Comparative Study of QSAR/QSPR Correlations Using Support Vector Machines, Radial Basis Function Neural Networks, and Multiple Linear Regression

X. J. Yao,[†,‡] A. Panaye,[†] J. P. Doucet,[†] R. S. Zhang,[‡] H. F. Chen,[†] M. C. Liu,[‡] Z. D. Hu,[‡] and B. T. Fan*[,†]

Université Paris 7-Denis Diderot, ITODYS-CNRS UMR 7086, 1, Rue Guy de la Brosse, 75005 Paris, France, and Department of Chemistry, Lanzhou University, Lanzhou 730000, China

Received January 16, 2004

Support vector machines (SVMs) were used to develop QSAR models that correlate molecular structures to their toxicity and bioactivities. The performance and predictive ability of SVM are investigated and compared with other methods such as multiple linear regression and radial basis function neural network methods. In the present study, two different data sets were evaluated. The first one involves an application of SVM to the development of a QSAR model for the prediction of toxicities of 153 phenols, and the second investigation deals with the QSAR model between the structures and the activities of a set of 85 cyclooxygenase 2 (COX-2) inhibitors. For each application, the molecular structures were described using either the physicochemical parameters or molecular descriptors. In both studied cases, the predictive ability of the SVM model is comparable or superior to those obtained by MLR and RBFNN. The results indicate that SVM can be used as an alternative powerful modeling tool for QSAR studies.

INTRODUCTION

Quantitative structure property/activity relationship (QSPR/QSAR) represents an attempt to correlate structural descriptors of compounds with their physicochemical properties and biological activities. It is now widely used for the prediction of physicochemical properties and biological activities in chemical, environmental, and pharmaceutical areas.[1,2] The main steps involved in this method include the following: data collection, molecular descriptor selection and obtaining, correlation model development, and finally model evaluation. The main problems encountered in this kind of research are still the description of the molecular structure using appropriate molecular descriptors and selection of suitable modeling methods. At present, many types of molecular descriptors such as topological indices and quantum chemical parameters have been proposed to describe the structural features of molecules.[3−5] Many different chemometrics methods, such as multiple linear regression (MLR), partial least squares regression (PLS), different types of artificial neural networks (ANN), genetic algorithms (GAs), and support vector machine (SVM) can be employed to derive correlation models between the molecular structures and properties.

ANNs are useful tools in QSAR/QSPR studies, and particularly in cases where it is difficult to specify an exact mathematical model for describing a given structure−property relationship. Most of these works used neural networks based on the back-propagation learning algorithm, which has some disadvantages such as local minimum, slow convergence, time-consuming nonlinear iterative optimization, difficulty in explicit optimum network configuration, etc. In contrast, the parameters of radial basis function neural networks (RBFNNs) can be adjusted by fast linear methods. It has advantages of short training times and is guaranteed to reach the global minimum of error surface during training. The optimization of its topology and learning parameters are easy to be implemented.[6]

As a new and powerful modeling tool, support vector machine (SVM) has gained much interest in pattern recognition and function approximation applications recently. In bioinformatics, SVMs have been successfully used to solve classification and correlation problems, such as cancer diagnosis,[7−10] identification of HIV protease cleavage sites,[11] protein class prediction,[12] etc. SVMs have also been applied in chemistry, for example, the prediction of retention index of protein, and other QSAR studies.[13−21] Compared with traditional regression and neural networks methods, SVMs have some advantages, including global optimum, good generalization ability, simple implementation, few free parameters, and dimensional independence.[22−24] The flexibility in classification and ability to approximate continuous function make SVMs very suitable for QSAR and QSPR studies.

In the present paper, we present the applications of support vector regression (SVR) for correlation problems in QSAR and compare its performance with MLR and RBFNN methods. Two data sets were selected for this study, the toxicities of 153 phenols and the activities of 85 cyclooxygenase 2 (COX-2) inhibitors.

MATERIALS AND METHODS

Radial Basis Function Neural Networks. RBFNNs can be described as a three-layer feed-forward structure. RBFNNs consist of three layers: input layer, hidden layer, and output layer. The input layer does not process the information; it

* Corresponding author e-mail: fan@paris7.jussieu.fr.
[†] Université Paris 7-Denis Diderot.
[‡] Lanzhou University.

10.1021/ci049965i CCC: $27.50 © 2004 American Chemical Society
Published on Web 04/23/2004

only distributes the input vectors to the hidden layer. The hidden layer of RBFNNs consists of a number of RBF units (n_h) and bias (b_k). Each hidden layer unit represents a single radial basis function, with associated center position and width. Each neuron on the hidden layer employs a radial basis function as a nonlinear transfer function to operate on the input data. The most often used RBF is a Gaussian function that is characterized by a center (c_j) and a width (r_j). The RBF functions by measuring the Euclidean distance between the input vector ($\mathbf{x}$) and the radial basis function center (c_j) and performs the nonlinear transformation with RBF in the hidden layer as given below

$$h_j(\mathbf{x}) = \exp(-\,||\mathbf{x} - c_j||^2/r_j^2) \tag{1}$$

in which h_j is the notation for the output of the jth RBF unit. For the jth RBF c_j and r_j are the center and the width, respectively. The operation of the output layer is linear, which is given in eq 2

$$y_k(\mathbf{x}) = \sum_{j=1}^{n_h} w_{kj} h_j(\boldsymbol{x}) + b_k \tag{2}$$

where y_k is the kth output unit for the input vector $\mathbf{x}$, w_{kj} is the weight connection between the kth output unit and the jth hidden layer unit, and b_k is the bias.

From eqs 1 and 2, one can see that designing a RBFNN involves selecting centers, number of hidden layer units, width, and weights. There are various ways for selecting the centers, such as random subset selection, ***K***-means clustering, orthogonal least squares learning algorithm, RBF−PLS, etc. The widths of the radial basis function networks can either be chosen the same for all the units or can be chosen differently for each units. In this paper, considerations were limited to the Gaussian functions with a constant width, which was the same for all units. A forward subset selection routine[25,26] was used to select the centers from training set samples. The adjustment of the connection weight between hidden layer and output layer is performed using a least-squares solution after the selection of centers and width of radial basis functions.

The overall performance of RBFNs is evaluated in terms of a root-mean-squared error (RMS) according to the equation below

$$\mathrm{RMS} = \sqrt{\frac{\sum_{i=1}^{n_s} (y_k - \hat{y}_k)^2}{n_s}} \tag{3}$$

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

The performance of RBFNN is determined by the values of following parameters:

- The number n_h of radial basis functions
- The center c_j and the width r_j of each radial basis function
- The connection weight w_{kj} between the jth hidden layer unit and the kth output unit.

The centers of RBFNN are determined with the forward subset selection method proposed by Orr.[25,26] The advantages of this method over others are simultaneous determination of the centers and the number of hidden layer units and without the need to fix the number of hidden layer units in advance.

The optimal width was determined by experiments with a number of trials by taking into account the leave-one-out (LOO) cross-validation error. The one which gives a minimum LOO cross-validation error is chosen as the optimal value. After the selection of the centers and number of hidden layer units, the connection weights can be easily calculated by linear least-squares methods.

SUPPORT VECTOR REGRESSION METHOD

SVM is a new and very promising classification and regression method developed by Vapnik et al.[22] A detailed description of the theory of SVM can be referred in several excellent books and tutorials.[23,24] SVMs are originally developed for classification problems; they can also be extended to solve nonlinear regression problems by the introduction of ϵ-insensitive loss function. In support vector regression, the input $\mathbf{x}$ is first mapped into a higher dimensional feature space by the use of a kernel function, and then a linear model is constructed in this feature space. The kernel functions often used in SVM include linear, polynomial, radial basis function, and sigmoid function. The linear model $f(\mathbf{x},\omega)$ in the feature space is given by

$$f(\mathbf{x},\omega) = \sum_{j=1}^{m} \omega_j g_j(\mathbf{x}) + b \tag{4}$$

where $g_j(\mathbf{x})$, $j = 1,..., m$ represents a set of nonlinear transformations, and b is the “bias” term.

The quality of estimation is measured by the loss functionL$(y,f(\mathbf{x},\omega))$. SVM regression uses a new type of loss function called ϵ-insensitive loss function proposed by Vapnik:[22]

$$L_\epsilon(y,f(\mathbf{x},\omega)) = \begin{cases} 0 & \textit{if } |\mathrm{y} - f(\mathbf{x},\omega)| \le \epsilon \\ |y - f(\mathbf{x},\omega)| - \epsilon & \text{otherwise} \end{cases} \tag{5}$$

The empirical risk is

$$R_{\mathrm{emp}}(\omega) = \frac{1}{n}\sum_{i=1}^{n} L_\epsilon(y_i,f(\mathbf{x}_i,\omega)) \tag{6}$$

SVM regression performs linear regression in the high-dimension feature space using ϵ-insensitive loss and, at the same time, tries to reduce model complexity by minimizing $||\omega||^2$. This can be described by introducing (non-negative) slack variables ξ_i,ξ_i^* $i = 1,...n$, to measure the deviation of training samples outside the ϵ-insensitive zone. Thus SVM regression is formulated as a minimization of the following functional:

$$\min \frac{1}{2}||\omega||^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$

$$\text{s.t.} \begin{cases} y_i - f(\mathbf{x}_i,\omega) \le \epsilon + \xi*_i \\ f(\mathbf{x}_i,\omega) - y_i \le \epsilon + \xi_i \\ \xi_i,\xi_i^* \ge 0, i = 1,...,n \end{cases} \tag{7}$$

This optimization problem can be transformed into a quadratic programming problem,[22] and its solution is given by

$$f(\mathbf{x}) = \sum_{i=1}^{n_{SV}} (\alpha_i - \alpha_i^*) K(\mathbf{x}_i, \mathbf{x}) \quad \text{s.t. } 0 \le \alpha_i^* \le C, 0 \le \alpha_i \le C \qquad (8)$$

where n_{SV} is the number of Support Vectors (SVs) and the kernel function

$$K(\mathbf{x}, x_i) = \sum_{j=1}^{m} g_j(\mathbf{x}) g_j(\mathbf{x}_i) \qquad (9)$$

The generalization performance of SVR depends on a good setting of parameters: C, ϵ and the kernel type and corresponding kernel parameters. The selection of the kernel function and corresponding parameters is very important because they define the distribution of the training set samples in the high dimensional feature space. Parameter C is a regularization constant which determines the trade-off between the model complexity and the degree to which deviations larger than ϵ are tolerated in an optimization formulation.

Similar with other multivariate statistical models, the performance of SVM for regression depends on the combination of several factors. They are kernel function type, capacity parameter C, ϵ of ϵ-insensitive loss function, and its corresponding parameters. To get the best generalization ability, some strategies are needed for optimizing these factors. The selection of the kernel function and corresponding parameters is very important because they implicitly define the distribution of the training set samples in the high-dimensional feature space and also the linear model constructed in the feature space. There are four possible choices of kernel functions available in the LibSVM package i.e., linear, polynomial, radial basis function, and sigmoid function. For regression tasks, the radial basis function kernel is often used because of its effectiveness and speed in training process. It was also used for all SVR models in our study. For the RBF kernel, the most important parameter is the width γ of the radial basis function. C is a regularization parameter that controls the trade-off 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. 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. The value of ϵ can affect the number of support vectors used to construct the regression function. The bigger ϵ, the fewer support vectors are selected. To select the proper values for the regulation parameter C, width and ϵ different values for these parameters have been tried; the set of values with the best leave-one-out cross-validation performance is selected for further analysis.

All calculation programs implementing RBFNNs were written in an M-file based on the MATLAB script for radial basis function neural networks.[25,26] All SVM models in our present study were implemented using the software LibSVM that is an efficient software for classification and regression developed by Chih-Chung Chang and Chih-Jen Lin.[27,28]

DATA SETS

Data Set 1. This data set is extracted from a recent work reported by Aptula et al.[29] The data set includes 221 phenols, for which toxicity data to the ciliate *Tetrahymena pyriformis* are available. In our QSAR study, we only use the 153 compounds grouped into polar narcotics. The molecular descriptors used include hydrophobicity (log K_{ow}), acidity constant (pK_a), frontier orbital energies (E_{homo} and E_{lumo}), and hydrogen bond donor/acceptor counts (N_{hdon}). The compounds and their corresponding toxicity are shown in Table 1. Three different modeling methods, i.e., MLR, RBFNNs, and SVM, were used to develop the correlation models.

Data Set 2. This data set is taken from a recent review contributed by Hansch et al.[30] The data set includes 85 COX-2 inhibitors with their activity IC_{50} values. The structures of the compounds and their corresponding activities are listed in Table 2. Two different sets of molecular descriptors were used to describe the structures of these compounds: one is from the cited review, and another is from CODESSA analysis.[31,32] The calculation process of CODESSA is described as below: molecules were drawn into Hyperchem[33] and then preoptimized using MM+ molecular mechanics force field. A more precise optimization is done with the semiempirical AM1 method in MOPAC6.0.[34] All calculations are carried out at a restricted Hartree−Fock level with no configuration interaction. The molecular structures were optimized using the Polak-Ribiere algorithm until the root-mean-square gradient reached 0.01. The resulting geometry was transferred into software CODESSA that can calculate constitutional, topological, electrostatic, and quantum chemical descriptors. After the calculation of molecular descriptors, the best multilinear regression method in CODESSA is used to select the structural descriptors that are correlated with the biological activity. Three different types of modeling methods, i.e., MLR, RBFNN, and SVM, were used to the QSAR study. This data set is also used to stress the importance of the choice of structural descriptors.

RESULTS AND DISCUSSION

Data Set 1. To compare the performance of MLR, RBFNN and SVM, we first used leave-one-out (LOO) cross-validation based on all the compounds to all these three modeling methods. The detailed description of the linear model is listed in Table 3. For RBFNN, the most important parameter that influences its performance is the width. For this data set, the optimal value for the width was determined as 0.937. For the SVM model with the RBF kernel function, there are three parameters, ϵ, γ, and C, to be determined. For this data set, the γ, ϵ, and C for this data set were fixed to 4.5, 0.1, and 6, respectively.

The comparison of the leave-one-out cross-validation results obtained with SVM, RBFNN, and MLR is summarized in Table 4. It is very clear from Table 4 that SVM and RBFNN models show similar correlation performance and outperform the MLR model. The performance of SVM is a bit better than that obtained by RBFNN.

Table 1. Observed and Calculated Toxicity Values by MLR, RBF, and SVM

no.	name	toxicity	MLR	SVM	RBFNN
1	1,3,5-trihydroxybenzene	−1.26	−1.11	−1.16	−1.25
2	2-(tert)-butyl-4-methylphenol	1.30	1.22	0.97	1.20
3[a]	2,3,5-trichlorophenol	2.37	1.46	1.67	1.83
4	2,3,5-trimethylphenol	0.36	0.65	0.56	0.54
5	2,3,6-trimethylphenol	0.28	0.67	0.55	0.57
6	2,3-dichlorophenol	1.28	0.81	1.01	1.09
7	2,3-dimethylphenol	0.12	0.31	0.22	0.19
8	2,4,5-trichlorophenol	2.10	1.48	1.66	1.68
9	2,4,6-tribromophenol	2.03	1.72	1.93	2.05
10[a]	2,4,6-tribromoresorcinol	1.06	2.06	2.10	1.32
11	2,4,6-trichlorophenol	1.41	1.27	1.55	1.62
12	2,4,6-trimethylphenol	0.28	0.69	0.54	0.55
13	2,4,6-tris(dimethylaminomethyl)phenol	−0.52	−0.80	−0.73	−0.52
14	2,4-dibromophenol	1.40	1.18	1.35	1.44
15	2,4-dichlorophenol	1.04	0.89	1.07	1.10
16	2,4-difluorophenol	0.60	0.34	0.33	0.36
17[a]	2,4-dimethylphenol	0.07	0.32	0.22	0.17
18	2,5-dichlorophenol	1.13	0.93	1.12	1.18
19	2,5-dimethylphenol	0.08	0.34	0.29	0.19
20	2,6-di-(tert)-butyl-4-methylphenol	1.80	2.46	1.90	1.75
21	2,6-dichloro-4-fluorophenol	0.80	0.97	1.12	1.15
22	2,6-dichlorophenol	0.74	0.65	0.90	0.96
23	2,6-difluorophenol	0.47	0.14	0.23	0.35
24[a]	2,6-dimethoxyphenol	−0.60	−0.54	−0.58	−0.46
25	2-allylphenol	0.33	0.38	0.33	0.30
26	2-bromo-4-methylphenol	0.60	0.73	0.87	0.80
27	2-bromophenol	0.33	0.44	0.46	0.48
28	2-chloro-4,5-dimethylphenol	0.69	0.88	1.02	0.91
29	2-chloro-5-methylphenol	0.39	0.56	0.72	0.60
30	2-chlorophenol	0.18	0.22	0.36	0.28
31[a]	2-cyanophenol	0.03	0.07	0.15	0.30
32	2-ethoxyphenol	−0.36	−0.08	−0.13	−0.21
33	2-ethylphenol	0.16	0.32	0.29	0.22
34	2-fluorophenol	0.19	−0.03	0.06	0.12
35	2-hydroxy-4,5-dimethylacetophenone	0.71	0.63	0.56	0.50
36	2-hydroxy-4-methoxyacetophenone	0.55	0.38	0.23	0.27
37	2-hydroxy-4-methoxybenzophenone	1.42	1.63	1.52	1.22
38[a]	2-hydroxy-5-methylacetophenone	0.31	0.74	0.51	0.46
39	2-hydroxyacetophenone	0.08	0.39	0.18	0.18
40	2-hydroxybenzyl alcohol	−0.95	−0.88	−0.85	−0.90
41	2-hydroxyethylsalicylate	−0.08	0.34	0.24	0.04
42	2-isopropylphenol	0.80	0.58	0.52	0.51
43	2-methoxy-4-propenylphenol	0.75	0.75	0.74	0.49
44	2-methoxyphenol	−0.51	−0.42	−0.46	−0.40
45[a]	2-phenylphenol	1.09	1.02	1.06	0.99
46	2-(tert)-butylphenol	1.30	0.88	0.69	0.86
47	3,4,5-trimethylphenol	0.93	0.59	0.54	0.45
48	3,4-dichlorophenol	1.75	1.06	1.17	1.24
49	3,4-dimethylphenol	0.12	0.28	0.22	0.12
50	3,5-dibromosalicylaldehyde	1.64	1.56	1.58	1.79
51	3,5-dichlorophenol	1.57	1.13	1.26	1.38
52[a]	3,5-dichlorosalicylaldehyde	1.55	1.32	1.34	1.45
53	3,5-diiododsalicylaldehyde	2.34	1.83	1.88	2.13
54	3,5-dimethoxyphenol	−0.09	−0.31	−0.24	−0.32
55	3,5-dimethylphenol	0.11	0.31	0.28	0.20
56	3,5-di-(tert)-butylphenol	1.64	2.00	1.92	2.05
57	3-acetamidophenol	−0.16	−0.78	−0.59	−0.49
58	3-bromophenol	1.15	0.62	0.68	0.72
59[a]	3-chloro-4-fluorophenol	1.13	0.83	0.84	0.86
60	3-chloro-5-methoxyphenol	0.76	0.48	0.59	0.55
61	3-chlorophenol	0.87	0.47	0.54	0.55
62	3-cyanophenol	−0.06	0.20	0.05	0.06
63	3-ethoxy-4-hydroxybenzaldehyde	0.02	0.32	0.36	0.32
64	3-ethoxy-4-methoxyphenol	−0.30	−0.06	−0.10	−0.22
65	3-ethylphenol	0.23	0.31	0.30	0.22
66[a]	3-fluorophenol	0.38	0.10	0.14	0.23
67	3-hydroxy-4-methoxybenzyl alcohol	−0.99	−1.00	−0.84	−0.69
68	3-hydroxyacetophenone	−0.38	0.14	−0.06	−0.08
69	3-hydroxybenzaldehyde	0.09	0.16	−0.06	−0.14
70	3-hydroxybenzoicacid	−0.81	0.06	−0.69	−0.83
71	3-hydroxybenzyl alcohol	−1.04	−0.83	−0.82	−0.84
72	3-iodophenol	1.12	0.77	0.87	0.91
73[a]	3-isopropylphenol	0.61	0.56	0.54	0.48
74	3-methoxyphenol	−0.33	−0.26	−0.25	−0.31
75	3-phenylphenol	1.35	1.16	1.16	1.22
76	3-(tert)-butylphenol	0.73	0.81	0.80	0.79
77	4-(tert)-octylphenol	2.10	2.01	2.00	2.05
78	4-(tert)-butylphenol	0.91	0.81	0.75	0.76
79	4,6-dichlororesorcinol	0.97	0.40	0.44	1.04
80[a]	4-allyl-2-methoxyphenol	0.42	0.38	0.32	0.24
81	4-benzyloxyphenol	1.04	1.06	0.96	0.97

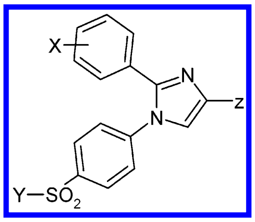
Table 1 (Continued)

no.	name	toxicity	MLR	SVM	RBFNN
82	4-bromo-2,6-dichlorophenol	1.78	1.38	1.65	1.76
83	4-bromo-2,6-dimethylphenol	1.17	1.32	1.29	1.53
84	4-bromo-3,5-dimethylphenol	1.27	1.23	1.26	1.44
85	4-bromo-6-chloro-2-cresol	1.28	1.35	1.53	1.63
86	4-bromophenol	0.68	0.61	0.65	0.65
87[a]	4-butoxyphenol	0.70	0.87	0.77	0.73
88	4-chloro-2-isopropyl-5-methylphenol	1.85	1.75	1.75	2.14
89	4-chloro-2-methylphenol	0.70	0.80	0.82	0.83
90	4-chloro-3,5-dimethylphenol	1.20	1.12	1.16	1.27
91	4-chloro-3-ethylphenol	1.08	1.12	1.18	1.29
92	4-chloro-3-methylphenol	0.80	0.78	0.83	0.82
93	4-chlorophenol	0.55	0.47	0.51	0.46
94[a]	4-chlororesorcinol	0.13	−0.05	0.00	0.39
95	4-cyanophenol	0.52	0.10	0.10	0.21
96	4-ethoxyphenol	0.01	0.18	0.08	−0.03
97	4-ethylphenol	0.21	0.30	0.29	0.17
98	4-fluorophenol	0.02	0.14	0.09	0.12
99	4-heptyloxyphenol	2.03	1.91	1.80	1.91
100	4-hexyloxyphenol	1.64	1.56	1.45	1.58
101[a]	4-hexylresorcinol	1.80	1.06	1.38	1.16
102	4-hydroxy-2-methylacetophenone	0.19	0.43	0.19	0.16
103	4-hydroxy-3-methoxyacetophenone	−0.12	0.06	−0.18	−0.07
104	4-hydroxy-3-methoxybenzonitrile	−0.03	0.08	0.07	0.11
105	4-hydroxy-3-methoxybenzyl alcohol	−0.70	−0.92	−0.84	−0.77
106	4-hydroxy-3-methoxybenzylamine	−0.97	−0.95	−0.87	−0.68
107	4-hydroxy-3-methoxyphenethyl alcohol	−0.18	−0.85	−0.68	−0.63
108[a]	4-hydroxyacetophenone	−0.30	0.02	0.01	0.11
109	4-hydroxybenzaldehyde	0.27	0.02	0.02	0.12
110	4-hydroxybenzamide	−0.78	−0.70	−0.85	−0.86
111	4-hydroxybenzoic	−1.02	0.02	−0.67	−0.85
112	4-hydroxybenzophenone	1.02	1.19	1.12	1.05
113	4-hydroxybenzylcyanide	−0.38	−0.56	−0.57	−0.17
114	4-hydroxyphenethyl alcohol	−0.83	−0.86	−0.79	−0.84
115[a]	4-hydroxyphenylacetic	−1.50	−0.69	−1.42	−1.07
116	4-hydroxypropiophenone	0.05	0.39	0.32	0.33
117	4-iodophenol	0.85	0.76	0.83	0.86
118	4-isopropylphenol	0.47	0.57	0.52	0.47
119	4-methoxyphenol	−0.14	−0.17	−0.24	−0.29
120	4-phenylphenol	1.39	1.14	1.19	1.08
121	4-propylphenol	0.64	0.66	0.61	0.58
122[a]	4-(sec)-butylphenol	0.98	0.91	0.85	0.89
123	4-(tert)-pentylphenol	1.23	1.16	1.11	1.22
124	5-bromo-2-hydroxybenzyl alcohol	0.34	0.07	0.10	0.19
125	5-bromovanillin	0.62	0.46	0.52	0.45
126	5-fluoro-2-hydroxyacetophenone	0.04	0.76	0.39	0.00
127	5-methylresorcinol	−0.39	−0.33	−0.21	−0.28
128	5-pentylresorcinol	1.31	1.02	1.28	1.14
129[a]	6-(tert)-butyl-2,4-dimethylphenol	1.16	1.62	1.13	1.47
130	α,α,α-trifluoro-4-cresol	0.62	0.89	0.89	0.98
131	ethyl-3-hydroxybenzoate	0.48	0.80	0.68	0.60
132	ethyl-4-hydroxy-3-methoxyphenylacetate	−0.23	−0.05	−0.13	−0.11
133	ethyl-4-hydroxybenzoate	0.57	0.74	0.67	0.68
134	isovanillin	−0.14	0.07	−0.09	−0.08
135	3-cresol	−0.06	−0.02	−0.05	−0.08
136[a]	methyl-3-hydroxybenzoate	−0.05	0.49	0.30	0.20
137	methyl-4-hydroxybenzoate	0.08	0.41	0.29	0.29
138	methyl-4-methoxysalicylate	0.62	0.71	0.64	0.65
139	nonylphenol	2.47	2.73	2.60	2.29
140	2-cresol	−0.30	−0.01	−0.06	−0.09
141	2-vanillin	0.38	0.23	0.24	0.22
142	4-cresol	−0.18	−0.02	−0.08	−0.12
143[a]	4-cyclopentylphenol	1.29	0.98	0.98	0.98
144	phenol	−0.21	−0.37	−0.36	−0.23
145	resorscinol	−0.65	−0.69	−0.61	−0.65
146	salicylaldehyde	0.42	0.29	0.23	0.27
147	salicylaldoxime	0.25	−0.14	0.01	0.41
148	salicylamide	−0.24	−0.17	−0.34	−0.47
149	salicylhydrazide	0.18	−0.22	−0.00	0.19
150[a]	salicylhydroxamic acid	0.38	−0.20	−0.33	0.09
151	salicylicacid	−0.51	0.34	−0.50	−0.61
152	syringaldehyde	0.17	−0.14	−0.17	0.01
153	vanillin	−0.03	−0.01	−0.00	0.07

[a] Test set compounds.

To further compare the performance of the different methods, the compounds were divided into a training set (131 compounds) and a test set (22 compounds). The training set was used to adjust the parameters of the models. The test set was used to evaluate and compare the performance of different methods. A detailed description of the linear model based on compounds in the training set is summarized in Table 5. The predicted versus experimental toxicity based

Table 2. IC50 Activities of Imidazole Derivatives



	substituent					substituent			
no.	X	Y	Z	Log(1/IC50)	no.	X	Y	Z	Log(1/IC50)
1	4-Cl	Me	CF_3	6.96	44	3-Me-4-F	Me	CF_3	6.77
2	4-F	Me	CF_3	7.00	45	3-Me-4-Cl	Me	CF_3	7.05
3[a]	H	Me	CF_3	6.92	46[a]	3-OMe-4-Cl	Me	CF_3	6.60
4	4-Me	Me	CF_3	6.80	47	3-NMe_2−4-Cl	Me	CF_3	5.98
5	4-OMe	Me	CF_3	6.24	48	3,4-OCH_2O	Me	CF_3	6.77
6	4-NHMe	Me	CF_3	5.83	49	3,4-F_2	Me	CF_3	6.92
7	4-NMe_2	Me	CF_3	6.16	50	3,4-Me_2	Me	CF_3	6.48
8	4-SMe	Me	CF_3	6.80	51	3-Me-5-Cl	Me	CF_3	7.10
9	4-SO_2Me	Me	CF_3	5.24	52	3-Me-5-F	Me	CF_3	6.96
10[a]	4-Cl	NH_2	CF_3	8.00	53[a]	3-OMe-5-Cl	Me	CF_3	6.02
11	4-F	NH_2	CF_3	8.00	54	3,5-Cl_2	Me	CF_3	6.77
12	H	NH_2	CF_3	7.40	55	3-F-4-OMe	NH_2	CF_3	7.52
13	4-Me	NH_2	CF_3	7.40	56	3-Cl-4-OMe	NH_2	CF_3	7.70
14	3-Cl	Me	CF_3	7.22	57	3-Br-4-OMe	NH_2	CF_3	7.52
15	3-F	Me	CF_3	6.92	58	3-Cl-4-SMe	NH_2	CF_3	8.00
16	3-Br	Me	CF_3	7.10	59	3-Cl-4-Me	NH_2	CF_3	8.52
17[a]	3-Me	Me	CF_3	7.22	60[a]	3-OMe-4-Cl	NH_2	CF_3	7.70
18	3-CF_3	Me	CF_3	6.68	61	3,4-F_2	NH_2	CF_3	7.52
19	3-OMe	Me	CF_3	6.46	62	3-Me-5-Cl	NH_2	CF_3	7.40
20	3-SMe	Me	CF_3	6.46	63	3-Me-5-F	NH_2	CF_3	7.52
21	3-CH_2OMe	Me	CF_3	4.17	64	3-OMe-5-F	NH_2	CF_3	6.34
22	3-NMe_2	Me	CF_3	5.50	65	3,5-F_2−4-OMe	Me	CF_3	6.77
23	3-NHMe	Me	CF_3	6.04	66	3,5-Cl_2−4-OMe	Me	CF_3	6.85
24	3-NH_2	Me	CF_3	5.23	67	3,5-Br_2−4-OMe	Me	CF_3	7.05
25[a]	3-NO_2	Me	CF_3	6.24	68[a]	3,5-Me_2−4-OMe	Me	CF_3	6.14
26	3-Cl	NH_2	CF_3	8.10	69	2,5-Me_2−4-OMe	Me	CF_3	4.91
27	3-F	NH_2	CF_3	7.52	70	3,5-Cl_2−4-NMe_2	Me	CF_3	6.85
28	3-Br	NH_2	CF_3	8.16	71	3,5-F_2−4-OMe	NH_2	CF_3	7.52
29	3-Me	NH_2	CF_3	7.52	72	4-Cl	Me	Me	6.62
30	2-Cl	Me	CF_3	6.05	73	4-Cl	Me	CF_3	6.96
31	2-F	Me	CF_3	6.40	74	4-Cl	Me	CHF_2	6.22
32[a]	2-Me	Me	CF_3	6.10	75[a]	4-Cl	Me	CH_2F	6.39
33	2-OMe	Me	CF_3	4.00	76	4-Cl	Me	CHO	5.80
34	2-F	NH_2	CF_3	7.00	77	4-Cl	Me	CN	6.64
35	2-Me	NH_2	CF_3	6.70	78	4-Cl	Me	$COOC_2H_5$	5.24
36	3-F-4-OMe	Me	CF_3	6.82	79	4-Cl	Me	C_6H_5	6.62
37	3-Cl-4-OMe	Me	CF_3	6.89	80	4-Cl	Me	$CH_2OC_6H_4$−4-Cl	7.52
38	3-Cl-4-SMe	Me	CF_3	7.40	81	4-Cl	Me	$CH_2SC_6H_4$−4-Cl	7.30
39[a]	3-Cl-4-NMe_2	Me	CF_3	6.50	82[a]	4-Cl	Me	CH_2OMe	5.43
40	3-F-4-NMe_2	Me	CF_3	6.48	83	4-Cl	Me	CH_2OH	5.08
41	3-Cl-4-NHMe	Me	CF_3	6.18	84	4-Cl	Me	CH_2SMe	6.50
42	3-Cl-4-Me	Me	CF_3	7.52	85	4-Cl	Me	CH_2CN	5.81
43	3-F-4-Me	Me	CF_3	6.96					

[a] Test set compounds.

Table 3. Descriptors, Ceofficient, Standard Error, and *t*-Test Values for the Linear Model[a]

no.	descriptor	coefficient	SE	*t*-test
0	intercept	−0.98	1.45	−0.676
1	$\log K_{ow}$	0.657	0.028	23.529
2	pK_a	0.062	0.028	2.255
3	E_{LUMO}	−0.687	0.131	−5.245
4	E_{HOMO}	0.085	0.151	0.56
5	N_{Hdon}	0.069	0.071	0.965

[a] $R = 0.911$, RMS = 0.335, $N = 153$, $F = 143.452$.

Table 4. Performance Comparison between MLR, RBFNN, and SVM (LOO Cross-Validation)

	MLR	RBFNN	SVM
R	0.911	0.945	0.947
RMS	0.352	0.260	0.257

on MLR was shown in Table 1 and Figure 1. The same set of descriptors was also employed to develop the nonlinear correlation models based on RBFNN and SVM. To obtain better results, the parameters that influence the performance of RBFNN and SVM were also optimized. The selection of the optimal width value for RBFNN was performed by systemically changing its value in the training step. The value which gives the best leave-one-out cross-validation result was used in the model. Figure 2 shows the detail of this selection process. For this data set, the optimal value was determined as 0.5. The corresponding number of centers (hidden layer

Table 5. Descriptors, Ceofficient, Standard Error, and *t*-Test Values for the Linear Model (Training Set)[a]

no.	descriptor	coefficient	SE	*t*-test
0	intercept	−0.237	1.434	−0.165
1	logK_{ow}	0.651	0.029	22.829
2	pK_a	0.049	0.029	1.717
3	E_{LUMO}	−0.704	0.131	−5.361
4	E_{HOMO}	0.149	0.15	0.99
5	N_{Hdon}	0.047	0.077	0.618

[a] $R = 0.924$, RMS = 0.31, $n = 131$, $F = 146.05$, prob > $F < 0.0001$.

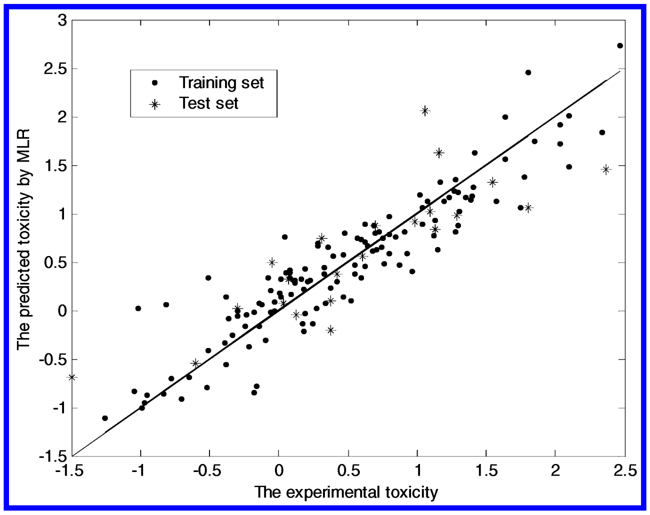


Figure 1. Predicted vs experimental toxicity by MLR.

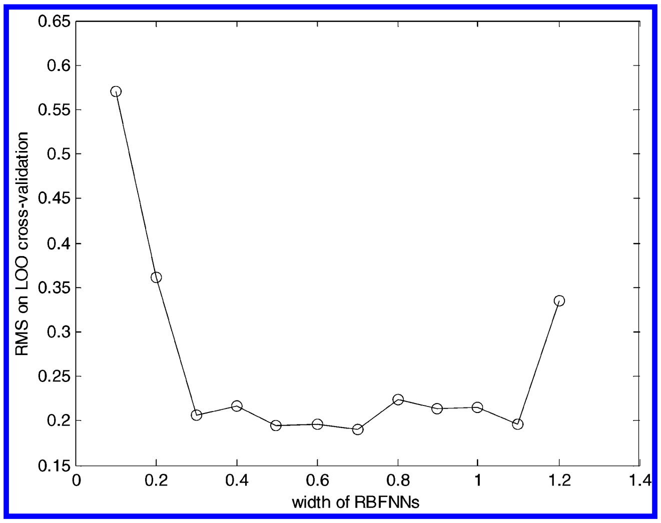


Figure 2. The selection of the optimal width for RBFNN.

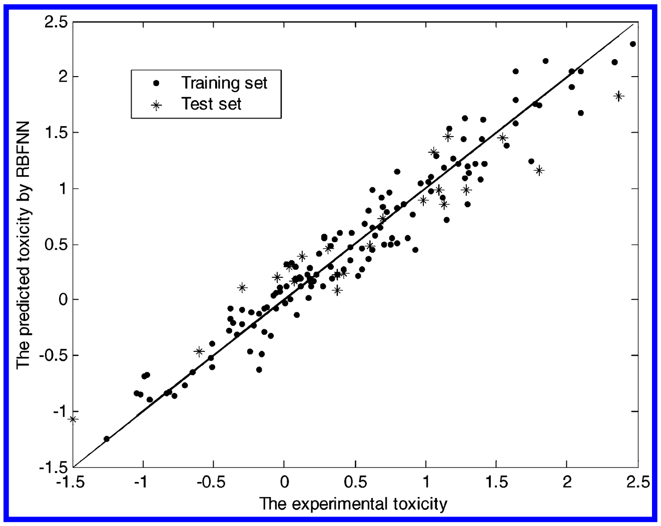


Figure 3. Predicted vs experimental toxicity by RBFNN.

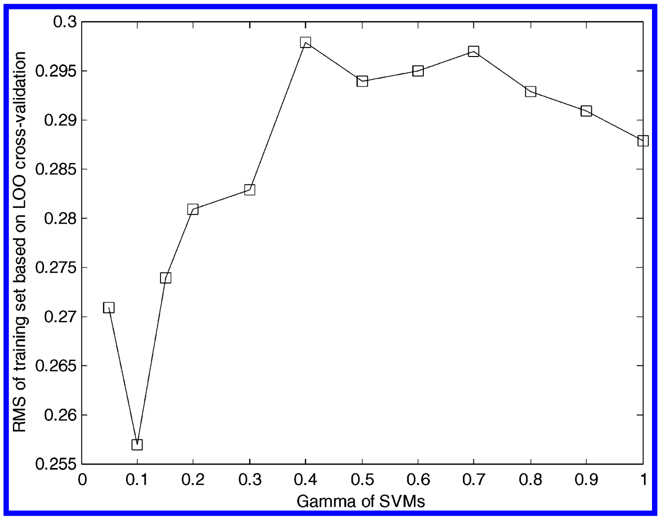


Figure 4. The selection of the optimal gamma for SVM ($\epsilon = 0.1$, $C = 1000$).

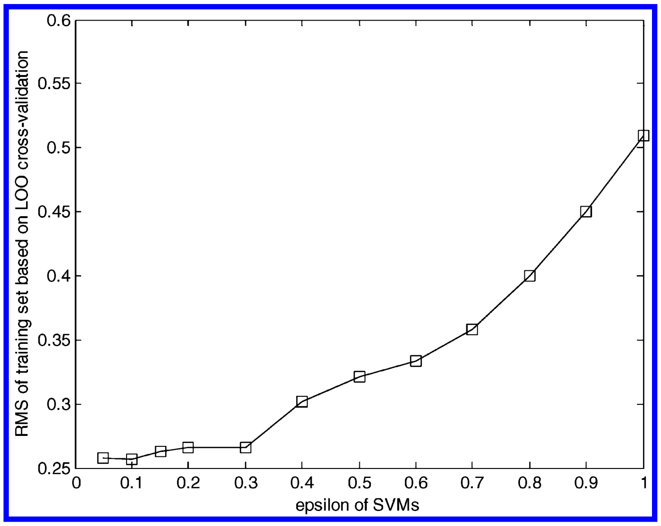


Figure 5. The selection of the optimal epsilon for SVM ($\gamma = 0.1$, $C = 1000$).

nodes) of RBFNN is 18. For the SVM model with an RBF kernel function, there are three parameters, ϵ, γ, and C, to be determined. Detailed descriptions of the process for selecting parameters and their effects on the generalization performance have been described in our previous works.[10,19] Their influences on the performance are shown in Figures 4−6. The γ, ϵ, and C for this data set were fixed to 0.1, 0.1, and 1200, respectively. The corresponding number of support vectors is 90.

The predicted results of the nonlinear models are shown in Table 1 and Figures 3 and 7. The comparison of the correlation models obtained with SVM, RBFNN, and MLR are summarized in Table 6. It is very clear that SVM and RBFNN models show similar correlation performance and outperform the MLR model. The performance of RBFNN is a bit better than that obtained by SVM.

Data Set 2. For this data set, two types of molecular descriptors were used to build QSAR models based on three modeling methods. In the first case, all the molecular

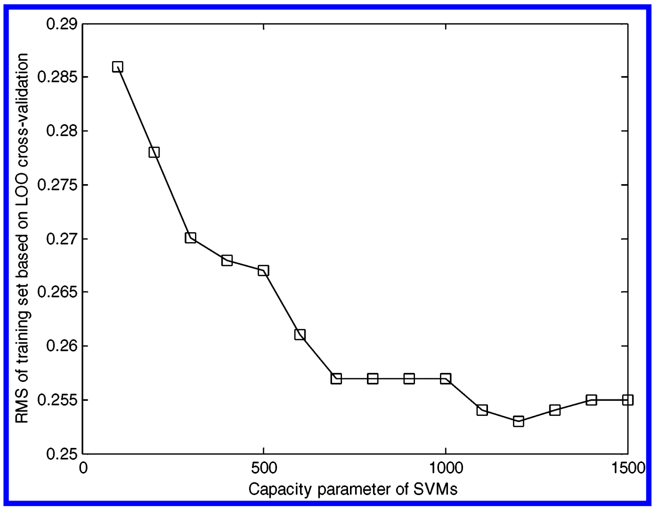


Figure 6. The selection of the optimal capacity factors for SVM ($\gamma = 0.1$, $\epsilon = 0.1$).

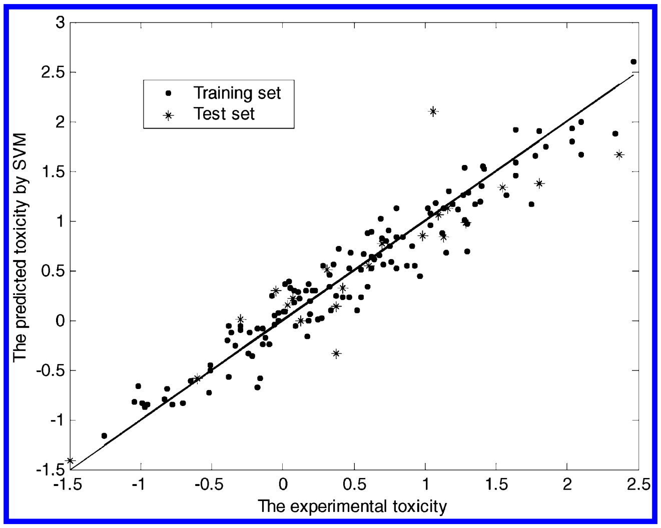


Figure 7. Predicted vs experimental toxicity by SVM.

Table 6. Performance Comparison between MLR, RBFNN, and SVM

method		training set	test set	all
mlr	R	0.924	0.834	0.911
	RMS	0.30	0.46	0.33
rbfnn ($N_c = 18$)	R	0.969	0.952	0.965
	RMS	0.19	0.29	0.21
svm ($n_{SV} = 90$)	R	0.961	0.902	0.952
	RMS	0.22	0.36	0.24

descriptors were extracted directly from the review work of Hansch et al. In the second case, six parameters were calculated using CODESSA. These descriptors include one constitutional descriptor, one geometrical descriptor, one topological index, one electrostatic descriptor, and two quantum chemical descriptors. The constitutional descriptor is the relative number of C atoms, which is related to the constitution and size of a molecule. The geometrical descriptor is YZ shadow/YZ rectangle, which describes the size and shape of a molecule. The topological descriptor is the average information content (order 0) which describes the size, branching, and composition of a molecule and relates to the dispersion interaction among molecules. The electrostatic descriptor is the fractional hydrogen bond surface area (FHBSA). The two quantum descriptors are the maximum nucleophilic reaction index (nucleoph.react.index) for a C atom and the max total interaction for O−S bond. These two descriptors are correlated with the electrostatic and hydrogen bonding interactions among molecules.

Table 7. Linear Model between Structure and Activity: (A) 85 Compounds and 4 Parameters, (B) 83 Compounds and 4 Parameters, and (C) 81 Compounds and 4 Parameters

descriptor	coefficient	error	t-test value
	(A)[a]		
intercept	10.215	0.785	13.014
ClogP	0.849	0.078	10.821
I_Y	0.861	0.109	7.868
MgVol	−2.07	0.316	−6.546
$L_{X,2}$	−0.857	0.144	−5.947
	(B)[b]		
intercept	10.001	0.666	15.014
ClogP	0.78	0.068	11.537
I_Y	0.915	0.095	9.667
MgVol	−1.843	0.271	6.805
$L_{X,2}$	−0.891	0.122	−7.287
	(C)[c]		
intercept	9.816	0.774	12.69
ClogP	0.791	0.073	10.845
I_Y	0.863	0.099	8.674
MgVol	−1.864	0.293	−6.352
$L_{X,2}$	−0.801	0.172	−4.659

[a] $N = 85$, $R = 0.887$, $F = 73.981$, $s = 0.401$, prob > F < 0.0001. [b] $N = 83$, $R = 0.913$, $F = 97.003$, $s = 0.34$, prob > F < 0.0001. [c] $N = 81$, $R = 0.887$, $F = 70.03$, $s = 0.361$, prob > F < 0.0001.

In Hansch's review work, there are two outliers (compounds 21 and 64) during the MLR analysis. In the Codessa analysis, we found four outliers (compounds 21, 33, 48, and 73) in the data set. To compare the performance of different models and the influence of different molecular descriptors, for each case, QSAR models with 85 (total data set), 83 (removing compounds 21 and 64 from data set), and 81-(removing compounds 21, 33, 48, and 73 from data set) compounds were investigated and compared for three different modeling methods. The detailed description of the six linear models is listed in Tables 7 and 8. The leave-one-out cross-validation results of different models are shown in Table 9. To further evaluate and compare the predictive ability of different models, the 83 compounds from Hansch's review were divided into a training set (71 compounds) and a test set (12 compounds). The linear correlation model between the structures and activity for the training set is shown in Table 10. The results of different models are gathered in Table 11. As can be seen from Table 11, the results of nonlinear models are better than those obtained by linear methods. By contrast, the results of SVM are comparable with those of RBFNN.

CONCLUSION

In the present work, we have compared the performance of MLR, RBFNN, and SVM in QSAR and QSPR studies with two data sets. The obtained results show that SVM and RBFNN can be used to derive statistical models with better qualities and better generalization capabilities than linear regression methods. SVM can give similar results compared with other nonlinear methods such as neural network. The optimization process of RBFNN and SVM is relatively easy to be implemented. They can be used as alternative nonlinear

Table 8. Linear Model between Structure and Activity: (A) 85 Compounds and 6 Parameters, (B) 83 Compounds and 6 Parameters, and (C) 81 Compounds and 6 Parameters

descriptor	coefficient	error	*t*-test value
(A)[a]			
intercept	350.217	39.388	8.892
max nucleoph react. index for a C atom	−41.401	13.912	−2.976
max total interaction for a O−S bond	−10.011	1.091	−9.173
relative number of C atoms	21.974	4.146	5.299
FHBSA fractional HBSA	−6.923	1.442	−4.802
YZ shadow/YZ rectangle	6.222	1.246	4.993
average information content (order 0)	2.218	0.57	3.891
(B)[b]			
intercept	335.217	36.042	9.301
max nucleoph react. index for a C atom	−48.004	13.126	−3.657
max total interaction for a O−S bond	−9.553	0.999	−9.559
relative number of C atoms	19.967	3.761	5.308
FHBSA fractional HBSA	−6.855	1.299	−5.277
YZ shadow/YZ rectangle	5.892	1.145	5.147
average information content (order 0)	2.091	0.514	4.065
(C)[c]			
intercept	363.142	25.967	13.985
max nucleoph react. index for a C atom	−41.151	9.073	−4.536
max total interaction for a O−S bond	−10.28	0.718	−14.317
relative number of C atoms	19.02	2.762	6.887
FHBSA fractional HBSA	−6.793	0.94	−7.225
YZ shadow/YZ rectangle	4.299	0.841	5.113
average information content (order 0)	1.961	0.381	5.154

[a] $N = 85$, $R = 0.854$, $F = 34.998$, $s = 0.457$, prob > F < 0.0001. [b] $N = 83$, $R = 0.871$, $F = 39.965$, $s = 0.412$, prob > F < 0.0001. [c] $N = 81$, $R = 0.927$, $F = 75.124$, $s = 0.297$, prob > F < 0.0001.

Table 9. Performance Comparison between MLR, RBFNN, and SVM: (A) LOO and 4 Parameters and (B) LOO and 6 Parameters

method		85	83	81
		(A)		
MLR	R^2	0.752	0.802	0.742
	RMS	0.420	0.358	0.384
RBFNN	R^2	0.739	0.810	0.724
	RMS	0.432	0.351	0.401
SVM	R^2	0.763	0.820	0.764
	RMS	0.411	0.342	0.369
		(B)		
MLR	R^2	0.672	0.703	0.815
	RMS	0.483	0.438	0.324
RBFNN	R^2	0.662	0.703	0.814
	RMS	0.490	0.440	0.322
SVM	R^2	0.701	0.732	0.825
	RMS	0.462	0.417	0.316

Table 10. Linear Correlation Model between Structure and Activity (71 Compounds and 4 Parameters)[a]

descriptor	coefficient	error	*t*-test value
intercept	9.727	0.708	13.748
ClogP	0.766	0.071	10.844
I_Y	0.909	0.102	8.919
MgVol	−1.692	0.288	−5.871
$L_{X,2}$	−0.917	0.128	−7.139

[a] $N = 71$, $R = 0.914$, $F = 84.412$, $s = 0.343$, prob > F < 0.0001.

modeling tools in QSAR and QSPR. As for SVM, only support vectors (a fraction of training samples) are used in the generalization process, the SVM adapts particularly to the problem with a great deal of data in cheminformatics. Furthermore the proposed approach can also be extended in other QSPR/QSAR investigations. The study of second data set also illustrates the importance of molecular descriptors and their selection in all the modeling tools.

Table 11. Performance Comparison between MLR, RBFNN, and SVM (83 Compounds and 4 Parameters)

method		training set	test set	all
MLR	R	0.915	0.890	0.912
	RMS	0.33	0.32	0.33
RBFNN ($N_c = 10$)	R	0.932	0.942	0.932
	RMS	0.30	0.26	0.29
SVM ($n_{SV} = 28$)	R	0.930	0.925	0.929
	RMS	0.30	0.29	0.30

ACKNOWLEDGMENT

The authors thank the Association Franco-Chinoise pour la Recherche Scientifique & Technique (AFCRST) (Program PRA SI 02-03) and University Paris7-Denis Diderot (No. 2003-088 AD) for supporting this study. The authors also thank Dr. Chih-Jen Lin for using his LibSVM software in the present study.

REFERENCES AND NOTES

(1) Katritzky, A. R.; Fara, D. C.; Petrukhin, R. O.; Tatham, D. B.; Maran, U.; Lomaka, A.; Karelson, M. The present utility and future potential for medicinal chemistry of QSAR/QSPR with whole molecule descriptors. *Curr. Top. Med. Chem.* **2002**, *2*, 1333−1356.

(2) Katritzky, A. R.; Maran, U.; Lobanov, V. S.; Karelson, M. Structurally Diverse Quantitative Structure−Property Relationship Correlations of Technologically Relevant Physical Properties. *J. Chem. Inf. Comput. Sci*. **2000**, *4*, 1−18.

(3) Karelson, M. *Molecular Descriptors in QSAR/QSPR*; John Wiley & Sons: New York, 2000.

(4) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, Germany, 2000.

(5) *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach: Amsterdam, The Netherlands, 1999.

(6) Walczak, B.; Massart, D. L. Local modeling with radial basis function networks. *Chemom. Intell. Lab. Syst.* **2000**, *50*, 179−198.

(7) Furey, T. S.; Cristianini, N.; Duffy, N.; Bednarski, D. W.; Schummer, M. and Haussler, D. Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data. *Bioinformatics* **2000**, *16*, 906−914.

(8) Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene Selection for Cancer Classification Using Support Vector Machines. *Mach. Learn.* **2002**, *46*, 389−422.

(9) Ramaswamy, S.; Tamayo, P.; Rifkin, R.; Mukherjee, S.; Yeang, C. H.; Angelo, M.; Ladd, C.; Reich, M.; Latulippe, E.; Mesirov, J. P.; Poggio, T.; Gerald, W.; Loda, M.; Lander, E. S.; Golub, T. R. Multiclass Cancer Diagnosis Using Tumor Gene Expression Signatures. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 15149−15154.

(10) 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.

(11) Cai, Y. D.; Liu, X. J.; Xu, X. B.; Chou, K. C. Support Vector Machines for Predicting HIV Protease Cleavage Sites in Protein. *J. Comput. Chem.* **2002**, *23*, 267−274.

(12) Hua, S. J.; Sun, Z. R. Novel Method of Protein Secondary Structure Prediction with High Segment Overlap Measure: Support Vector Machine Approach. *J. Mol. Biol.* **2001**, *308*, 397−407.

(13) 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.

(14) Tugcu, N.; Song, M.; Breneman, C. M.; Sukumar, N.; Bennett, K. P.; Cramer, S. M. Prediction of the Effect of Mobile-Phase Salt Type on Protein Retention and Selectivity in Anion Exchange Systems. *Anal. Chem.* **2003**, *75*, 3563−3572.

(15) 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.

(16) Kramer, S.; Frank, E.; Helma, C. Fragment Generation and Support Vector Machines for Inducing SARs. *SAR QSAR Environ Res.* **2002**, *13*, 509−523.

(17) Czerminski, R.; Yasri, A.; Hartsough, D. Use of support vector machine in pattern classification: application to QSAR studies. *Quant*. *Struct*. *Act*. *Relat*. **2001**, *20*, 227−240

(18) Warmuth, M. K.; Liao, J.; Ratsch, G.; Mathieson, M.; Putta, S.; Lemmen, C. Active Learning with Support Vector Machines in the Drug Discovery Process. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 667−673.

(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-κB Mediated Gene Expression Based on Support Vector Machines. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 1288−1296.

(20) Ivanciuc, O. Support vector machine identification of the aquatic toxicity mechanism of organic compounds. *Internet Electron. J. Mol. Des.* **2002**, *1*, 151−172.

(21) Ivanciuc, O. Support vector machine classification of the carcinogenic activity of polycyclic aromatic hydrocarbons. *Internet Electron. J. Mol. Des.* **2002**, *1*, 203−218.

(22) Vapnik V. N. *Statistical Learning Theory*; John Wiley & Sons: New York, 1998.

(23) Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines*; Cambridge University Press: Cambridge, UK, 2000.

(24) Schölkopf, B.; Smola, A. *Learning with Kernels*; MIT Press: Cambridge, MA, 2002.

(25) Orr, M. J. L. *Introduction to Radial basis function networks*; Centre for Cognitive Science, Edinburgh University, 1996.

(26) Orr, M. J. L. *MATLAB routines for subset selection and ridge regression in linear neural networks*; Centre for Cognitive Science, Edinburgh University, 1996.

(27) Hsu, C. W.; Lin, C. J. A comparison of methods for multi-class support vector machines. *IEEE Trans. Neural Networks* **2002**, *13*, 415−425.

(28) Chang, C. C.; Lin, C.-J. LIBSVM- -A Library for Support Vector Machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm/

(29) Aptula, A. O.; Netzeva, T. I.; Valkova, I. V.; Cronin, M. T. D.; Schultz, T. W.; Kühne, R.; Schüürmann, G. Multivariate discrimination between modes of toxic action of phenols. *Quant*. *Struct*.−*Act*. *Relat*. **2002**, *21*, 12−22.

(30) Garg, R.; Kurup, A.; Mekapati, S. B.; Hansch, C.; Cyclooxygenase (COX) Inhibitors: A Comparative QSAR Study. *Chem. Rev.* **2003**, *103*, 703−732.

(31) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. CODESSA: Training Manual; University of Florida, Gainesville, 1995.

(32) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. CODESSA: Reference Manual; University of Florida, Gainesville, 1994.

(33) HyperChem 4.0, Hypercube, Inc., 1994.

(34) *MOPAC*, v.6.0 Quantum Chemistry Program Exchange, Program 455; Indiana University: Bloomington, IN.

CI049965I