Quantitative Prediction of Liquid Chromatography Retention of N-Benzylideneanilines Based on Quantum Chemical Parameters and Radial Basis Function Neural Network

Y. H. Xiang, M. C. Liu, X. Y. Zhang, R. S. Zhang,* and Z. D. Hu

Department of Chemistry, Lanzhou University, Lanzhou, Gansu 73000, P.R. China

B. T. Fan, J. P. Doucet, and A. Panaye

Universite Paris 7-denis Diderot, ITODYS 1, rue Guy de La Brosse 75005 Paris, France

Received July 21, 2001

Based on quantum chemical parameters and a simple numerical coding, the liquid chromatography retention of bifunctionally substituted N-benzylideneaniles (NBA) has been predicted using a radial basis function neural network (RBFNN) model. The quantum chemical parameters involved in the model are dipole moment (m), energies of the highest occupied and lowest unoccupied molecular orbitals ($E_{homo,}$ E_{lumo}), net charge of the most negative atom (Q_{min}), sum of absolute values of the charges of all atoms in two given functional groups (△), total energy of the molecule (E_T), weight of the molecule (W), and numerical coding (N). N was used to indicate the different positions of two substituents. The predictive values are consistent with the experimental results. The mean relative error of the testing set is 1.6%, and the maximum relative error is less than 5.0%. In this work the success of the whole modeling process only depends on the optimization of the spread parameter in network.

1. INTRODUCTION

The retention of liquid chromatographic is determined by the distribution of a solute between a mobile and a stationary chromatographic phase,[1] which depends on the forces existing between the solute molecules and those of each phase. When the stationary and mobile phases are given, the retention behavior is determined by solute structural and property parameters. A number of investigators have reported the correlation between observed retention value and hydrophobicity constant,[2−4] Hammett's constant,[5] and solubility parameters.[6] Nevertheless it is difficult to collect these parameters, and sometimes they lack comparability for the parameters which may be obtained from different sources. Unlike experimental measurements, there are no statistical errors in quantum parameters. And in principle, they can express all of the electronic and geometric properties of molecules and their interactions;[7] therefore, they can be used as parameters for establishing correlation models and to predict the retention of the substituted N-benzylideneaniles (NBA). In this work, the general formula of the substituted NBA can be expressed with $X-C_6H_4-CH{=}N-C_6H_4-Y$, where the X and Y indicate respectively the substitutions on positions 3 or 4 of left and right aromatic rings. A numerical coding has been put forward to describe the influence of different positions of two substituents in NBA.

In recent years artificial neural networks (ANNs) have become an important modeling technique in the field of quantitative structure activity/property relationship (QSAR/QSPR). The advantage of ANNs is in their inherent ability to incorporate nonlinear and cross product terms into the model. They do not require the mathematical function to be known previously. In many reported works a back-propagation (BP) neural network had been developed for predicting chromatographic retention.[8−10] However the process of obtaining an optimum BP working modeling is time-consuming. Moreover it is difficult to guarantee the realization of explicit optimum network configurations, because of the random weight initialization used for each model during the start of training.[11] Like the back-propagation neural network, the radial basis function (RBF) neural network is a feed-forward network, but it is a local adjustment network. Therefore its training rate is faster than BP neural networks, and its training and optimization procedures are relatively simpler compared to BP neural networks.[11] In this paper, we report the results of the investigation on prediction of the liquid chromatography retention for bifunctionally substituted NBA by using the RBF neural network model.

2. METHOD

2.1. RBFNN Model. The theory of RBF neural network has been extensively presented in Derks' paper.[12] Figure 1 shows basic network architecture. It consists of an input layer, a hidden layer, and an output layer.

The activation function of RBFNNs is a probability density function. In this work, Gaussian function was used. The output of the hidden layer can be expressed as below

$$h = e^{-(||w - p||b)^2} \qquad (1)$$

where w is the input weight matrix, p is the input, and b is the bias of hidden layer, which is reciprocal to radius.

In this paper, a special radial basis function neural network has been used. It is a memory-based RBFNN: all training cases must be stored in the trained network. Let $\boldsymbol{u}_i$ and v_i (i

[*] Corresponding author phone: +86-931-891-2578; fax: +86-931-863-5376; e-mail: ruison@public.lz.gs.cn.

10.1021/ci010067l CCC: $22.00 © 2002 American Chemical Society
Published on Web 04/26/2002

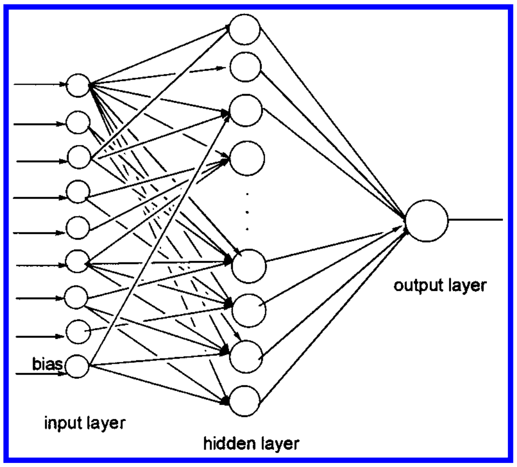

Figure 1. Structure of RBF neural network.

$= 1,...,k$) denote a scaled n-dimensional training and m-dimensional target vector, respectively. With each training sample, a new hidden neuron is added. The weight between the new hidden neuron and output unit is set to be the target value for that output. That is, $w_{ij} = v_{ji}$ is the weight from the ith hidden neuron to the jth output unit. Thus the neural network has n input units, k hidden neurons, and m outputs.[13] The hidden neuron centers correspond to the number of training cases; the hidden-to-output weights are just the target values. Therefore the output is simply a weighted average of the target values of training cases close to the given input cases. To express the output of hidden layer uniformly with the back-propagation neural network, eq 1 was used. In this equation, b is a crucial and sensitive parameter in the Gaussian function, and it is a single adjustable parameter to modify the function spread. The special operation between the hidden layer and the output layer is the dot product of output weight and h, and the output layer operation is a linear function.

2.2. Quantum Chemical Parameters and Numerical Coding. The choice of suitable structure descriptors is guided by fundamental theories of the intermolecular interactions governing the retention of solutes in liquid chromatography. It is generally agreed that three types of intermolecular interactions are the main factors influencing the solute retention: (a) polar forces resulting from permanent or induced dipoles from solute, stationary-phase, and mobile-phase molecules; (b) nonpolar forces resulting from dispersive interactions; and (c) hydrogen bond. Considering that amino bonded stationary phase is liable to form hydrogen bond with the oxygen, nitrogen, and halogen in substituents, the hydrogen bond must be taken into account. The polarity of solutes is expressed in terms of electronic parameters such as dipole moment (global polarity) and charge of functional groups[7] (local polarity). The nonpolar forces are described by parameters that reflect the size of molecule, such as the total energy of molecule[7] and molecular weight.[14] The hydrogen bond acceptor ability is relatively dependent upon the net charge of the most negative atoms and energies of the highest occupied, and the hydrogen donor ability can be measured by the lowest unoccupied molecular orbital.[15] Based on the above analysis, the following related quantum chemical parameters were obtained from the semiempirical quantum calculations using the AM1 method: (1) dipole moment (μ); (2) sum of absolute values of the charges of all the atoms in two given functional groups ($\triangle$); (3) net charge of the most negative atoms (Q_{min}); (4) energies of the highest occupied and lowest unoccupied molecular orbital ($E_{homo,}$ E_{lumo}); and (5) total energy of molecule (E_T) and molecular weight (W).

The numerical coding N was considered according to the located positions of two (or one) substituents on two aromatic rings (X: 3 or 4 and Y: 3 or 4). We use a four-dimensional binary vector to mark the positions of the substituents with the order X3, X4, Y3, and Y4. If one position is occupied by a substituent, the corresponding site of the vector is settled to 1, otherwise to 0. We use this binary coding scheme as the distributed representation. The distributed vector is also treated as a whole binary number and converted into the decimalization number N, and it is used as the “numerical coding”. All data including the numerical coding were shown in Table 1.

The quantum parameters and numerical coding were standardized. Using the following equations:

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{i,j} \tag{2}$$

$$V_j = \frac{1}{n\text{-}1}\sum_{i}^{n}(x_{i,j} - \bar{x}_j)^2 \tag{3}$$

$$X_{i,j} = \frac{x_{i,j} - \bar{x}_j}{\sqrt{V_j}} \tag{4}$$

where $x_{i,j}$ is the any data in the jth column of training set, $\bar{x}_j$ is the average value of jth column, V_j is the deviation of the jth column, and $X_{i,j}$ is the standarized result, which was used as neural network input.

3. EXPERIMENTS

All the retention values in this work were extracted from literature.[16] They were obtained in normal phase liquid chromatography with an amino bonded stationary phase, the eluents consisting of solvent mixtures based on heptan plus tetrahydrofuran.

To investigate the possible correlation of input parameters and retention values, multilinear regression (MLR) was used. The MLR results were shown in Table 2.

From Table 2 it can be seen that there is no simple linear correlation between the retention of NBA and the input parameters, nonlinear model RBF neural network was then applied to predict the retention values. In general, construction of RBF neural networks involves three main steps i.e., selection of radial basis function centroids; choice of suitable radius (spread) value, and determination of the hidden layer node number of radial basis function network.[11]

In this work a special RBF neural network was used. It consists of an input layer, the number of hidden layer nodes is equal to that of training cases (i.e. 52 nodes), and an output layer with one normalized output node. In the first phase, we use distributed representation and quantum chemical parameters as inputs. The input layer has 11 input nodes (seven quantum parameters and four-dimension distributed vector). The structure of the neural network is 11−52−1. In the second phase, we use numerical coding and quantum

Table 1. Studied Compounds and the Data Used in This Work

no.	solute X−Y	m	Q_{min}	E_{homo}	E_{lumo}	E_T	Δ	W	binary code	N	retention (exp)
1	H-4Br	2.464	−0.1722	−8.9993	−0.7418	−86.761	0.1870	260.0	0001	1	121.08
2	4CN-4F	2.238	−0.1650	−9.1428	−1.1704	−103.383	0.2320	224.0	0101	5	269.59
3	H−H	1.381	−0.1528	−8.8983	−0.5014	−74.2805	0.2660	181.8	0000	0	116.27
4	H-4F	2.396	−0.1639	−8.9049	−0.6927	−74.2805	0.2390	199.0	0001	1	119.34
5	H-4CL	2.244	−0.1523	−8.9391	−0.6937	−87.5143	0.2390	213.5	0001	1	117.30
6	H-3CL	2.658	−0.1550	−9.0753	−0.6603	−87.5137	0.1490	213.5	0010	2	117.30
7	H-3CN	4.708	−0.1540	−9.2306	−0.8270	−86.0586	0.2650	206.0	0010	2	226.37
8	H-4CN	4.417	−0.1638	−9.2171	−0.9700	−86.0593	0.2680	206.0	0001	1	241.46
9	H-3NO$_2$	4.507	−0.1533	−9.2493	−0.8329	−104.662	0.4590	226.0	0010	2	200.00
10	H-4NO$_2$	6.772	−0.3615	−9.5688	−1.3826	−104.815	1.4290	226.0	0001	1	200.00
11	4NO$_2$−4OCH$_3$	6.483	−0.3582	−8.9372	−1.4442	−122.301	1.8250	256.0	0101	5	287.22
12	4NO$_2$−4CH$_3$	6.439	−0.3579	−9.1718	−1.4656	−110.542	1.7240	240.0	0101	5	191.71
13	4NO$_2$−3CH$_3$	6.175	−0.3577	−9.3222	−1.4664	−110.541	1.7230	240.0	0110	6	184.30
14	4NO$_2$−3F	5.331	−0.3562	−9.5881	−1.6024	−122.136	1.3780	244.0	0110	6	219.31
15	4NO$_2$−4F	4.380	−0.3565	−9.3506	−1.5907	−122.137	1.3780	244.0	0101	5	230.31
16	4NO$_2$−4CL	4.617	−0.3564	−9.3736	−1.5937	−118.047	1.2840	260.5	0101	5	229.32
17	4NO$_2$−4Br	4.377	−0.3560	−9.4300	−1.6225	−117.293	1.3380	305.0	0101	5	233.59
18	4NO$_2$−3CL	5.440	−0.3564	−9.5511	−1.5832	−118.046	1.2850	260.5	0110	6	228.44
19	4OCH$_3$−H	2.744	−0.2145	−8.6786	−0.4576	−91.7687	0.6720	211.0	0100	4	182.38
20	4OCH$_3$−4F	3.600	−0.2146	−8.7122	−0.6458	−109.093	0.6470	229.0	0101	5	186.35
21	4OCH$_3$−4CL	3.489	−0.2148	−8.7417	−0.6490	−105.003	0.5570	245.5	0101	5	186.53
22	4OCH$_3$−3CL	3.998	−0.2150	−8.8219	−0.6145	−105.002	0.5570	245.5	0110	6	181.45
23	4OCH$_3$−3NO$_2$	8.042	−0.3595	−9.0990	−1.1281	−122.302	1.8270	256.0	0110	6	285.49
24	4CH$_3$−H	1.519	−0.1832	−8.8048	−0.4908	−80.0088	0.5750	195.0	0100	4	112.29
25	4CH$_3$−4F	2.751	−0.1838	−8.8231	−0.6771	−97.3330	0.5690	213.0	0101	5	114.33
26	4CH$_3$−4CL	2.587	−0.1839	−8.8560	−0.6802	−93.2426	0.4600	229.5	0101	5	110.19
27	4CH$_3$−3CL	2.887	−0.1840	−8.9684	−0.6743	−93.2421	0.4610	229.5	0110	6	112.29
28	4CH$_3$−4CN	4.833	−0.1849	−9.1055	−0.9505	−91.7877	0.5830	220.0	0101	5	230.58
29	4CH$_3$−3NO$_2$	4.716	−0.1847	−9.1186	−0.8144	−110.391	0.6720	240.0	0110	6	192.92
30	4CH$_3$−4NO$_2$	7.230	−0.3619	−9.4044	−1.3599	−110.543	1.7440	240.0	0101	5	192.17
31	3OCH$_3$−H	1.106	−0.2117	−8.8697	−0.5438	−91.7679	0.4680	211.0	1000	8	185.71
32	3OCH$_3$−4F	1.043	−0.2111	−8.8954	−0.7343	−109.092	0.4400	229.0	1001	9	103.54
33	4F-4CH$_3$	2.233	−0.1799	−8.7957	−0.6970	−97.3330	0.5380	213.0	0101	5	104.68
34	4F−H	1.976	−0.1778	−8.9477	−0.7095	−97.3330	0.2360	199.0	0100	4	112.29
35	4F-4F	1.384	−0.1779	−8.9589	−0.8769	−108.929	0.2060	217.0	0101	5	121.08
36	4F-4CL	1.385	−0.1781	−8.9903	−0.8795	−104.838	0.1150	233.5	0101	5	119.21
37	4F-3CL	2.413	−0.1781	−9.1142	−0.8522	−104.838	0.1160	233.5	0110	6	120.94
38	4F-4CN	2.870	−0.1785	−9.2520	−1.1212	−103.383	0.2320	224.0	0101	5	271.92
39	4F-3NO$_2$	5.989	−0.3578	−9.4643	−1.2843	−122.137	1.3760	244.0	0110	6	228.16
40	4F-4NO$_2$	5.153	−0.3605	−9.5684	−1.4971	−122.139	1.3860	244.0	0101	5	227.95
41	4CL-4CH$_3$	2.037	−0.1800	−8.8117	−0.7049	−93.2424	0.4470	229.5	0101	5	102.36
42	4CL-H	1.797	−0.1512	−8.9662	−0.7174	−87.5141	0.1450	215.5	0100	4	112.29
43	4CL-4F	1.436	−0.1643	−8.9747	−0.8839	−104.838	0.1130	233.5	0101	5	120.12
44	4CL-4CL	1.399	−0.1485	−9.0057	−0.8861	−100.748	0.0220	250.0	0101	5	115.33
45	4CL-3CL	2.357	−0.1535	−9.1320	−0.8589	−100.747	0.0220	250.0	0110	6	119.08
46	4CL-3NO$_2$	6.071	−0.3578	−9.4799	−1.2875	−118.046	1.2830	260.5	0110	6	231.20
47	4CL-−4NO$_2$	5.407	−0.3604	−9.5822	−1.4993	−118.048	1.2910	260.5	0101	5	225.74
48	3CL-3OCH$_3$	1.777	−0.2086	−8.6409	−0.6478	−105.001	0.5550	245.5	1001	9	178.01
49	3CL-H	2.475	−0.1509	−9.0193	−0.6780	−87.5136	0.1470	215.5	1000	8	121.08
50	3CL-4F	2.622	−0.1645	−9.0175	−0.8512	−104.838	0.1140	233.5	1001	9	130.72
51	3CL-4CL	2.565	−0.1462	−9.0490	−0.8532	−100.747	0.0220	250.0	1001	9	130.44
52	3CL-3CL	3.366	−0.1532	−9.1950	−0.8233	−100.747	0.0230	250.0	1010	10	128.99
53	3CL-4CN	4.100	−0.1619	−9.3354	−1.1019	−99.2921	0.1400	240.5	1001	9	280.80
54	4CF$_3$−4OCH$_3$	4.725	−0.2066	−8.8219	−0.9976	−149.489	1.4970	279.0	0101	5	172.01
55	4CF$_3$−4CH$_3$	4.242	−0.1813	−9.0343	−1.0105	−137.729	1.3940	263.0	0101	5	200.00
56	4CF$_3$−H	3.861	−0.1658	−9.2284	−1.0261	−132.001	1.0950	249.0	0100	4	108.02
57	4CF$_3$−4F	2.323	−0.1653	−9.2110	−1.1705	−149.325	1.0600	267.0	0101	5	115.33
58	4CF$_3$−4CL	2.533	−0.1653	−9.2381	−1.1730	−145.234	0.9670	283.5	0101	5	114.33
59	4CF$_3$−4CN	1.394	−0.1646	−9.5483	−1.3769	−143.779	1.0870	274.0	0101	5	286.78
60	4CF$_3$−4NO$_2$	3.443	−0.3593	−9.8383	−1.4865	−162.533	2.2350	294.0	0110	6	248.02
61	4CN-4CH$_3$	4.144	−0.1810	−8.9714	−1.0192	−91.7868	0.5700	220.0	0101	5	225.79
62	4CN−H	3.770	−0.1489	−9.1506	−1.0354	−86.0586	0.2660	206.0	0100	4	236.18
63	3NO$_{2-4}$OCH$_3$	4.327	−0.3599	−8.8836	−1.1834	−122.302	1.8270	256.0	1001	9	300.00
64	3NO$_2$−4CH$_3$	4.816	−0.3594	−9.1032	−1.1956	−110.542	1.7250	240.0	1001	9	212.84
65	3NO$_2$−3CH$_3$	4.483	−0.3594	−9.2324	−1.2002	−110.542	1.7250	240.0	1010	10	206.17
66	3NO$_2$−H	4.512	−0.3592	−9.3051	−1.2144	−104.814	1.4190	226.0	1000	8	223.41
67	3NO$_2$−4F	3.507	−0.3506	−9.2813	−1.3171	−122.137	1.3860	244.0	1001	9	255.74
68	3NO$_2$−4CL	3.588	−0.3585	−9.3065	−1.3188	−118.047	1.2900	260.5	1001	9	260.82
69	3NO$_2$−3CL	4.612	−0.3579	−9.4455	−1.3036	−118.047	1.2930	260.5	1010	10	254.43
70	4Br−H	1.951	−0.1530	−9.0087	−0.7753	−86.7605	0.1900	260.0	0100	4	115.33

Table 2. MLR Results on the Correlation between Input Parameters and the Retention Values

item	degrees of freedom	sum of square	mean square	F statistic	R
model	8	131286.24	16410.78	8.25213	0.7209
error	61	121308.97	1988.67		
total	69	252595.20			

Table 3. Training Validation and Testing Sets

set	compound numbers
training set	1, 2, 3, 6, 7, 8, 9, 10, 13, 14, 15, 17, 18, 19, 20, 22, 23, 24, 25, 27, 29, 30, 31, 32, 33, 34, 35, 37, 39, 41, 42, 43, 44, 45, 46, 48, 50, 51, 52, 53, 54, 55, 56, 57, 59, 60, 62, 63, 64, 66, 68, 69.
validation set	4, 12, 16, 28, 36, 40, 49, 65, 70
testing set	5, 11, 21, 26, 38, 47, 58, 61, 67

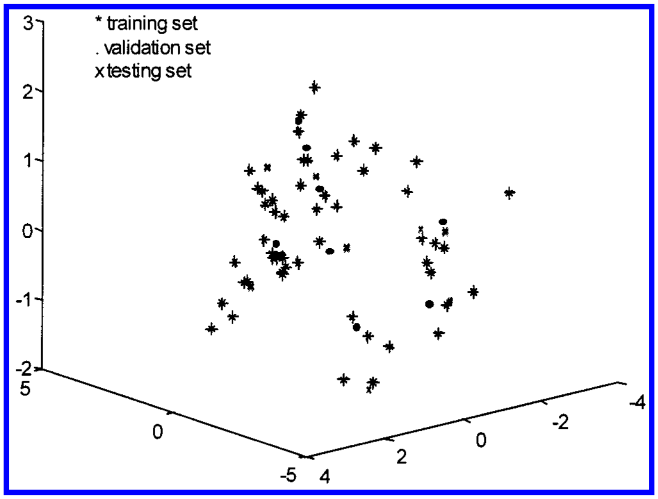

Figure 2. Location of compressed data in 3D space.

chemical parameters as inputs. The neural network has eight inputs (seven quantum parameters and a numerical coding), and the architecture of the neural network is 8−52−1.

The models of neural network only can be well generalized when the population of training cases is sufficiently large and representative for all cases. It is obvious that interpolation generalization is more reliable than extrapolation generalization (Interpolation is applied to cases that are more or less surrounded by nearby training cases. Everything else is extrapolation.).[17] Therefore the selection of training set is important.

To show spatial location of samples, the principal component analysis (PCA) method was used to reduce the data dimensions. And the first three maximum principal component values PC1, PC2, and PC3 were used to plot the data projection in a three-dimensional space. As shown in Figure 2, it gives us a clear view of the training samples. Obviously, the training samples constitute a representative subset of all cases. We divide the data into three sets as shown in Table 3.

4. RESULTS AND DISCUSSION

4.1. ANNs Structure Optimization. In our study, each of the training samples acts as the centroid of the radial basis function. Therefore the optimization of the model only depends on the choice of the spread parameter. This value selection is crucial, because it determines the shape of the Gaussian function, a large radius possesses a smooth shape and has the advantage of interpolation, and a small radius leads to a sharp shape and reduces the overlap between adjacent samples.[18] But too small a spread cannot generalize well, because unknown samples only lie in the region that Gaussian function enclosing can be generalized. We have to optimize the spread and find the optimal radius. To optimize the radius, nine samples were used as a validation set. A trial and error method was used to find the best radius. The MSE was used as an error function, and it is computed according to the following equation

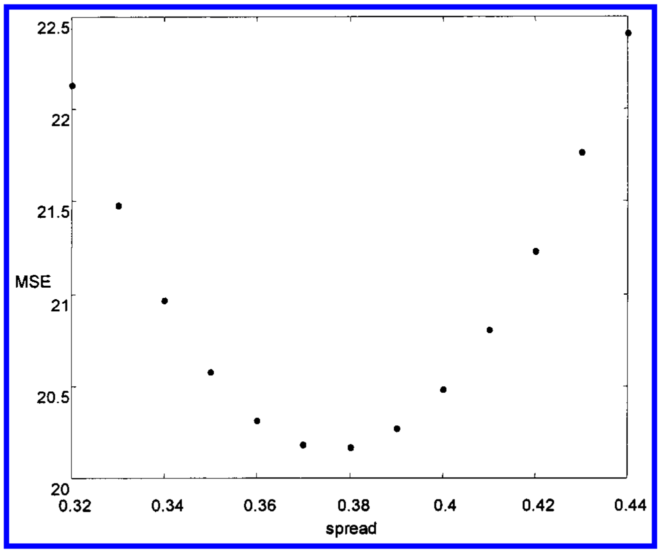

Figure 3. The spread versus MSE error on validation set.

$$MSE = \frac{\sum_{i}^{n}(d_i - o_i)^2}{n} \tag{5}$$

where d_i are the teaching outputs (desired outputs) in the validation set, o_i are the actual outputs, and n is the number of samples in validation set.

To obtain the optimal radius, the neural networks with different radii were trained, the spread varying from 0.32 to 0.44. We calculated the MSE on different radii, according to the generalization ability on the validation set in order to determine the optimal radius. The curve of MSE versus the radius was shown in Figure 3. The optimal radius was found as 0.38.

4.2. The Influence of Numerical Coding. S. Ounnar has proved that the same functional group in the X or Y position and the meta or para position has a different contribution to retention values because of the asymmetry of two aromatic rings in NBA.[14] Using a distributed four-dimensional binary vector reflects the position of the functional group. When the quantum chemical parameters and this 4-bits binary numerical code were used as neural network inputs (architecture 11−52−1), some results are shown in Table 4. The MSE of validation set and testing set were 97.08 and 403.90, respectively, which are even worse than the results without numerical coding as inputs. In the absence of this binary numerical code, the use of only quantum parameters (the optimal radius was found as 0.32 in this case) gives the MSE of testing set 20.37. But from the chemical point of view,

Table 4. Relative Errors (%) between Experimental and Predicted Retention Values of Testing Set Using 11−52−1 Network

no.	5	11	21	26	38	47	58	61	67
exptl retention	117.30	287.22	186.53	110.19	271.92	225.74	114.33	225.79	255.74
pred retention	114.59	251.10	174.42	110.61	260.77	227.18	123.09	181.41	255.92
rel error	2.38	12.58	6.52	−0.30	4.11	−0.62	−7.58	19.66	−0.061

Table 5. Relative Errors (%) between Experimental and Predicted Retention Values Using 8−52−1 Network

	retention		rel error		retention		rel error
no.	exptl	pred	(%)	no.	exptl	pred	(%)
1	121.08	121.16	−0.071	36[a]	119.21	120.14	−0.714
2	269.59	269.31	0.103	37	120.94	121.12	−0.146
3	116.27	116.36	−0.080	38[b]	271.92	268.63	1.214
4[a]	119.34	116.75	2.240	39	228.16	228.30	−0.063
5[b]	117.30	117.37	0.015	40[a]	227.95	220.50	3.280
6	117.30	117.38	−0.067	41	102.36	103.02	−0.642
7	226.37	228.19	−0.797	42	112.29	112.37	−0.069
8	241.46	239.70	0.734	43	120.12	120.07	0.039
9	200.00	200.05	−0.026	44	115.33	115.75	−0.363
10	200.00	200.05	−0.024	45	119.08	119.16	−0.072
11[b]	287.22	284.44	0.972	46	231.20	231.03	0.074
12[a]	191.71	184.77	3.650	47[b]	225.74	228.06	−1.011
13	184.30	184.58	−0.150	48	178.01	178.06	−0.027
14	219.31	219.88	−0.258	49[a]	121.08	124.93	−3.110
15	230.31	230.33	−0.009	50	130.72	130.77	−0.040
16[a]	229.32	229.94	−0.255	51	130.44	130.49	−0.038
17	233.59	233.63	−0.017	52	128.99	129.18	−0.144
18	228.44	227.98	0.202	53	280.80	280.74	0.021
19	182.38	182.07	0.171	54	172.01	172.07	−0.033
20	186.35	186.30	0.029	55	200.00	199.98	0.010
21[b]	186.53	182.85	2.002	56	108.02	108.16	−0.133
22	181.45	181.54	−0.050	57	115.33	115.42	−0.074
23	285.49	285.49	−0.001	58[b]	114.33	115.43	−0.895
24	112.29	112.38	−0.081	59	286.78	286.78	0.001
25	114.33	110.98	3.016	60	248.02	248.04	−0.007
26[b]	110.19	105.47	4.354	61[b]	225.79	229.44	−1.6032
27	112.29	112.22	0.059	62	236.18	236.21	−0.014
28[a]	230.58	233.45	−1.235	63	300.00	299.98	0.007
29	192.92	192.97	−0.027	64	212.84	212.90	−0.031
30	192.17	192.07	0.052	65[a]	206.17	212.98	−3.280
31	185.71	185.75	−0.023	66	223.41	223.44	−0.015
32	103.54	103.65	−0.110	67[b]	255.74	260.73	−1.941
33	104.68	108.13	−3.192	68	260.82	260.74	0.029
34	112.29	112.35	−0.053	69	254.43	254.54	−0.045
35	121.08	121.13	−0.039	70[a]	115.33	115.99	−0.498

[a] The compounds in validation set. [b] The compounds in testing set.

the position of substitution really plays an important role of influencing the retention values, and this influence should be taken into account globally. When the 4-bits binary code is injected into the network via four input neurons, they are in fact considered as four separated parameters. As the coding order for four different substitution positions (3X, 4X, 3Y, and 4Y) is chosen arbitrarily, it is difficult for the network to determine the influence of each single substitution position. Therefore, we tried to convert this 4-bits binary code into a single code N by decimalization (the rules of conversation have been explained in section 2.2). The architecture of the network is then simplified as 8−52−1. The results obtained are gathered in Table 5. The MSE reduces now to 11.10 (the optimal radius was found as 0.38 with parameter N). It can be seen that a decimalization number indeed can reflect the influence of positions. In fact, the decimal code N is only a parameter describing the substitution positions. The code similarity of two molecules reflects only their situations in substitution position. Moreover, the activation function used in network is not linear, and the comparison of their absolute values is not really meaningful.

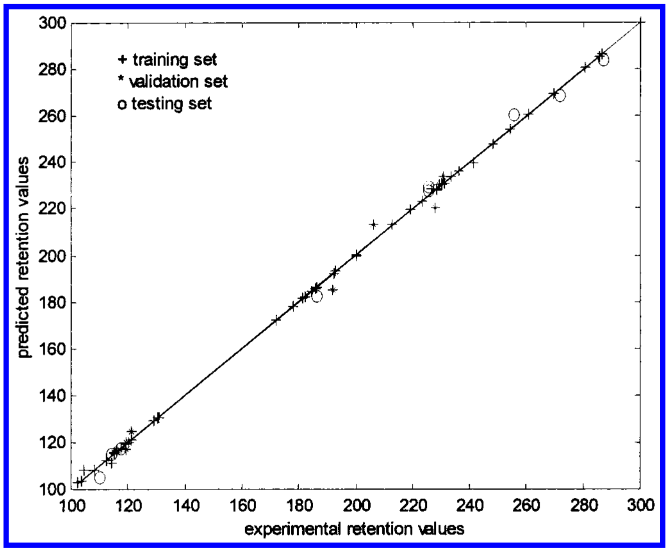

Figure 4. Predicted retention results versus experimental data.

4.3. The Results with RBF Neural Network. From the above discussion, the radius of hidden layer nodes was fixed to 0.38. The predicted results of the optimal neural network are shown in Table 5 and Figure 4. (The mean relative errors of the training set, the validation set, and the testing set are 0.20%, 2.03%, and 1.56% respectively, and the corresponding correlation coefficients (r) are 1.000, 0.996, and 0.999. The mean-absolute errors are respectively 0.29, 3.62, and 2.94.)

In Table 5, we can remark that the relative errors of nos. 12, 25, 26, 33, and 65 solutes are relatively important. It may be due to CH_3, since all substituents used in this work could form a hydrogen bond with the amino bonded stationary phase except CH_3. This is probably the reason of the increase of relative errors for these compounds, the training of the network is probably dominated by the solutes that can form hydrogen-bonded with the stationary phase.

5. CONCLUSION

This study of QSAR model of NBA shows that the RBFNN is a very promising tool for the nonlinear approximation. The training and optimization are easier and faster compared with BP neural networks, because there is only one adjustable parameter. The predictive results are consistent with the experimental data. The mean relative error is 1.6%. Therefore it is a good approach for predicting the expected physicochemical parameters of molecules.

ACKNOWLEDGMENT

This work was supported by the Ministry of Sciences and Technology of China and the Ministry of Foreign Affairs of France (AFCRST PRA SI00-05).

REFERENCES AND NOTES

(1) Ośmialowski, K.; Halkiewicz, J.; Radecki, A.; Kalizan, R. Quantum Chemical Parameters in correlation Analysis of Gas−Liquid Chromatographic Retention Indices of Amines. *J. Chromatogr*. **1985**, *346*, 53−60.

(2) Koopmans, R. E.; Rekker, R. F. Relationship with Lipophilicities as Determined From Octanol−Water Partition Coefficients or Calculated From Hydrophobic Fragmental Data and Connectivity Indices; Lipophilicity Predictions for Polyaromatics. *J. Chromatogr*. **1984**, *285*, 267−279.

(3) Jandera, P. Correlation of Retention and Selectivity of separation in Reversed-Phase High-Performance Liquid Chromatography with Interaction Indices and with Lipophilic and polar Structural Indices. *J. Chromatogr*. **1993**, *656*, 437−467.

(4) Braumann, T.; Weber, G.; Grimme, L. H. Reversed-Phase Liquid Chromatographic Retention Parameter, log k_w, versus Liquid−Liquid Partition Coefficient as a Model of Hydrophobicity of Phenylureas, *s*-Triazines and Phenoxycarbonic Acid Derivatives. *J. Chromatogr*. **1983**, *261*, 329−343.

(5) Exner, O. *Correlation Analysis of Chemical Data*; Plenum Press: New York, 1988.

(6) Hanai, T. Structure-Retention correlation in Liquid Chromatography. *J. Chromatogr*. **1991**, *550*, 313−324.

(7) Karelson, M.; Lobanov, V. S. Quantum-Chemical Descriptors in QSAR/QSPR Studies. *Chem. Rev*. **1996**, *96*, 1027−1043.

(8) Zhang, R. S.; Yan, A. X.; Liu M. C.; Hu Z. D. Application of Artificial Neural Networks for Prediction of the Retention Indices of Alkylbenzenes. *Chemom. Intell. Lab. Syst*. **1999**, *45*, 113−120.

(9) Yan, A. X.; Zhang, R. S.; Liu, M. C.; Liu, H.; Hu, Z. D.; Hooper, M. A.; Zhao, Z. F. Large Artificial Neural Networks Applied to the Prediction of Acyclic and cyclic Alkanes, Alkenes, Alcohols, Ester, Ketones and Esters. *Comput. Chem*. **1998**, *22* (5), 405−412.

(10) Guo, W. Q.; Lu, Y.; Zheng, X. M. The Prediction Study for Chromatographic Retention Index of Saturated Alcohols by MLR and ANN. *Talanta* **2000**, *51*, 479−488.

(11) Tetteh, J.; Suzuki, T.; Metcalfe, E.; Howells, S. Quantitative Structure−Property Relationship for the Estimation of Boiling Point and Flash Point Using a Radial Basis Function Neural Network. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 491−507.

(12) 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.

(13) Wasserman, P. D. *Advanced Methods in Neural Computing*; Van Nostrand Reinhold: New York, 1993.

(14) Kaliszan, R. *Quantitative Structure-Chromatographic Retention Relationships*; John Wiley & Sons: New York, 1987.

(15) Dearden J. C.; Ghafourian T. Hydrogen Bonding Parameters for QSAR: Comparison of Indicator Variables, Hydrogen Bond Counts, Molecular Orbital and Other Parameters. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 231−235.

(16) Ounnar, S.; Righezza, M.; Chrétien, J. R. Factor Analysis in Normal Phase Liquid Chromatography of N-benzylideneanilines. *J. Liq. Chrom., Relat. Technol*. **1998**, *21*(13), 2017−2037.

(17) Sarle, W. S. Ai-FAQ/Neural-nets/Part3 ftp://ftp.sas.com/pub/neural/FAQ3.html.

(18) Shaffer, R. E.; Rose-Pehrsson, S. L. Improved Probabilistic Neural Network Algorithm for Chemical Sensor Array Pattern Recognition. *Anal. Chem*. **1999**, *71*, 4263−4271.

CI010067L