Study of the Quantitative Structure-Mobility Relationship of Carboxylic Acids in Capillary Electrophoresis Based on Support Vector Machines

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

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

Received December 2, 2003

The support vector machines (SVM), as a novel type of learning machine, were used to develop a quantitative structure-mobility relationship (QSMR) model of 58 aliphatic and aromatic carboxylic acids based on molecular descriptors calculated from the structure alone. Multiple linear regression (MLR) and radial basis function neural networks (RBFNNs) were also utilized to construct the linear and the nonlinear model to compare with the results obtained by SVM. The root-mean-square errors in absolute mobility predictions for the whole data set given by MLR, RBFNNs, and SVM were 1.530, 1.373, and 0.888 mobility units ($*10^{-5}$ cm^2 S^{-1} V^{-1}), respectively, which indicated that the prediction result agrees well with the experimental values of these compounds and also revealed the superiority of SVM over MLR and RBFNNs models for the prediction of the absolute mobility of carboxylic acids. Moreover, the models we proposed could also provide some insight into what structural features are related to the absolute mobility of aliphatic and aromatic carboxylic acids.

1. INTRODUCTION

Capillary electrophoresis provides high efficiency separations of samples of a very diverse nature (pharmaceutical, biological, environmental...). Its speed, resolving power, efficiency, analyte solubility and stability, minimal reagent and solvent consumption, compatibility with mass spectrometry, and availability of several modes has made CE a very popular technique and an alternative to other analytical methods such as high performance liquid chromatography (HPLC). Electrophoretic mobility is the most important parameter governing the separation of solutes in capillary electrophoresis. According to Max Born's model,[1] the mobility (μ) of an ion can be expressed by

$$\mu = q/(f_h + f_{dl}) \quad (1)$$

where q is the effective charge on the ion and f_h and f_{dl} are hydrodynamic (size- and shape-related) and dielectric (charge-induced) frictional drag. The hydrodynamic friction associates with moving the solute through a continuum solvent of finite viscosity. The dielectric friction is due to the interaction between the moving ion and the adjacent solvent dipoles. As an ion migrates through the solvent, it causes the adjacent solvent dipoles to orient. After passage, the solvent dipoles relax to their normal random orientation. However, this relaxation takes a finite period of time (0.82×10^{-11} s in water at 25 °C)[2] and so imposes a retarding force on the migrating ion. In essence, the dielectric friction can be considered as an effective increase in the local viscosity around the ion. Absolute mobility is a constant characteristic of an ion. Typically, absolute mobility (μ_0) is measured experimentally either by extrapolating the mobilities observed over a range of ionic strength to infinite dilution or by measuring their limiting equivalent conductance.

During method development in CE to develop an optimized separation, the analysts generally have to employ a large number of experiments, which is often costly and time-consuming. The basic mechanism in electrophoresis is the differences in the analytes' mobilities and any attempt to provide a computation method to calculate the mobility in certain practical conditions could provide a useful tool for a faster method optimization process in CE. Therefore, developing theoretical models to predict the electrophoretic behavior of analytes is necessary. However, only a few reports have investigated the quantitative correlation between the molecular parameters and the responses obtained in CE.

The computational methods used to calculate/predict electrophoretic mobility can be classified into two categories. One approach is to use a mathematical equation to correlate electrophoretic mobility with the molecular parameters.[2−7] The other methods are more empirically based on QSPR approaches using multiple linear regression (MLR) and artificial neural networks (ANN) techniques.[8−13]

Of those previous studies that aimed at predicting the electrophoretic mobility, the most promising method is to use the QSPR approach. QSPR methods have been successfully used to predict many physicochemical properties. The advantage of this approach over other methods lies in the fact that the descriptors used can be calculated from the structure alone and are not dependent on any experimental properties. Once the structure of a compound is known, any descriptor can be calculated no matter whether it is found or not. So once a reliable model is established, we can use this method to predict the property of compounds. Therefore, quantitative structure-mobility relationship (QSMR) is a

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

10.1021/ci034280o CCC: $27.50
Published on Web 02/20/2004

useful tool to predict the electrophoretic mobilities avoiding long and tedious separation optimization. The QSMR study can also tell us which of the structural factors may play an important role in the determination of mobility.

After the calculation of 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 QSPR analysis since the late 1980s, mainly in response to increased accuracy demands. The most popular neural networks model is the back-propagation (BP) neural networks due to its simple architecture yet powerful problem-solving ability. However, the BP neural networks 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 QSPR analysis.

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

In this work, SVM were used for the prediction of absolute mobility of 58 carboxylic acids in capillary electrophoresis using descriptors calculated by the software CODESSA.[21] The aim was to establish a QSMR model that could be used for the prediction of electrophoretic mobilities of carboxylic acids from their molecular structures alone and to show the flexible modeling ability of SVM and, at the same time, to seek the important structural features related to the absolute mobility of carboxylic acids. MLR and RBFNNs methods were also utilized to establish quantitative linear and nonlinear relationship to compare with the results obtained by SVM.

2. EXPERIMENTAL SECTION

2.1. Data Preparation. The values of absolute mobilities of 58 carboxylic acids studied were taken from ref 3. Table 1 contained the mobilities of the data set, in $*10^{-5}$ cm^2 S^{-1} V^{-1}. The compounds contain of aliphatic and aromatic monofuntional carboxylic acids with various groups, heteroatoms and structural isomers. The electrophoretic mobilities of these compounds were obtained in the same conditions. The data set was randomly divided into two subsets: a training set of 43 compounds and a test set of 15 compounds. The training set was used to adjust the parameters of the models and the test set was used to evaluate its prediction ability. Leave-one-out (LOO) cross-validation was performed to evaluate the modeling ability of the model.

Table 1. Compounds and Electrophoretic Mobilities ($*10^{-5}$ cm^2 S^{-1} V^{-1})

no.	name	exptl[a]	MLR[b]	RBFNNs[c]	SVM[d]
1[e]	fluoroacetic acid	43.9	42.2	44.5	43.4
2	3-iodopropionic acid	34.9	34.8	34.4	34.4
3	benzoic acid	34.4	34.0	34.7	33.8
4	gallic acid	34.4	31.6	32.8	34.3
5[e]	phenoxyacetic acid	27.8	30.3	30.2	28.6
6	*o*-aminoenzoic acid	31.6	31.9	32.7	31.8
7	2-hydroxybutyric acid	34.2	33.4	34.7	33.2
8	bromoacetic acid	38.8	40.0	40.0	39.4
9[e]	3, 5-dinitrobenzoic acid	29.1	29.5	29.1	29.9
10	*p*-hydroxybenzoic acid	34.0	33.4	34.2	33.1
11	vanillic acid	27.1	29.1	27.8	27.4
12	chloroacetic acid	41.9	41.0	41.3	40.2
13[e]	*p*-fluorobenzoic acid	33.4	34.7	34.8	34.3
14	pyruvic acid	40.4	38.3	39.0	39.8
15	2-nitro-3-chlorobenzoic acid	31.3	30.3	30.4	30.5
16	trichloroacetic acid	36.2	35.7	35.0	36.3
17[e]	glycolic acid	42.3	39.9	41.2	42.3
18	*p*-nitrobenzoic acid	32.1	31.6	32.5	31.9
19	nicotinic acid	34.6	35.5	34.9	35.2
20	2-nitro-3-bromobenzoic acid	28.2	29.1	28.7	29.8
21[e]	glucutonic acid	26.6	28.6	24.0	26.7
22	4-bromobutyric acid	32.8	33.9	32.9	33.5
23	3, 4-dibromofluoroacetic acid	36.9	36.7	37.2	37.0
24	*o*-isopropylbenzoic acid	24.7	22.8	24.7	24.5
25[e]	trifluoroacetic acid	42.5	41.7	42.0	41.4
26	cinnamic acid	28.3	29.7	28.5	29.7
27	*p*-methoxybenzoic acid	28.3	29.6	28.5	30.1
28	2-chlorobutyric acid	32.8	36.6	33.8	33.2
29[e]	gloconic acid	27.2	27.2	21.7	27.2
30	*p*-bromobenzoic acid	31.5	32.1	31.6	32.0
31	iodoacetic acid	40.2	38.4	38.5	39.7
32	salicylic acid	35.4	33.2	34.0	35.0
33[e]	lactic acid	36.5	38.7	39.1	38.3
34	dichloroacetic acid	39.4	39.1	40.2	38.6
35	2, 3-dimethylbenzoic acid	27.1	27.8	26.3	28.0
36	*p*-chloeobenzoic acid	33.4	33.3	33.1	33.1
37[e]	5-bromovaleric acid	30.8	31.2	29.5	30.7
38	trichloroacetic acid	34.2	35.7	35.0	34.3
39	*p*-*tert*-butylbenzoic acid	23.2	23.3	23.6	23.9
40	5-iodovaleric acid	30.8	29.4	30.4	29.0
41[e]	2-bromobutyric acid	30.8	32.2	31.0	32.3
42	3, 4-dihydroxybenzoic acid	34.4	32.4	33.2	34.2
43	chlorodibromoacetic acid	34.9	34.8	34.7	35.4
44	*p*-toluic acid	29.1	30.8	30.5	30.9
45[e]	glyoxalic acid	37.8	43.3	34.2	38.1
46	tribromoacetic acid	34.9	34.6	35.3	34.7
47	glyceric acid	36.3	37.2	35.8	36.9
48	2-bromopropionic acid	33.4	34.9	35.8	35.1
49[e]	3-chloropropionic acid	36.8	37.8	35.3	37.0
50	2, 3-dibromopropionic acid	32.3	32.0	31.5	32.9
51	4-iodobutyric acid	32.9	32.2	34.0	31.7
52	2-chloro-3-hydroxybutyric acid	32.9	32.8	31.9	33.0
53[e]	*p*-ethylbenzoic acid	26.5	27.7	25.3	28.1
54	2,4,6-trimethylbenzoic acid	24.7	24.9	24.6	25.2
55	2, 4-dihydroxybenzoic acid	32.0	32.6	33.7	32.3
56	*p*-ethoxybenzoic acid	26.6	27.6	26.2	28.0
57[e]	5-chlorovaleric acid	30.8	32.3	29.5	31.7
58	phenylacetic acid	31.7	30.8	30.7	31.0

[a] Experimental absolute mobility. [b−d] Predicted mobility by MLR, RBFNNs, and SVM, respectively. [e] Test set.

2.2. Descriptor Calculation. All structures of the molecules were drawn with the HYPERCHEM program (Hypercube, 1994).[22] The final geometries were obtained with the semiempirical PM3 method. All calculations were carried out at the 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 resulted geometry was transferred into software CODESSA that can calculate constitutional, topological, geometrical, electrostatic, and quantum-chemical

descriptors.[21] The constitutional descriptors reflect the molecular composition of the compound without using the geometry or electronic structure of the molecule. The topological descriptors describe the atomic connectivity in the molecule. The geometrical descriptors describe the size of the molecule and require 3D-coordinates of the atoms in the given molecule. The electrostatic descriptors reflect characteristics of the charge distribution of the molecule. The quantum-chemical descriptors add important information to the conventional descriptors. Additionally, some physicochemical descriptors include lopP, refractivity, etc. were calculated by software HYPERCHEM.

3. METHODOLOGY

3.1. Feature Selection and Regression Analysis. Once descriptors were generated, in this work, correlation analysis of descriptors was performed first. In the process of correlation analysis, pairwise correlations between descriptors were examined so that only one descriptor was retained from a pair contributing similar information (correlation coefficients greater than 0.85). After correlation analysis of the descriptors, descriptor-screening methods were used to select the most relevant descriptor to establish the models for prediction of the molecular property. Here, the forward stepwise regression method was used to choose the subset of the molecular descriptors. Forward stepwise regression starts with no model terms, and at each step it adds the most statistically significant term (the one with the highest *F*-statistic or lowest *P*-value) until there are none left.

After the descriptor was selected, multiple linear regression was used to develop the linear model of the property of interest, which takes the form below:

$$Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_nX_n \qquad (2)$$

In this equation, Y is the property, that is, the dependent variable, X_1-X_n represent the specific descriptor, while b_1-b_n represent the coefficients of those descriptors, and b_0 is the intercept of the equation. The statistical evaluation of the data was obtained by the software SPSS.

3.2. Radial Basis Function Neural Networks Theory. The theory of RBFNNs has been extensively presented in Derks' paper.[23] Here, only a brief description of the RBFNNs principle was given. The RBFNNs consist of three layers: input layer, hidden layer, and output layer. The input layer does not process the information, it only distributes the input vectors to the hidden layer. Each neuron on the hidden layer employs a radial basis function (RBF) as nonlinear transfer function to operate on the input data. The most often used RBF is the Gaussian function that is characterized by a center (c_j) and width (r_j). In this study, Gaussian was selected as the radial basis function. The operation of the output layer is linear, which is given in eq 14

$$y_k(x) = \sum w_{kj}h_j(x) + bk \qquad (3)$$

where y_k is the kth output unit for the input vector x, w_{kj} is the weight connection between the kth output unit and the jth hidden layer unit, and h_j is the notation for the output of the jth RBF unit.

The training procedure when using RBF involves selecting centers, width, and weights. In this paper, the forward subset selection routine was used to select the centers from training set samples.[24,25] The adjustment of the connection weight between the hidden layer and the output layer was performed using a least-squares solution after the selection of centers and width of radial basis functions.

3.3. Support Vector Machines.[26,27] The foundation of Support Vector Machines (SVM) has been developed by Vapnik, and they are gaining popularity due to many attractive features and promising empirical performance.[28,29] The formulation embodies the Structural Risk Minimization (SRM) principle,[26,27] 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 SVM with good generalization performance, which is the goal in statistical learning. Originally, SVM were developed for classification problems,[30] and now, with the introduction of ϵ-insensitive loss function, SVM have been extended to solve nonlinear regression estimation.[31]

Compared to other neural network regressors, there are three distinct characteristics when SVM are used to estimate the regression function. First of all, SVM estimate the regression using a set of linear functions that are defined in a high dimensional space. Second, SVM carry out the regression estimation by risk minimization where the risk is measured using Vapnik's ϵ-insensitive loss function. Third, SVM use 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 approximate the function using the following

$$y = f(x) = w\Phi(x) + b \qquad (4)$$

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

$$R_{SVMs}(C) = C\frac{1}{n}\sum_{i=1}^{n}L_\epsilon(d_i,y_i) + \frac{1}{2}||w||^2 \qquad (5)$$

$$L_\epsilon(d,y) = \begin{cases} |d-y| - \epsilon & |d-y| \geq \epsilon \\ 0 & otherwise \end{cases} \qquad (6)$$

In eq 5, R_{SVMs} 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 6. This loss function provides the advantage of enabling one to use sparse data points to represent the decision function given by eq 4. 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.

Table 2. Descriptors, Coefficients, Standard Error, and T-Values for the Linear Model[a]

descriptor	chemical meaning	coefficient	SE	beta	*t*-test	sig
intercept	intercept	28.929	4.551		6.357	0
ABIC0	average bonding information content (order 0)	6.838	1.882	0.301	3.633	0.001
ZXS/ZXR	ZX shadow/ZX rectangle	18.902	4.606	0.281	4.104	0
CHDS	count of H-donors sites [Zefirov's PC]	−0.330	0.091	−0.207	−3.606	0.001
REF	refractivity	−0.383	0.032	−0.764	−11.829	0

[a] $R = 0.952$; Standard error of the estimate = 1.379; $RMS = 1.296$; $n = 43$; $F = 91.290$.

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:

$$f(x,a_i,a_i^*) = \sum(a_i - a_i^*)K(x,x_i) + b \qquad (7)$$

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 7, 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 the 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 RBFNNs and SVM were evaluated in terms of the root-mean-square (*RMS*) error which was defined as below:

$$RMS = \sqrt{\frac{\sum_{i=1}^{n_s}(y_k - \hat{y}_k)^2}{n_s}} \qquad (8)$$

To compare the predicted mobility with the corresponding experimental value, the absolute average relative deviation (*AARD*) as an accuracy criterion was computed by

$$AARD = \frac{100}{n_s}\sum_{i=1}^{n_s}\left(\frac{|y_k - \hat{y}_k|}{\hat{y}_k}\right) \qquad (9)$$

In eqs 8 and 9, y_k is the desired output, y_k is the actual output of the model, and n_s is the number of compounds in analyzed set.

3.4. RBFNNs and SVM Implementation and Computation Environment. All calculation programs implementing RBFNNs were written in M-file based on the basis MATLAB script for RBFNNs. All calculation programs implementing SVM were written in R-file based on the R script for SVM.[32] The scripts were compiled using an R 1.7.1 compiler running operating system on a Pentium IV PC with 256M RAM.

4. RESULTS AND DISCUSSION

4.1. Results of MLR. About 170 descriptors were calculated by the CODESSA program. After the correlation analysis of the descriptors, the pool of descriptors was reduced to 102. The stepwise regression routine was used to develop the linear model for the prediction of the absolute mobilities of carboxylic acids in capillary electrophoresis using calculated structural descriptors. The best linear model contained 4 molecular descriptors. The regression coefficients of the descriptors and their physical-chemical meaning were listed in Table 2, and the corresponding values were listed in Table 3, respectively. The predicted results were given in Table 1. This model produced a *RMS* error of 1.296 electrophoretic mobility units for the training set, 2.057 for the test set, and 1.530 for the whole set, the corresponding correlation coefficients (R) were 0.952, 0.951, and 0.947, and the corresponding *AARD* were 3.140, 4.758, and 3.558%, respectively. Figure 1 showed these predicted versus experimental electrophoretic mobilities.

4.2. Results of RBFNNs. After the establishment of a linear model, RBFNNs were used to develop a nonlinear model based on the same subset of descriptors. Such RBFNNs can be designed as $4 - n_h - 1$ net to indicate the number of units in the input, the hidden layer, and the output layer, respectively. The optimal width was selected by experimenting with a number of trials and selecting the one most favored by the model selection criterion: a width <1 gives poor prediction ability, and varying the width indicates the width has little effect on the performance of RBFNNs, if the width exceeds 3. So we selected the optimum width from 1 to 3, every 0.1. Each minimum error on LOO cross-validation was plotted versus the width (Figure 2) and the minimum was chosen as the optimal conditions. In this case $r = 1.9$ and $n_h = 19$.

Through the above process, the best number of hidden layer units and the optimum width were selected as 19 and 1.9, respectively. From the best network, the inputs in the test set were presented with it, and the results with RBFNNs were obtained. They were shown in Table 1 and Figure 3. The network gave an *RMS* error of 0.918 for the training set, 2.206 for the test set, and 1.373 electrophoretic mobility unit for the whole set, the corresponding correlation coefficients (R) were 0.976, 0.951, and 0.960, and the corresponding *AARD* were 2.240, 5.467, and 3.075%, respectively.

4.3. Results of SVM. 4.3.1. Selection of the Kernel Function and Parameters of SVM. After the establishment of models by MLR and RBFNNs, the support vector machines were used to develop an accurate nonlinear model 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. C is a regularization parameter that controls the tradeoff between maximizing the margin and minimizing the training error. If C is too small, then

Table 3. Values of the 4 Descriptors

no.	ABIC0	ZXS/ ZXR	CHDS	REF	no.	ABIC0	ZXS/ ZXR	CHDS	REF
1	0.8569	0.6994	3	12.65	30	0.4666	0.8321	1	40.44
2	0.6689	0.7722	5	30.43	31	0.8569	0.7616	3	25.73
3	0.4000	0.8070	1	32.82	32	0.4194	0.8078	2	34.51
4	0.4250	0.8220	4	37.90	33	0.6694	0.6943	4	17.09
5	0.4344	0.7517	3	38.61	34	0.8905	0.7075	2	22.60
6	0.4344	0.8110	3	37.52	35	0.4051	0.7845	7	42.90
7	0.4876	0.6708	8	23.36	36	0.4666	0.8401	1	37.62
8	0.8569	0.7412	3	20.38	37	0.4847	0.8004	9	34.44
9	0.4888	0.7897	1	45.46	38	0.8569	0.6369	1	28.16
10	0.4194	0.8187	2	34.51	39	0.3694	0.7869	10	51.66
11	0.4400	0.7665	5	40.97	40	0.4847	0.8129	9	39.78
12	0.8569	0.7317	3	17.40	41	0.5578	0.6885	7	29.40
13	0.4666	0.8228	1	33.03	42	0.4256	0.8126	3	36.20
14	0.6694	0.6848	4	17.99	43	0.9796	0.6593	1	33.78
15	0.5336	0.7893	1	43.94	44	0.4184	0.7856	4	37.86
16	0.8569	0.6369	1	28.16	45	0.7544	0.8043	2	13.51
17	0.6863	0.6893	4	14.35	46	0.8569	0.7445	1	36.58
18	0.4817	0.7786	1	39.14	47	0.5663	0.7237	6	19.05
19	0.4912	0.8089	1	30.60	48	0.6689	0.6615	5	24.87
20	0.5336	0.7838	1	46.76	49	0.6689	0.7550	5	22.10
21	0.4201	0.7347	10	35.79	50	0.7130	0.6339	4	32.45
22	0.5578	0.7886	7	29.83	51	0.5578	0.8040	7	35.18
23	0.9796	0.6596	1	28.94	52	0.5638	0.6900	7	27.85
24	0.3873	0.6276	8	47.01	53	0.4051	0.7494	6	42.46
25	0.8569	0.6647	1	13.65	54	0.3873	0.7929	10	47.94
26	0.3492	0.8380	3	43.06	55	0.4256	0.8220	3	36.20
27	0.4344	0.7424	4	39.28	56	0.4203	0.7745	6	44.03
28	0.6421	0.7475	4	24.95	57	0.4847	0.7986	9	31.45
29	0.4009	0.7493	12	38.27	58	0.4184	0.7577	3	37.37

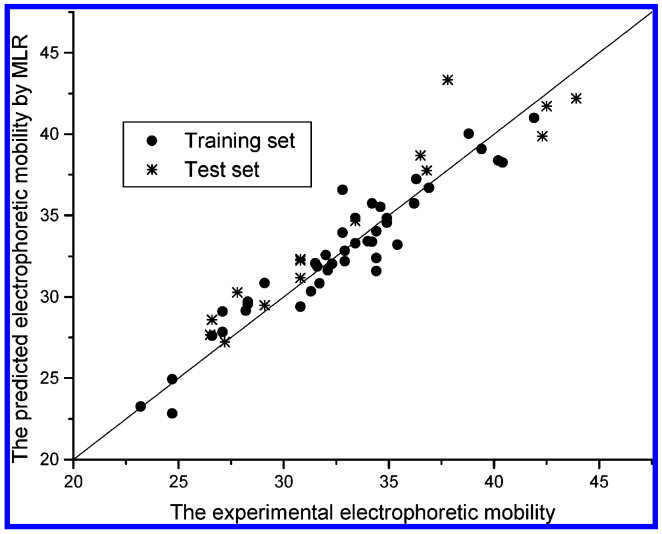

Figure 1. Predicted vs experimental electrophoretic mobilities (MLR).

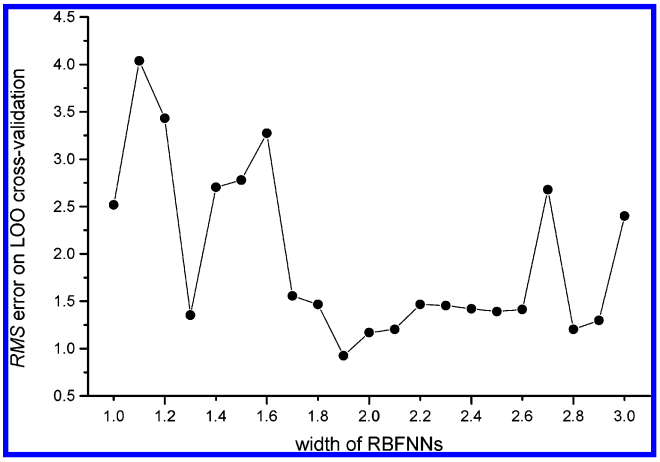

Figure 2. The width of RBFNNs vs *RMS* error on LOO cross-validation.

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 represented as follows

$$exp(-\gamma^*|u - v|^2) \tag{10}$$

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 LOO cross-validation was plotted versus γ (Figure 4), and the minimum was chosen as the optimal conditions. In this case $\gamma = 0.0004$.

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 of ϵ is critical from theory. To find an optimal ϵ, the *RMS* error on LOO cross-validation on different ϵ was calculated. The curve of *RMS* error versus the epsilon (ϵ) was shown in Figure 5. The optimal ϵ was found as 0.12.

The last important parameter is regularization parameter C, which effect on the *RMS* was shown in Figure 6. From Figure 6, the optimal C was found as 100.

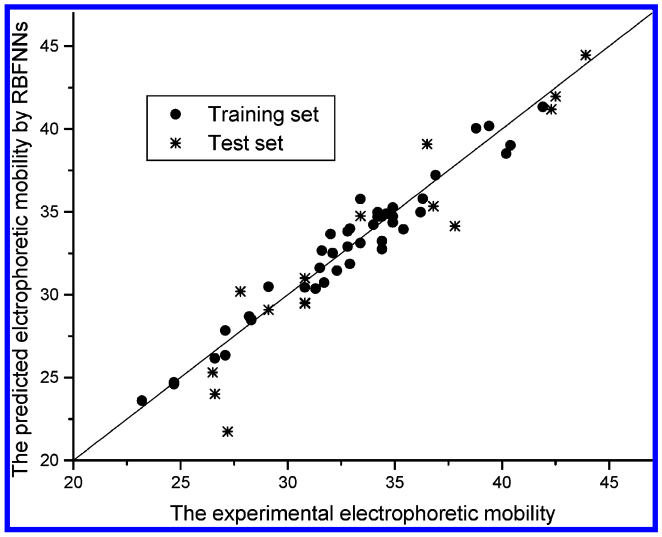

Figure 3. Predicted vs experimental electrophoretic mobilities (RBFNNs).

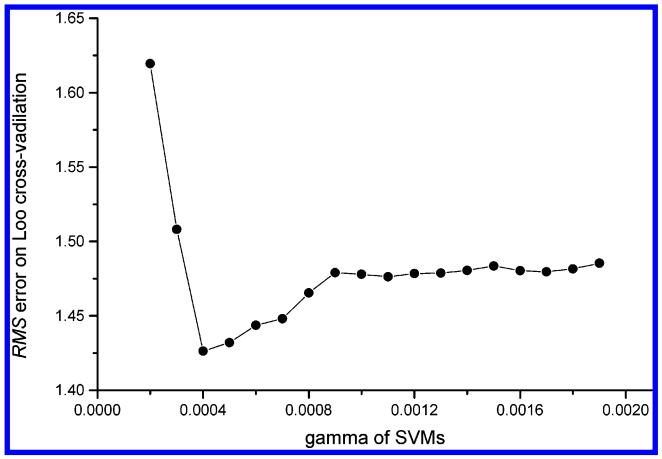

Figure 4. The gamma vs *RMS* error on LOO cross-validation (C=100, ϵ=0.1)

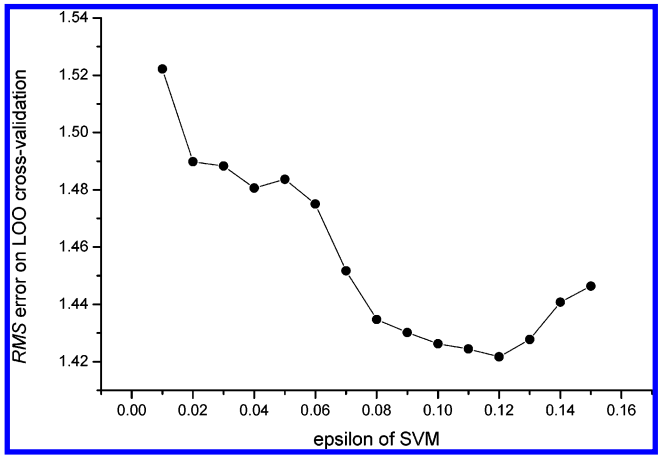

Figure 5. The epsilon vs *RMS* error on LOO cross-validation (C=100, γ=0.0004).

4.3.2. The Predicted Results of SVM. From the above discussion, the γ, ϵ, and C were fixed to 0.0004, 0.12, and 100, respectively, and the support vector number of the SVM model was 28. From the optimal model, the inputs in the test set were presented with it, and the results with SVM were obtained. They were shown in Table 1 and Figure 7. The model gave an *RMS* error of 0.859 for the training set, 0.917 for the test set, and 0.888 for the whole set, the

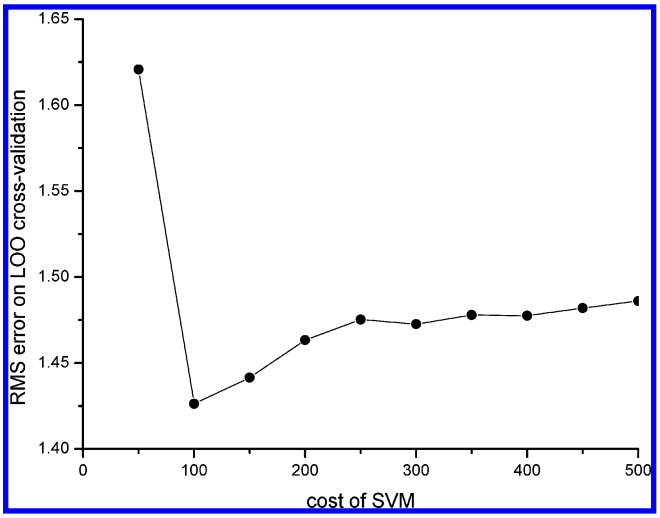

Figure 6. The cost (C) vs *RMS* error on LOO cross-validation (γ=0.0004, ϵ=0.12).

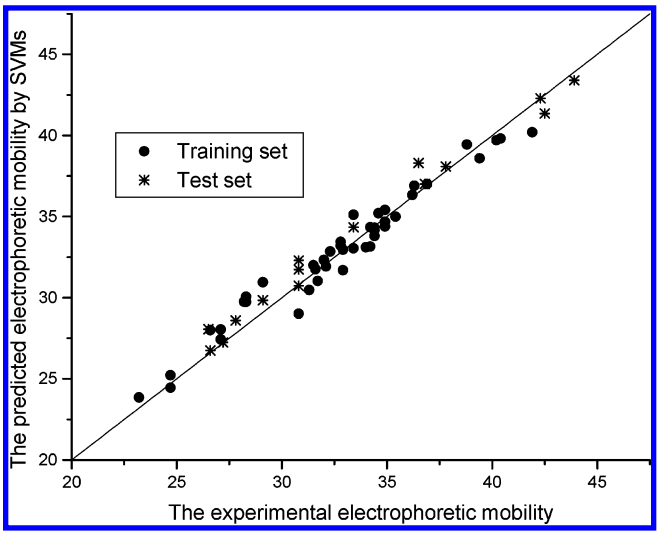

Figure 7. Predicted vs experimental electrophoretic mobilities (SVM).

corresponding correlation coefficients (R) were 0.980, 0.993, and 0.984, and the corresponding *AARD* were 2.235, 2.202, and 2.227%, respectively.

4.4. Discussion of the Input Parameters and the Results. By interpreting the descriptors in the regression model, it is possible to gain some insight into factors that are likely to govern the absolute mobilities of the carboxylic acids in capillary electrophoresis. Of the four descriptors, ABIC0 is topological, ZXS/ZXR is geometrical, CHDS is electrostatic, and REF is physicochemical descriptor. These descriptors encoded different aspects of the molecular structure. As mentioned in the Introduction two fundamental frictional factors are found to be important in the electrophoretic mobility of a solute in capillary electrophoresis. One is the hydrodynamic friction factor, which is related to the molecular size and/or mass of solute, and the other is the dielectric friction factor, which is related to the charge distribution within the solute. The descriptors in the present model can account for these friction factors. The topological descriptor (ABIC0), which encode the size, shape, and degree of branching in the compound, gives some information about the hydrodynamic friction factors. The geometrical descriptor (ZXS/ZXR) describes the size of the molecules and also has

some correlation with the hydrodynamic friction. In organic acid, the charge distribution of the carboxylate anion significantly influences the acid dissociation constant (pK_a). Hence the pK_a value is an effective measure of the charge distribution within a fully deprotonated carboxylate ion, so each parameter that affected the pK_a value can influence the electrophoretic mobility of the solute. The inclusion of electrostatic descriptors, CHDS, can influence the pK_a values of solutes and can affect the dielectric friction term and play important roles in the migration behavior of ions. Refractivity is a combined measure of the size and polarizability of a molecule and can explain the hydrodynamic and dielectric friction contribution in determination of the electrophoretic mobility.[33] According to the beta values (Table 2), the more relevant descriptor is the refractivity of the molecule.

Analysis of the results obtained indicated that the models we proposed correctly represent the structure-mobility relationships of carboxylic acids and that the molecular descriptors calculated solely from structures can represent the structural features of the compounds responsible for their absolute mobility in capillary electrophoresis. Moreover, the performance of SVM is much better than MLR and RBFNNs models. The root cause that SVM can obtain the best results is that SVM adopts the Structural Risk Minimization principle.

5. CONCLUSION

An attempt to summarize the calculation/prediction of electrophoretic mobility in capillary electrophoresis was made to illustrate how the various theoretical models actually reflect the experimental facts in the past 50 years. Then, SVM, as a novel type of learning machine, for the first time, were used to develop a QSMR model for the prediction of absolute mobility of 58 carboxylic acids in capillary electrophoresis based on descriptors calculated from the molecular structure alone. MLR and RBFNNs were also utilized to establish quantitative linear and nonlinear relationships to compare with the results obtained by SVM. Very satisfactory results were obtained with the proposed method. The models proposed could identify and give some insight into factors that are likely to govern the absolute mobilities of the carboxylic acids in capillary electrophoresis. Additionally, nonlinear models using SVM based on these same sets of descriptors produced even better models with a good predictive ability than the two other MLR and RBFNNs models. This study of the QSMR model shows that SVM are very promising tools in the prediction of electrophoretic mobility and exhibit a high speed of leaning when compared with RBFNNs. The training procedure is also simple when using SVM because there are fewer parameters having to be optimized, and only support vectors (only a fraction of all data) are used in the generalization process. Besides, SVM exhibit 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. Furthermore, the proposed approach can also be extended to other QSPR investigations.

ACKNOWLEDGMENT

The authors thank the Association Franco-Chinoise pour la Recherche Scientifique & Technique (AFCRST) for supporting this study (Program PRA SI 02-03). The authors also thank the R Development Core Team for affording the free R1.7.1 software.

REFERENCES AND NOTES

(1) Kay, R. L. The current state of our understanding of ionic mobilities. *Pure Appl. Chem.* **1991**, *63*, 1393−1399.
(2) Offord, R. E. Electrophoretic mobilities of peptides on paper and their use in the determination of amide groups. *Nature* **1966**, *211*, 591−593.
(3) Wroński, M. Concept of effective mass and hidden mass for calculation of mobility of organic anions and peptides. *J. Chromatogr. A* **1993**, *657*, 165−173.
(4) Fu, S. L.; Lucy, C. A. Prediction of electrophoretic mobilities. 1. monoamines. *Anal. Chem.* **1998**, *70*, 173−181.
(5) Fu, S. L.; Li, D. M.; Lucy, C. A. Prediction of electrophoretic mobilities. Part 2. Effect of acid dissociation constants on the intrinsic mobilities of aliphatic carboxylates and amines. *Analyst* **1998**, *123*, 1487−1492.
(6) Li, D. M.; Fu, S. L.; Lucy, C. A. Prediction of Electrophoretic Mobilities. 3. Effect of Ionic Strength in Capillary Zone Electrophoresis. *Anal. Chem.* **1999**, *71*, 687−699.
(7) Li, D. M.; Lucy, C. A. Prediction of Electrophoretic Mobilities. 4. Multiply charged aromatic carboxylates in capillary zone electrophoresis. *Anal. Chem.* **2001**, *73*, 1324−1329.
(8) Liang, H. R.; Vuorela, H.; Vuorela, P.; Riekkola, M. L.; Hiltunen, R. Prediction of migration behaviour of flavonoids in capillary zone electrophoresis by means of topological indices. *J. Chromatogr. A* **1998**, *798*, 233−242.
(9) Jalali-Heravi, M.; Garkani-Nejad, Z. Prediction of electrophoretic mobilities of sulfonamides in capillary zone electrophoresis using artificial neural networks. *J. Chromatogr. A* **2001**, *927*, 211−218.
(10) Jalali-Heravi, M.; Garkani-Nejad, Z. Prediction of electrophoretic mobilities of alkyl- and alkenylpyridines in capillary electrophoresis using artificial neural networks. *J. Chromatogr. A* **2002**, *971*, 207−215.
(11) Li, Q. F.; Dong, L. J.; Jia, R. P.; Chen, X. G.; Hu, Z. D.; Fan, B. T. Development of a quantitative structure−property relationship model for predicting the electrophoretic mobilities. *Comput. Chem.* **2002**, *26*, 245−251.
(12) Jouyban, A.; Yousefi, B. H.; A quantitative structure property relationship study of electrophoretic mobility of analytes in capillary zone electrophoresis. *Comput. Biol. Chem.* **2003**, *27*, 297−303.
(13) Wang, Y. W.; Gao, S. L.; Gao, Y. H.; S Liu, S. H.; Liu, M. C.; Hu, Z. D.; Fan, B. T. Study of the relationship between the structure and the relative mobility of chlorophenols in different buffers modified by different organic additives by capillary zone electrophoresis. *Anal. Chim. Acta* **2003**, *486*, 191−197.
(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 leaning: 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 Isoelectric Point of Amino Acid Based on GA-PLS and SVMs. *J. Chem. Inf. Comput. Sci.* **2004**, in press.
(21) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. CODESSA Version 2.0 Reference Manual, 1995−1997.
(22) HyperChem, Release 4.0 for Windows, Hypercube, Inc., 1995.
(23) Derks, E. P. P. A.; Sanchez Pastor, M. S.; Buydens, L. M. C. Robustness Analysis of Radial Basis Function and Multi-Layered Feed-Forward Neural Network Models. *Chemom. Int. Lab. Sys.* **1995**, *28*, 49−60.

(24) Orr, M. J. L. *Introduction to Radial basis function networks;* Center for Cognitive Science, Edinburgh University, 1996.
(25) Orr, M. J. L. *MATLAB routines for subset selection and ridge regression in linear neural networks*; Center for Cognitive Science, Edinburgh University, 1996.
(26) Burges, C. J. C. A tutorial on support vector machines for pattern recognition. *Data Mining Knowledge Discovery* 2 **1998**, *2*, 1−47.
(27) Vapnik,V. *Estimation of Dependences Based on Empirical Data*; Springer, Berlin, 1982.
(28) Smola, A. J.; Schölkopf, B. *A tutorial on support* V*ector regression*; NeuroCOL2 Technical report series, NC2-TR-1998-030; October, 1998.
(29) Vapnik, V. *The Nature of Statistical Learning Theory*; Springer: New York, 1995.
(30) Burges, C. J. C. A tutorial of support vector machines for pattern recognition, http://svm.research.bell-labs.com/SVMdoc.html, 1998.
(31) 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.
(32) Venables, W. N. D.; Smith, M.; the R Development Core Team. R manuals 2003.
(33) Rybolt, T. R.; Hooper, D. N.; Stensby, J. B.; Thomas, H. E.; Baker, M. L. Molar Refractivity and Connectivity Index Correlations for Henry's Law Virial Coefficients of Odorous Sulfur Compounds on Carbon and for Gas-Chromatographic Retention Indices. *J. Colloid Interface Sci.* **2001**, *231*, 168−177.

CI034280O