

# 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- $\kappa$ B Mediated Gene Expression Based on Support Vector Machines

H. X. Liu,<sup>†</sup> R. S. Zhang,<sup>\*,†,‡</sup> X. J. Yao,<sup>†,||</sup> M. C. Liu,<sup>†</sup> Z. D. Hu,<sup>†,§</sup> and B. T. Fan<sup>||</sup>

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

Received February 25, 2003

The support vector machine, as a novel type of learning machine, for the first time, was used to develop a QSAR model of 57 analogues of ethyl 2-[(3-methyl-2,5-dioxo(3-pyrrolinyl))amino]-4-(trifluoromethyl)pyrimidine-5-carboxylate (EPC), an inhibitor of AP-1 and NF- $\kappa$ B mediated gene expression, based on calculated quantum chemical parameters. The quantum chemical parameters involved in the model are Kier and Hall index (order3) (KHI3), Information content (order 0) (IC0), YZ Shadow (YZS) and Max partial charge for an N atom (MaxPCN), Min partial charge for an N atom (MinPCN). The mean relative error of the training set, the validation set, and the testing set is 1.35%, 1.52%, and 2.23%, respectively, and the maximum relative error is less than 5.00%.

## 1. INTRODUCTION

In this section, the special function of a series of analogues of EPC was described simply first. Then, the application of quantitative structure–activity relationship (QSAR) on drug design was introduced. Finally, the advantages and disadvantages of the techniques commonly used in QSAR were displayed, and new techniques need to be introduced in order to design drugs better.

**1.1. Special Function of the Analogues of EPC.** T-lymphocytes (T-cell) orchestrate both the initiation and the propagation of various immune responses through the secretion of protein mediators termed cytokines. These cytokines play a significant role in a number of inflammatory diseases such as asthma, psoriasis, rheumatoid arthritis, and transplant rejection. Several studies have shown that T-cell driven immune responses appear to overreact in these disease states.<sup>1</sup> In activated T-cells, transcription factors such as the activator protein-1 (AP-1) regulate IL-2 and matrix metalloproteinases production, while the nuclear factor- $\kappa$ B (NF- $\kappa$ B) is essential for the transcriptional regulation of the proinflammatory cytokines IL-1, IL-6, IL-8, and TNF- $\alpha$ . Based on these findings, it appears that inhibition of AP-1 and NF- $\kappa$ B transcriptional activation in T-cells may represent an attractive target in the development of novel antiinflammatory drugs.

Very few compounds are known to inhibit both the AP-1 and NF- $\kappa$ B mediated transcriptional activation.<sup>2</sup> Recently, Moorthy S. S. Palanki et al. synthesized a series of novel compounds (Figure 1, Table 1) and tested them in Jurkat

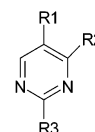


Figure 1.

T-cells stably transfected with promoter-reporter gene constructs driven by an AP-1 binding site and a NF- $\kappa$ B binding site.<sup>3</sup> However, no quantitative structure–activity relationship (QSAR) model has been reported to date.

To further design new drugs with high activity, it is very necessary and urgent to investigate quantitative structure–activity relationships of the series of compounds.

**1.2. Application of QSAR on Drug Design.** The advent of combinatorial chemistry in the mid-1980s has allowed the synthesis of hundreds, thousands, and even millions of new molecular compounds at a time. Nevertheless, even this level of compound production will fall short of exhausting the trillions of potential combinations within a few thousand years. The need for a more refined search than simply producing and testing every single molecular combination possible has meant that statistical methods and, more recently, intelligent computation have become an integral part of the drug production process. Structure–activity relationship (SAR) analysis is one technique used to reduce the search for new drugs. It also presents an extremely challenging problem to the field of intelligent systems. A successful solution to this problem has the potential to provide significant economic benefit via increased process efficiency.<sup>4</sup>

The underlying assumption behind SAR analysis is that there is a relationship between the variation of biological activity within a group of molecular compounds with the variation of their respective structural and chemical features. The analyst searches for a rule or function that predicts a

\* Corresponding author phone: +86-931-891-2578; fax: +86-931-891-2582; e-mail: ruison@public.lz.gs.cn.

<sup>†</sup> Department of Chemistry, Lanzhou University.

<sup>‡</sup> Department of Computer Science, Lanzhou University.

<sup>§</sup> State Key Laboratory of Applied Organic Chemistry, Lanzhou University.

<sup>||</sup> Université Paris 7-Denis Diderot.

**Table 1.** Structures of the Analogues of EPC

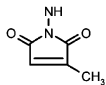

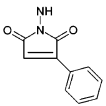

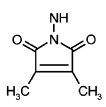
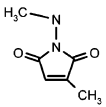
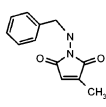
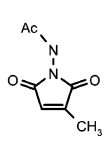
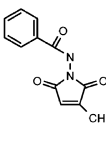
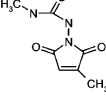
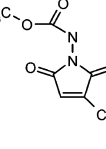



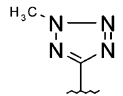



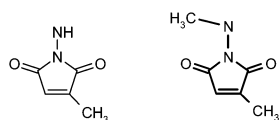
No.	R1	R2	R3	No.	R1	R2	R3
1		CF <sub>3</sub>	CO <sub>2</sub> Et	30	a	CF <sub>3</sub>	CONMe <sub>2</sub>
2		CF <sub>3</sub>	CO <sub>2</sub> Et	31	a	CF <sub>3</sub>	COMe
3	NH <sub>2</sub>	CF <sub>3</sub>	CO <sub>2</sub> Et	32	a	CF <sub>3</sub>	COPh
4		CF <sub>3</sub>	CO <sub>2</sub> Et	33	a	CF <sub>3</sub>	
5		CF <sub>3</sub>	CO <sub>2</sub> Et	34	a	CH <sub>2</sub> CH <sub>3</sub>	COPh
6		CF <sub>3</sub>	CO <sub>2</sub> Et	35	a	CH <sub>2</sub> CH <sub>3</sub>	COCH <sub>2</sub> CH <sub>2</sub> CH <sub>3</sub>
7		CF <sub>3</sub>	CO <sub>2</sub> Et	36	a	CH <sub>2</sub> CH <sub>3</sub>	Cyclopropyl ester
8		CF <sub>3</sub>	CO <sub>2</sub> Et	37	a	CH <sub>2</sub> CH <sub>3</sub>	CH <sub>2</sub> OH
9		CF <sub>3</sub>	CO <sub>2</sub> Et	38	a	CH <sub>2</sub> CH <sub>3</sub>	CN
10		CF <sub>3</sub>	CO <sub>2</sub> Et	39	a	CH <sub>3</sub>	COCH <sub>3</sub>
11		CF <sub>3</sub>	CO <sub>2</sub> Et	40	a	CH <sub>2</sub> CH <sub>3</sub>	
12	a	H	CO <sub>2</sub> Et	41	a	CH <sub>2</sub> CH <sub>3</sub>	
13	a	CH <sub>3</sub>	CO <sub>2</sub> Et	42	a	CH <sub>2</sub> CH <sub>3</sub>	
14	a	CH <sub>2</sub> CH <sub>3</sub>	CO <sub>2</sub> Et	43	a	CH <sub>2</sub> CH <sub>3</sub>	

Table 1 (Continued)

No.	R1	R2	R3	No.	R1	R2	R3
15	a	CF <sub>2</sub> CF <sub>3</sub>	CO <sub>2</sub> Et	44	a	CH <sub>2</sub> CH <sub>3</sub>	
16	a	Ph	CO <sub>2</sub> Et	45	a	2-Thienyl	
17	a	CH <sub>2</sub> Ph	CO <sub>2</sub> Et	46	a	2-Thienyl	CO <sub>2</sub> -tBu
18	a	CH <sub>2</sub> OCH <sub>3</sub>	CO <sub>2</sub> Et	47		CH <sub>2</sub> CH <sub>2</sub> CH <sub>3</sub>	CO <sub>2</sub> Et
19	a	2-Furanyl	CO <sub>2</sub> Et	48	b	CF <sub>2</sub> CF <sub>3</sub>	CO <sub>2</sub> Et
20	a	2-Thienyl	CO <sub>2</sub> Et	49	b	CH <sub>2</sub> CH <sub>3</sub>	CO <sub>2</sub> Et
21	a	3-Thienyl	CO <sub>2</sub> Et	50	b	2-Beno[b]thienyl	CO <sub>2</sub> Et
22	a	2-(5-Methylthienyl)	CO <sub>2</sub> Et	51	b	2-Thiazolyl	CO <sub>2</sub> Et
23	a	2-(5-Chlorothieryl)	CO <sub>2</sub> Et	52	b	CH <sub>2</sub> CH <sub>3</sub>	CO <sub>2</sub> Et
24	a	2-Beno[b]thienyl	CO <sub>2</sub> Et	53	b	CF <sub>2</sub> CF <sub>3</sub>	CO <sub>2</sub> Et
25	a	2-Thiazolyl	CO <sub>2</sub> Et	54	b	2-Thienyl	CO <sub>2</sub> Et
26	a	Cyclopropyl	CO <sub>2</sub> Et	55	b	3-Thienyl	CO <sub>2</sub> Et
27	a	CF <sub>3</sub>	CO <sub>2</sub> -tBu	56	b	2-(5-Methylthienyl)	CO <sub>2</sub> Et
28	a	CF <sub>3</sub>	CO <sub>2</sub> H	57	b	CH <sub>2</sub> CH <sub>3</sub>	CH <sub>2</sub> OCH <sub>3</sub>
29	a	CF <sub>3</sub>	CONH <sub>2</sub>				



Note: a b

molecule's activity from the values of its structural descriptors. The aim of SAR analysis is to discover such general rules and equations. QSAR involves modeling a continuous activity for quantitative prediction of the activity of previously unseen compounds. The advances in quantitative structure-activity relationship (QSAR) studies have widened the scope of rationalizing drug design and the search for the mechanisms of drug actions.

**1.3. The Techniques Commonly Used in QSAR.** Artificial intelligence techniques have been applied to SAR analysis since the late 1980s, mainly in response to increased accuracy demands. Machine learning techniques have, in general, offered greater accuracy than have their statistical forebears, but there exist accompanying problems for the SAR analyst to consider. Neural networks, for example, offer high accuracy in most cases but can suffer from overfitting

the training data.<sup>5</sup> Other problems with the use of neural networks concern the reproducibility of results, due largely to random initialization of the network and variation of stopping criteria, and lack of information regarding the classification produced.<sup>5</sup> 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. Owing to the reasons outlined above, there is a continuing need for the application of more accurate and informative techniques to SAR analysis.<sup>4</sup>

The support vector machine (SVM) is a new algorithm from the machine learning community. Due to its remarkable generalization performance, the SVM has attracted attention and gained extensive application.<sup>4,6,7-11</sup>

In this paper, we built the model of the 2D-QSAR model based on support vector machines which recently developed from the machine learning community, with structural descriptors calculated by the software CODESSA, to explore the correlations of the molecular structure and the activity of a series of novel compounds.

## 2. DATA DESCRIPTION

**2.1. Data Set.** The target compounds are the derivatives of 1,3-diazine (pyrimidine) with a similar skeleton (see Figure 1, Table 1). The biological activity expressed by IC<sub>50</sub> (the 50% inhibitory concentration to AP-1 and NF- $\kappa$ B mediated transcriptional activation in Jurkat T-cells) was taken from ref 3. In this article, the biological activity was expressed by PIC<sub>50</sub> (that is  $-\log IC_{50}$ ). The data set was divided into a training set of 45 compounds, a validation set of 6 compounds, and a test set of 6 compounds by random.

**2.2. Descriptor Calculation.** The three-dimensional structures of the molecules were drawn with the ISIS DRAW program. The final geometries were obtained with the semiempirical AM1 method in the HYPERCHEM program. All calculations were carried out at restricted Hartree-Fock level with no configuration interaction. The molecular structures were optimized using the Polak-Ribiere algorithm until the root-mean-square gradient was 0.001. The resulting geometry was transferred into software CODESSA, developed by the Katritzky group,<sup>12,13</sup> that can calculate constitutional, topological, electrostatic, and quantum chemical descriptors and has been successfully used in various QSPR and QSAR researches. Constitutional descriptors are related to the number of atoms and bonds in each molecule. Topological descriptors include valence and nonvalence molecular connectivity indices calculated from the hydrogen-suppressed formula of the molecule, encoding information about the size, composition, and the degree of branching of a molecule. The quantum chemical descriptors include information about binding and formation energies, partial atom charge, dipole moment, and molecular orbital energy levels. A full list of 74 calculated descriptors and their chemical meaning is given in the Supporting Information.

**2.3. Selection of Descriptors.** Since it is not possible to know a priori which descriptors are most relevant to the problem at hand, a comprehensive set of descriptors is usually employed, chosen based on experience, software availability, and computational cost. However, it is well-

**Table 2.** Correlation Matrix of the Five Parameters

	KHI3	IC0	YZS	MaxPCN	MinPCN
KHI3	1.000	0.629	0.775	-0.318	0.220
IC0		1.000	0.804	0.127	0.301
YZS			1.000	-0.142	0.125
MaxPCN				1.000	-0.011
MinPCN					1.000

known, both in the chemical and statistical fields, that the accuracy of classification and regression techniques is not monotonic with respect to the number of features employed by the model. Depending on the nature of the regression technique, the presence of irrelevant or redundant features can cause the system to focus attention on the idiosyncrasies of the individual samples and lose sight of the broad picture that is essential for generalization beyond the training set. This problem is compounded when the number of observations is also relatively small. If the number of variables is comparable to the number of training patterns, the parameters of the model may become unstable and unlikely to replicate if the study were to be repeated. So, selection of descriptors is very necessary in order to remedy this situation by identifying a small subset of relevant features and using only them to construct the actual model. Generally, the number of the samples is five times the descriptors at least. Although there is more strong capacity of suffering redundancy for SVMs, the better it is from application, the less the number of descriptors is on the condition that the same generality is reached.

In this article, correlation analysis of descriptors was performed first. In the process of correlation analysis, either parameter which correlation coefficient is more than 0.85 was discarded. To decide to which should be discarded, chemical experience is very important. For example, the correlation coefficient between the polarity parameter (Qmax-Qmin) and Max partial charge (Qmax) is 0.942. Which parameter should be discarded? It is generally agreed that the biological activities (in vitro) of drugs are mainly determined by their steric and electrostatic features from chemistry and biology. From the chemistry, the polarity parameter can express the electrostatic feature better on the whole, and it was selected for the following analysis. The new data set of 35 parameters after correlation analysis was dealt with stepwise regression analysis. Five parameters given in Table 2 were selected into the equation with the forward stepwise regression analysis. The correlation matrix of the five parameters was displayed in Table 2.

## 3. METHODOLOGY

**3.1. Support Vector Machines.** The support vector machine (SVM), developed by Vapnik,<sup>14</sup> as a novel type of learning machine, is gaining popularity due to many attractive features and promising empirical performance. Comparing with traditional neural networks, SVM possesses the following prominent advantages: (1) Strong theoretical background provides SVM with high generalization capability and can avoid local minima. (2) SVM always has solution, which can be quickly obtained by a standard algorithm (quadratic programming). (3) SVM need not determine network topology in advance, which can be automatically

obtained when training process ends. (4) SVM builds a result based on a sparse subset of training samples, which reduce the workload. It can solve high-dimension problems and therefore avoid the “curse of dimensionality”. The root cause that SVM attracts more and more attention is that SVM adopts the structure risk minimization (SRM) principle, which has been shown to be superior to the traditional empirical risk minimization (ERM) principle (Vapnik, 1998), employed by conventional neural networks.<sup>15</sup> SRM minimizes an upper bound of the generalization error on Vapnik-Chernoverkis (VC) dimension, as opposed to ERM that minimizes the training error. It is the difference that equips SVM with good generalization performance, which is the goal of learning problems. Originally, SVMs are developed for pattern recognition problems. And now, with the introduction of  $\epsilon$ -insensitive loss function, SVMs have been extended to solve nonlinear regression estimation and time-series prediction,<sup>16</sup> and excellent performances have been obtained.<sup>10,16</sup>

Theories of support vector classification and SRM can be found in the tutorials for SVMs<sup>18</sup>. Here, only the theory of support vector machines for regression was introduced simply.

**3.2. Theory of SVMs for Regression.**<sup>16,17,19</sup> SVMs can also be applied to regression by the introduction of an alternative loss function and results appear to be very encouraging. In SVR, the basic idea is to map the data  $x$  into a higher-dimensional feature space  $F$  via a nonlinear mapping  $\Phi$  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  $G = \{(x_i; d_i)\}_{i=1}^l$  ( $x_i$  is input vector,  $d_i$  is the desired value). SVMs approximate the function in the following form

$$y = \sum_{i=1}^l w_i \Phi_i(x) + b \quad (1)$$

where  $\{\Phi_i(x)\}_{i=1}^l$  are the features of inputs, and  $\{w_i\}_{i=1}^l$  and  $b$  are coefficients. They are estimated by minimizing the regularized risk function (2)

$$R(C) = C \frac{1}{N} \sum_{i=1}^N L_\epsilon(d_i, y_i) + \frac{1}{2} \|w\|^2 \quad (2)$$

where

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

and  $\epsilon$  is a prescribed parameter.

In eq 2,  $C(1/N)\sum_{i=1}^N L_\epsilon(d_i, y_i)$  is the so-called empirical error (risk), which is measured by  $\epsilon$ -insensitive loss function  $L_\epsilon(d, y)$ , which indicates that it does not penalize errors below  $\epsilon$ . The second term,  $1/2\|w\|^2$ , is used as a measurement of function flatness.  $C$  is a regularized constant determining the tradeoff between the training error and the model flatness. Introduction of slack variables “ $\xi$ ” leads eq 2 to the following constrained function:

$$\text{Max } R(w, \xi^*) = \frac{1}{2} \|w\|^2 + C^* \sum_{i=1}^n (\xi_i + \xi_i^*) \quad (4)$$

$$\begin{aligned} \text{s.t. } w\Phi(x_i) + b - d_i &\leq \epsilon + \xi_i \\ d_i - w\Phi(x_i) - b_i &\leq \epsilon + \xi_i \\ \xi_i, \xi_i^* &\geq 0 \end{aligned} \quad (5)$$

Thus, decision function (1) becomes the following form

$$f(x, \alpha_i, \alpha_i^*) = \sum_{i=1}^l (\alpha_i^* - \alpha_i) K(x, x_i) + b \quad (6)$$

In function (6),  $\alpha_i, \alpha_i^*$  are the introduced Lagrange multipliers. They satisfy the equality  $\alpha_i \alpha_i^* = 0$ ,  $\alpha_i \geq 0$ ,  $\alpha_i^* \geq 0$ ;  $i = 1, \dots, l$ , and are obtained by maximizing the dual form of function (4), which has the following form

$$\begin{aligned} \Phi(\alpha_i, \alpha_i^*) &= \sum_{i=1}^l d_i (\alpha_i - \alpha_i^*) - \epsilon \sum_{i=1}^l (\alpha_i - \alpha_i^*) \\ &- \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l (\alpha_i - \alpha_i^*) (\alpha_j - \alpha_j^*) K(\alpha_i, \alpha_j) \end{aligned} \quad (7)$$

with the following constraints:

$$\begin{aligned} \sum_{i=1}^l (\alpha_i - \alpha_i^*) &= 0 \\ 0 &\leq \alpha_i \leq C, \quad i = 1, \dots, l \\ 0 &\leq \alpha_i^* \leq C, \quad i = 1, \dots, l \end{aligned} \quad (8)$$

Based on the Karush-Kuhn-Tucker (KKT) conditions of quadratic programming, only a number of coefficients  $(\alpha_i - \alpha_i^*)$  will assume nonzero values, and the data points associated with them could be referred to as support vectors.

In eq 6,  $K(x_i, x_j)$  is the kernel function. The value is equal to the inner product of two vectors  $x_i$  and  $x_j$  in the feature space  $\Phi(x_i)$  and  $\Phi(x_j)$ . That is,  $K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$ . 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. Any function that satisfies Mercer’s condition can be used as the kernel function. In support vector regression, the Gaussian kernel  $K(x, y) = \exp(-(x - y)^2/\delta^2)$  is commonly used.

**3.3. SVM Implementation and Computation Environment.** All calculation programs implementing SVM were written in R-file based on R script for SVM. All scripts were compiled using R1.5.1 compiler running operating system on a Pentium IV with 256M RAM.

## 4. RESULTS AND DISCUSSION

**4.1. Result of MLR.** To investigate the possible correlation of input parameters and the biology activity values, multilinear regression (MLR) was used. The MLR results were shown in Table 3. From Table 3, it can be seen that there is no simple linear correlation between the biology activity values and the input parameters; nonlinear model SVM was then applied to predict the biology activity values.

**4.2. Result of SVM. 4.2.1. Selection of the Kernel Function and Parameters of the SVM.** Similar to other multivariate statistical models, the performances of SVM for



**Table 3.** MLR Results on the Correlation between Input Parameters and the  $-\log IC_{50}$ 

item	degrees of freedom	sum of square	mean square	F statistic	R
model	5	15.179	3.306	9.702	0.698
error	51	15.957	0.313		
total	56	31.136			

**Table 4.** Training Validation and Testing Sets

set	compound numbers
training set	2,3,4,5,6,7,8, 11,12,13,14,15,16,17,18,19,20,21, 22,23,25,26,27,28,29,30,32,33,35,36,38,40, 41,42,43,44,46,47,49,50,51,53,54,55,56
validation set	10,24,31,34,52,57
testing set	1,9,37,39,45,48

regression depend on the combination of several parameters. They are capacity parameter  $C$ ,  $\epsilon$  of  $\epsilon$ -insensitive loss function, the kernel type  $K$ , and its corresponding parameters.  $C$  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. But, ref 16 indicated that the prediction error was scarcely influenced by  $C$ . To make the learning process stable, a large value should be set up for  $C$  (e.g.,  $C = 100$ ).

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

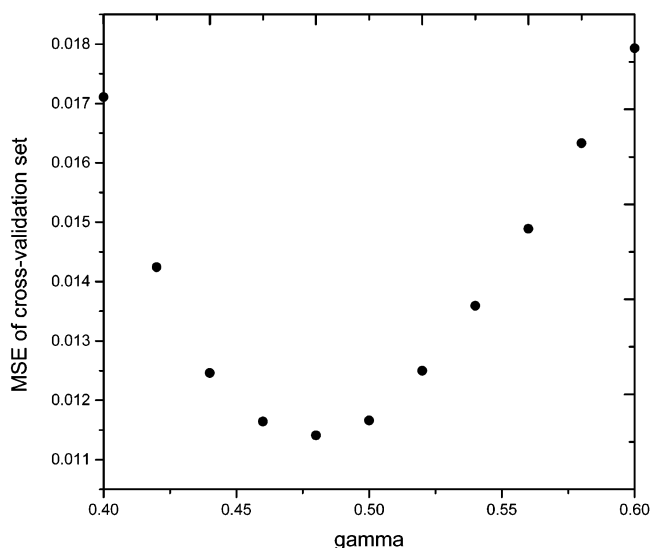
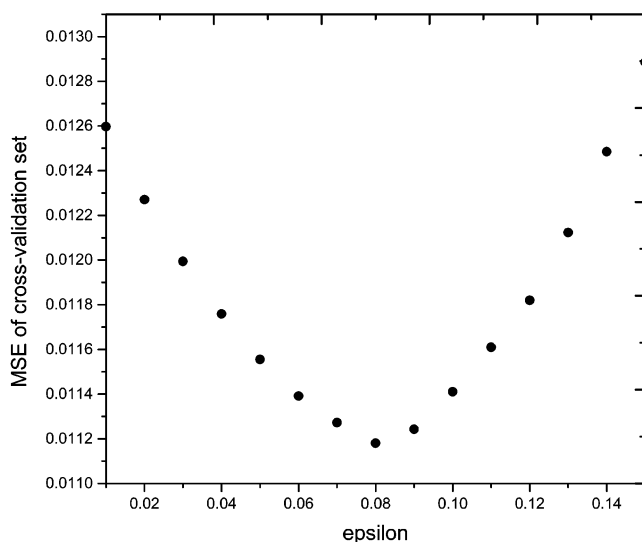
$$\exp(-\gamma*|u - v|^2)$$

where  $\gamma$  is a constant, the parameter of the kernel;  $u$  and  $v$  are two independent variables; and  $\gamma$  controls the amplitude of the Gaussian function and, therefore, controls the generalization ability of SVM. We have to optimize  $\gamma$  and find the optimal one. To optimize the  $\gamma$ , six samples, whose number was listed in Table 4, were used as a validation set. A trial and error method was used to find the best  $\gamma$ . The MSE was used as an error function, and it is computed according to the following equation

$$MSE = \frac{\sum_{i=1}^n (d_i - o_i)^2}{n}$$

where  $d_i$  are the teaching outputs (desired outputs) in the validation set,  $o_i$  are the actual outputs, and  $n$  is the number of samples in validation set. To obtain the optimal  $\gamma$ , the support vector learning machines with different  $\gamma$  were trained, the  $\gamma$  varying from 0.4 to 0.6. We calculated the MSE on different  $\gamma$ , according to the generalization ability on the validation set in order to determine the optimal one. The curve of MSE versus the gamma was shown in Figure 2. The optimal  $\gamma$  was found as 0.48.

The optimal value for  $\epsilon$  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  $\epsilon$ , there is the practical consideration of the number of

**Figure 2.** The gamma versus MSE error on validation set ( $C = 100$ ,  $\epsilon = 0.1$ ).**Figure 3.** The epsilon versus MSE error on validation set ( $C = 100$ ,  $\gamma = 0.48$ ).

resulting support vectors.  $\epsilon$ -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 of  $\epsilon$  is critical from theory.<sup>16</sup> To find an optimal  $\epsilon$ , the MSE of the cross-validation set on different  $\epsilon$  was calculated. The curve of MSE versus the epsilon was shown in Figure 3. The optimal  $\epsilon$  was found as 0.08.

**4.2.2. The Predicted Result of SVMs.** From the above discussion, the  $\gamma$ ,  $\epsilon$ , and  $C$  were fixed to 0.48, 0.08, and 100, respectively, when the support vector number of the SVM model is 43. The predicted results of the optimal SVMs are shown in Table 5 and Figure 4. The mean relative errors of the training set, the validation set, and the testing set are 1.35%, 1.52%, and 2.23% respectively, and the corresponding correlation coefficients ( $r$ ) are 0.997, 0.963, and 0.969. The mean-absolute errors are respectively 0.079, 0.093, and 0.129.

**4.3. Discussion.** For the biological activity values with high noise, it can be said that the predicted values are in

**Table 5.** Used Parameters and Predicted Results Using SVMs

no.	KHI3	IC0	YZS	MaxPCN	MinPCN	ACT	PRED	rel err (%)
1 <sup>a</sup>	2.8652	91.4321	41.1006	-0.0506	-0.0704	5.699	5.918	3.85
2	2.5737	84.6767	39.4605	-0.0506	-0.0704	5.796	5.876	-1.38
3	1.5712	62.9197	35.4404	-0.0542	-0.1012	4.523	4.603	1.76
4	3.7709	101.5036	42.8206	-0.0506	-0.0701	5.432	5.512	1.48
5	3.3747	97.6668	43.3206	-0.0506	-0.0704	5.409	5.489	1.48
6	3.3747	97.6668	46.2407	-0.0506	-0.0704	6.523	6.443	-1.23
7	4.0085	108.4169	65.9611	-0.0505	-0.0689	6.347	6.267	-1.26
8	3.3061	103.1876	55.2209	-0.0487	-0.0668	6.387	6.332	-0.86
9 <sup>a</sup>	4.2209	113.7074	60.721	-0.0485	-0.0666	6.081	6.049	-0.53
10 <sup>b</sup>	3.414	108.8357	60.901	-0.0481	-0.0938	6.229	6.050	-2.88
11	3.4485	114.1277	64.1611	-0.0408	-0.0653	6.420	6.340	-1.25
12	2.4756	74.9414	39.1605	-0.0573	-0.0734	4.886	4.966	1.64
13	2.7873	80.9377	39.1205	-0.0573	-0.0735	5.721	5.802	1.41
14	3.0105	86.5274	42.8806	-0.0573	-0.0734	6.398	6.318	-1.25
15	3.2501	100.864	43.6006	-0.0528	-0.0706	6.699	6.619	-1.19
16	3.6439	90.3642	54.4209	-0.0572	-0.073	5.921	6.001	1.35
17	3.89	96.7165	55.3609	-0.0573	-0.0733	5.420	5.500	1.48
18	2.9567	91.2368	47.1207	-0.0572	-0.0725	5.222	5.302	1.53
19	3.3513	89.2122	48.6607	-0.0571	-0.0721	6.000	6.079	1.32
20	4.2279	91.2122	50.3408	-0.0572	-0.0727	6.854	6.774	-1.17
21	4.0746	91.2122	49.5408	-0.0572	-0.0729	6.699	6.779	-1.19
22	4.5779	97.6167	59.661	-0.0572	-0.0727	7.347	7.267	-1.09
23	4.6412	96.4095	55.7809	-0.0572	-0.0725	6.284	6.204	-1.27
24 <sup>b</sup>	5.3784	100.1367	61.961	-0.0572	-0.0727	6.097	6.051	-0.75
25	4.0114	92.0313	51.3608	-0.0571	-0.0844	5.658	5.738	1.42
26	3.3016	89.3367	45.0807	-0.0573	-0.0733	5.824	5.904	1.37
27	2.9887	103.5074	42.7206	-0.0506	-0.0704	6.678	6.598	-1.20
28	2.6003	77.1909	39.6205	-0.0505	-0.0703	4.523	4.603	1.77
29	2.6353	76.4127	40.5805	-0.0515	-0.0969	4.523	4.603	1.77
30	3.0065	90.3038	45.6407	-0.0516	-0.1017	4.523	4.603	1.77
31 <sup>b</sup>	2.7492	78.2568	40.1205	-0.0523	-0.0706	5.357	5.402	0.85
32	3.6505	87.2544	46.7207	-0.052	-0.0706	7.009	6.929	-1.14
33	3.1858	89.7154	40.3005	-0.0519	-0.0783	5.194	5.274	1.53
34 <sup>b</sup>	3.7958	83.5465	49.0207	-0.0574	-0.0748	6.187	6.048	-2.25
35	3.2927	85.0245	44.6006	-0.0574	-0.075	6.347	6.427	1.26
36	3.2445	89.3367	44.2406	-0.0573	-0.0734	6.921	6.841	-1.16
37 <sup>a</sup>	2.772	75.162	41.0606	-0.0574	-0.0761	5.310	5.535	4.24
38	2.6356	74.3981	41.1206	-0.0566	-0.0753	5.959	5.878	-1.35
39 <sup>a</sup>	2.6712	68.5214	36.7405	-0.0574	-0.0751	5.957	6.048	1.50
40	3.4086	86.0092	41.7806	-0.0434	-0.075	5.000	4.920	-1.60
41	3.2716	86.2275	42.3006	-0.0423	-0.0741	5.000	5.081	1.61
42	3.3646	86.0092	42.5206	-0.0574	-0.0745	6.081	6.001	-1.32
43	3.2644	84.5529	42.0806	-0.0334	-0.0745	5.000	5.080	1.60
44	4.3096	94.6499	47.7807	-0.0574	-0.0745	5.553	5.633	1.44
45 <sup>a</sup>	4.5819	88.6667	52.0208	-0.0573	-0.0739	6.119	6.067	-0.85
46	4.3513	103.6045	53.9209	-0.0572	-0.0727	7.301	7.221	-1.09
47	3.6033	102.7745	59.221	-0.0524	-0.0714	6.367	6.286	-1.26
48 <sup>a</sup>	3.5927	107.1503	56.7209	-0.0533	-0.0748	6.347	6.193	-2.42
49	3.3531	91.807	51.7008	-0.0585	-0.0767	7.456	7.376	-1.08
50	5.6843	106.7778	65.8611	-0.0584	-0.0764	6.886	6.806	-1.16
51	4.354	98.5269	56.1809	-0.0583	-0.0844	5.770	5.849	1.38
52 <sup>b</sup>	4.1384	89.7981	51.2208	-0.0586	-0.077	6.387	6.456	1.08
53	3.5927	107.1503	58.0009	-0.0533	-0.0748	6.446	6.366	-1.24
54	4.5704	97.6167	55.5809	-0.0584	-0.0764	7.027	6.947	-1.14
55	4.4171	97.6167	57.9409	-0.0584	-0.0765	6.921	6.841	-1.16
56	4.9204	103.6045	60.361	-0.0584	-0.0764	6.456	6.376	-1.25
57 <sup>b</sup>	3.2441	85.5623	48.2207	-0.0586	-0.0772	6.201	6.119	-1.32

<sup>a</sup> The compounds in the test set. <sup>b</sup> The compounds in the validation set.

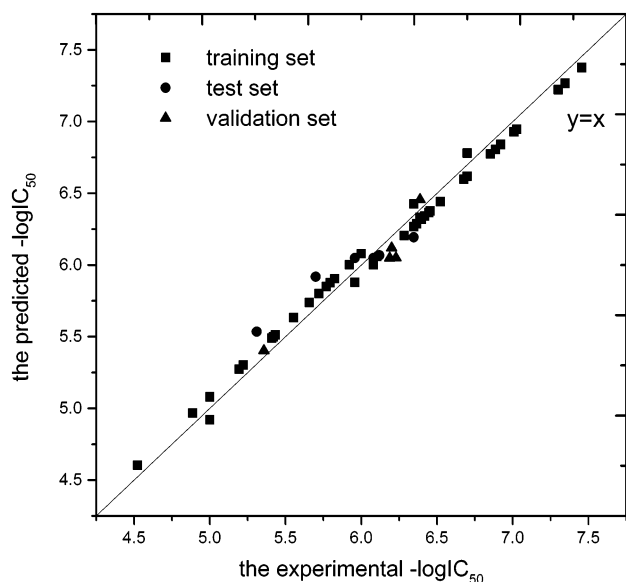
very good agreement with the experimental values from the above results. So it can be concluded that (1) the support vector machine is a very promising tool for the nonlinear approximation and (2) the selected parameters can account for the structural features of the compounds related to biological activity. The biological activity (in vitro) of drugs is mainly determined by their steric and electrostatic features. From the predicted results, the descriptors in the present model should be able to account for these features. The topological descriptor, Kier and Hall index of three order *KHI3*, which represents the size of the hydrophobic segment

and contains group contributions from all non-hydrogen atoms in the fragment, is defined as

$$KHI3 = \sum_{i=1}^N (\delta_{i1} \delta_{i2} \delta_{i3} \delta_{i4})^{-1/2}$$

where

$$\delta_i = \frac{Z_i^v - H_i}{Z_i - Z_i^v - 1}$$



**Figure 4.** The predicted values of  $-\log IC_{50}$  results versus the experimental data.

Here  $Z_i$  is the total number of electrons in the  $i$ th atom,  $Z_i^v$  is the number of valence electrons, and  $H_i$  is the number of hydrogen directly attached to the  $i$ th atom. Valence contributions are summed for all atoms in the fragment, with the exception of the hydrogen atoms ( $N = N_{total} - N_H$ ). The second topological descriptor, Information Content Index  $IC_0$ , is equal to the average information content, which is defined on the basis of the Shannon information theory, multiplied by the total number of atoms. The average information content is calculated as follows

$${}^0IC = -\sum_i \frac{n_i}{n} \log_2 \frac{n_i}{n}$$

where  $n_i$  is a number of atoms in the  $i$ th class and  $n$  is a total number of atoms in the molecule. The division of atoms into different classes depends on the coordination sphere taken into account. The geometrical descriptor YZ Shadow YZS is the area of the shadows of the molecule as projected on the YZ planes by the orientation of the molecule in the space along the axes of inertia. The above three descriptors describe the size and branching information of molecules and give some information about steric structure comprehensively. The maximum partial charge of N atom and the minimum partial charge of N atom represent the partial charge of the two N atoms nearest to the skeleton of the group R1 except for the molecule 17. Thus, they describe the electronegativity of the R1 group and state the electrostatic features of the compounds from some aspect.

Analysis of the results obtained indicates that the model we proposed correctly represents structural-activity relationships of these compounds and that molecular descriptors calculated solely from structures can describe the structural features of the compounds responsible for their biological activity.

## 5. CONCLUSION

This study of QSAR model shows that the SVMs is a very promising tool for the nonlinear approximation. The training

and optimization are easier and faster compared with other machines learning techniques, because there are fewer free parameters and only support vectors (only a fraction of all data) 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 of converging to the global optimum and not to a local optimum. The predictive results are consistent with the experimental data. The mean relative error is 1.46%. Therefore it is a good approach for predicting the expected activity of drugs and aiding drug design. At the same time, the models proposed could identify and provide some insight into what structural features are related to the biological activity of these compounds and afford some instruction for further designing the new compounds of inhibiting AP-1 and NF- $\kappa$ B mediated gene expression.

## ACKNOWLEDGMENT

The authors thank the Association Franco-Chinoise pour la Recherche Scientifique & Technique (AFCRST) for supporting this study (Program PRA SI 00-05). The authors also thank the R Development Core Team for affording the free R1.5.1 software.

**Supporting Information Available:** A full list of 74 calculated descriptors and their chemical meaning. This material is available free of charge via the Internet at <http://pubs.acs.org>.

## REFERENCES AND NOTES

- (1) Manning, A. M. Transcription factors: a new frontier for drug discovery. *Drug Discovery Today* **1996**, *1*, 151–160.
- (2) Palanki, M. S. S. Inhibitors of AP-1 and NF- $\kappa$ B Mediated Transcriptional Activation: Therapeutic Potential in Autoimmune Diseases and Structural Diversity. *Curr. Med. Chem.* **2002**, *9*, 219–227.
- (3) Palanki, M. S. S.; Gayo-Fung, L. M.; Shevlin, G. I.; Erdman, P.; Sato, M.; Goldman, M.; Ransone, L. J.; Spooner C. Structure–Activity Relationship Studies of Ethyl 2-[(3-Methyl-2,5-dioxo(3-pyrrolyl)) amino]-4-(trifluoromethyl) pyrimidine-5-carboxylate: An Inhibitor of AP-1 and NF- $\kappa$ B Mediated Gene Expression. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 2573–2577.
- (4) Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. *Comput. Chem.* **2002**, *26*, 5–14.
- (5) 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.
- (6) Bao, L.; Sun, Z. R. Identifying genes related to drug anticancer mechanisms using support vector machine. *FEBS Lett.* **2002**, in press.
- (7) Belousov, A. I.; Verzakov, S. A.; Von Frese, J. A flexible classification approach with optimal generalization performance: support vector machines. *Chemometrics Intelligent Lab. Systems* **2002**, *64*, 15–25.
- (8) Cai, Y. D.; Liu, X. J.; Xu, X. B.; Chou, K. C. Prediction of protein structural classes by support vector machines. *Comput. Chem.* **2002**, *26*, 293–296.
- (9) Colin, W. M.; Autret, A.; Boddy, L. Support vector machines for identifying organisms- a comparison with strongly partitioned radial basis function networks. *Ecological Modelling* **146** 2001 57–67.
- (10) Song, M.; Breneman, C. M.; Bi, J.; Sukumar, N.; Bennett, K. P.; Cramer, S.; Tugcu, N. Prediction of Protein Retention Times in Anion-Exchange Chromatography Systems Using Support Vector Regression. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1347–1357.
- (11) 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**, in press.
- (12) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. *CODESSA: Training Manual*; University of Florida, Gainesville, FL, 1995.
- (13) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. *CODESSA: Reference Manual*; University of Florida, Gainesville, FL, 1994.
- (14) Cortes, C.; Vapnik, V. Support-Vector Networks. *Machine Learning* **1995**, *20*, 273–297.



- (15) Gunn, S. R.; Brown, M.; Bossley, K. M. Network performance assessment for neurofuzzy data modeling. *Lecture Notes Comput. Sci.* **1997**, 1280, 313–323.
- (16) Wang, W. J.; Xu, Z. B.; Lu, W. Z.; Zhang, X. Y. Determination of the spread parameter in the Gaussian kernel for classification and regression. *Neurocomputing*, in press.
- (17) Tay, F. E. H.; Cao, L. J. Modified support vector machines in financial time series forecasting. *Neurocomputing* **2002**, 48, 847–861.
- (18) Burges, C. J. C. A tutorial on support vector machines for pattern recognition. *Data Mining Knowledge Discovery* **2** **1998**, 2, 1–47.
- (19) Smola, A. J.; Schölkopf, B. *A tutorial on support vector regression*; NeuroCOL2 Technical report series, NC2-TR-1998-030; October, 1998.

CI0340355