Asian Journal of Chemistry Vol. 21, No. 4 (2009), 2583-2592

Quantitative Structure-Mobility Relationship Prediction of Electrophoretic Mobilities of Some Organic Acids, Amino Acids and Carbohydrates

M.H. Fatemi*, M.R. Hadjmohammadi and F. Bararjanian
Department of Chemistry, Mazandaran University, Babolsar, Iran
Tel/Fax: (98)(112)5242002; E-mail: mhfatemi@umz.ac.ir

Based on theoretical calculated molecular structural descriptors from the solute's structure alone, the electrophoretic mobility of 67 solutes including 23 organic acids, 18 amino acids and 26 carbohydrates in capillary electrophoresis were predicted. In order to find the best model, heuristic method was used to build several multivariable linear models using different numbers of molecular descriptors. This model gave the following statistical values; the square of correlation coefficient R^2 was 0.968, standard error was 0.0778 and the statistical F-value of 380.82. Descriptors which appeared in the selected model can account the hydrodynamic and dielectric friction forces which affected on the electrophoretic mobility of solute. Also in order to evaluate the credibility of model the leave one cross-validation test and Y-scrambling method were employed. The statistically results obtained by these tests reveals the reliability of constructed model.

Key Words: Capillary zone electrophoresis, Electrophoretic mobility, Quantitative structure-mobility relationship, Heuristic method, Multiple linear regressions, Molecular descriptor.

INTRODUCTION

Capillary electrophoresis (CE) is a powerful technique for the separation of varieties of analytes owing to the advantages of high efficiency, high resolution, rapid analysis and very small value of sample[1-3]. Ionic analytes can be separated by both capillary zone electrophoretic (CZE) and micellar electrokinetic chromatography (MEKC), while MEKC is commonly employed for the separation of non-ionic compounds. However, the separation of non-ionic compounds can occasionally be achieved by CZE when using a background electrolyte containing an ionic surfactant at concentration below the critical micelle concentration, provided that the selective interactions between neutral analytes and ionic surfactant monomers occur[4,5]. During method development in CE to develop an optimized separation conditions, the analysts generally have to employ a large number of experiments, which is often costly and time-consuming. Therefore, developing theoretical models to predict the electrophoretic behaviour of analytes are interesting and necessary. Today, more and more investigators have paid attention to this problem and some papers have contributed to study of quantitative relationship between molecular structures and electrophoretic mobilities[6]. Based on the published reports, two principal methods can be summarized, *i.e.*, the mechanistic and the statistical methods. The mechanistic model is closely related to the mechanism of electrophoretic separation. The basic expression of such method is Max Born's model[7]:

$$\mu = \frac{q}{f_h + f_{dl}} \tag{1}$$

where q is the effective charge on the ion, f_h is hydrodynamic friction force, f_{dl} is the dielectric friction forces and μ is the electrophoretic mobility of solute.

The statistical models are based on the quantitative structure-mobility relationship (QSMR). This approach aims to get high predictive performance with relatively less consideration to the mechanism of separation. One of the most important factors governing the quality of the QSMR model is the quantitation of structural features, *i.e.*, the extraction of molecular descriptors. Both new descriptors developed by oneself and existing descriptors embodied in commercial special software can be used to build linear or nonlinear models by some techniques such as multiple linear regression (MLR), artificial neural network (ANN) and support vector machine (SVM)[6,8-12]. The advantage of QSMR approach over other methods lies in the fact that the descriptors which used in the model can be calculated from structure alone and are not dependent on any experimental properties. So once a reliable model is established, this model can be used to predict the property of interested compound. Therefore the quantitative structure-mobility relationship investigation is a useful method to predict the electrophoretic mobilities of solutes in CZE avoiding long and tedious separation optimization. QSMR results can also tell us which of the structural factors may play an important role in the determination of absolute mobility of the compound. There are some published reports about quantitative structure-mobility relationship investigations. In one of the early published QSMR studies, Liang *et al.*[13] used MLR technique to establish a model for predicting the mobilities of 13 flavonids from their topological descriptors. A comparative study between MLR and ANN has been carried out employing electrophoretic mobility of 13 sulfonamides by Jalali-Heravi and Garakani-Nejad[14]. The linear models they proposed were represented as follows:

$$\mu_e = C_0 + C_1 \Delta H_f + C_2 PPCH + C_3 SA \tag{2}$$

and

$$\mu_e = C_4 + C_5 \Delta H_f + C_6 PPCH + C_7 pK \tag{3}$$

where μ_e is the effective electrophoretic mobility, ΔH_f is the heat of formation of anions, PPCH denotes maximum positive partial charge on anions, SA represent the surface area, pK is *p*-function of dissociation constant and C_0-C_7 are the model constants. Then a non-linear 3-4-2 ANN model was generated using ΔH_f, PPCH and SA as inputs for prediction of the electrophoretic mobilities of anions sulfonamides.

The authors concluded that the ANN model shows the superiority over the MLR model. In the previous work, a QSMR model is constructed to estimate the electrophoretic mobilities of benzoic acid derivatives by means of a multi-layer neural network using back-propagation training algorithm[15]. The standard error of training, validation and test sets for ANN model are 0.402, 0.952 and 0.716, respectively. In another work, Wang *et al.*[16] studied relationship between the relative mobility of a group of 19 chlorophenols in different buffers modified by 8 kinds of different organic additives in CZE by means of MLR and radial basis function neural network (RBFNN). They used approximate molecular surface area, hydration energy, dipole moment, highest occupied molecular orbital energy level and polarity of organic additives as inputs for generated ANN. Their obtained model gives the correlation coefficient (R) of 0.986 for the training set and 0.980 for the prediction set. In the present study, QSMR modeling based on theoretical derived molecular descriptor combined with multiple linear regression analysis was explored to drive a simple model for the reliable prediction of the electrophoretic mobility of some amino acids, organic acids and carbohydrates. Resulting model was evaluated for its reliability by some validation tests.

EXPERIMENTAL

Data set: The electrophoretic mobilities of 67 compounds including 23 organic acids, 18 amino acids and 26 carbohydrates were taken[17]. The names and corresponding electrophoretic mobility of these compounds are shown in Table-1. The effective mobility, μ_e, for each compound was calculated using the following equation:

$$\mu_e = \frac{lL}{T_a V} - \frac{lL}{T_{EOF} V} \quad [cm^2\ V^{-1}\ s^{-1}] \tag{4}$$

where l and L are the length of the capillary to the detector and the total length of capillary, respectively, V is the applied potential, T_a is the migration time of anion and T_{EOF} is migration time of a neutral marker.

Separations were carried out on fused-silica capillaries with 112.5 cm (104 cm effective length) × 50 µm i.d. with indirect UV detection using 2,6-pyridne dicarboxylic acid as background electrolyte. Highly alkaline conditions were used in order to confer the existence of a negative charge not only on organic anions but also on amino acids and carbohydrates and to promote their migration toward anode. Electrophoretic mobilities of these compounds ranged from -5.784 to -0.064 for oxalate and inositol, respectively.

Descriptors calculation: Since the electrophoretic mobility represents the migration of solute under a certain applied electric field, it would be strongly related to the charge, size and topological structure of the corresponding solute. Therefore, one must calculate some structural descriptors for each solute which can encode these features of molecule numerically. The build model performance and the

TABLE-1
DATA SET OF EXPERIMENTAL AND PREDICTED
VALUES OF ELECTROPHORETIC MOBILITIES

Number	Name	μ_{exp}	μ_{calk}	Error
1	Oxalate	-5.784	-5.136	-0.65
2	Ascorbate	-5.409	-5.358	-0.05
3	Malonate	-5.093	-5.112	0.02
4	Formate	-4.911	-4.814	-0.10
5	Citrate	-4.775	-5.275	0.50
6	Tartarate	-4.584	-4.073	-0.51
7	Succinate	-4.565	-4.697	0.13
8	Malate	-4.520	-4.925	0.41
9	α-Ketoglutarate	-4.513	-4.817	0.30
10	Asp	-4.418	-4.335	-0.08
11	Glutarate	-4.196	-4.214	0.02
12	Glu	-4.084	-4.219	0.14
13	Adipate	-3.934	-4.885	0.95
14	Acetate	-3.589	-3.741	0.15
15	Pyruvate	-3.540	-3.985	0.45
16	Cys-Cys	- 3.514	-3.402	-0.11
17	Glycolate	-3.495	-3.495	0.00
18	Tyr	-3.493	-2.686	-0.81
19	Gly	-3.260	-3.236	-0.02
20	*n*-Propioonate	-3.111	-3.141	0.03
21	Lactate	-3.041	-3.142	0.10
22	*n*-Butyrate	-2.781	-2.679	-0.10
23	Levulinate	-2.729	-2.800	0.07
24	Mannuronic acid	-2.647	-2.097	-0.55
25	Pyroglutamate	-2.631	-3.045	0.41
26	*n*-Pentanoate	-2.578	-2.423	-0.16
27	Thr	-2.542	-2.409	-0.13
28	Glucuronic acid	-2.497	-2.200	-0.30
29	Pro	-2.450	-2.523	0.07
30	Val	-2.444	-2.440	0.00
31	Met	-2.389	-2.296	-0.09
32	*n*-Hexanoate	-2.385	-2.203	-0.18
33	Galacturonic acid	-2.337	-1.990	-0.35
34	His	-2.310	-2.423	0.11
35	Leu	-2.300	-2.212	-0.09
36	Ilu	-2.300	-2.505	0.21
37	Phe	-2.22	-2.413	0.19
38	*n*-Heptanonate	-2.149	-1.897	-0.25
39	Gluconate	-2.060	-1.901	-0.16
40	Lys	-2.026	-2.058	0.03
41	Trp	-1.910	-2.413	0.50
42	NGNA	-1.719	-1.439	-0.28
43	*n*-Octanoate	-1.707	-1.551	-0.16
44	NANA	-1.675	-1.621	-0.05

Number	Name	μ_{exp}	μ_{calk}	Error
45	N-Acetylmannosamine	-1.221	-1.246	0.03
46	Ribose	-1.037	-1.219	0.18
47	Ribose	-0.983	-0.803	-0.18
48	N-Acetylglucoseamine	-0.975	-1.235	0.26
49	Mannose	-0.966	-0.899	-0.07
50	Mannose	-0.935	-1.129	0.19
51	N-Acetylgalactosamine	-0.919	-1.121	0.20
52	Ramnose	-0.904	-0.816	-0.09
53	Glucosamine	-0.846	-0.589	-0.26
54	Mannosamine	-0.832	-0.561	-0.27
55	Lactose	-0.774	-0.673	-0.10
56	Arabinose	-0.764	-1.244	0.48
57	Glucose	-0.761	-0.800	0.04
58	Galactosamine	-0.725	-0.798	0.07
59	Lactulose	-0.676	-0.504	-0.17
60	Galactose	-0.659	-0.806	0.15
61	Fucose	-0.485	-0.840	0.36
62	Sucrose	-0.291	-0.414	0.12
63	Trhalose	-0.121	-0.149	0.03
64	Galactitol	-0.104	-0.061	-0.04
65	Xylitol	-0.086	-0.069	-0.02
66	Erythritol	-0.076	-0.044	-0.03
67	Inositol	-0.064	-0.047	-0.02

accuracy of the results are strongly depends on the way that the structural representation was performed. The calculation process of molecular descriptors in the present work is described as below: the three-dimensional structures of molecules were drawn using HYPERCHEM 7.0 program[18] and exported in a file format suitable for MOPAC program[19]. The geometry optimization was performed with the semi empirical quantum method AM$_1$[20] using the MOPAC 6.0. All Geometry had been fully optimized without symmetry restrictions. In all case frequency calculation have been performed in order to ensure that all calculated geometries correspond to true minima. The HYPERCHEM and MOPAC output files were used by CODESSA program. This software developed by Katritzky group enables the calculation of a large number of quantitative descriptors based on the molecular structural informations and codes this chemical information into mathematical[21]. CODESSA can calculate 5 classes of descriptors which are: constitutional (number of various types of atoms and bonds, number of rings, molecular weight, *etc.*); topological (winner index, randic indices, Kier-Hall shape indices, *etc.*); geometrical (moment of inertia, molecular volume, molecular surface area, *etc.*); electrostatic (minimum and maximum of partial charges, polarity parameters, charged partial surface area descriptors, *etc.*) and quantum chemical (reactivity indices, dipole moment, HOMO and LUMO energies, *etc.*).

Selection of the descriptors: Since it is not possible to know a priority which descriptors are most relevant to the problem at hand, a comprehensive set of descriptors is usually employed, chosen based on experience, software availability and computational cost. The heuristic multi-linear regression procedures available in the framework of CODESSA program were used to perform a complete search for the best multi linear correlations with a multitude of descriptors. This procedures provide colinearity control (*i.e.,* any two descriptors inter-correlated above 0.80 are never involved in the same model) and implement heuristic algorithm for the rapid selection of the best correlation, without testing all possible combinations of the available descriptors. The heuristic method of 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 structure, (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 correlation coefficient. 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 heuristic method usually produces correlations 2-5 times faster than other methods with comparable quality[22]. The rapidity of calculations from the heuristic method renders it the as a suitable method of choice in practical research.

RESULTS AND DISCUSSION

Table-1 shows the observed and calculated electrophoretic mobility of all compounds studied in this work. For selection of the best MLR model, the credibility and goodness of models are tested by calculation of coefficient of multiple correlation regression (R^2), the F-test value (F) and the standard error of the model (SE). The stability of the correlations was tested against the cross validated coefficient, R^2_{cv}. The R^2_{cv} value describes the stability of a regression model obtained by focusing on of the sensitivity of 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 of this point and the set of estimated values calculated in this way is correlated with the experimental values. In this way a variety of subset size was investigated to determine the optimum number of descriptors in a model. Figs. 1 and 2 shows the influences of the number of descriptors in the model on the R^2_{cv} (Q^2 for cross - validation test) and standard error of models, respectively. From these figures, it can be seen that 5 descriptors appear to be sufficient for a successful regression model. The specification of selected model summarized in Table-2. Also the correlation matrix of these descriptors was shown in Table-3. The linear correlation coefficient values of each two

TABLE-2
SPECIFICATION OF MULTIPLE LINEAR REGRESSION MODELS

Descriptor	Notation	Coefficient	SE	t-test
DPSA-1 Difference, in PSAs(PPSA1-PNSA1)	DPSA1	0.005	±0.001	4.608
HACA-2/TMSA[zefirov's pc]	HACA2	241.050	±12.664	23.015
FNSA-2 Fractional PNSA	FNSA2	43.563	±1.279	7.589
Minimum atomic orbital electronic population	MAOEP	-6.781	±2.189	-3.864
Minimum valency of a C atom	MVCA	12.601	±1.567	9.930
Constant		-45.184	±10.300	

n = 67; R^2 = 0.968; SE = 0.078; F = 380.

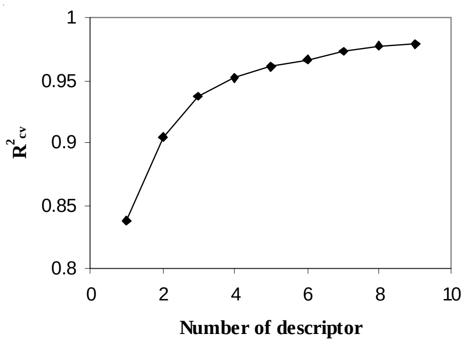

Fig. 1. Influence of the number of descriptors on the R^2_{cv}

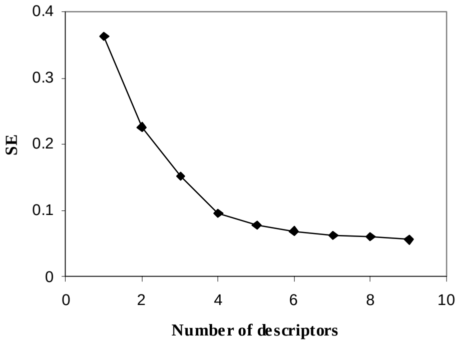

Fig. 2. Influence of the number of descriptors on the standard error

TABLE-3
CORRELATION MATRIX FOR DESCRIPTORS APPLYING IN THIS WORK*

	DPSA1	MAOEP	HACA2	MVCA	FNSA2
DPSA1	1				
MAOEP	0.467	1			
HACA2	0.279	0.163	1		
MVCA	0.321	0.517	-0.025	1	
FNSA2	0.573	0.419	0.578	0.187	1

*The definition of notations are given in Table-2.

descriptors are less than 0.60 (Table-3), which means that selected descriptors were independent in this multi-linear analysis. The obtained model has a correlation coefficient R^2 = 0.968, F = 380.82, with SE = 0.0778 and the cross-validated correlation coefficient R^2_{cv} = 0.961. The risk of chance correlation in the obtained model is verified also by Y-scrambling procedure. In this method, the dependent variable vector, Y-vector, is randomly shuffled and a new QSMR model is developed using the original independent variable matrix. This process is repeated several times. The obtained R^2 for less than 1 % of all scrambled Y-vectors have a correlation with the original Y-values that is higher than R^2 = 0.2 (95 % below R^2 = 0.15). These results reveals that the proposed model is well founded and not just the result of chance correlation.

By interpreting descriptors in the regression model it is possible to gain some insight into factors that are likely to govern electrophpretic mobility of a solute. As mentioned in the introduction section two fundamental frictional forces are 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 and caused by the orientation of the solvent dipole in response to the ionic charges. As can be seen from Table-2, there are five descriptors in the obtained MLR model. The first descriptor is DPSA1, which is the difference between PPSA1 (atom charged weighted partial positive surface area) and PNSA1 (atom charged weighted partial negative surface area) and is related to the positive and negative charge distribution and also to the respective surface area in a molecule[23]. The second descriptor was HACA-2/TMSA [Zerfirov,s PC] (HACA2) which is CPSA (charged partial surface area) descriptors and have been proposed by Jurs *et al.*[24]. This descriptor is the ratio of hydrogen acceptors charged surface area to the total molecular surface area and denote the sum of solvent-accessible surface area of the H-bonding acceptor atoms. As the HACA2 increasing, the proportion of the H-donors surface area among the total molecular surface is increase and the formation of the H-bond became easier. The third descriptor which appeared in the selected model was fractional negative surface area (FNSA2). FNSA2 is defined as the ratio of the atomic charge weighted partial negative surface area, (which is obtained by summation of products of the individual atomic partial charges and the atomic solvent-accessible surface area) to the total molecular surface area[25]. This descriptor is expected to encode the features responsible for polar interactions between molecules. The forth descriptor was minimum atomic orbital electronic population (MAOEP) and the last one was minimum valency of carbon atoms (MVCA). As can be seen all of these molecular parameters are electronic type descriptors which can account for dielectric friction factor in the whole. On the other hand the since first three descriptors (DPSA1, HACA2 and FNSA2) are related to the molecular surface area and charge distribution among of this surface, therefore they can encode some information about the hydrodynamic friction forces.

Fig. 3 represents the plot of cross-validation predicted electrophoretic mobility of interested molecules against their experimental values. Also the residual of calculated electrophoretic mobilities from their experimental values are shown in Fig. 4. The propagation of residuals at both side of the zero line indicates that there aren't any systematic errors in obtained model.

Conclusion

The heuristic method was used to construct a linear quantitative structure-mobility relationship model for the prediction of electrophoretic mobility in capillary zone electrophoretic (CZE) of a set of organic acid, carbohydrate and amino acids. The performance of model was evaluated by cross-validation test and Y-scrambling methods. The obtained results of these tests reveal that the constructed linear model

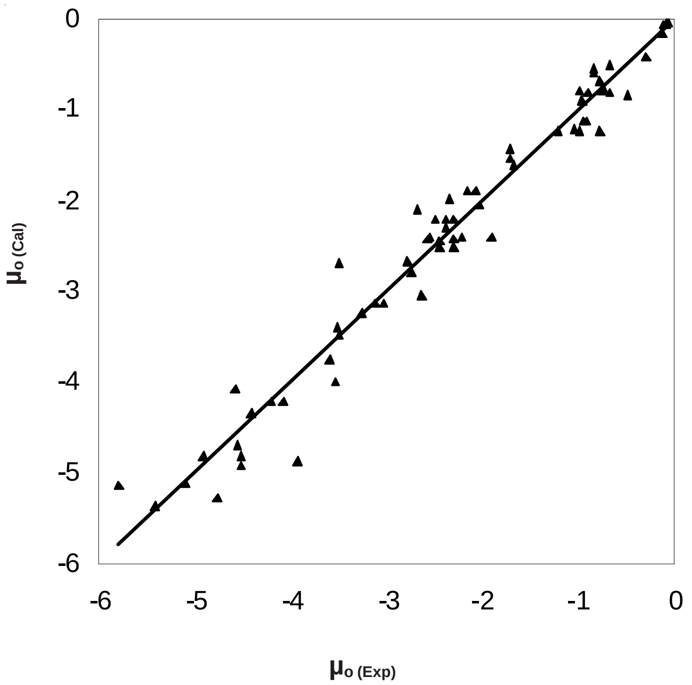

Fig. 3. Plot of the cross-validation calculated mobility against the experimental values

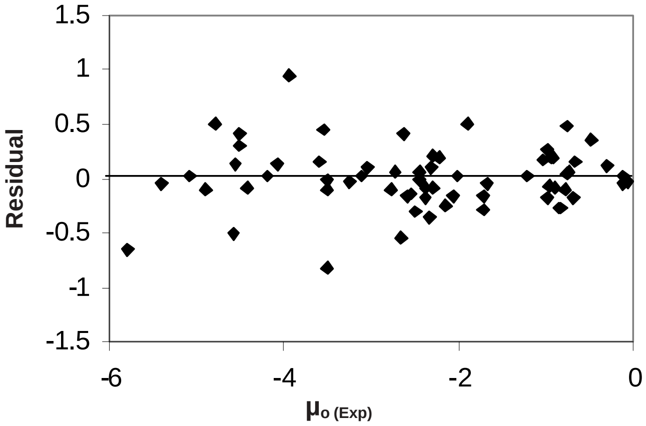

Fig. 4. Plot of the residuals versus experimental values of mobilities

was satisfactory. Descriptors appeared in the model can account the hydrodynamic and dielectric friction force which affected on the electrophoretic mobility of solute in capillary electrophoresis. Furthermore, the proposed approach also can be extended in other QSMR investigation.

REFERENCES

1. P. Camillari, Capillary Electrophoresis: Theory and practice, CRC, Boca Raton, FL, USA, pp. 235-248 (1993).
2. Z. EL. Rassi and R.W. Giese, Selectivity and Optimization in Capillary Electrophoresis, Elsevier, Amsterdam (1997).
3. M.G. Khaledi, High-Performance Capillary Electrophoresis: Theory, Techniques and Applications, Wiley, New York, p. 142 (1998).
4. J.L. Beckers and P. Bocek, *Electrophoresis*, **23**, 1947 (2002).
5. C.E. Lin, T.Z. Wang, H.C. Huang, C.C. Hsueh and Y.C. Liu, *J. Chromatogr. A*, **878**, 137 (2000).
6. A. Jouyban and B.H. Yousefi, *Comput. Biol. Chem.*, **27**, 297 (2003).
7. R.L. Kal, *Pure Appl. Chem.*, **63**, 1393 (1991).
9. M. Jalali-Heravi and Z. Garakani-Nejad, *J. Chromatogr. A*, **971**, 207 (2002).
10. S.F. Wang, C.X. Xue, X.G. Chen, M.C. Liu and Z.D. Hu, *J. Chromatogr. A*, **1033**, 153 (2004).
11. Y. Cheng and H. Yuan, *Anal. Chim. Acta*, **565**, 112 (2006).
12. M.H. Fatemi, *J. Chromatogr. A*, **1038**, 231 (2004).
13. H. Golmohammadi and M.H. Fatemi, *Electrophoresis*, **26**, 3438 (2005).
14. H.R. Liang, H. Vuorela, P. Vuorela, M.L. Riekkola and R. Hiltunen, *J. Chromatogr. A*, **798**, 233 (1996).
15. M. Jalali-Heravi and Z. Garakani-Nejad, *J. Chromatogr. A*, **927**, 211 (2001).
16. M.H. Fatemi and N. Goudarzi, *Electrophoresis*, **26**, 2968 (2005).
17. Y.W.Wang, S.L. Gao, Y.H. Gao, S.H. Liu, M.C. Liu, Z.D. Hu and B.T. Fan, *Anal. Chim. Acta*, **486**, 191 (2003).
18. T. Soga and G.A. Ross, *J. Chromatogr. A*, **837**, 231 (1999).
19. HyperChm. re. 4 for Windows, AutoDesk, Sausalito, CA (1995).
20. J.J.P. Stewart, MOPAC, Semi Empirical Molecular Orbital Program, QCPE, p. 455 (1983), version 6 (1990).
21. M.J.S. Dewar, E.G. Zoebisch, E.F. Healy and J.J.P. Stewart, *J. Am. Chem. Soc.*, **107**, 3902 (1985).
22. K.V.S. Atritzky, V.S. Lovanov and M. Karelson, Comprehensive Descriptors for Structural and Statis tical Analysis, Reference Manual, Version 2.0 (1994).
23. A.R. Katritzky, R. Petruhin, R. Jain and M. Karelson, *J. Chem. Inf. Comput. Sci.*, **41**, 1521 (2001).
24. A.R. Katritzky and D.B. Tatham, *J. Chem. Inf. Comput. Sci.*, **41**, 1162 (2001).
25. D.T. Stanto, L.M. Egolf, P.C. Jurs and M.G. Hicks, *Chem. Inf. Comput. Sci.*, **32**, 306 (1992).
26. R. Todeschini and V. Consonni, Handbook of Molecular Descriptors, Wiley-VCH, Veinheim, p. 425 (2000).

(*Received*: 24 January 2008; *Accepted*: 7 January 2009) AJC-7095