error (s^2) were shown in Figure 1. From Figure 1, it can be seen that seven descriptors appear to be sufficient for a successful regression model. The multilinear analysis of the binding affinity values for the 84 compounds of the training set resulted in the seven-parameter model were summarized in Table 2, and the correlation matrix of these descriptors was shown in Table

Table 3. Correlation Matrix of the 7 Descriptors Used in This Work[a]

	HDCA-2	MSA	NO	RNR	RNN	BI	RNCS
HDCA-2	1.000	0.434	0.689	0.094	0.158	−0.320	−0.198
MSA		1.000	0.608	0.011	−0.208	−0.700	−0.394
NO			1.000	0.019	−0.278	−0.394	−0.193
RNR				1.000	0.098	−0.432	−0.023
RNN					1.000	0.197	−0.004
BI						1.000	0.311
RNCS							1.000

[a] The definitions of the descriptors were given in Table 2.

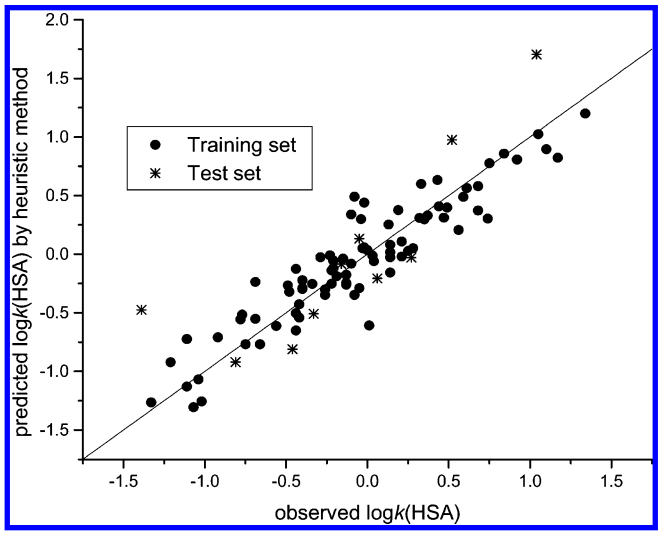

Figure 2. Plot of predicted log*k*(HSA) versus experimental values for the training and test sets by heuristic method.

3. The linear correlation coefficient value of each two descriptors is < 0.80 (Table 3), which means the descriptors were independent in this multilinear analysis. The obtained model had a correlation coefficient $R^2 = 0.86$, $F = 63.89$, with a squared standard error (s^2) of 0.050, and the cross-validated coefficient (R^2_{cv}) of 0.63. This model gave an rms error of 0.212 binding affinity units for the training set.

With the test set (Table 1), the prediction results were obtained, confirming the predictive capability of the model. The statistical parameters were $R^2 = 0.71$; $F = 19.27$; and $s^2 = 0.216$. The heuristic model produced an rms error of 0.430 binding affinity units for the test set and 0.245 for the whole data set. Figure 2 showed a plot of the calculated versus experimental binding affinities for all of the 94 compounds studied, the training set and the test set.

By interpreting the descriptors in the regression model, it is possible to gain some insight into factors that are likely to govern the binding affinities of the drugs to HSA. Generally, small molecules are bound to macromolecules through several types of interactions, such as hydrogen bonding, van der Waals, electrostatic, and hydrophobic interactions.[35,36] Extensive biochemical studies by Sudlow in the 1970s resulted in the proposition of two main drug-binding sites in HSA, denoted as I, or warfarin site, and II, or indole-benzodiazepine site.[37,38] These sites were afterward localized at subdomains IIA and IIIA, when the crystal structures of HSA with ligands were available. Due to the diversity of the molecules studied in this work, the binding affinities of the drugs related to the molecular structure in a complex way. Of the 7 descriptors, 3 are constitutional, 1 is topological, 1 is geometrical, 1 is electrostatic, and 1 is quantum-chemical descriptors. These descriptors encode different aspects of the molecular structure.

HA dependent HDCA-2 [Zefirov's PC] (HDCA-2), an electrostatic descriptor, is a hydrogen bonding acceptor dependent hydrogen bonding donor surface area, and this descriptor describes the hydrogen bonding acceptor properties of the compounds. The number of O atoms (NO) and the relative number of N atoms (RNN) are two constitutional descriptors. RNN is calculated as the number of N atoms divided by the number of atoms. The NO and RNN also partially account for the hydrogen bonding acceptor ability of the compounds. The three descriptors, HDCA-2, NO, and RNN, have a negative coefficient in the linear model, which indicates that these structural features make a negative contribution to the extent of protein binding. The larger the descriptors value is, the lower the calculated log*k*(HSA) is. Hence, the hydrogen bonding might not be favorable in protein binding, and we speculated that it is probably due to the hydrogen bonding weakening other factors which are important in determining HSA binding extent. Hydrogen bonding is formed when the distance between the hydrogen donor and the hydrogen acceptor and the angle made by covalent bonds to the donor and acceptor atoms are under certain conditions. The formation of the hydrogen bonding affects the space-matching between the protein and the drug, and this might weaken other interactions between the drug and HSA.

The relative negative charged surface area (RNCS), a quantum-chemical descriptor, represents or depends directly on the quantum-chemically calculated charge distribution in the molecules and can account for the electrostatic interaction between drugs and HSA. The descriptors of NO and RNN, on the other hand, also describe the electron accessibility of the molecules and give some information about the electrostatic interaction. NO, RNN, and RNCS have negative coefficients in the linear model, which indicates that log*k*(HSA) is inversely proportional to these descriptors. This might due to the electrostatic repulsion. The larger the descriptors value is, the larger the electrostatic repulsion is. Thus, an increase of these descriptors leads to a decrease of the calculated log*k*(HSA).

The molecular surface area (MSA) is a geometrical descriptor, which calculation requires 3D-coordinates of the atoms in the given molecule and gives information about the hydrophobic interaction. The positive coefficient in the linear model indicates that log*k*(HSA) is proportional to this descriptor; therefore, binding is favored for the molecules with large molecular surface area; and we concluded that an increasing of hydrophobicity increases drug binding to HSA. According to the *t*-test values (Table 2), the more relevant descriptor is MSA, and this indicates that hydrophobic interaction plays a prevailing role in the binding. This coincides with the conclusion in the previous work[2] and is supported by the X-ray structures of HSA, both alone and bound to different ligands.[39−41]

The relative number of rings (RNR), a constitutional descriptor, is calculated as the number of rings divided by the number of atoms. A wide variety of five- and six-membered rings are encountered in this data set including saturated, unsaturated rings, various di- and triazo rings, and various systems with more than one heteroatom. This

descriptor has a positive coefficient in the linear model and therefore indicates that molecules with a larger number of rings are expected to bind more tightly to HSA.

The Balaban index (BI), a topological descriptor, describes the atomic connectivity and branching information in the molecule and has some correlation with the hydrophobic interaction of the molecules. Because of its positive coefficient in the linear model, increasing this descriptor also increases the calculated log*k*(HSA) values, indicating that the large degree of branching for molecules is in favor of the binding. This echoes the importance of hydrophobicity in binding.

From the above discussion, it can be seen that all the descriptors involved in the model have explicit physical meaning, and these descriptors can account for the structural features responsible for the drug protein binding. According to the analysis of the corresponding regression coefficient (Table 2), molecular surface area, relative number of rings, and the Balaban index present positive contributions for binding affinity, whereas HA dependent HDCA-2 [Zefirov's PC], number of O atoms, relative number of N atoms, and relative negative charges surface area present negative contribution.

4.2. Result of SVM. 4.2.1. Selection of the Parameters of the SVM. From Table 1 and Figure 2, it can be seen that the model of the heuristic method was not sufficiently accurate and the prediction ability was not satisfactory (the rms error for the test set was 0.430), showing the factors influencing the binding affinities of these compounds were complex and not all of them were linear correlations with the binding affinity. So, after the establishment of the linear model by HM, we built the nonlinear prediction model by SVM to further discuss the correlation between the molecular structure and the binding affinity based on the same subset of descriptors.

Similar to other multivariate statistical models, the performances of SVM for regression depend on the combination of several parameters. They are capacity parameter C, ϵ of ϵ-insensitive loss function, the kernel type K, and its corresponding parameters. In this work, LOO cross-validation was performed for parameters selection,[42,43] which probably is the current best-performing approach to the SVM design problem.[44] C is a regularization parameter that controls the tradeoff between maximizing the margin and minimizing the training error. If C is too small, then insufficient stress will be placed on fitting the training data. If C is too large, then the algorithm will overfit the training data. To make the learning process stable, a large value should be set up for C.

The kernel type is another important parameter. For regression tasks, the Gaussian kernel is commonly used. The form of the Gaussian function is as follows

$$\exp(-\gamma^*|u - v|^2)$$

where γ is a constant, the parameter of the kernel, and u and v are two independent variables. γ controls the amplitude of the Gaussian function and, therefore, controls the generalization ability of SVM. Each rms error on the LOO cross-validation was plotted versus γ (Figure 3), and the minimum was chosen as the optimal conditions. In this case: $\gamma = 0.010$.

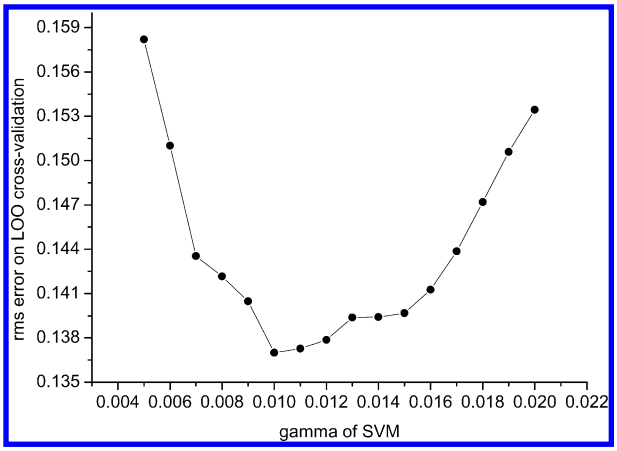

Figure 3. The gamma versus rms error on LOO cross-validation ($C = 100$, $\epsilon = 0.1$).

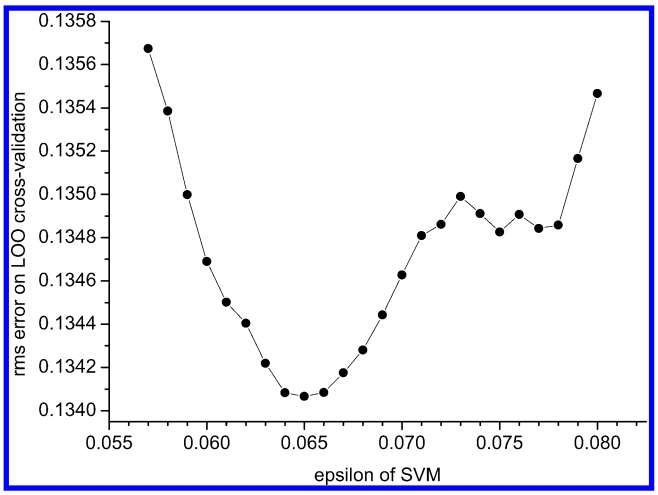

Figure 4. The epsilon versus rms error on LOO cross-validation ($C = 100$, $\gamma = 0.008$).

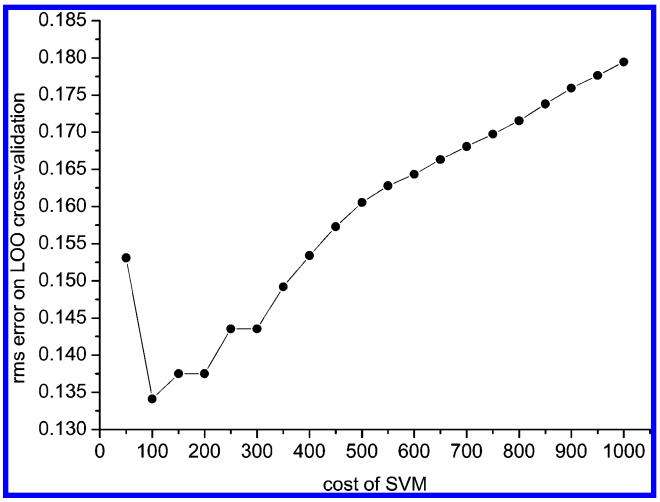

Figure 5. The C versus rms error on LOO cross-validation ($\gamma = 0.010$, $\epsilon = 0.065$).

The optimal value for ϵ depends on the type of noise present in the data, which is usually unknown. Even if enough knowledge of the noise is available to select an optimal value for ϵ, there is the practical consideration of the number of resulting support vectors. ϵ-insensitivity prevents the entire training set meeting boundary conditions and so allows for the possibility of sparsity in the dual formulation's solution. So, choosing the appropriate value

Table 4. Comparison of Different QSAR Models of Binding Affinity Prediction

	training set		test set		whole set					
approach	rms	R^2	rms	R^2	rms	R^2	*F*-test	Sig	*t*-test	Sig
ref 2	*c*	0.83	*c*	0.82	*c*	*c*	*c*	*c*	*c*	*c*
ref 9	0.269	0.77	0.387	0.74	0.284	0.76	294.882	0.000	17.172	0.000
HM[a]	0.212	0.86	0.430	0.71	0.245	0.83	434.274	0.000	20.839	0.000
SVM[b]	0.134	0.94	0.222	0.89	0.146	0.94	1392.176	0.000	37.312	0.000

[a] Model of HM. [b] Model of SVM. [c] Not reported.

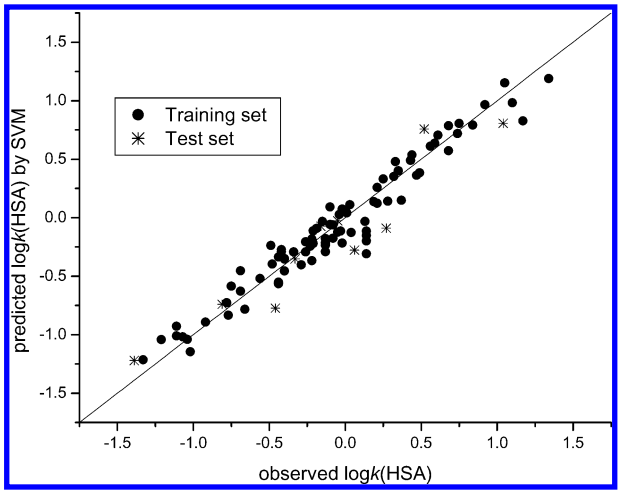

Figure 6. Plot of predicted log*k*(HSA) versus experimental values for the training and test sets by SVM ($\gamma = 0.010$, $\epsilon = 0.065$, $C = 100$).

of ϵ is critical from theory. To find an optimal ϵ, the rms on LOO cross-validation on different ϵ was calculated. The curve of rms versus the epsilon was shown in Figure 4. The optimal ϵ was found as 0.065.

The last important parameter is the regularization parameter *C*, of which the effect on the rms was shown in Figure 5. From Figure 5, the optimal *C* was found as 100.

4.2.2. The Predicted Results of SVM. Through the above process, the γ, ϵ, and *C* were fixed to 0.010, 0.065, and 100, respectively, when the support vector number of the SVM model was 70, the predicted results of the optimal SVM were shown in Table 1 and Figure 6. The model gave an rms error of 0.134 for the training set, 0.222 for the prediction set, and 0.146 for the whole set, and the corresponding correlation coefficients (R^2) were 0.94, 0.89, and 0.94, respectively. Figure 6 proved that the SVM model was statistically stable and fitted the data well.

4.3. Compare the Results Obtained by Different QSAR Approaches. To test the suitability of the QSAR approach constructed by SVM, the obtained binding affinities were compared with those calculated in refs 2 and 9 and the heuristic method. Table 4 showed the statistical parameters of the results obtained from the three studies for the same set of compounds. The rms errors of the SVM model for the training, the test, and the whole data set were much lower than that of the models proposed in ref 9 and the heuristic method. The correlation coefficient (R^2) given by the SVM model was higher than that of the models in refs 2 and 9 and the heuristic method. Through a regression analysis on the experimental and the calculated binding affinity obtained by different methods for the whole data set, the results of *F*-test and *t*-test were obtained and also shown in Table 4. From Table 4, it can be seen that the SVM model gives the highest *F* and *t* values, so this model gives the most satisfactory results, compared with the results obtained from ref 2, ref 9, and the heuristic methods. Consequently, this SVM approach currently constitutes the most accurate method to predict the binding affinity of drugs.

5. CONCLUSION

The heuristic method and the support vector machine were used to construct the linear and nonlinear quantitative relationships for the prediction of the affinity of a diverse set of 94 drugs binding to human serum albumin based on the descriptors calculated from the molecular structure alone. Both the linear and nonlinear models provided the satisfactory results, and, at the same time, the nonlinear SVM models produced better results with good predictive ability than that of the linear model, so we can conclude that (1) the linear model constructed by the heuristic method could correctly represent the relationship between the binding affinities and the molecular descriptors calculated solely from the molecular structures, moreover, the 7 selected descriptors can represent the features of the compounds responsible for their binding behavior. So, the heuristic linear model could identify and provide some insight into which structural features are related to the drug albumin binding affinity. The linear model indicates that an increase of hydrophobicity is expected to result in an increased drug HSA binding. (2) By comparison of the results from the different QSAR approaches, it can be seen that the nonlinear model can describe accurately the relationship between the structural parameter and the drug HSA binding affinity. (3) SVM proved to be a very promising tool in the prediction of the affinity of drugs binding to HSA. The training procedure is also simple when using SVM because there are fewer parameters having to be optimized, and only support vectors are used in the generalization process. Besides, the SVM exhibits the better whole performance due to embodying the Structural Risk Minimization principle and some advantages over the other techniques. Furthermore, the proposed approach can also be extended to other QSPR or QSAR investigations.

ACKNOWLEDGMENT

The authors thank the Association Franco-Chinoise pour la Recherche Scientifique & Technique (AFCRST) for supporting this study (Program PRA SI 02-02).

REFERENCES AND NOTES

(1) Hall, L. M.; Hall, L. H.; Kier, L. B. QSAR modeling of β-lactam binding to human serum proteins. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 103−118.
(2) Colmenarejo, G.; Alvarez-Pedraglio, A.; Lavandera, J.-L. Cheminformatic models to predict binding affinities to human serum albumin. *J. Med. Chem.* **2001**, *44*, 4370−4378.

(3) Blake, J. F. Chemoinformatics-predicting the physicochemical properties of ''drug-like" molecules. *Curr. Opin. Biotechnol.* **2000**, *11*, 104−107.
(4) Butina, D.; Segall, M. D.; Frankcombe, K. Predicting ADME properties in silico: Methods and models. *Drug Discovery Today* **2002**, *7*, S83−S88.
(5) Prentis, R. A.; Lis, Y.; Walker, S. R. Pharmaceutical innovation by the seven UK-owned pharmaceutical companies (1964−1985). *Br. J. Clin. Pharmacol.* **1988**, *25*, 387−396.
(6) Ekins, S.; Obach, R. S. Three-dimensional quantitative structure activity relationship computational approaches for prediction of human in vitro intrinsic clearance. *J. Pharmacol. Exp. Ther.* **2000**, *295*, 463−473.
(7) Cronin, M. T. D. Computational methods for the prediction of drug toxicity. *Curr. Opin. Drug Discovery Dev.* **2000**, *3*, 292−297.
(8) Yoshida, F.; Topliss, J. G. QSAR model for drug human oral bioavailability. *J. Med. Chem.* **2000**, *43*, 2575−2585.
(9) Hall, L. M.; Hall, L. H.; Kier, L. B. Modeling drug albumin binding affinity with E-state topological structure representation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2120−2128.
(10) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. Comprehensive descriptors for structural and statistical analysis, Reference Manual, Version 2.13; 1995−1997.
(11) Katritzky, A. R.; Lobanov, V.; Karelson, M. QSPR: The correlation and quantitative prediction of chemical and physical properties from structure. *Chem. Soc. Rev.* **1995**, *24*, 279−287.
(12) Oblak, M.; Randic, M.; Solmajer, T. Quantitative structure−activity relationship of flavonoid analogues. 3. Inhibition of $p56^{lck}$ protein tyrosine kinase. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 994−1001.
(13) Katritzky, A. R.; Tatham, D. B. Theoretical descriptors for the correlation of aquatic toxicity of environmental pollutants by quantitative structure-toxicity relationships. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1162−1176.
(14) Manallack, D. T.; Livingstone, D. J. Neural networks in drug discovery: have they lived up to their promise? *Eur. J. Med. Chem.* **1999**, *34*, 95−208.
(15) Goldberg, D. *Genetic Algorithms in Search, Optimization and Machine Learning*; Addison-Wesley: Reading, MA, 1989.
(16) Byvatov, E.; Fechner, U.; Sadowski, J.; Schneider, G. Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1882−1889.
(17) Liu, H. X.; Zhang, R. S.; Luan, F.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T. Diagnosing breast cancer based on support vector machines. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 900−907.
(18) 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.
(19) Liu, H. X.; Zhang, R. S.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T. QSAR study of ethyl 2-[(3-Methyl-2,5-dioxo(3-pyrrolinyl))amino]-4-(trifluoromethyl) pyrimidine-5-carboxylate: an inhibitor of AP-1 and NF−KB mediated gene expression based on support vector machines. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1288−1296.
(20) Liu, H. X.; Zhang, R. S.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T. Prediction of the isoelectric point of an amino acid based on GA-PLS and SVMs. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 161−167.
(21) Xue, C. X.; Zhang, R. S.; Liu, H. X.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T. An accurate QSPR study of O−H bond dissociation energy in substituted phenols based on support vector machines. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 669−677.
(22) Xue, C. X.; Zhang, R. S.; Liu, M. C.; Hu, Z. D.; Fan, B. T. Study of the quantitative structure-mobility relationship of carboxylic acids in capillary electrophoresis based on support vector machines. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, *950−957*.
(23) Xue, C. X.; Zhang, R. S.; Liu, H. X.; Liu, M. C.; Hu, Z. D.; Fan, B. T. Support vector machines-based quantitative structure−property relationship for the prediction of heat capacity. *J. Chem. Inf. Comput. Sci*. **2004**, *44*, 1267−1274.
(24) HyperChem, Release 4.0 for Windows, Hypercube, Inc., 1995.
(25) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.
(26) Stewart, J. P. P. *MOPAC 6.0, Quantum Chemistry Program Exchange;* QCPE, No. 455, Indiana University, Bloomington, IN, 1989.
(27) Katritzky, A. R.; Petrukhin, R.; Jain, R.; Karelson, M. QSPR analysis of flash point. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1521−1530.
(28) Burges, C. J. C. A tutorial on support vector machines for pattern recognition. *Data Mining Knowledge Discovery* **1998**, *2*(2), 1−47.
(29) Vapnik, V. *Estimation of Dependences Based on Empirical Data*; Springer: Berlin, 1982.
(30) Smola, A. J.; Schölkopf, B. *A tutorial on support vector regression*; NeuroCOL2 Technical report series, NC2-TR-1998-030; October, 1998.
(31) Vapnik, V. *The Nature of Statistical Learning Theory*; Springer: New York, 1995.
(32) Burges, C. J. C. A tutorial of support vector machines for pattern recognition. http://svm.research.bell-labs.com/SVMdoc.html, 1998.
(33) Vapnik, V.; Golowich, S.; Smola, A. Support vector method for function approximation, regression estimation, and signal processing. *Adv. Neural Inform. Process. Systems* **1997**, *9*, 281−287.
(34) Venables, W. N. D.; Smith, M. and the R Development Core Team. R manuals 2003.
(35) Liu, J. Q.; Tian, J. N.; Tian, X.; Hu, Z. D.; Chen, X. G. Interaction of isofraxidin with human serum albumin. *Bioorg. Med. Chem.* **2004**, *12*, 469−474.
(36) Liu, J. Q.; Tian, J. N.; Li, Y.; Yao, X. J.; Hu, Z. D.; Chen, X. G. Binding of the bioactive component daphnetin to human serum albumin demonstrated using tryptophan fluorescence quenching. *Macromol. Biosci.* **2004**, *4*, 520−525.
(37) Sudlow, G.; Birkett, D. J.; Wade, D. N. Characterization of two specific drug binding sites on human serum albumin. *Mol. Pharmacol.* **1975**, *11*, 824−832.
(38) Sudlow, G.; Birkett, D. J.; Wade, D. N. Further characterization of specific drug binding sites on human serum albumin. *Mol. Pharmacol.* **1976**, *12*, 1052−1061.
(39) Carter, D. C.; He, X.-M.; Munson, S. H.; Twigg, P. D.; Gernert, K. M.; Broom, M. B.; Miller, T. Y. Three-dimensional Structure of Human Serum Albumin. *Science* **1994**, *244*, 1195−1198.
(40) Carter, D. C.; He, X.-M. Structure of Human Serum Albumin. *Science* **1990**, *249*, 302−303.
(41) He, X. M.; Carter, D. C. Atomic structure and chemistry of human serum albumin. *Nature* **1992**, *358*, 209−215.
(42) Schölkopf, B.; Burges, J.; Smola, A. *Advances in kernel methods: Support vector machine*; MIT Press: Cambridge, MA, 1999.
(43) Cherkassky, V.; Mulier, F. *Learning from data: Concepts, theory, and methods*; Wiley: New York, 1998.
(44) Anguita, D.; Ridella, S.; Rivieccio, F.; Zunino, R. Hyperparameter design criteria for support vector classifiers. *Neurocomputing* **2003**, *55*, 109−134.

CI049820B