همایش ملی انسان، محیط زیست و توسعه پایدار
باشگاه پژوهشگران جوان دانشگاه آزاد اسلامی واحد همدان
۱۹و۲۰ اسفندماه ۱۳۸۸

Correlation ranking procedure for factor selection in PC-ANN modeling and application to aqueous Solubility evaluation

A. Najafi[1], S. Sobhan Ardakani*[2]

Abstract:

A correlation ranking procedure is proposed for selection of factors in principal component-artificial neural network (PC-ANN). The model was applied in the aqueous Solubility (-logS) evaluation of diverse Organic molecules. Experimental values for the observed -logS values for organic molecules can range from about -0.380 (oxalic acid) to 10.410 (2,2',3,3',4,5,5',6,6'-PCB) -log units. Ten different Sh indices were calculated for each molecule. Principal component analysis of the Sh data matrix showed that the seven PCs could explain 99.97% of variances in the Sh data matrix. The extracted PCs were used as the predictor variables (input) for PCR and ANN models. The ANN model could explain 97.63% of variances in the solubility data, while the value obtained from PCR procedures were 84.27%. For the PCR studies, the data set was divided into a training set of 320 compounds for model building and an external prediction set of 60 compounds for model validation. Both subsets were chosen to ensure that a diverse set of compounds was present. For the ANN studies, a cross-validation set of 50 compounds was chosen, leaving 270 compounds in the training set, and the prediction set remained the same. Models to predict the solubility is constructed using PCR and PC-ANN with errors comparables to the experimental errors of the solubility data. The root mean-square-errors (RMS-error) associated with the calibration, prediction, and validation set compounds used for the PC-ANN model were 0.314, 0.450, and 0.314 -logS units, respectively.

Keywords: QSPR, Topological Indices, Aqueous Solubility, PCR, PC-ANN.

1- Islamic Azad University- Hamedan Branch, Member of Young Researchers Club (YRC), Hamedan, Iran.
2- Islamic Azad University- Hamedan Branch, Hamedan, Iran.

1. Introduction

The aqueous solubility of organic compounds is an important molecular property, playing a large role in the behavior of compounds in many areas of interest. In modeling the environmental impact of a contaminant, along with the soil-water absorption coefficient, the solubility is a key term in the understanding of transport mechanisms and distribution in groundwater [7, 10, 11].

Quantitative structure property relationships (QSPR), mathematical equations relating chemical structure to the physicochemical properties, have information that is useful for environment chemistry [17-19]. A major step in constructing the QSAR/QSPR models is to find one or more molecular descriptors that represent variation in the structural property of the molecules by a number. Topological Indices (TIs) are a convenient means of translating chemical constitution into numerical values, which can be used for correlation with physical properties and biological activities.

2. Material and Methods

2.1. Aqueous Solubility Data

The data set of aqueous solubility of diverse organic compounds in the present study was recompiled from several literature sources. The final set of 380 diverse organic compounds was representative for all classes of organic compounds containing C, H, O, N, Cl, Br, and I, and included saturated and unsaturated hydrocarbons, halogenated hydrocarbons, polychlorinated biphenyls (PCBs), esters, aldehydes, organic acids, alcohols, ethers, amines, and aromatic compounds. In this list, the experimental -logS values for organic compounds can range from about -0.380 (oxalic acid) to 10.410 (2,2',3,3',4,5,5',6,6'-PCB) log units [12, 14, 15, 20, 21].

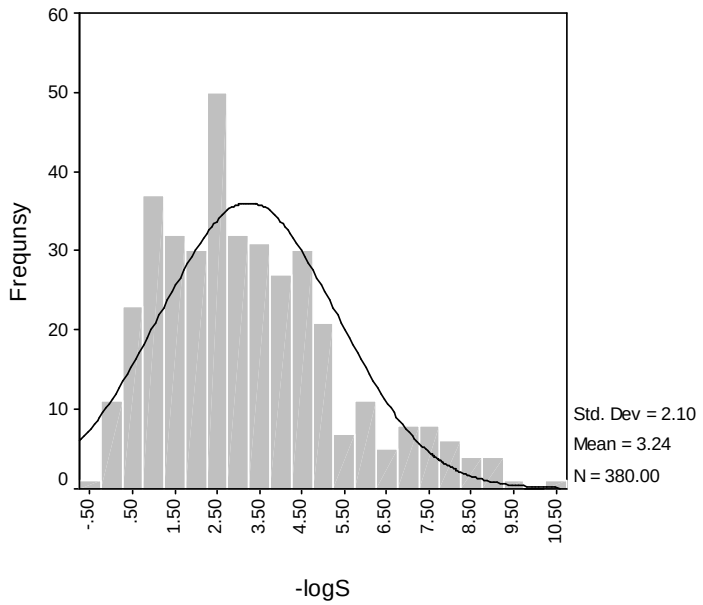

Figure 1. Histogram of the distribution of the experimental -logS for the total data set of 380 organic compounds used in this study. The solid curve is the fitting of the -logS data to the normal distribution.

2.2. Sh Topological Indices

Ten different Sh topological indices (Sh1 – Sh10) were calculated for each molecule based on the different combinations of the distance sum and connectivity vectors. The theoretical basis for calculation of these indices is found in our previous papers [8, 13, 22]. A home-made program (written in MATLAB environment) calculated the Sh indices. The calculated indices were collected in a data matrix with 380×10 dimension. Each chemical is now a point in the 10-dimensional space, $\mathbf{X}^{10}$

2.3. Linear Modeling: Principal Component Regression

Due to the some co-linearity between the Sh topological indices, orthogonal transformation of the Sh indices by principal component analysis was performed. The score and loading matrices were calculated by singular value decomposition (SVD) procedure [2]:

$$\mathbf{D} = \mathbf{U}\,\mathbf{S}\,\mathbf{V}^{T} \qquad (1)$$

where **U** and **V** are the orthonormal matrices spanned the respective row and column spaces of the data matrix (**D**). **S** is a diagonal matrix whose elements are the squared root of the eigen-values. The superscript "T" denotes the transpose of the matrix. The eigen-vectors included in **U** are named as principal components (PC). The PCs of the validation ($\mathbf{D}_v$) and Prediction ($\mathbf{D}_p$) sets were calculated by the equation:

$$\mathbf{U}_{p/v} = \mathbf{D}_{p/v}\,\mathbf{S}^{-1}\,\mathbf{V} \qquad (2)$$

Application of the PCA on the Sh indices data matrix resulted in 10 factors or principal components (PC_1-PC_{10}). A linear regression model was build between the solubility and resulted factors. The best set of factors was selected by the eigen-value ranking (EV) and correlation ranking (CR) procedures. In the EV-PCR procedure, the PCs were entered to the PCR model consecutively based on their decreasing eigen-value. Once each new factor was entered to the model, the model performances were evaluated by the leave-one-out cross-validation (LOO-CV). In the CR-PCR, the correlation between each one of the extracted PC's with the solubility data was determined first. The stepwise entrance of the PCs to the PCR model was based on their decreasing correlation with the solubility.

2.4. Nonlinear Modeling: PC-ANN

To model the -logS-Sh indices more accurate, artificial neural network was employed to process the nonlinear relationships between the selected PCs in the previous section and solubility data. The PC-ANN model was the same as we reported previously [3, 9]. The totals of 380 compounds were randomly divided to 270 calibration (or training) samples, 60 prediction samples and 50 validation samples. The PCs of the calibration samples were calculated by equation 1 and those of prediction and validation samples were calculated by equation 2. The prediction set is a subset of compounds used to help find an optimal set of weights and biases during ANN calibrating, and it is also used to avoid overtraining of the ANN. The ANNs used in this study were fully connected, three layer, feed-forward ANN. The number of neurons in the input layer is equal to the number of PCs selected for the model. The PC's used here were those selected by the CR-PCR and EV-PCR models.

The transformed values are then passed to the hidden layer. The input value of a hidden layer neuron is the summation of the products of the weights (neuron connections) times the corresponding outputs of the previous input layer plus a bias term. The ANN model confined to a single hidden layer, because the network with more than one hidden layer would be harder to train.

3. Results and discussion

3.1. PCR Modeling

For each subset of molecules separate PCR models based on the eigen-value ranking and correlation ranking were obtained. The results obtained by the correlation ranking procedure are shown in Table 3.

Table 3. Linear multivariate regression models and statistical parameters of compounds properties using PC indices.

Subset	N	Equation	R^2	SE	RMS	REP	F	R^2_{CV}
CH	133	$-\log S = 4.112 + 1.306\ PC_1 - 0.641\ PC_3 - 0.354\ PC_7 - 0.276\ PC_7 + 0.235\ PC_4 - 0.146\ PC_9 +$	0.9255	0.455	0.440	10.70	192	0.9068
O	64	$-\log S = 1.313 + 0.927\ PC_1 - 0.168\ PC_7 - 0.166\ PC_2 + 0.102\ PC_6 - 0.073\ PC_4 + 0.071\ PC_5 +$	0.9816	0.142	0.132	10.04	367	0.9709
N	9	$-\log S = 1.918 + 0.851\ PC_1 + 0.0.190\ PC_3 - 0.263\ PC_4$	0.9278	0.321	0.240	12.48	21	0.7614
Halogen	124	$-\log S = 3.970 + 2.121\ PC_1 + 0.518\ PC_3 - 0.323\ PC_9 - 0.265\ PC_8 + 0.182\ PC_5 + 0.118\ PC_{10}$	0.9413	0.572	0.555	6.10	312	0.9284
Overall	50	$-\log S = 1.792 + 1.026\ PC_1 + 0.358\ PC_3 - 0.304\ PC_2 - 0.180\ PC_6 - 0.177\ PC_7$	0.8565	0.500	0.469	26.147	52	0.7974
Total	380	$-\log S = 3.237 + 1.344\ PC_1 + 0.993\ PC_3 - 0.636\ PC_6 + 0.498\ PC_{10} + 0.403\ PC_7 - 0.296\ PC_9 -$	0.8580	0.797	0.790	27.37	321	0.8477

As can be seen, the number of PCs, used in the QSPR model of each subset was similar but the set of PCs are different. The least number of factors (i.e. 3 factors) is used for modeling the solubility of subset of nitrogen containing compounds, while the higher number of factors (i.e. 8 factors) is used for by CH and oxygen subsets of compounds. For all subsets, the factors selected by the correlation ranking procedures are different from those of eigen-value ranking.

To further check the prediction ability and overfitting of the resulting models, the leave-one-out cross validation (LOO-CV) procedure was applied. In LOO-CV procedure, *n-1* sample from a total data set of each subset were used to construct a calibration set (assessment set) and to build a QSPR model between the PCs and the examined solubility, and the solubility property of the left out sample was estimated by the designed model. This procedure was repeated until every sample in the total data set for each subset was used for a prediction. Then, PRESS (the predicted residual sum of squares) and SSD (the sum of the squared deviation from the mean) were calculated for each regression equation. The squared correlation coefficient for cross validation (R^2_{CV}) was then calculated by the following equation $R^2_{CV} = 1 - (\text{PRESS/SSD})$. The results of LOO-CV examination for each subset of organic

compounds are listed in column 8 in Table 3. The cross-validation results show that all models (regression expressions) presented in the Table 3 have R^2_{CV} values greater than 0.90 excepted for the subset of nitrogen that it is due to small number of molecules in this class. Thus, the cross-validation test indicates that the Sh indices can model the liquid solubility of some subsets of organic compounds were used in this studies, perfectly. [1].

In the last row of Table 3 the CR-PCR model obtained for the solubility of entire set of compounds by the correlation ranking procedure is listed. The trend of the PCs in order of decreasing their correlation is PC1 > PC3 > PC6 > PC10 > PC7 > PC9 > PC8 which was not in the same direction as their decreasing eigen-value. The resulting correlation equation had correlation coefficient R^2 = 0.8580, RMS = 0.790, F = 321, R^2_{CV} = 0.8477. The seven factors used in this equation can explain 85.80 % of the variance in the -logS of all data set of solubility organic compounds. Further attempts were made to examine the quality of the resulted model by splitting the data set into the calibration set (320 molecules) and prediction set (60 molecules). The resulted CR-PCR model was the same as that obtained for entire set of molecules. The R^2 value and RMS error for the validation set are 0.9147, and 0.769, respectively.

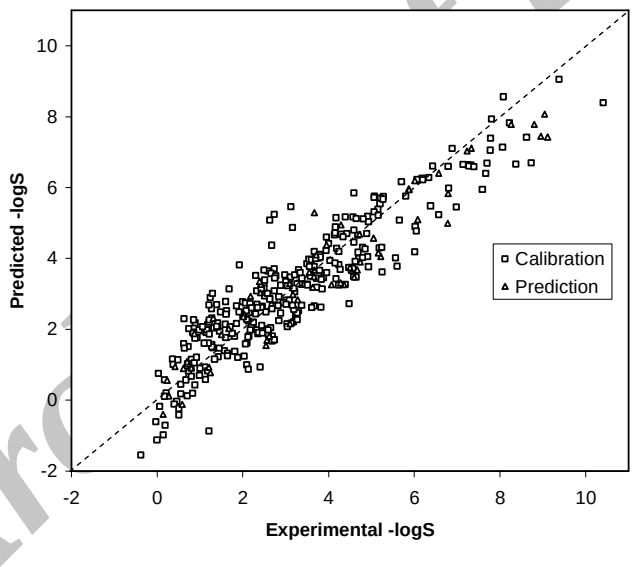

Figure 2. Plot of the predicted -logS by CR-PCR against the experimental values. The dash line is the ideal fit to the straight line.

3.2. PC-ANN Modeling

Once valid linear models were found using PCR, steps were taken to see if prediction results could be improved by the use of artificial neural networks (ANNs). Typically, superior models can be found using ANN because they implement nonlinear relationships and because they have more adjustable parameters than the linear models. Therefore, in this study we suggested the use of ANN as the nonlinear model. A fully connected, three-layered feed-forward ANN model with back-propagation [16] learning algorithm is developed for nonlinear modeling between the selected PCs by

the CR-PCR model. The seven PCs were test with several ANN architectures, the ANN model was confined to a single hidden layer and a sigmoid transfer function, as a more versatile transfer function, was used in this layer. Because of the large number of adjustable parameters, it is possible to over-train the network. If over-training does occur, contributions of a small subset of the training set compounds may be considered as a major contribution, thus hindering the ability of the network to accurately predict the physical property in question. To avoid over-training, the data set is split into a calibration set, a prediction set and a validation set. Each connection in the network is made up of a weighting factor and a bias term. The weights and biases are changed during training based on the RMS error of the validation set; the corresponding values are then calculated for the validation set for each of configuration. The convergence criterion was the least RMS error in the prediction set. The number of iterations for convergence was between 15000 and 20000. In each ANN, the neuron architecture (i.e., the number of nodes in hidden layer; n_H) and parameters (i.e., learning rate and momentum) were optimized to reach the lowest the RMS error of the validation set as the performances of the resulted models, because it is believed that overtraining occurs when the RMS error begins to rise. At this point, the values of the weights and biases are not changed further. A plot of RMS error as a function of linear rate and momentum in three different numbers of nodes in hidden layer is shown in Figure 3. The results indicate that an ANN with eight PCs as input variables, 6 nodes in its hidden layer, learning rate of 0.15, and momentum of 0.65 resulted in the optimum network model. A comparison between the results obtained by the eigen-value ranking and correlation ranking-based PC-ANN models revealed that the latter produced accurate results, which is in accordance with previous findings [4-6, 22].

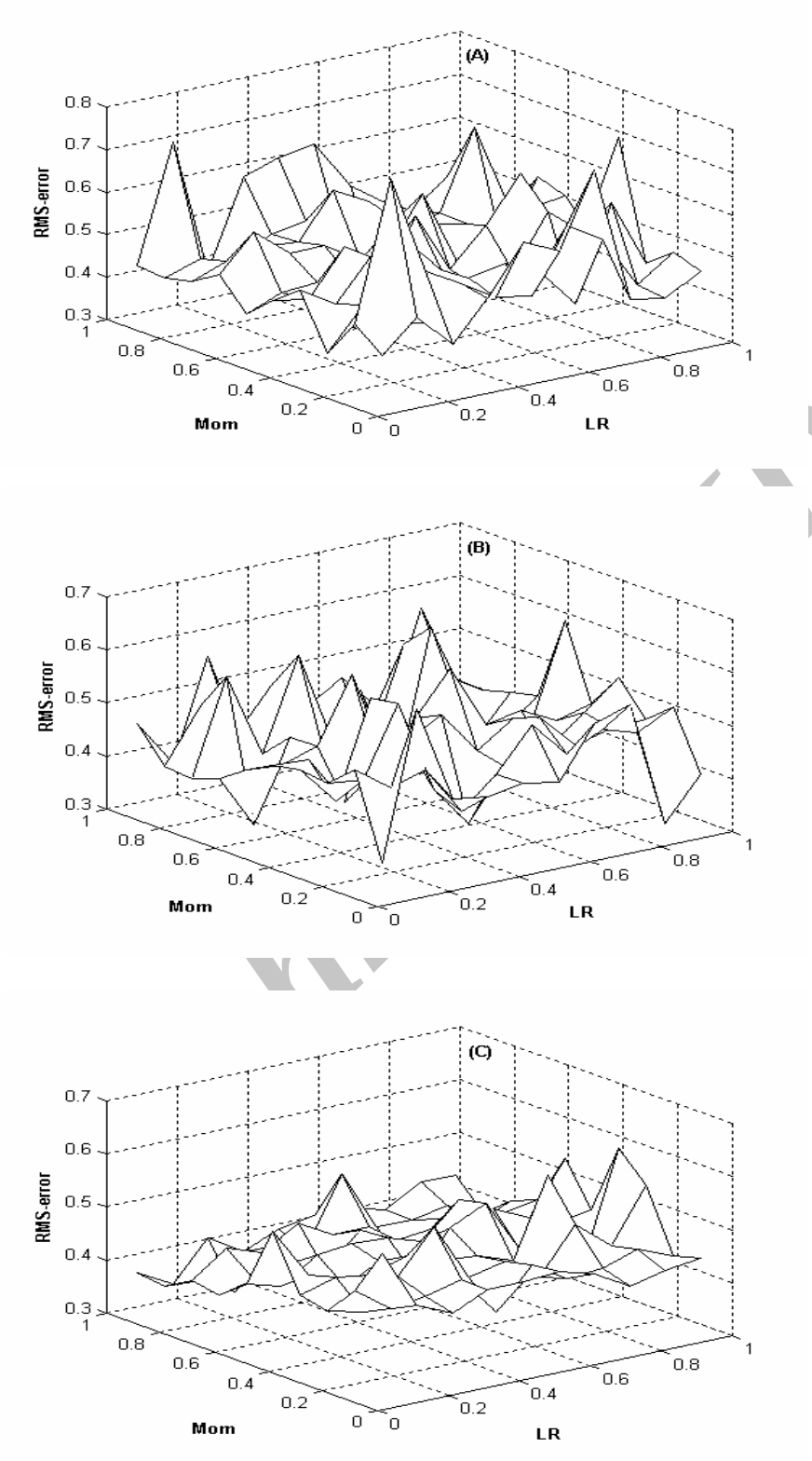

Figure 3. Optimization of linear rate (LR), momentum (Mom) and number of hidden layer nodes (n_H) for ANN modeling; (**A**) $n_H = 5$; (**B**) $n_H = 6$ and (**C**) $n_H = 7$.

The predicted values of -logS resulted from application correlation ranking ANN procedures model (CR-ANN) are plotted in Figure 4 against the corresponding experimental values, and the statistical parameters for the best-fitted model are represented in Table 4.

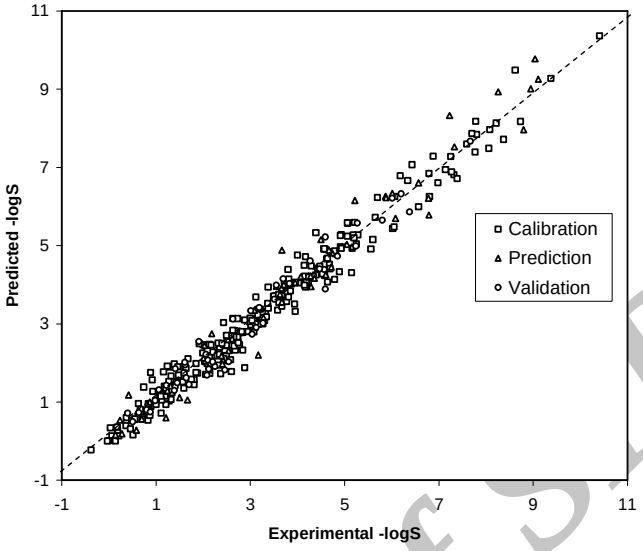

Figure 4. Plot of the predicted -logS by PC-CR-ANN against the experimental values. The dash lines are the ideal fit to the straight line.

Table 4. Statistics of principal component regression models and artificial neural network models with one hidden-layer neurons for calculating aqueous solubility.

	CR-PCR		CR-PC-ANN		
	Calibration set	Prediction set	Calibration set	Prediction set	Validation set
-					
N	320	60	270	60	50
SE	0.806	0721	0.315	0.438	0.318
RMS	0.796	0.769	0.314	0.450	0.314
REP	25.456	20.095	9.864	11.767	11.054
R^2	0.8427	0.9147	0.9763	0.9685	0.9700
F	239	-	11047	1781	1551
Bias	0.0000	-0.2360	-0.0040	-0.0350	-0.0300
error range	(-2.52)-(2.08)	(-1.78)-(1.62)	(-1.00)-(0.94)	(-1.00)-(1.21)	(-0.70)-(0.64)

As it is observed, the models obtained by the PC-ANN have superior qualities relative to those obtained by PCR. This means that there are nonlinear relationships between the proposed Sh topological indices and the –logS of the organic molecules used in this study. A comparison between the results obtained by the eigen-value ranking and correlation ranking-based PC-ANN models

revealed that the latter produced accurate results, which is in accordance with our previous findings [4-6, 22].

4. Conclusions

The usefulness of the some newly proposed topological indices (Sh indices) in quantitative structure-aqueous solubility (-logS) relationship analysis were use to predict the aqueous solubility of different subsets of organic compounds containing various heteroatoms for a wide variety of 380 organic compounds by using the principal component regression and principal component-artificial neural network modeling methods. The PCs were entered to the models based on their decreasing eigen-values (EV) and their correlation ranking coefficients (CR) with the -logS, in which the latter produced better results. Successful correlation equations were developed for the aqueous solubilities of different five subsets of organic compounds. The resulting individual QSPR correlation equations involve three to eight PCs and have RMS-errors ranging from 0.132 for oxygen containing compounds to 0.555 –log units for halogenated compounds. Proceeding from the correlation equations for the subsets of compounds, a general seven-PC correlation model was developed for the prediction of solubility of any organic compound containing C, H, O, N, S, Cl, Br, and I atoms. This correlation model covers a large diversity of organic structures and offers a RMS-error of 0.790 –log unit. In conclusion, we applied both linear and nonlinear models to performances of the prediction of aqueous solubility by using these seven PCs. PCR analysis of the data showed that proposed Sh indices could explain about 91.47% of variations in the solubility data; while the variations explained by the ANN modeling were more than 96.85%. These results demonstrated that aqueous solubilities for a wide range of compounds could be predicted accurately based solely on molecular structure, with no corrective factor for physical state or the use of other data and was easy to use.

References

1. Gramatica P., Papa E., *QSAR Comb. Sci.* **2003**, *22*, 374-385.
2. Huuskonen J., *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773-777.
3. Huibers P. D. T., Katritzky A. R., *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 283-292.
4. Hemmateenejad B., Shamsipur M., *Internet Electronic Journal of Molecular Design*, **2004**, *3*, 316-334.
5. Hemmateenejad B., *Chemom. Intell, Lab. Syst.* **2005**, *75*, 231-245.
6. Hemmateenejad B., Safarpour M. A., Miri R., Nesari N., *J. Chem. Inf. Model.* **2005,** *45*, 190-199.
7. Kartritzky A. R., Lobanov V. S., Karelson M., *Chemical Society Reviews*, **1995**, 279-287.
8. Kier L. B., Hall L. H., in, *Molecular Connectivity in Chemistry and Drug Research*, Academic Press, New York, **1976**.
9. Koprinarov I. N., Hitchcock A. P., McCrory C. T., Childs R. F., *J. Phys. Chem. B.* **2002**, *106*, 5358-5364.

10. Luan F., Liu H.T., Ma W.P., Fan B.T., *Ecotoxicology and Environmental Safety*, **2008**, 71, 731-739.
11. Murugan R., Grendze M. P., Toomey J. E., Kartritzky A. R., Karelson M., Lobanov V. S., Rachwal P., *Chemtech*, **1994**, 17-23.
12. Mitchell B. E., Jurs P. C., *J. Chem. Inf. Comput. Sci.* **1998,** *38*, 489–496.
13. Miller M. M., Ghodbane S., Wasik S. P., Tewari Y. B., Martire D. E., *J. Chem. Eng. Data*, **1984**, *29*, 184-190.
14. Pan Y. , Jiang J., Wang R., Cao H., Zhao J., *Journal of Hazardous Materials*, **2008**, 157, 510-517.
15. Ruelle P., Kesselring U. W., *J. Pharm. Sci.* **1997**, *86*, 179-186.
16. Rumelhart D. E., Hinton G. E., R. Williams J., *Nature*, **1986**, *323*, 533-539.
17. Shamsipur M., Hemmateenejad B., Akhond M., *Bull. Korean Chem. Soc.* **2004**, *25*, 1-7.
18. Shamsipur M., Ghavami R., Hemmateenejad B., Sharghi H., *QSAR & Comb. Sci.* **2004,** *23*, 734-753.
19. Shamsipur M., Ghavami R., Hemmateenejad B., Sharghi H., *Internet Electronic Journal of Molecular Design*, **2005**, *4*, 882-910.
20. Shamsipur M., Ghavami R., Sharghi H., Hemmateenejad B., *Journal of Molecular Graphics and Modelling*, **2008**, 27, 506-511
21. Sun L., Zhou L.,Yu Y., Lan Y., Li Z., *Chemosphere*, **2007**, 66, 1039-1051
22. Sutter J. M., Jurs P. C., *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 100-107.