Prediction of the Isoelectric Point of an Amino Acid Based on GA-PLS and SVMs

H. X. Liu,[†] R. S. Zhang,*[,†,‡] X. J. Yao,[†,§] 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 August 10, 2003

The support vector machine (SVM), as a novel type of a learning machine, for the first time, was used to develop a QSPR model that relates the structures of 35 amino acids to their isoelectric point. Molecular descriptors calculated from the structure alone were used to represent molecular structures. The seven descriptors selected using GA-PLS, which is a sophisticated hybrid approach that combines GA as a powerful optimization method with PLS as a robust statistical method for variable selection, were used as inputs of RBFNNs and SVM to predict the isoelectric point of an amino acid. The optimal QSPR model developed was based on support vector machines, which showed the following results: the root-mean-square error of 0.2383 and the prediction correlation coefficient $R = 0.9702$ were obtained for the whole data set. Satisfactory results indicated that the GA-PLS approach is a very effective method for variable selection, and the support vector machine is a very promising tool for the nonlinear approximation.

1. INTRODUCTION

The isoelectric point is the pH at which a molecule carries no net electrical charge, and a substance in a solution is electrically neutral and has its unique properties. The isoelectric point of an amino acid is a very important physical-chemical property of amino acids, which is necessary to separate amino acids because of a special neutral property of amino acids at the isoelectric point. Due to a small quantity of amino acids extracted from nature, consumption of reagents determining the isoelectric point of an amino acid causes a lot of inconvenience to the following analysis. At the same time, determination of the isoelectric point of an amino acid by experiments is also time-consuming and expensive. Alternatively, the quantitative structure−property relationship (QSPR) provides a promising method for the estimation of the isoelectric point of an amino acid based on descriptors derived solely from the molecular structure to fit experimental data. The advantage of this approach over other methods lies in the fact that it requires only the knowledge of the chemical structure and is not dependent on any experimental properties.[1]

The QSPR approach has become very useful in the prediction of physical and chemical properties. This approach is based on the assumption that the variation of the behavior of the compounds, as expressed by any measured physical or chemical properties, can be correlated with changes in molecular features of the compounds termed descriptors.[2] The main steps involved in QSPR include the following: data collection, molecular geometry optimization, molecular descriptor generation, descriptor selection, model development, and finally model performance evaluation.[3] This study can develop a method for the prediction of the property of new compounds that have not been synthesized or found. It can also identify and describe important structural features of molecules that are relevant to variations in molecular properties, thus gaining some insight into structural factors affecting molecular properties. Although QSPR methods have been successfully used to predict many physicochemical properties, their use in predicting the isoelectric point of an amino acid only from the knowledge of the chemical structure has not been reported to this day.

To develop a QSPR, molecular structures are often represented using molecular descriptors which encode much structural information. In recent years there has been a shift from empirical parameters to purely calculated descriptors, such as topological indices and quantum chemical descriptors. The advantage of these calculated descriptors over other empirical descriptors is the possibility of calculating descriptors solely from the molecular structure and then applying them to the sets of structurally diverse compounds.

After the calculation of molecular descriptors, the following step is to reduce the data by selecting pertinent descriptors from a large set that faithfully describes the activity of interest. Choosing the adequate descriptors for QSAR/QSPR studies is difficult because there are no absolute rules that govern this choice. However, it is well-known, both in the chemical and the 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

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

10.1021/ci034173u CCC: $27.50
Published on Web 12/02/2003

important 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 of the descriptors at least. To deal with this issue, variable selection techniques were introduced. Frequent optimization search algorithms such as stepwise forward and stepwise backward MLR depend on an assumed linear relationship between the dependent variable and one or more descriptors, while there exists a nonlinear relationship between the physical-chemical property of compounds and their structural descriptors generally. Recently, some published papers suggested that genetic algorithms (GA) might be useful in data analysis, especially in the task of reducing the number of features for regression models.[4−6] Rogers and Hopfinger first applied this method in QSAR analysis and proved GA to be a very effective tool with many merits that other methods did not have. In this paper, we choose to use GA-PLS, which is a sophisticated hybrid approach that combines GA as a powerful optimization method with PLS as a robust statistical method for variable selection, to choose the adequate descriptors for QSPR studies.[7]

The last steps in a QSPR study are the data modeling and prediction. Artificial intelligence techniques have been applied to the data modeling and prediction of QSPR/QSAR 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.[8] 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.[8] Owing to the reasons outlined above, there is a continuing need for the application of more accurate and informative techniques to QSPR study. 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.[9−15] Based on the Structural Risk Minimization principle which seeks to minimize an upper bound of the generalization error rather than minimize the empirical error commonly implemented in other neural networks, SVMs achieve a higher generalization performance than traditional neural networks in solving these machine learning problems. Another key property is that unlike the training of other networks, which requires nonlinear optimization with the danger of getting stuck in local minima, training SVMs is equivalent to solving a linearly constrained quadratic programming problem. Consequently, the solution of SVM is always unique and globally optimal.[16]

In this paper, we built the 2D-QSAR model based on support vector machines which recently were developed from the machine learning community, with structural descriptors calculated by the software HYPERCHEM and selected using the GA-PLS approach, to explore the correlations of the molecular structure and the isoelectric point of an amino acid. Radial Basis Function Neural Networks (RBFNNs) were also applied to predict the isoelectric point of an amino acid in order to identify the reliability of the support vector machines.

Table 1. Amino Acids and Their Isoelectric Points

no.	amino acid	pI	no.	amino acid	pI
1	alanine	6.11	19	tyrosine	5.63
2	arginine	10.76	20	valine	6.02
3	asparagine	5.43	21	α-aminoadipic acid	3.18
4	aspartic acid	2.98	22	α-aminobutyric acid	6.06
5	cysteine	5.15	23	γ- aminobutyric acid	7.30
6	glutamic acid	3.08	24	α-amino isobutyric acid	5.72
7	glutamine	5.65	25	canavanine	7.93
8	glyeine	6.06	26	citrulline	5.92
9	histidine	7.64	27	2,4-diaminobutyric acid	9.27
10	isoleucine	6.04	28	homocysteine	5.55
11	leucine	6.04	29	homoserine	6.17
12	lysine	9.47	30	3-hydroxyglutamic acid	3.28
13	methionine	5.71	31	δ-hydroxylysine	9.15
14	phenylalanine	5.76	32	hydroxyproline	5.74
15	proline	6.3	33	norleucine	6.09
16	serine	5.7	34	ornithine	9.73
17	threonine	5.6	35	6-amino caproic acid	7.29
18	tryptophan	5.88			

2. METHODS

2.1. Data Set. The isoelectric point data of 20 familiar amino acids were taken from ref 17. Other isoelectric point data were obtained by calculating as follows

$$pH = (pK_n + pK_{n+1})/2$$

where n is the number of groups with a positive charge when an amino acid combines with most protons, and pK is the ionization constant of an amino acid, which was taken from ref 18. Its calculating step is as follows: (1) Sort pK of an amino acid from small to large. (2) Determine the value of n: first, judging the kind (the acidic, neutral or alkaline) of an amino acid; then for an acidic or a neural acid, n is equal to 1 and for an alkaline amino acid, n is determined as 2.

2.2. Descriptor Calculation. To obtain a QSPR model, compounds must be represented using molecular descriptor. Here, the quantum-chemical descriptors were used, and their calculation was described as follows: 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 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 was 0.001. In addition, the number of nitrogen atoms (NN) and the number of carboxyl (NC) atoms, the difference between the number of oxygen atoms and the number of nitrogen atoms (NONN) were also used to express molecular structural information. A full list of 23 calculated descriptors and their chemical meanings has been given in Table 2.

2.3. Selection of Descriptors Based on the GA-PLS Approach. GA-PLS is a sophisticated hybrid approach that combines GA as a powerful optimization method with PLS as a robust statistical method for variable selection. GA is a novel optimization technique that mimics the natural selection in nature. The natural selection in nature is that species having a high fitness under some environmental conditions can prevail in the next generation, and the best species may be reproduced by crossover together with random mutations of chromosomes in surviving ones. In GA-PLS, the chromosome and its fitness in the species correspond to a set of

Table 2. A Full List of 35 Descriptors and Their Chemical Meaning

symbol	meaning
TE	total energy
BE	binding energy
IAE	isolated atomic energy
EE	electronic energy
CCI	core−core interaction
HF	heat of formation
CN1	the maximum of the net atomic charge on the N atom
CN2	the submaximum of the net atomic charge on the N atom
CO1	the minimum of the net atomic charge on the O atom of carboxyl
CO2	the subminimum of the net atomic charge on the O atom of carboxyl
HOMO	energy of highest occupied molecular orbital
LUMO	energy of lowest unoccupied molecular orbital
SAA	molecular surface area (approximately)
SAG	molecular surface area (Grid)
VOL	molecular volume
HYE	hydration energy
LOGP	the octanol/water partition coefficient
REF	refractivity
POL	polarizability
NC	number of carboxyl
NN	number of the N atoms
NONN	difference between the number of oxygen atoms and the number of nitrogen atoms
LUMO1	energy of sublowest unoccupied molecular orbital

variables and internal prediction of the derived PLS model, respectively.[7]

The selection of descriptors based on GA-PLS contains five fundamental steps: (1) The initial population of chromosomes is created by setting all bits in each chromosome to a random value. Bit "1" denotes a selection of the corresponding variable, and bit "0" denotes a nonselection. The number of the population (N_p) is dependent on dimensions of the application problem. (2) The fitness of each chromosome which is evaluated by the internal prediction of the model is expressed as follows

$$\text{fitness} = 1 - \frac{(n-1)(1-q^2)}{(n-c)}$$

$$q^2 = 1 - \frac{\text{PRESS}}{\text{SSY}}$$

where q^2 is a cross-validated r^2 value (hereafter, denoted by q^2) by the leave-one-out procedure; SSY is the sum of the squared deviation of the dependent variable values from their mean; and PRESS is the predicted sum of squares obtained from the leave-one-out cross-validation method; n is the number of compounds; and c is the number of selected variables. Not only the reliability of the model expressed by q^2 but also the number of selected variables were considered by this fitness function. (3) The chromosomes with the higher fitness are selected from the population in an arbitrary proportion. The other necessary chromosomes, which make up the population in the next generation, are created by the following crossover and mutation steps in order to ensure diversity of the population. (4) In a crossover, a pair of randomly selected chromosomes is individually divided, mutually exchanged, and merged with a predefined frequency (crossover frequency: F_c). In a mutation, a binary bit pattern in each chromosome is changed with a small probability (mutation rate: P_m). (5) The reinsertion of offspring in a population replacing parents is preformed according to the fitness. The cycle of the above four steps (from steps 2 to 5) is repeated until the number of generations reaches the given maximum (maximum number of generations: N_g).

The values of empirical parameters affecting the performance of GA-PLS are defined as follows: the number of populations (N_p) is 200, the maximum number of generations (N_g) is 1000, the generation gap (GGAP) is 0.9, the crossover frequency (F_c) is 0.5, and the mutation rate (P_m) is 0.01. These values were empirically determined by experience from the series of the GA-PLS studies. The GA-PLS program was written in m-file and was compiled using a Matlab 6.1 compiler running operating system on a Pentium IV with 256M RAM.

2.4. Support Vector Regression.[19−21] In recent years, there has been a lot of interest in studying support vector machines (SVMs) in the field of machine learning due to many attractive features and promising empirical performances of SVMs. SVMs are a class of supervised learning algorithms initially proposed by Vapnik. To date, SVMs have been applied successfully to a wide range of pattern recognition problems, such as image recognition,[22] microarray gene expression classification,[23] protein folding recognition,[24] protein structural class prediction,[25] identification of protein cleavage sites,[26] QSAR, and other pharmaceutical data analysis.[23,27] Although SVMs were originally developed for classification, Vapnik enabled them to solve regression problems by choosing a suitable cost function (ϵ-insensitive loss function).

In 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 $G = \{(x_i; d_i)\}_{i=1}^{l}$ (x_i is the input vector and d_i is the desired value). SVMs approximate the function in the following form

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

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

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

where

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

and ϵ is a prescribed parameter.

In eq 2

$$C\frac{1}{N}\sum_{i=1}^{N} L_\epsilon(d_i, y_i)$$

is the so-called empirical error (risk), which is measured by ϵ-insensitive loss function $L_\epsilon(d,y)$, which indicates that it does not penalize errors below ϵ. The second term, $1/\|w\|^2$, 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 "ξ" leads eq 2 to the following constrained function:

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

$$\text{s.t. } w\Phi(x_i) + b - d_i \le \epsilon + \xi_i,$$

$$d_i - w\Phi(x_i) - b_i \le \epsilon + \xi_i,$$

$$\xi,\xi^* \ge 0. \quad (5)$$

Thus, decision function (1) becomes the following form

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

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

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

with the following constraints:

$$\sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0$$

$$0 \le \alpha_i \le C, i = 1, \cdots, l$$

$$0 \le \alpha_i^* \le C, i = 1, \cdots, l \quad (8)$$

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, $K(x_i,x_j) = \Phi(x_i)\bullet\Phi(x_j)$. 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. 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.

2.5. SVM Implementation and Computation Environment. All calculation programs implementing SVM were written in R-file using libsvm based on R script for SVM. All scripts were compiled using an R1.7.1 compiler running operating system on a Pentium IV with 256M RAM.

3. RESULTS AND DISCUSSION

3.1. The Results of GA-PLS. GA-PLS was applied to the data set of amino acids in order to reduce the number of field variables and increase the prediction power of the model. The model with the highest fitness value is considered as the best model. At the same time, the model with more than 7 parameters was not considered because the number of compounds should be 5 times at least the number of the parameters generally. Through the procedure of GA-PLS, the best model was found, which contained 7 parameters (see Table 3) with the fitness value of 0.5528, when the leave-one-out cross-validation q^2 was 0.6317. The 7 selected parameters were used to build the following model of RBFNNs and SVMs.

Table 3. Selected Parameters and Corresponding Values

no.	HF	CN2	CO1	LUMO	SAG	HYE	REF
1	−98.39	0	−0.2882	0.9421	245.72	48.28	20.5
2	−82.14	−0.0103	−0.2873	0.7619	391.66	37.57	43.51
3	−133.86	−0.0299	−0.2956	0.6927	293.09	102.51	28.35
4	−183.81	0	−0.311	0.6231	289.03	100.36	26.53
5	−87.61	0	−0.302	0.043	271.53	47.47	28.17
6	−186.87	0	−0.3142	0.4158	308.31	102.33	31.29
7	−139.22	−0.0205	−0.2788	0.7269	325.75	102.02	33.11
8	−93.6	0	−0.3114	0.9059	216.77	44.42	16
9	−62.73	−0.0303	−0.2914	0.5029	330.41	43.54	39.7
10	−113.49	0	−0.3011	0.9711	312.34	55.12	34.09
11	−116.08	0	−0.3057	0.9754	320.72	54.44	34.17
12	−108.59	−0.0351	−0.3054	0.9305	355.07	48.27	37.81
13	−100.22	0	−0.301	0.1339	337.31	51.13	37.83
14	−69.45	0	−0.2842	0.1596	356.67	48.29	45.12
15	−99.95	0	−0.3075	1.0517	277.59	53.32	28.06
16	−138.37	0	−0.3094	0.7261	252.07	42.95	22.04
17	−143.3	0	−0.3194	0.7145	276.49	46.61	26.46
18	−50.9	−0.0251	−0.2827	−0.0211	397.47	43.6	56.5
19	−114.54	0	−0.2833	0.1599	369.61	40.27	46.81
20	−105.76	0	−0.2763	0.9823	289.47	54.05	29.49
21	−196.01	0	−0.3044	0.793	350.63	102.33	35.89
22	−102.9	0	−0.2845	0.9494	273.88	50.91	25.02
23	−104.53	0	−0.3081	0.9139	278.03	47.95	25.46
24	−106.4	0	−0.2948	0.7833	267.21	49.36	25.21
25	−79.65	0.0369	−0.2975	0.5235	376.65	31.81	40.48
26	−145.26	−0.0181	−0.2965	0.6378	385.03	98.26	41.33
27	−98.59	−0.0362	−0.2961	0.7144	294.67	44.81	28.56
28	−95.85	0	−0.2966	0.1575	305.7	48.6	33.03
29	−144.27	0	−0.2973	0.7001	287.05	43.75	26.91
30	−231.88	0	−0.3305	0.4985	324.53	99.67	32.5
31	−149.79	−0.0367	−0.2968	0.6594	360.2	43.53	39.17
32	−143.12	0	−0.3015	0.8304	284.39	48.31	29.38
33	−116.64	0	−0.2983	0.7349	333.56	53.75	34.22
34	−103.85	−0.0328	−0.2992	0.7089	325.99	46.54	33.21
35	−110.23	0	−0.3131	0.9424	308.38	48.99	30.06

3.2. Results of the RBFNNs. Recently, there is a growing interest in the use of neural networks for QSAR/QSPR due to its flexibility in modeling a nonlinear problem. Neural networks are particularly useful in cases where it is difficult to specify an exact mathematical model, which describes a specific structure−property relationship. Most of these works used neural networks based on the back-propagation learning algorithm, which has some disadvantages such as the following: 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 (RBFFNs) 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 implement. So, we applied the RBFFNs to build the nonlinear model predicting the isoelectric point of an amino acid.

In the RBFNNs, the spread and the number of the radial basis function (the hidden layer units) are the two important parameters influencing the performances of the RBFNNs.

Table 4. Predicted Results Using SVMs and RBFNNs

		results of SVMs		results of RBFNNs	
no.	exp.	pred	residue	pred	residue
1[a]	6.11	6.10	−0.01	5.64	−0.47
2	10.76	9.42	−1.34	10.71	−0.05
3	5.43	5.5	0.07	5.50	0.07
4	2.98	3.48	0.5	3.43	0.45
5	5.15	4.67	−0.48	4.91	−0.24
6	3.08	2.54	−0.54	3.30	0.22
7	5.65	5.15	−0.5	4.87	−0.78
8	6.06	6.53	0.47	6.10	0.04
9	7.64	8.5	0.86	7.96	0.32
10[a]	6.04	5.70	−0.34	6.63	0.59
11	6.04	6.33	0.29	6.80	0.76
12	9.47	10.18	0.71	9.94	0.47
13	5.71	6.07	0.36	5.84	0.13
14	5.76	5.58	−0.18	6.77	1.01
15	6.3	5.6	−0.7	5.80	−0.50
16	5.7	5.59	−0.11	5.83	0.13
17	5.6	5.91	0.31	5.61	0.01
18	5.88	5.51	−0.37	4.90	−0.98
19	5.63	5.83	0.2	5.73	0.10
20	6.02	6.4	0.38	5.90	−0.12
21	3.18	3.66	0.48	3.50	0.32
22[a]	6.06	6.46	0.4	6.10	0.04
23[a]	7.3	6.75	−0.55	6.45	−0.85
24	5.72	5.81	0.09	6.23	0.51
25	7.93	8.01	0.08	7.50	−0.43
26	5.92	6.2	0.28	5.84	−0.08
27	9.27	9.06	−0.21	9.49	0.22
28[a]	5.55	5.58	0.03	5.68	0.13
29	6.17	6.09	−0.08	6.34	0.17
30	3.28	2.83	−0.45	2.74	−0.54
31	9.15	8.95	−0.2	8.78	−0.37
32[a]	5.74	4.97	−0.77	5.94	0.20
33[a]	6.09	6.92	0.83	7.54	1.45
34	9.73	9.18	−0.55	9.51	−0.22
35	7.29	6.96	−0.33	6.77	−0.52

[a] The compounds in the test set.

To find the optimized values of two parameters and prohibit the overfitting of the model, the data set was separated into a training set of 28 compounds and a test set of 7 compounds randomly, and leave-one-out cross-validation of the whole training set was performed. The leave-one-out (LOO) procedure consists of removing one example from the training set, constructing the decision function on the basis only of the remaining training data, and then testing on the removed example. In this fashion one tests all examples of the training data and measures the fraction of errors over the total number of training examples. The MSE was used as an error function, and it is computed according to the following equation

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

where d_i are the teaching outputs (desired outputs) in the training set, o_i are the actual outputs obtained from the leave-one-out cross-validation method, and n is the number of samples in the training set.

The selected parameters were as follows: the number of hidden layers is 10 and the optimal spread is 2.9. The obtained results of the optimal RBFNNs were given in Table 4. The root-mean-square error of the training set and the testing set are 0.2124 and 0.4963, respectively, and the prediction correlation coefficient R = 0.9583 were obtained for the whole data set.

3.3. Results of SVM. 3.3.1. SVM Parameters Optimization. Similar with 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 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. So, choosing the appropriate value of ϵ is critical from theory.[19]

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 γ is a constant, the parameter of the kernel; u and v are two independent variables; γ controls the amplitude of the Gaussian function and, therefore, controls the generalization ability of SVM. We have to optimize γ and find the optimal one.

To find the optimized combination of several parameters, leave-one-out cross-validation of the training set was performed using the mean squared error as the error function.

The detailed process of selecting parameters and the effects of every parameter on the generalization performance of the corresponding model were shown in Figures 1−3. To obtain the optimal γ, the support vector learning machines with different γ were trained, the γ varying from 0.001 to 0.01. We calculated the MSE on different γ, according to the generalization ability of the model based on LOO cross-validation for the training set in order to determine the optimal one. The curve of MSE versus the gamma was shown in Figure 1. The optimal γ was found as 0.003.

To find an optimal ϵ, the MSE on different ϵ was calculated. The curve of the MSE versus the epsilon was shown in Figure 2. The performance of the SVM is insensitive first and then better, finally worse, and unstable as ϵ increases from Figure 2. The optimal ϵ was found as 0.04.

The last important parameter is the regularization parameter C, which the effect on the MSE was shown in Figure 3. From Figure 3, the performance of the model becomes better first and then worse as C increases and which its optimal value was 800.[18]

3.3.2. The Predicted Results of SVMs. From the above discussion, the γ, ϵ, and C were fixed to 0.003, 0.04, and 800, respectively. The predicted results of the optimal SVMs

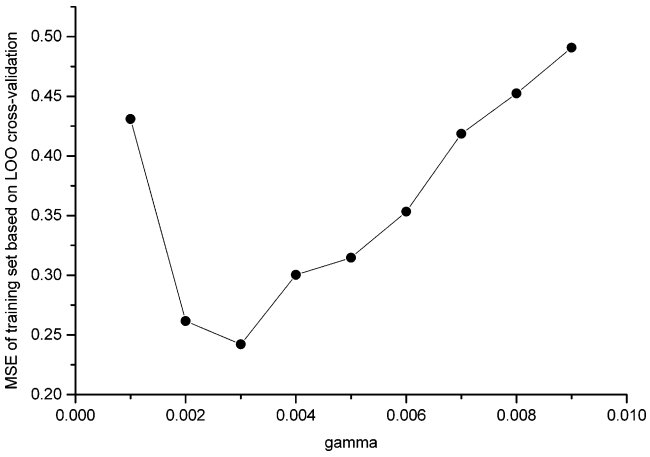

Figure 1. The gamma versus MSE error of the training set based on LOO cross-validation (C = 1000, $\epsilon = 0.01$).

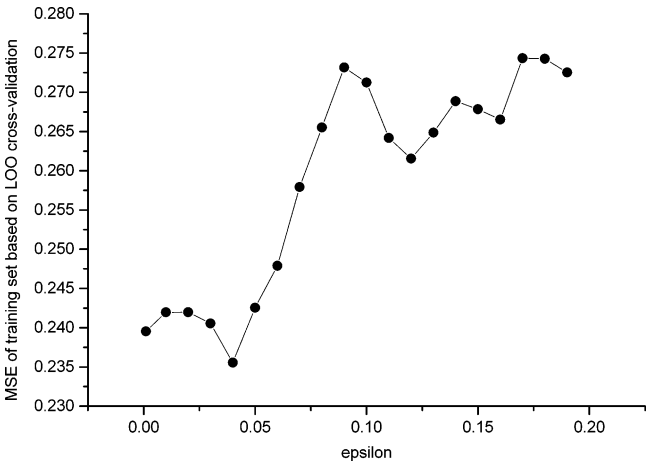

Figure 2. The epsilon versus MSE error of the training set based on LOO cross-validation (C = 1000, $\gamma = 0.003$).

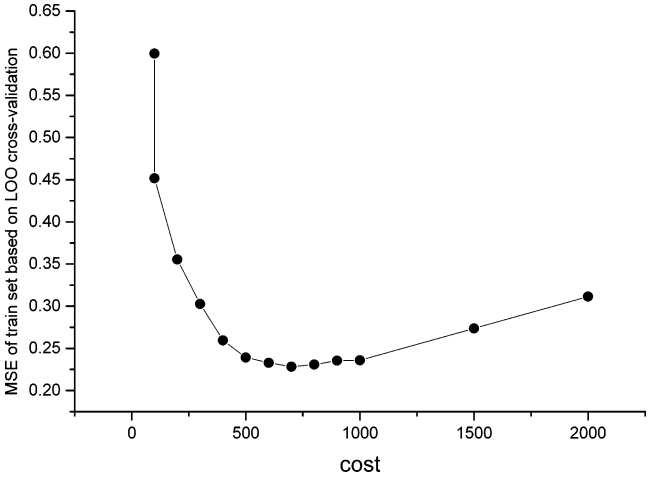

Figure 3. The epsilon versus MSE error of the training set based on LOO cross-validation ($\epsilon = 0.04$, $\gamma = 0.003$).

are shown in Table 4. The root-mean-square error of the training set and the testing set are 0.2359 and 0.2822, respectively, and the prediction correlation coefficient R = 0.9702 were obtained for the whole data set. For the isoelectric point values of amino acids with high noise, it can be concluded that the predicted values are in very good agreement with the experimental values from the above results. By comparing the results from the RBFNNs model, it can be seen that the results of SVMs are almost as good as the RBFNNs on the LOO cross-validation of the training set, but for the testing results of the test set, the SVM is much better, which showed practically that SVM has a better generalization performance than the traditional neural networks in solving this kind of nonlinear problem and can avoid the overfitting effectively.

3.4. The Analysis of Results and the Discussion of the Input Parameters. From the above results, the following can be concluded: (1) The selected parameters by the GA-PLS approach can account for the structural features of the compounds related to the isoelectric point, which indicates that the GA-PLS approach is a very effective method for variable selection. (2) The support vector machine is a very promising tool for the nonlinear approximation, which has a better generalization ability than the RBFNNs. Generally, there exists a nonlinear relationship between the property of the compounds and their structure, and we cannot express this kind of relationship simply. So the best way to solve this kind of problem is by nonlinear approximation. The heat of formation HF is the enthalpy gained in forming a molecule from its constituent atoms. It is a measure of the relative thermal stability of a molecule and reaction activity, and the more negative which a value is, the more stable and weak reactivity it has.[28] REF is another thermodynamic descriptor. The REF (Refractivity) index of a compound is a combined measure of its size and polarizability.[29] SAG (surface area) is mainly the size of a molecule. The above three parameters indicate that the isoelectric point of an amino acid is related to the polarity, the reactivity, and the bulkiness of the molecule. The HYE (calculated hydration energy) is the enthalpy gained when the molecule is combined with water, which expresses the reactivity of the molecule in water. The descriptor LUMO is an electronic parameter, which measures the electrophilicity of the molecules. When a molecule acts as a Lewis acid (an electron pair acceptor) in a bond formation, incoming electrons are received in its LUMO. Molecules with low-lying LUMO are more able to accept electrons than those with high energy LUMO.[28] CO1 (the minimum of the net atomic charge on the O atom of carboxyl) and CN2 (the submaximum of the net atomic charge on the N atom) are two other electronic parameters. The more negative value that CO1 is, the more difficult it is to break away from the molecule proton on the hydroxy. For the molecule with only one N atom, the CN2 is prescribed as zero. At the same time, for the molecule with more than one N atom, the higher its CN2 value is, the easier it is for its second amino-group to release the proton. From the above discussion, it can be seen that the isoelectric point of an amino acid is determined by several aspects of the structural factor such as polarity, reactivity, electrophilicity and their steric features, etc.

CONCLUSION

Support vector machines, a novel machine learning method, were applied to build the QSPR model for predicting the isoelectric point of an amino acid, based on descriptors calculated from the molecular structure alone and selected by the GA-PLS approach. Satisfactory results were obtained with the proposed method. From the analysis of the results obtained, the following can be concluded: (1) GA-PLS is very powerful for variable selecting in QSPR analysis, which

combines GA as a powerful global optimization method with PLS as a robust statistical method and offers a new approach to build effective QSPR models. (2) The models proposed could identify and provide some insight into what structural features are related to the isoelectric point of an amino acid of these compounds. (3) Additionally, nonlinear models using SVMs produced better models with good predictive ability than the RBFNNs. SVMs proved to be a useful tool in the prediction of the isoelectric point of an amino acid. It has some advantages over the other techniques of converging to the global optimum and not to a local optimum. Besides, as the only support vectors (only a fraction of all data) used in the generalization process, the SVM adapts particularly to the problem with a great deal of data in cheminformatics. At last, there are fewer free parameters to be adjusted in the SVM. Then the model selecting process is easily controlled. Therefore, the SVM is a very promising machine learning technique from many aspects and will gain a more extensive application. Furthermore the proposed approach can also be extended in other QSPR/QSAR 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 providing the free R1.7.1 software.

REFERENCES AND NOTES

(1) Yao, X. J.; Wang, Y. W.; Zhang, X. Y.; Zhang, R. S.; Liu, M. C.; Hu, Z. D.; Fan B. T. Radial basis function neural network-based QSPR for the prediction of critical temperature. *Chemom. Intell. Lab. Syst.* **2002**, *62*, 217−225.

(2) Yao, X. J.; Liu, M. C.; Zhang, X. Y.; Hu, Z. D.; Fan, B. T. Radial basis function network-based quantitative structure−property relationship for the prediction of Henry's law constant. *Anal. Chim. Acta* **2002**, *462*, 101−117.

(3) Yasri, A.; Hartsough, D. Toward an Optimal Procedure for Variable Selection and QSAR Model Building. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1218−1227.

(4) Hou, T. J.; Wang, J. M.; Liao, N.; Xu, X. J. Application of Genetic algorithms on the structure−activity relationship analysis of some cinnamaides. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 775−781.

(5) Hasegawa, K. GA Strategy for Variable Selection in QSAR Studies: Application of GA-Based Region Selection to a 3D-QSAR Study of Acetylcholinesterase Inhibitors. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 9, 112−120.

(6) Goicoechea, H. C.; Olivieri, A. C. Wavelength Selection for Multivariate Calibration Using a Genetic Algorithm: A Novel Initialization Strategy.*J. Chem. Inf. Comput. Sci.* **2001**, *42*, 1146−1153.

(7) Hasegawa, K.; Kimura, T.; Funatsu, K. GA strategy for variable selection in QSAR studies: Enhancement of comparative molecular binding energy analysis by GA-based PLS method. *Quant. Struct. − Act. Relat.* **1999**, *18*, 262−272.

(8) 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.

(9) Belousov, A. I.; Verzakov, S. A.; Von Frese J. A flexible classification approach with optimal generalization performance: support vector machines. *Chemom. Intell. Lab. Syst.* **2002**, *64*, 15−25.

(10) 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.

(11) Morris, C. W.; Autret, A.; Boddy, L. Support vector machines for identifying organisms- a comparison with strongly partitioned radial basis function networks. *Ecological Modelling* **2001**, *146*, 57−67.

(12) Song, M. H.; Breneman, C. M.; Bi, J. B.; Sukumar, N.; Bennett, K. P.; Cramer, S. and 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.

(13) 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.

(14) Liu, H. X.; Zhang, R. S.; 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.

(15) Warmuth, M. K.; Liao, J.; Rätsch, 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.

(16) Cao, L. J. Support vector machines experts for time series forecasting. *Neurocomputing* **2003**, *51*, 321−339.

(17) Wu, G. Y.; Pan, H. Z.; Wu, H. *Handbook of Commonly Used Experimental Data in Biochemistry and Molecular Biology*; Science Press: Beijing, 1999.

(18) Tan, P. X.; Tao, Z. P.; Qi, G. R.; Xu, J. J. *Handbook of Modern Chemical Reagent/3rd fascicule: Biochemical Reagent (one)*; Chemical Industry Press: Beijing, 1990.

(19) 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* **2003**, *55*, 643−663.

(20) Tay, F. E. H.; Cao, L. J. Modified support vector machines in financial time series forecasting. *Neurocomputing* **2002**, *48*, 847−861.

(21) Smola, A. J.; Schölkopf, B. A tutorial on support vector regression. *NeuroCOL2 Technical report series, NC2-TR-1998-030*, October, 1998.

(22) Zhang, L.; Zhou, W. D.; Jiao, L. C. Support vector machine for 1-D image recognition. *J. Infrared Millimeter Waves* **2002**, *21*, 119−123.

(23) 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.

(24) Ding, C. H. Q.; Dubchak, I. Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics* **2001**, *17*, 349−358.

(25) Karchin, R.; Karplus, K.; Haussler, D. Classifying G-protein coupled receptors with support vector machines. *Bioinformatics* **2002**, *18*, 147−159.

(26) 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.

(27) 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.

(28) Karki, R. G. and Kulkarni, V. M. Three-Dimensional Quantitative Structure−Activity Relationship (3D-QSAR) of 3-Aryloxazolidin-2-one Antibacterials. *Bioorg. Med. Chem.* **2001**, *9*, 3153−3160.

(29) 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**, *234*, 168−177.

CI034173U