Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Chromatography A, 1048 (2004) 233–243
JOURNAL OF CHROMATOGRAPHY A
www.elsevier.com/locate/chroma

Study of quantitative structure–mobility relationship of carboxylic and sulphonic acids in capillary electrophoresis

Chunxia Xue[a], Huanxiang Liu[a], Xiaojun Yao[a,b], Mancang Liu[a], Zhide Hu[a,*], Botao Fan[b]

[a] *Department of Chemistry, Lanzhou University, Lanzhou, Gansu Province 730000, China*
[b] *Université Paris 7-Denis Diderot, ITODYS 1, Rue Guy de la Brosse, 75005 Paris, France*

Received 16 April 2004; received in revised form 2 July 2004; accepted 15 July 2004

Abstract

A quantitative structure–mobility relationship (QSMR) was developed for the absolute mobilities of 115 carboxylic and sulphonic acids in capillary electrophoresis based on the descriptors calculated from the structure alone. The heuristic method (HM) and radial basis function neural networks (RBFNN) were utilized to construct the linear and nonlinear prediction models, respectively. The prediction results were in agreement with the experimental values. The HM model gave an root-mean-square (RMS) error of 3.76 electrophoretic mobility units for the training set, 5.59 for the test set, and 4.19 for the whole data set, while the RBFNN gave an RMS error of 1.78, 2.04, and 1.83, respectively. The heuristic linear model could give some insights into the factors that are likely to govern the mobilities of the compounds, however, the prediction results of the RBFNN model seem to be better than that of the heuristic method.

Keywords: Heuristic method; Quantitative structure-mobility relationship; Radial basis function neural networks; Electrophoretic mobility; Carboxylic acids; Sulphonic acids

1. Introduction

1.1. Summary of the calculation/prediction of electrophoretic mobility in CE

Capillary electrophoresis (CE) has become an important separation technique in analytical chemistry. This technique has been used to separate analytes ranging from small inorganic and organic ions to macromolecular species such as DNA and proteins. 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 alternative to other analytical methods like high performance liquid chromatography (HPLC). Electrophoretic mobility is the most important parameter governing the separation of solutes in CE. According to Born's model (see [1]), the mobility of an ion (μ) can be expressed by:

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

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 [1]. 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 [2].

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

* Corresponding author. Tel.: +86 931 891 2578; fax: +86 931 891 2582.
E-mail address: snowmoun@21cn.com (Z. Hu).

0021-9673/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2004.07.043

differences in the analyte's mobility and any attempt to provide a computation method to calculate the mobility in a certain practical conditions could provide a useful tool for faster method optimization process in CE. Therefore, developing theoretical models to predict the electrophoretic behaviour of analytes are necessary. However, only a few reports have investigated the quantitative correlation between the molecular parameters and the responses obtained in CE. Because the number of such studies is limited and to provide a summary of them, a brief review was presented here.

The computational methods used in the previous studies to calculate/predict electrophoretic mobility can be classified into two categories. One approach is to use mathematical equation to correlate electrophoretic mobility with the molecular parameters [2–7].

An approximation of mobility was proposed for peptides in Offord's equation [3]:

$$\mu = \frac{Cq}{M^{2/3}} \tag{2}$$

where C is a proportionality constant, q the charge on the analyte and M is the molecular mass of the ion.

Based on three assumptions: (1) the molecular mass M can be treated as composed of effective mass E and hidden mass H, $M = E + H$; (2) the mobility μ is proportional to the charge q and inversely proportional to the effective mass to the power 2/3; and (3) both E and H are additive functions of composition, Wroński proposed the equation [4]:

$$\mu = 668qE^{-2/3} \tag{3}$$

The mobilities of about 200 carboxylic, sulphonic, and amino acids calculated by Eq. (3) yields an average relative error of ±1.5% for organic acids and ±0.8% for peptides.

Fu and Lucy predicted the absolute mobility of 34 aliphatic amines and other 20 monoamines by the equation [2]:

$$\mu_0 = \frac{6.39 \times 10^{-3}}{M_r^{0.620} + 0.221H_w - 0.157S} \tag{4}$$

where M_r is the molecular mass of the analyte, H_w represents the mean water of hydration calculated by McGowan's fragment addition method and S is an empirical shape index. The best model yields an average prediction error of 4.1% for 34 aliphatic amines, 7.2% for 7 monoamines, and 3.3% for 13 monoamines, respectively.

Fu et al. then developed an equation based on the Max Born's model for calculating the absolute mobility of 34 amines and 15 aliphatic carboxylates employing molar volume (V) and acid/base dissociation constant (pK) [5]. The equation is formulated as:

$$\mu_0 = \frac{C_0}{V^{C_1} + C_2 pK} \tag{5}$$

where C_0–C_2 are the model constants. The obtained average errors for amines and aliphatic carboxylates are 4.1 and 3.7%, respectively based on correlative equations.

Li et al. used Pitts equation to account for the effect of ionic strength on the mobilities in capillary zone electrophoresis using 55 carboxylates, phenols, sulfonates, and monoamines as examples [6]. The influence was represented as follows:

$$\mu = \mu_0 - Cq\frac{\sqrt{I}}{1 + 2.4\sqrt{I}} \tag{6}$$

where μ_0 is the infinite dilution mobility of the ion, C a constant, q the charge on the solute anion, and l is the ionic strength of the buffer.

Li and Lucy correlated the absolute electrophoretic mobility of 39 aromatic carboxylates and sulfonates with models incorporating both hydrodynamic and dielectric frictions represented as follows [7]:

$$\mu_0 = \frac{17.3 \times 10^{-4}q}{V^{1/3} + 44.6\,(z^2/V)} \tag{7}$$

and

$$\mu_0 = \frac{18.9 \times 10^{-4}q}{f/f_0 V^{1/3} + 40.8\,(z^2/V)} \tag{8}$$

where q is the charge number, V the molecular volume, and f/f_0 is a frictional correction ratio. The obtained percentage errors for Eqs. (7) and (8) are 4.4 and 4.0%, respectively.

The other methods are more empirically based on quantitative structure-property relationship (QSPR) approaches using the techniques such as multiple linear regression, artificial neural network [8–13], and support vector machine [14].

Multiple linear regression (MLR) technique was used to establish the models for predicting the mobilities of 13 flavonoids by means of the topological indices and 23 models of two-indices were generated with similar statistical results ($R \geq 0.93$, relative standard error $\leq 10\%$) in the work of Liang et al. [8].

A comparative study between MLR and artificial neural networks (ANNs) has been carried out employing electrophoretic mobility of 13 sulfonamides by Jalali-Heravi and Garkani-Nejad [9]. The linear models they proposed were represented as follows:

$$\mu_e = C_0 + C_1\Delta H_f + C_2\,\mathrm{PPCH} + C_3\,\mathrm{SA} \tag{9}$$

and

$$\mu_e = C_4 + C_5\Delta H_f + C_6\,\mathrm{PPCH} + C_7 pK \tag{10}$$

where μ_e is the effective electrophoretic mobility, ΔH_f the heat of formation of anions, PPCH denotes maximum positive partial charge on the anions, SA represents the surface area, pK is p-function of dissociation constant, and C_0–C_7 are the model constants. A non-linear 3-4-2 ANN was generated using the three descriptors of ΔH_f, PPCH and SA as inputs for the anionic sulfonamides and a 3-6-1 ANN was generated using ΔH_f, PPCH and pK for the cationic sulfonamides. The authors concluded that the ANN model shows the superiority over the MLR model for sulfonamide data.

In the same way, Jalali-Heravi and Garkani-Nejad developed models for the prediction of the electrophoretic mobilities of 31 isomeric alkyl- and alkenylpyridines in capillary electrophoresis [10]. The three descriptors reciprocal of Van der Waals radius of the molecules ($RVDW^{-1}$), principal moment of inertia of the molecules around the x axis (MO_x), and dipole moment of the molecules (DIMO) which are selected by the MLR technique were used to as inputs for the ANN. The neural network is a fully connected back-propagation model with a 3-6-1 architecture. Standard error of training and prediction are 6.28 and 5.11%, respectively, for the MLR model and 1.03 and 1.20%, respectively, for the ANN model.

In the previous work of our laboratory, Li et al. constructed a model to estimate the electrophoretic mobilities of 56 aliphatic carboxylates and amines by means of a multilayer neural network using extended delta-bar-delta (EDBD) algorithm [11]. The molecular mass (W), molecular volume (V), the code (+1 represent basic solute and −1 represent acid solute) of acid and base and pK value were used as input parameters of the neural networks. The average absolute prediction errors in the training, validation and test sets are 2.19, 3.65 and 3.22%, respectively.

Jouyban proposed a QSPR model to calculate the logarithm of the electrophoretic mobility of five data sets using structural descriptors computed by HyperChem software [12]. The proposed model is:

$$\ln \mu = C_0 + C_1\,\mathrm{PQ} + C_2 V^{2/3} + C_3\,\mathrm{TE} + C_4\,\Delta H_\mathrm{f} + C_5\,\mathrm{MR} \quad (11)$$

where PQ is partial charge, $V^{2/3}$ denotes surface area, TE stands for total energy, ΔH_f represents heat of formation, MR is molecular refractivity and C_0–C_5 are the model constants which are calculated using a least squares analysis. The absolute average relative deviation values for predicting of electrophoretic mobilities of 10 beta-blockers, 26 benzoates, 11 non-steroidal anti-inflammatory drugs, 13 sulfonamides, and 18 amines are 1.0, 2.1, 0.8, 0.6, and 2.7%, respectively.

Wang et al. in our laboratory studied the relationship between the relative mobility of a group of 19 chlorophenols in different buffers modified by eight kinds of different organic additives in capillary zone electrophoresis by means of MLR and radial basis function neural networks [13]. The linear relationship was represented as follow:

$$\mu_\mathrm{r} = -19.718 + 0.208\mathrm{GS} + 1.276\mathrm{HE} - 1.377\mathrm{DIP} - 0.183\mathrm{AP} + 2.316E_\mathrm{HOMO};$$
$$n = 152,\ F = 207.26,\ \mathrm{S.E.} = 1.38 \quad (12)$$

where μ_r is the relative mobility, GS approximate molecular surface area, HE hydration energy, DIP dipole moment, AP the polarity of the organic additives, and E_HOMO is the energy of the highest occupied molecular orbital. The nonlinear RBFNN model gives the correlation coefficient R of 0.986 for the training set and 0.980 for the test set.

In our latest work in this area, MLR, RBFNN, and the support vector machines (SVM) were used to develop a predictive model for the absolute mobility of 58 aliphatic and aromatic carboxylic acids based on four molecular descriptors calculated from the structure alone [14]. The selected four descriptors are average bonding information content (order 0) (ABIC0), ZX Shadow/ZX Rectangle (ZXS/ZXR), count of H-donors sites [Zefirov's PC] (CHDS), and refractivity (REF). The correlation coefficient in absolute mobility predictions for the whole data set given by MLR, RBFNN, and SVM are 0.947, 0.960, and 0.984, respectively.

Of those previous studies that aimed at predicting the electrophoretic mobility, the most promising method is to use QSPR approach. QSPR methods have been successfully used to predict a variety of physical, chemical, and biological properties of compounds. The advantage of this approach over other methods lies in the fact that the descriptors used can be calculated from 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 they are 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 useful tool to predict the electrophoretic mobilities avoiding long and tedious separation optimization. QSMR study can also tell us which of the structural factors may play an important role in the determination of the absolute mobility of the compound.

1.2. Techniques used and the aims of the present work

One of the important problems for the QSPR applications is the numerical representation (often called molecular descriptor) of the chemical structure. The built model performance and the accuracy of the results are strongly dependent on the way the structural representation was performed. Various numerical representations of the compounds were proposed in QSPR studies: constitutional and topological descriptors; numerical code; quantum chemistry descriptors, etc. The Software CODESSA, developed by Katritzky group, enables the calculation of a large number of quantitative descriptors based solely on the molecular structural information [15–17] and codes this chemical information into mathematical form. CODESSA combines diverse methods for quantifying the structural information about the molecule with advanced statistical analysis to establish molecular structure–property/activity relationships. CODESSA has been applied successfully in a variety of QSPR analyses [18–21].

Neural networks have been applied to QSPR analysis since the late 1980s due to its flexibility in modeling nonlinear problem, mainly in response to increased accuracy demands. They have been widely used to predict many physicochemical properties. There exist many models of neural networks, which have different approaches both in architecture and in learning algorithms. The most popular neural network model

Table 1
The compounds and the predicted electrophoretic mobilities (10^{-5} cm^2 s^{-1} V^{-1})

No.	Compounds	Experimental[a]	Calculated$_{HM}$[b]	Abs error[c]	Calculated$_{RBFNN}$[d]	Abs error[c]
1	Fluoroacetic acid	43.9	36.2	7.7	43.8	0.1
2	Trifluoroacetic acid	42.5	42.4	0.1	42.2	0.3
3	Chloroacetic acid	41.9	39.8	2.1	39.4	2.5
4	Dichloroacetic acid	39.4	39.7	0.3	39.7	0.3
5[e]	Trichloroacetic acid	36.2	40.7	4.5	35.5	0.7
6	3-Chloropropionic acid	36.8	37.0	0.2	39.2	2.4
7	2-Chlorobutyric acid	32.8	35.0	2.2	32.8	0.0
8	5-Chlorovaleric acid	30.8	32.2	1.4	31.0	0.2
9	Bromoacetic acid	38.8	38.0	0.8	40.7	1.9
10[e]	2-Bromopropionic acid	33.4	38.3	4.9	36.7	3.3
11	2-Bromobutyric acid	30.8	34.8	4.0	32.4	1.6
12	4-Bromobutyric acid	32.8	32.3	0.5	31.9	0.9
13	5-Bromovaleric acid	30.8	31.7	0.9	30.8	0.0
14	2,3-Dibromopropionic acid	32.3	33.7	1.4	31.9	0.4
15[e]	Tribromoacetic acid	34.9	36.1	1.2	35.5	0.6
16	Iodoacetic acid	40.2	38.8	1.4	39.9	0.3
17	3-Iodopropionic acid	34.9	37.7	2.8	35.4	0.5
18	4-Iodobutyric acid	32.9	32.6	0.3	32.0	0.9
19	5-Iodovaleric acid	30.8	32.0	1.2	30.9	0.1
20[e]	3,4-Dibromofluoroacetic acid	36.9	25.3	11.6	39.4	2.5
21	Chlorodibromoacetic acid	34.9	38.4	3.5	35.0	0.1
22	Glycolic acid	42.3	43.0	0.7	40.4	1.9
23	Lactic acid	36.5	41.3	4.8	40.7	4.2
24	2-Hydroxybutyric acid	34.2	36.6	2.4	34.6	0.4
25[e]	Glyceric acid	36.3	40.2	3.9	38.2	1.9
26	Glucuronic acid	26.6	32.3	5.7	26.2	0.4
27	Gluconic acid	27.2	31.5	4.3	27.7	0.5
28	2-Chloro-3-hydroxybutyric acid	32.9	34.5	1.6	34.6	1.7
29	Glyoxalic acid	37.8	48.5	10.7	38.2	0.4
30[e]	Pyruvic acid	40.4	51.3	10.9	40.9	0.5
31	Trichlorolactic acid	34.2	36.6	2.4	34.6	0.4
32	Maleic acid	62.0	64.0	2.0	62.5	0.5
33	Fumaric acid	61.2	64.7	3.5	62.6	1.4
34	Tartaric acid	60.5	57.5	3.0	59.9	0.6
35[e]	Citric acid	70.8	60.7	10.1	71.0	0.2
36	2-Ketoglutaric acid	59.0	54.6	4.4	58.9	0.1
37	Malic acid	59.0	52.7	6.3	60.2	1.2
38	Thiomalic acid	58.5	49.2	9.3	54.2	4.3
39	2,3-Dimercaptopropanesulphonic acid	34.4	34.3	0.1	35.2	0.8
40[e]	2-Hydroxyethanesulphonic acid	39.6	41.6	2.0	41.0	1.4
41	Cyclobutane-1,1-dicarboxylic acid	51.1	55.0	3.9	56.1	5.0
42	Cyclopentane-1,1-dicarboxylic acid	50.0	53.8	3.8	51.1	1.1
43	Cyclohexane-1,1-dicarboxylic acid	48.0	53.2	5.2	48.3	0.3
44	Methylmalonic acid	58.5	54.1	4.4	58.2	0.3
45[e]	Methylethylmalonic acid	50.0	49.9	0.1	52.3	2.3
46	Propylmalonic acid	52.0	45.7	6.3	48.9	3.1
47	Diethylmalonic acid	49.5	48.5	1.0	48.6	0.9
48	Ethylpropylmalonic acid	47.0	44.7	2.3	45.9	1.1
49	Dipropylmalonic acid	46.0	44.3	1.7	44.4	1.6
50[e]	Oxaloacetic acid	56.0	52.7	3.3	53.0	3.0
51	3-Propylglutaric acid	47.0	43.4	3.6	47.6	0.6
52	Benzoic acid	34.4	34.2	0.2	34.5	0.1
53	Benzenesulphonic acid	38.7	39.2	0.5	35.1	3.6
54	*p*-Toluenesulphonic acid	31.1	31.3	0.2	34.0	2.9
55[e]	*o*-Aminoenzoic acid	31.6	27.4	4.2	30.3	1.3
56	Sulphonic acid	33.7	33.1	0.6	35.2	1.5
57	*p*-Fluorobenzoic acid	33.4	29.5	3.9	33.3	0.1
58	*p*-Chlorobenzoic acid	33.4	28.5	4.9	33.7	0.3
59	*m*-Iodobenzoic acid	33.4	24.8	8.6	30.0	3.4
60[e]	*p*-Bromobenzoic acid	31.5	28.0	3.5	34.5	3.0
61	*p*-Nitrobenzoic acid	32.1	30.6	1.5	30.6	1.5
62	3,5-Dinitrobenzoic acid	29.5	31.9	2.4	30.6	1.1
63	*p*-Toluic acid	29.1	28.1	1.0	30.7	1.6

Table 1 (*Continued*)

No.	Compounds	Experimental[a]	Calculated$_{HM}$[b]	Abs error[c]	Calculated$_{RBFNN}$[d]	Abs error[c]
64	*p*-Ethylbenzoic acid	26.5	24.7	1.8	27.4	0.9
65[e]	2,3-Dimethylbenzoic acid	27.1	27.6	0.5	29.2	2.1
66	*o*-Isopropylbenzoic acid	24.7	27.5	2.8	28.0	3.3
67	2,4,6-Trimethylbenzoic acid	24.7	30.0	5.3	29.0	4.3
68	*p-tert*-Butylbenzoic acid	23.2	18.6	4.6	23.4	0.2
69	*p*-Hydroxybenzoic acid	34.0	33.6	0.4	34.7	0.7
70[e]	Salicylic acid	35.4	38.5	3.1	34.9	0.5
71	2,4-Dihydroxybenzoic acid	32.0	37.8	5.8	31.1	0.9
72	3,4-Dihydroxybenzoic acid	34.4	32.7	1.7	33.5	0.9
73	Gallic acid	34.4	35.7	1.3	34.5	0.1
74	*p*-Methoxybenzoic acid	28.3	27.7	0.6	30.1	1.8
75[e]	*p*-Ethoxybenzoic acid	26.6	23.7	2.9	27.7	1.1
76	2-Nitro-3-bromobenzoic acid	28.2	31.5	3.3	29.3	1.1
77	2-Nitro-3-chlorobenzoic acid	31.3	33.1	1.8	32.8	1.5
78	Phenol	34.4	39.4	5.0	31.5	2.9
79	*p*-Nitrophenol	33.4	30.7	2.7	34.2	0.8
80[e]	2,4-Dinitrophenol	31.3	40.6	9.3	29.5	1.8
81	Picric acid	31.5	35.9	4.4	31.1	0.4
82	*p*-Chlorophenol	33.4	33.1	0.3	33.1	0.3
83	2,4-Dichlorophenol	31.3	27.3	4.0	28.6	2.7
84	Vanillic acid	27.1	24.9	2.2	25.7	1.4
85[e]	Cinnamic acid	28.3	39.7	11.4	32.0	3.7
86	Phenylacetic acid	31.7	34.0	2.3	31.6	0.1
87	Phenoxyacetic acid	27.8	33.1	5.3	33.9	6.1
88	Nicotinic acid	34.6	38.8	4.2	33.9	0.7
89	2-Naphthalenesulphonic acid	31.3	25.4	5.9	29.2	2.1
90[e]	Acetic acid	42.4	43.8	1.4	45.7	3.3
91	Propionic acid	36.9	37.6	0.7	38.5	1.6
92	Butyric acid	33.7	33.4	0.3	32.4	1.3
93	Valeric acid	31.6	31.9	0.3	29.3	2.3
94	Hexanoic acid	30.2	31.2	1.0	27.6	2.6
95[e]	Heptanoic acid	28.4	31.4	3.0	26.6	1.8
96	Octanoic acid	27.4	30.8	3.4	26.3	1.1
97	Nonanoic acid	26.7	31.1	4.4	26.7	0.0
98	Oxalic acid	74.6	63.3	11.3	74.3	0.3
99	Malonic acid	66.0	62.3	3.7	62.9	3.1
100[e]	Succinic acid	60.3	62.2	1.9	61.4	1.1
101	Glutaric acid	55.6	54.3	1.3	55.9	0.3
102	Adipic acid	52.4	51.4	1.0	52.3	0.1
103	Pimelic acid	49.9	48.3	1.6	50.0	0.1
104	Suberic acid	47.2	45.6	1.6	47.5	0.3
105[e]	Azelaic acid	45.9	44.5	1.4	46.4	0.5
106	Seacic acid	44.9	41.6	3.3	44.4	0.5
107	Methanesulphonic acid	50.5	57.2	6.7	51.2	0.7
108	Ethanesulphonic acid	42.7	45.5	2.8	41.4	1.3
109	Propanesulphonic acid	37.5	36.9	0.6	34.9	2.6
110[e]	Butanesulphonic acid	33.9	35.9	2.0	33.0	0.9
111	Pentanesulphonic acid	31.4	32.5	1.1	30.7	0.7
112	Hexanesulphonic acid	29.4	27.4	2.0	28.5	0.9
113	Octanesulphonic acid	26.2	24.0	2.2	26.4	0.2
114	Nonanesulphonic acid	25.1	20.4	4.7	25.7	0.6
115[e]	Dodecanesulphonic acid	22.3	21.1	1.2	25.1	2.8

[a] Experimental mobility.
[b] Predicted mobility by HM.
[c] Absolute value of (calculated − experimental).
[d] Predicted mobility by RBFNN.
[e] Compounds in the test set.

is the back-propagation (BP) neural networks due to its simple architecture yet powerful problem-solving ability. However, the BP neural network suffers from a number of weaknesses, including the need for a large number of controlling parameters, difficulty in obtaining a stable solution. However, the radial basis function neural network (RBFNN) has some advantages such as short training times, few free parameters to be adjusted by fast linear methods. The optimization of its topology and learning parameters are easy to implement [22]. Many problems in chemistry and chemical engineering have been successfully solved by the use of RBFNN [23–28].

In the present work, the CODESSA program was used for the calculation of the descriptors and for the statistical analysis to obtain multi-parameter QSMR equations describing the absolute mobilities of 115 carboxylic and sulphonic acids. The heuristic method and RBFNN were utilized to establish quantitative linear and nonlinear relationship between the electrophoretic mobility and the molecular structure, respectively. The aim of the present study was to establish a QSMR model that could be used for the prediction of absolute mobilities of carboxylic and sulphonic acids from their molecular structures alone, and at the same time, to seek for the important structural features related to the electrophoretic mobility of these compounds. Compared with previous work, the data set used in our investigation is more diverse and the models developed are more general and practical.

2. Method

2.1. Data set

The values of the absolute mobilities of 115 carboxylic and sulphonic acids studied were taken from the work of Wroński [4]. Table 1 contained the absolute mobility of the data set, in 10^{-5} cm^2 s^{-1} V^{-1}. The compounds consist of 100 carboxylic and 15 sulphonic acids with various groups, heteroatoms and structural isomers. Through the comparison of experimental values of the absolute mobilities of the carboxylates (Nos. 90–95, 97) and sulphonates (Nos. 107–113), it could be seen that the absolute mobilities of the sulphonates are almost greater than that of the corresponding carboxylates, except for octanesulphonic acid (No. 113) and nonanoic acid (No. 97). Of 100 carboxylic acids, there are 73 monofuntional, 26 difunctional, and 1 trifunctional. The compounds contain 77 aliphatic and 38 aromatic acids. The data set was split randomly into a 92 member training set and an external prediction set of 23 compounds. Of the training set, there are 61 aliphatic, 31 aromatic acids, 70 monofunctional, 22 difunctional acids. Of the test set, there are 16 aliphatic, 7 aromatic acids, 18 monofunctional, 4 difunctional, and 1 trifunctional acids. 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 used to prevent the network from overfitting.

2.2. Descriptor calculation

The structures of the compounds were drawn with HyperChem 4.0 programme [29] and exported in a file format suitable for MOPAC. The geometry optimization was performed with the semiempirical quantum method PM3 [30] using the MOPAC 6.0 program [31]. All the geometries had been fully optimized without symmetry restrictions. In all cases frequency calculations have been performed in order to ensure that all the calculated geometries correspond to true minima. The MOPAC output files were used by the CODESSA program to calculate five classes of descriptors: constitutional (number of various types of atoms and bonds, number of rings, molecular weight, etc.); topological (Wiener index, Randic indices, Kier-Hall shape indices, etc.); geometrical (moments of inertia, molecular volume, molecular surface area, etc.); electrostatic (minimum and maximum partial charges, polarity parameter, charged partial surface area descriptors, etc.); and quantum chemical (reactivity indices, dipole moment, HOMO and LUMO energies, etc.).

2.3. The heuristic method

The heuristic multilinear regression procedures available in the framework of the CODESSA program were used to perform a complete search for the best multilinear correlations with a multitude of descriptors. These procedures provide collinearity control (i.e., any two descriptors intercorrelated above 0.8 are never involved in the same model) and implement heuristic algorithms for the rapid selection of the best correlation, without testing all possible combinations of the available descriptors. The heuristic method of the descriptor selection proceeds with a pre-selection of descriptors by eliminating (i) those descriptors that are not available for each structure, (ii) descriptors having a small variation in magnitude for all structures, (iii) descriptors that give a F-test's value below 1.0 in the one-parameter correlation, and (iv) descriptors whose t-values are less than the user-specified value, etc. This procedure orders the descriptors by decreasing correlation coefficient when used in one-parameter correlations. The next step involves correlation of the given property with (i) the top descriptor in the above list with each of the remaining descriptors and (ii) the next one with each of the remaining descriptors, etc. The best pairs, as evidenced by the highest F-values in the two-parameter correlations, are chosen and used for further inclusion of descriptors in a similar manner.

The goodness of the correlation is tested by the coefficient regression (R^2), the F-test (F), the standard deviation (s). The stability of the correlations was tested against the cross-validated coefficient, R^2_{cv}. The R^2_{cv} describes the stability of a regression model obtained by focusing on the sensitivity of the model to the elimination of any single data point. Briefly, for each data point, the regression is recalculated with the same descriptors but for the data set without this point. The obtained regression is used to predict the value

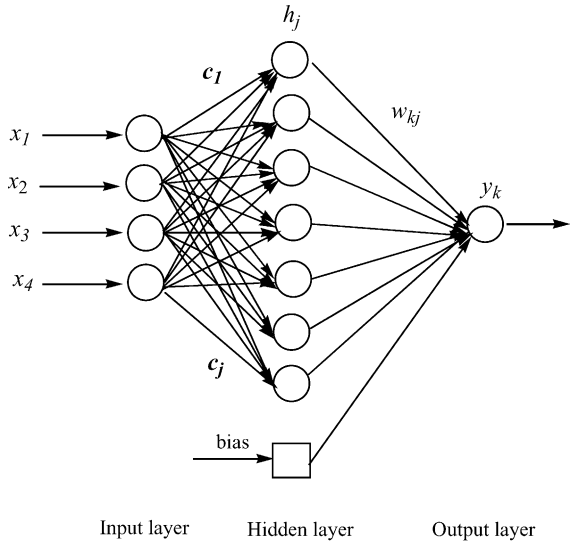

Fig. 1. The typical architecture of the RBFNN.

of this point, and the set of estimated values calculated in this way is correlated with the experimental values.

The Heuristic method usually produces correlations 2–5 times faster than other methods, with comparable quality [21]. The rapidity of calculations from the heuristic method renders it the first method of choice in practical research. Thus, we used this method for our calculations.

2.4. Radial basis function neural networks theory

The theory of RBFNN has been adequately described in Refs. [23–28]. Here only a brief description of RBFNN principle was given. RBFNN can be described as a three-layer feedforward structure, as presented schematically in Fig. 1.

The RBFNN consists of three layers: the input layer, the hidden layer and the 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 functions as a nonlinear transfer function to operate on the input data. In general, there are several radial basis functions (RBFs): linear, cubic, thin plate spline, Gaussian, multi-quadratic and inverse multi-quadratic. The most often used RBF is a Gaussian function that is characterized by a center (c_j) and width (r_j). In this study, the Gaussian was selected as the radial basis functions. The nonlinear transformation with RBF in the hidden layer is given as follow:

$$h_j(x) = \exp\left(\frac{-||x - c_j||^2}{r_j^2}\right) \tag{13}$$

in which h_j is the notation for the output of the jth RBF unit, c_j and r_j are the center and width of the jth RBF, respectively. The operation of the output layer is linear, which is given in Eq. (14):

$$y_k(x) = \sum w_{kj} h_j(x) + b_k \tag{14}$$

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, h_j is the notation for the output of the jth RBF unit, and b_k is the bias.

RBFNN are completely specified by choosing the number units n_k of radial basis functions, the center c_j and width r_j of each radial basis function, and the connection weights w_{kj} between jth hidden layer unit and kth output unit. In this paper, the forward subset selection routine proposed by Orr [32,33] was used to select the centers from training set samples. The advantages of this selection is that it can determine the number of the hidden layer units simultaneously and there is no need to fix the number of the hidden layer units in advance. The adjustment of the connection weight between the hidden layer and the output layer is performed using a least-squares solution after the selection of centers and width of radial basis functions.

The overall performance of RBFNN was evaluated in terms of root-mean-square (RMS) error which was defined as below:

$$\mathrm{RMS} = \sqrt{\frac{\sum_{i=1}^{n_s}(y_k - \hat{y}_k)^2}{n_s}} \tag{15}$$

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

$$\mathrm{AARD} = \frac{100}{n_s}\sum_{i=1}^{n_s}\left(\frac{|y_k - \hat{y}_k|}{\hat{y}_k}\right) \tag{16}$$

In Eqs. (15) and (16), y_k is the desired output, $\hat{y}$ the actual output of the network, and n_s is the number of compounds in analyzed set. To compare the results obtained by HM and RBFNN straightforwardly, the RMS and AARD errors was also calculated in the HM model.

All calculation programs implementing RBFNN were written in M-file based on basis MATLAB script for radial basis function neural networks [32,33]. The scripts were compiled using MATCOM compiler running on a Pentium IV personal computer with 256M RAM.

3. Results and discussion

3.1. Results of the heuristic method

About 600 descriptors were calculated by the CODESSA program for each of the compounds. After the heuristic reduction the pool of descriptors was reduced to 246. A variety of subset sizes was investigated to determine the optimum number of descriptors in a model. When adding another descriptor did not improve significantly the statistics of a model, it was determined that the optimum subset size had been achieved. The influences of the number of the descriptors on the correlation coefficient (R^2) and the standard deviation (s) were shown in Figs. 2 and 3, respectively. From Figs. 2 and 3, it

Table 2
Descriptors, coefficients, standard error, and *t*-values for the linear model[a,b]

Descriptor	Chemical meaning	Coefficient	Error	*t*-test
(Constant)	Intercept	125.18	10.64	11.76
NDB	Number of double bonds	8.37	0.94	8.94
WPSA-3	WPSA-3 weighted PPSA (PPSA3 TMSA/1000) [quantum-chemical PC]	−1.20	0.12	−10.04
AIC2	Average information content (order 2)	−11.72	1.24	−9.42
RNCG	RNCG relative negative charge (QMNEG/QTMINUS) [quantum-chemical PC]	−93.78	13.64	−6.88
HDCA2	HA dependent HDCA-2/TMSA [Zefirov's PC]	1783.00	448.28	3.98
ACIC0	Average complementary information content (order 0)	−9.01	1.79	−5.04
TMEER	Tot molecular 1-center E–E repulsion/no. of atoms	−0.17	0.04	−4.47

[a] R^2 = 0.88; s = 3.93; RMS = 3.76; AARD = 7.89%; n = 92; F = 87.22; R^2_{cv} = 0.83.
[b] Further discussion of the chemical meaning of the descriptors was given in Section 3.1 of the text.

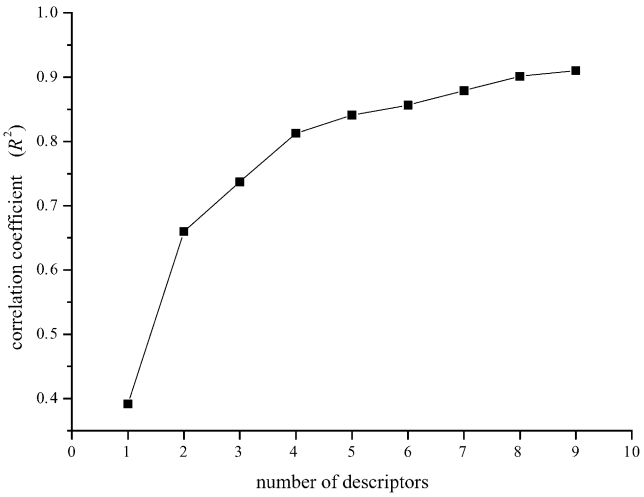

Fig. 2. Influence of the number of descriptors on the correlation coefficient (R^2) of the regression models.

can be seen that seven descriptors appear to be sufficient for a successful regression model. The multilinear analysis of the absolute mobilities values for the 92 compounds of the training set resulted in the seven-parameter model summarized in Table 2, and the correlation matrix of these descriptors was

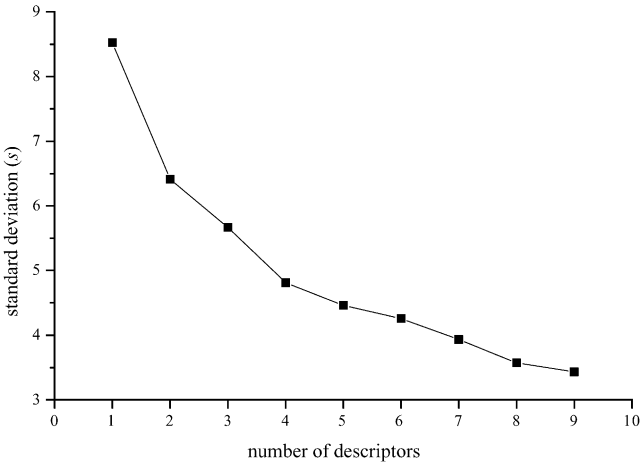

Fig. 3. Influence of the number of descriptors on the standard deviation (s) of the regression models.

Table 3
Correlation matrix of the seven descriptors used in this work[a]

	NDB	WPSA-3	AIC2	RNCG	HDCA2	ACIC0	TMEER
NDB	1.000	0.115	0.064	−0.558	−0.256	0.772	0.204
WPSA-3		1.000	−0.151	−0.677	−0.685	0.397	−0.708
AIC2			1.000	−0.208	0.135	−0.204	0.277
RNCG				1.000	0.588	−0.570	0.182
HDCA2					1.000	−0.354	0.433
ACIC0						1.000	−0.009
TMEER							1.000

[a] The definitions of the descriptors were given in Table 2.

shown in Table 3. The linear correlation coefficient value of each two descriptors is <0.80 (Table 3), which means the descriptors were independent in this multilinear analysis. The values of the 7 selected parameters of each compound are available as Supplementary Data via Sciencedirect (the electronic publication outlet). The obtained model has a correlation coefficient $R^2 = 0.88$, $F = 87.22$, with a standard deviation (s) of 3.93 (10^{-5} cm^2 s^{-1} V^{-1}), and the cross-validated coefficient R^2_{cv} = 0.83. This model gave an RMS error of 3.76 electrophoretic mobility units for the training set, and the corresponding AARD was 7.89%.

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 and sulphonic acids in CE. Due to the diversity of the molecules studied in this work, the electrophoretic mobility of the compounds related to molecular structure in a complex way. Of the seven descriptors, one is constitutional, two are topological, one is electrostatic and three are quantum-chemical descriptors. These descriptors encode different aspects of the molecular structure. As mentioned in Introduction two fundamental frictional factors are found to be important in the electrophoretic mobility of a solute in CE. One is hydrodynamic friction factor, which is related to the molecular size and/or mass of solute, and the other is 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. Average information content (order 2) (AIC2) and average complementary Information content (order 0) (ACIC0) belong to topological descriptors. AIC2 and ACIC0 describe the size, shape and branching information of

the molecules and give some information about the hydrodynamic friction factors. AIC2 and ACIC0 have negative coefficients in the linear model, which indicates that the absolute mobility is inversely proportional to these descriptors. The number of double bonds (NDB), a constitutional descriptor, describes the degree of the delocalization of electron and affects the charge distribution within the molecule. Thus, NDB has some correlation with the dielectric friction. NDB has a positive coefficient in the linear model, which indicates that this structural feature makes positive contribution to the absolute mobility. HA dependent HDCA-2/TMSA [Zefirov's PC] (HDCA2) is an electrostatic descriptor. HDCA2 is hydrogen donor charged solvent-accessible surface area, and this descriptor represents the sum of solvent-accessible surface area of the H-bonding donor atoms. HDCA2 reflects characteristics of the charge distribution of the molecule, so it can affect the dielectric friction term. This descriptor has a positive coefficient in the linear model, which indicates that the absolute mobility is proportional to this descriptor. The three quantum-chemical descriptors are WPSA-3 (weighted PPSA (PPSA3 TMSA/1000) [Quantum-Chemical PC]), RNCG (relative negative charge (QMNEG/QTMINUS) [quantum-chemical PC]), and TMEER1 (tot molecular 1-center E–E repulsion/# of atoms). WPSA-3 is equal to atomic charge weighted partial positive surface area (PPSA3) multiplied by total molecular surface area (TMSA). RNCG is the relative negative charge of the molecule. The electron–electron repulsion energy describes the electron repulsion driven processes in the molecule and may be related to the conformational (rotational, inversional) changes or atomic reactivity in the molecule [34]. The quantum-chemical descriptors such as WPSA-3, RNCG, and TMEER1 represent or depend directly on the quantum-chemically calculated charge distribution in the molecules, and therefore can also explain the dielectric friction contribution in determination of electrophoretic mobility. The three descriptors, WPSA-3, RNCG, and TMEER1, all have negative coefficient in the linear model, which indicates that the absolute mobility is inversely proportional to these descriptors.

From the above discussion, it can be seen that all the descriptors involved in the model have physical meaning, and these descriptors can account for the structural features responsible for the electrophoretic mobilities of the carboxylic and sulphonic acids. According to the *t*-test values (Table 2), the more relevant descriptors are a quantum-chemical descriptor (WPSA-3) and a topological descriptor (AIC2).

With the test set (Table 1) the prediction results were obtained, confirming the predictive capability of the model. The statistical parameters were R^2 = 0.77; F = 71.67; n = 23; s = 5.44. The heuristic model produced an RMS error of 5.59 mobility units for the test set, 4.19 for the whole data set, and the corresponding AARD was 11.84, 8.68%, respectively. Fig. 4 showed a plot of the calculated versus experimental electrophoretic mobilities for all the 115 acids studied, the training set and the test set.

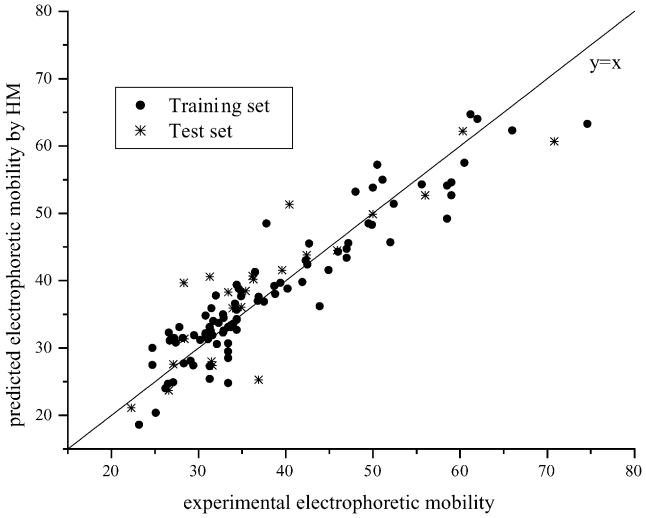

Fig. 4. Predicted vs. experimental electrophoretic mobilities (HM).

3.2. Result of RBFNN

From Table 1 and Fig. 4, it can be seen that the model of the heuristic method was not sufficiently accurate (RMS = 419, AARD = 8.68%) showed the factors influencing the electrophoretic moblility of these compounds were complex and not all of them were linear correlation with the absolute mobility. So, we built the nonlinear prediction model by RBFNN to further discuss the correlation between the molecular structure and the absolute mobility based on the same subset of descriptors. Such a RBFNN can be designed as 7-n_k-1 net to indicate the number of the units in the input, the hidden, and the output layer, respectively. The optimal width was determined by experiments with a number of trials by taking into account of the model selection criterion: a width <1 gives poor prediction ability, varying the width indicates width has little effect on the performance of RBFNN, if width exceeds 5. So we choicely computed the width from 1 to 5, every 0.1. Each minimum error on LOO cross-validation was plotted versus the width (Fig. 5) and the minimum was

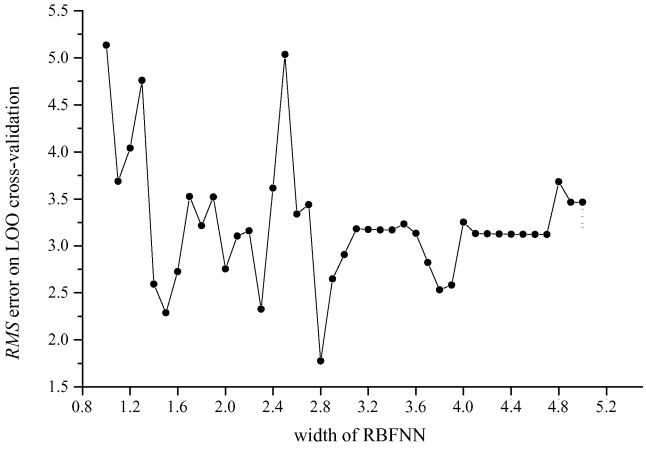

Fig. 5. The width of RBFNN vs. RMS error on LOO cross-validation.

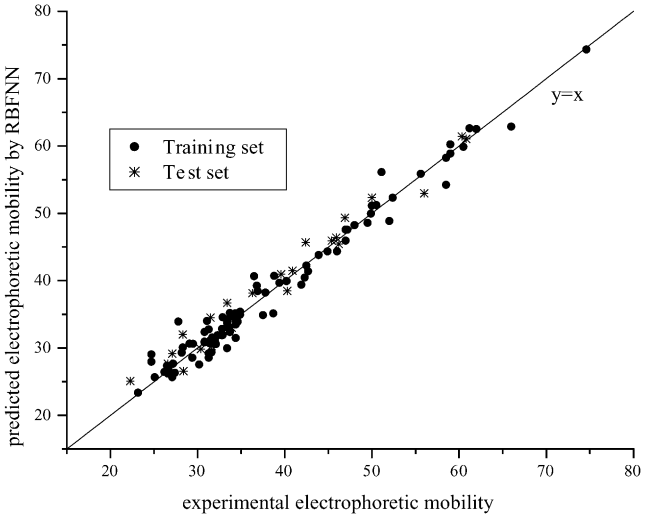

Fig. 6. Predicted vs. experimental electrophoretic mobilities (RBFNN).

chosen as the optimal condition. In this case: r = 2.8 and n_k = 19.

From the best network, the inputs in the test set were presented with it, and the results with RBFNN were obtained. They were shown in Table 1 and Fig. 6. The network gave an *RMS* error of 1.78 electrophoretic mobility units for the training set, 2.04 for the test set, and 1.83 for the whole set, the corresponding correlation coefficients (R^2) were 0.97, 0.97, and 0.97, and the corresponding AARD were 3.54, 5.01, and 3.84%, respectively. Fig. 6 proved that the RBFNN model was statistically stable and fitted the data well.

Analysis of the result obtained indicates that the models we proposed can correctly represent the relationship between the absolute mobilities of carboxylic and sulphonic acids and molecular descriptors calculated solely from molecular structures, moreover, the seven selected descriptors can represent the features of the compounds responsible for their electrophoretic mobility behavior. By comparison of the results from the heuristic method and RBFNN, it can be seen that the RMS and AARD errors of RBFNN are much lower than that of the HM. The performance of RBFNN models seems to be better than that of the heuristic method, which indicates that nonlinear model can simulate the relationship between the structural descriptors and the mobility of the acids more accurately.

4. Conclusion

The heuristic method and the radial basis function neural networks were used to construct the linear and nonlinear quantitative structure–mobility relationships for the prediction of absolute mobility in capillary electrophoresis of 115 carboxylic and sulphonic acids based on the descriptors calculated from the molecular structure alone. Both the linear and nonlinear models provided the satisfactory results, at the same time, the nonlinear RBFNN models produced better results with good predictive ability than linear model, so we can conclude that (1) the proposed linear model by the heuristic method could identify and provide some insight into what structural features are related to the absolute mobility of the carboxylic and sulphonic acids. (2) Nonlinear model can describe accurately the relationship between the structural parameter and the absolute mobilities of the 115 carboxylic and sulphonic acids. (3) RBFNN proved to be a very promising tool in the prediction of electrophoretic mobility. Nonlinear models using RBFNN based on these same sets of descriptors produced even better models with good predictive ability. The training procedure is also simple because there are only two parameters to be optimized: the number of units in the hidden layer and the width of radial basis function. If we add to the increasing accuracy of RBF networks, the lack of difficulty to find an optimum architecture and the almost instant training, it will be easily concluded that RBFNN can be a significant partner for the development of different quantitative structure–mobility relationship system. Additionally, the compounds studied in this work include carboxylic, sulphonic, aliphalic, aromatic, monofunctional, difunctional, and tridunctional acid, etc., which cover almost all types of organic acids. The models we developed are general and applicable for the prediction of the absolute mobility of the organic acids. Furthermore, the proposed approach can also be extended in other QSPR investigation.

Acknowledgements

The authors thank the Association Franco-Chinoise pour la Recherche Scientifique & Technique (AFCRST) for supporting this study (Programme PRA SI 02-02).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.chroma.2004.07.043.

References

[1] R. Kay, Pure Appl. Chem. 63 (1991) 1393.
[2] S.L. Fu, C.A. Lucy, Anal. Chem. 70 (1998) 173.
[3] R.E. Offord, Nature 211 (1966) 591.
[4] M. Wroński, J. Chromatogr. A 657 (1993) 165.
[5] S.L. Fu, D.M. Li, C.A. Lucy, Analyst 123 (1998) 1487.
[6] D.M. Li, S.L. Fu, C.A. Lucy, Anal. Chem. 71 (1999) 687.
[7] D.M. Li, C.A. Lucy, Anal. Chem. 73 (2001) 1324.
[8] H.R. Liang, H. Vuorela, P. Vuorela, M.L. Riekkola, R. Hiltunen, J. Chromatogr. A 798 (1998) 233.
[9] M. Jalali-Heravi, Z. Garkani-Nejad, J. Chromatogr. A 927 (2001) 211.
[10] M. Jalali-Heravi, Z. Garkani-Nejad, J. Chromatogr. A 971 (2002) 207.

[11] Q.F. Li, L.J. Dong, R.P. Jia, X.G. Chen, Z.D. Hu, B.T. Fan, Comput. Chem. 26 (2002) 245.
[12] A. Jouyban, B.H. Yousefi, Comput. Biol. Chem. 27 (2003) 297.
[13] Y.W. Wang, S.L. Gao, Y.H. Gao, S.H. Liu, M.C. Liu, Z.D. Hu, B.T. Fan, Anal. Chim. Acta 486 (2003) 191.
[14] C.X. Xue, R.S. Zhang, M.C. Liu, Z.D. Hu, B.T. Fan, J. Chem. Inf. Comput. Sci. 44 (2004) 950.
[15] A.R. Katritzky, V.S. Lobanov, M. Karelson, Comprehensive Descriptors for Structural and Statistical Analysis, Reference Manual, Version 2.0, 1994.
[16] A.R. Katritzky, V.S. Lobanov, M. Karelson, Chem. Soc. Rev. (1995) 279.
[17] A.R. Katritzky, M. Karelson, V. Lobanov, Pure Appl. Chem. 69 (1997) 245.
[18] E.J. Delgado, J.B. Alderete, G.A. Jaña, J. Chem. Inf. Comput. Sci. 43 (2003) 1928.
[19] R. Hiob, M. Karelson, J. Chem. Inf. Comput. Sci. 40 (2000) 1062.
[20] R. Bosque, J. Sales, J. Chem. Inf. Comput. Sci. 43 (2003) 637.
[21] A.R. Katritzky, R. Petrukhin, R. Jain, M. Karelson, J. Chem. Inf. Comput. Sci. 41 (2001) 1521.
[22] S. Chen, C.F.N. Cowan, P.M. Grant, IEEE Trans. Neur. Net. 2 (1991) 302.
[23] X.J. Yao, X.Y. Zhang, R.S. Zhang, M.C. Liu, Z.D. Hu, B.T. Fan, Comput. Chem. 25 (2001) 475.
[24] X.J. Yao, X.Y. Zhang, R.S. Zhang, M.C. Liu, Z.D. Hu, B.T. Fan, Comput. Chem. 26 (2002) 159.
[25] X.J. Yao, M.C. Liu, X.Y. Zhang, Z.D. Hu, B.T. Fan, Anal. Chim. Acta 462 (2002) 101.
[26] X.J. Yao, Y.W. Wang, X.Y. Zhang, R.S. Zhang, M.C. Liu, Z.D. Hu, B.T. Fan, Chemomet. Intell. Lab. Syst. 62 (2002) 217.
[27] X.J. Yao, X.Y. Zhang, R.S. Zhang, M.C. Liu, Z.D. Hu, B.T. Fan, Talanta 57 (2002) 297.
[28] X.J. Yao, M.C. Liu, X.Y. Zhang, R.S. Zhang, Z.D. Hu, B.T. Fan, Chin. J. Chem. 20 (2002) 722.
[29] HyperChem 4.0, Hypercube, 1994.
[30] M.J.S. Dewar, E.G. Zoebisch, E.F. Healy, J.J.P. Stewart, J. Am. Chem. Soc. 107 (1985) 3902.
[31] J.P.P. Stewart, MOPAC 6.0, Quantum Chemistry Program Exchange, QCPE, No. 455, Indiana University, Bloomington, IN, 1989.
[32] M.J.L. Orr, Introduction to Radial basis function networks, Center for Cognitive Science, Edinburgh University, 1996.
[33] M.J.L. Orr, MATLAB routines for subset selection and ridge regression in linear neural networks, Center for Cognitive Science, Edinburgh University, 1996.
[34] O. Strouf, Chemical Pattern Recognition, Wiley, New York, 1986.