Artificial Neural Network Prediction of Normalized Polarity Parameter for Various Solvents with Diverse Chemical Structures

Aziz Habibi-Yangjeh

Department of Chemistry, Faculty of Science, University of Mohaghegh Ardabili, P.O. Box 179, Ardabil, Iran
E-mail: habibiyangjeh@yahoo.com; ahabibi@uma.ac.ir
Received October 31, 2006

Artificial neural networks (ANNs) are successfully developed for the modeling and prediction of normalized polarity parameter (E_T^N) of 216 various solvents with diverse chemical structures using a quantitative-structure property relationship. ANN with architecture 5-9-1 is generated using five molecular descriptors appearing in the multi-parameter linear regression (MLR) model. The most positive charge of a hydrogen atom (q^+), total charge in molecule (q_t), molecular volume of solvent (V_m), dipole moment (μ) and polarizability term (π_l) are input descriptors and its output is E_T^N. It is found that properly selected and trained neural network with 192 solvents could fairly represent the dependence of normalized polarity parameter on molecular descriptors. For evaluation of the predictive power of the generated ANN, an optimized network is applied for prediction of the E_T^N values of 24 solvents in the prediction set, which are not used in the optimization procedure. Correlation coefficient (R) and root mean square error (RMSE) of 0.903 and 0.0887 for prediction set by MLR model should be compared with the values of 0.985 and 0.0375 by ANN model. These improvements are due to the fact that the E_T^N of solvents shows non-linear correlations with the molecular descriptors.

Key Words : Quantitative-structure property relationship, Normalized polarity parameter, Artificial neural networks, Theoretical descriptors

Introduction

The energetic level of molecules may be modified by interactions with surrounding molecules of solvents and it may be difficult to relate chemical properties to molecular structures.[1] The strong influence of solvent on chemical and physical processes (for example, reaction rates, selectivity, chemical equilibria, position and intensity of spectral absorption bands and liquid chromatographic separations, etc.) has well established.[1-8] The use of solvatochromic indicators is a suitable method for studing solute-solvent interactions, since the transition energy of the indicators depends on the solvation's sphere composition and properties.[1] The solvatochromic parameter for measuring empirically the polarity of solvents, $E_T(30)$, is calculated from the maxima of absorbance of the betaine dye as a solution in the solvent under investigation at 25 °C and at a pressure of 0.1 MPa expressed in wavenumber.[1] The solvatochromic parameter is demonstrated to be successful in correlating a wide range of chemical and physical properties involving solute-solvent interactions as well as biological activities of compounds.[1] Normalized polarity parameter (E_T^N) is a dimensionless "normalized" scale, defined by equation (1) in reference to tetramethylsilane (TMS) and water.[1,2]

$$E_T^N = \frac{E_T(30) - E_T(30)_{TMS}}{E_T(30)_{Water} - E_T(30)_{TMS}} \quad (1)$$

The macroscopic (bulk) properties of chemical compounds clearly depend on their microscopic (structural) characteristics. Because of importance of solvent effects, it has been of the highest interest to develop quantitative structure property/activity relationships (QSPR/QSAR), which reflect intermolecular interactions in dense media. Such QSPR/QSAR correlation equations are usually multi-parametric.[2-7] To obtain a significant correlation, it is crucial that appropriate descriptors be employed.[9] Famini *et al.* used theoretical linear solvation energy relationship (TLSER) methodology to correlate E_T^N of 30 solvents with molecular descriptors.[10] The authors concluded that by the TLSER could predict E_T^N values for various solvents and provide better understanding of E_T^N depend on molecular parameters. These descriptors have small cross-correlation, that is to say the descriptors reflect a particular microscopic property nearly without "mixing" or contamination from other descriptors.[10-18]

Table 1 demonstrates the molecular descriptors that have been used in this article. V_m is molecular volume of solvent that inversely proportional to the cohesion energy of molecules. The polarizability term (π_l) is obtained by dividing the polarizability volume by the molecular volume to produce a unitless, size independent quantity, which indicates the ease with which the electron cloud may be moved or polarized. Dipole moment (μ) and total charge in molecule (q_t) terms demonstrate dipole-dipole interactions. The hydrogen-bond donating ability is divided into two components: ε_A (the energy difference between the ε_{HOMO} of water and ε_{LUMO} of solvent) and q^+ (the most positive charge of a hydrogen atom) of solvent molecule. Analogously, the hydrogen-bond accepting ability is divided into two components: ε_B (the energy difference between the ε_{LUMO} of water and ε_{HOMO} of solvent) and q^- (the most negative

Table 1. The molecular descriptors used in the MLR and ANN models[a]

Symbol	Name	Definition	Units
V_m	Molecular volume	Molecular volume	Å^3
π_I	Polarizability index	Polarizability/V_m	none
ε_A	Covalent HB acidity	0.3-0.01($E_l - E_{hw}$)	heV
q+	Electrostatic HB acidity	Maximum(+) charge on an H atom	acu
ε_B	Covalent HB basicity	0.3-0.01($E_{lw} - E_h$)	heV
q^-	Electrostatic HB basicity	Maximum(−) charge on an atom	acu
q_t	Total charge	Total charge on molecule	acu
μ	Dipole moment	Dipole moment	D

[a] Hev = hecto-electron volt (1 heV = 100 ev = 9.6485×10^3 kJmol^{-1}); acu = atomic charge unit; D = debye; HB = hydrogen bond; E_l = LUMO energy and E_h = HOMO energy of the solvent; E_{lw} and E_{hw} refer to the LUMO and HOMO energy for water, respectively.

atomic charge) of solvent.[11-18]

Various methods for constructing QSAR/QSPR models have been used including multi-parameter linear regression (MLR), principal component analysis (PCA) and partial least-squares regression (PLS). In addition, artificial neural networks (ANNs) have become popular due to their success where complex non-linear relationships exist amongst data.[19-21] ANNs are biologically inspired computer programs designed to simulate the way in which the human brain processes information. ANNs gather their knowledge by detecting the patterns and relationships in data and learned (or rained) through experience, not from programming. There are many types of neural networks designed by now and new ones are invented every week.[22] The behavior of a neural network is determined by transfer functions of its neurons, by learning rule, and by the architecture itself. An ANN is formed from artificial neuron or processing elements (PE), connected with coefficients (weights), which constitute the neural structure and are organized in layers. The first layer is termed the input layer, and the last layer is the output layer. The layers of neurons between the input and output layers are called hidden layers. The wide applicability of ANNs stems from their flexibility and ability to model non-linear systems without prior knowledge of an empirical model. Neural networks do not need on explicit formulation of the mathematical or physical relationships of the handled problem. These give ANNs an advantage over traditional fitting methods for some chemical application. For these reason in recent years, ANNs have been used to a wide variety of chemical problems such as simulation of mass spectra, ion interaction chromatography, aqueous solubility and partition coefficient, simulation of nuclear magnetic resonance spectra, prediction of bioconcentration factor, solvent effects on reaction rate, prediction of normalized polarity parameter in mixed solvent systems and dissociation constant of acids.[23-39]

The main aim of the present work is to develop a QSPR model based on molecular descriptors using ANN for modeling and prediction of E_T^N values for various solvents (including 216 solvents) with diverse chemical structures. In the first step, a MLR model was constructed. Then for inspection of non-linear interactions/relation between different parameters of solvents in the model, an ANN model was generated for the prediction of E_T^N values and the results were compared with the experimental and calculated values using MLR model.

Theory

A detailed description of theory behind a neural network has been adequately described by different researchers.[19-21] There are many types of neural network architectures, but the type that has been most useful for QSAR/QSPR studies is the multilayer feed - forward network with back-propagation (BP) learning rule.[22] The number of neurons in the input and output layers are defined by system's properties. The number of neurons in the hidden layer could be considered as an adjustable parameter, which should be optimized. The input layer receives the experimental or theoretical information. The output layer produces the calculated values of dependent variable. The use of ANNs consists of two steps: "training" and "prediction". In the training phase the optimum structure, weight coefficients and biases are searched for. These parameters are found from a training and validation data sets. After the training phase, the trained network can be used to predict (or calculate) the outputs from a set of inputs. ANNs allow one to estimate relationships between input variables and one or several output dependent variables. Information from inputs is fed forward through the network to optimize the weights between neurons. Optimization of the weights is made by backward propagation of the error during training or learning phase. The ANN reads the input and target values in the training data set and changes the values of the weighted links to reduce the difference between the calculated output and target values. The error between output and target values is minimized across many training cycles until network reaches specified level of accuracy. If a network is left to train for too long, however, it will overtrain and will lose the ability to generalize.[34-37]

Methods and Procedure

Data set. As first step for developing the MLR and ANN models, the molecular descriptors should be generate. Normalized polarity parameter, and molecular volume of solvents are literature values.[1,40] In order to calculate the theoretical descriptors, the z-matrices (molecular models) were constructed with the aid of HyperChem 5.01 and molecular structures were optimized using AM1 algorithm. In order to calculate the theoretical descriptors and to find optimized geometries, the molecular geometries of molecules were further optimized with the same algorithm in MOPAC version 6. The molecules in the data sets are including: alkanes, alkenes, haloalkanes, haloalkenes, cycloalkanes,

cycloalkenes, alcohols, esters, ethers, ketones, amines, nitriles, amides, acids, phenols, hetrocyclic, nitro and aromatic compounds. The molecular descriptors were calculated for 216 solvents. The data set was randomly divided into three groups: a training set, a validation set and a prediction set consisting of 168, 24 and 24 molecules, respectively. The training and validation sets were used for the model generation and the prediction set was used for the evaluation of the generated model, because a prediction set is a better estimator of the ANN generalization ability than a validation (monitoring) set.[41]

Linear correlations. MLR model was developed for prediction of normalized polarity parameter by molecular descriptors. The method of stepwise multi-parameter linear regression was used to select the most important descriptors and to calculate the coefficients relating the E_T^N to the descriptors. The MLR models were generated using spss/pc software package. Quality of the equation was indicated by the root mean square error (RMSE), Fisher index of quality (F) and correlation coefficient (R).

Neural network generation. The specification of a typical neural network model requires the choice of the type of inputs, the number of hidden layers, the number of neurons in each hidden layer and the connection structure between the inputs and the output layers. The number of input nodes in the ANNs was equal to the number of molecular descriptors in the MLR model. A three-layer network with a sigmoidal transfer function was designed. The initial weights were randomly selected between 0 and 1. Before training, the input and output values were normalized between 0.1 and 0.9. The optimization of the weights and biases was carried out according to the resilient back-propagation algorithm.[42] For evaluation of the predictive power of the network, the trained ANN was used to predict E_T^N values of the molecules included in the prediction set. The performances of training, validation and prediction of ANNs are evaluated by RMSE, which is defined as follows:

$$\text{RMSE} = \sqrt{\sum_{i=1}^{N} \frac{(P_i^{\exp} - P_i^{cal})^2}{N}} \qquad (2)$$

Where $P_i^{\exp}$ and P_i^{cal} are experimental and calculated values of E_T^N with ANN model and N denote the number of data points.

The processing of the data was carried on Intel Pentium III processor, 800 MHz PC with 256 Mb of RAM in windows XP environment using Matlab 6.5.[42] The neural networks were implemented using Neural Network Toolbox Ver. 4.0 for Matlab.[43]

Results and Discussion

Multi-parameter linear correlation of E_T^N values *vs.* the molecular descriptors for 168 solvents in the training set gives equation (3).

$$E_T^N = 0.391(\pm 0.066) + 2.375(\pm 0.126)q^+ + 0.033(\pm 0.007)\,\mu + 0.0645(\pm 0.012)\,q_t - 0.115(\pm 0.024)V_m - 2.583(\pm 0.577)\,\pi_I \qquad (3)$$

$(n = 168, R = 0.874, RMSE = 0.1043\ F_{5,162} = 104.92)$

$\beta_{q^+} = 0.737$, $\beta_\mu = 0.196$, $\beta_{qt} = 0.245$, $\beta_{Vm} = -0.208$, $\beta\pi_I = -0.176$

It is clear that from eight descriptors in Table 1, five descriptors are important in correlation of E_T^N *vs.* the molecular descriptors. As can be seen, E_T^N of solvents increase with increasing q^+, μ and q_t and decrease with V_m and π_I. Also effects of q^+ and q_t are higher than that of the other descriptors, because standardized coefficients (β values) of q^+ and q_t are higher than that of the other descriptors. The equation is similar to the model obtained for 30 solvents.[10] With increasing V_m and π_I descriptors, normalized polarity parameter decrease. Because, both descriptors are indicative of dispersion effects.[10] Descriptor for electrostatic hydrogen-bond acidity is q^+. With increasing this descriptor, the hydrogen-bonding interactions between the solvent molecules and the betaine dye increases. Dipole-dipole interactions between the molecules of solvents and betaine dye increases with increasing μ and q_t descriptors.

The next step in this work was the generation of ANN model. There are no rigorous theoretical principles for choosing the proper network topology, so different structures were tested in order to obtain the optimal hidden neurons and training cycles.[37] Before training the network, the number of nodes in the hidden layer was optimized. In order to optimize the number of nodes in the hidden layer, several training sessions were conducted with different numbers of hidden nodes (from one to twelve). The root mean squared error of training (RMSET) and validation (RMSEV) sets were obtained at various iterations for different number of neurons at the hidden layer and the minimum value of RMSEV was recorded as the optimum value. Plot of RMSET and RMSEV *vs.* the number of nodes in the hidden layer has been shown in Figure 1. It is clear that the nine

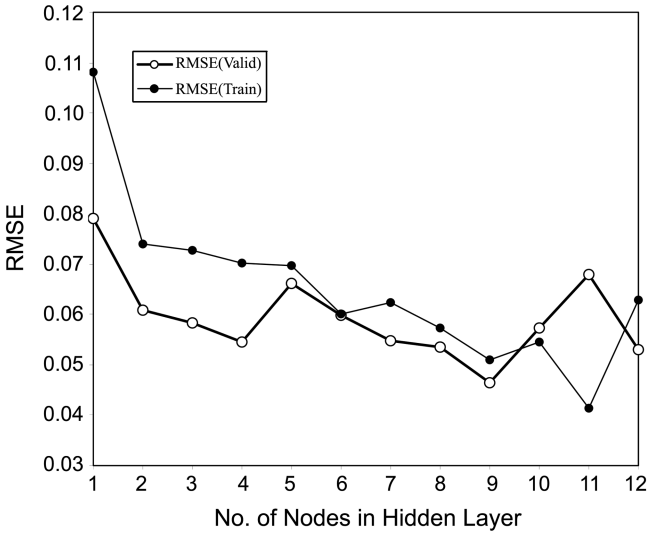

Figure 1. Plot of RMSE for training and validation sets *vs.* the number of nodes **i**n hidden layer.

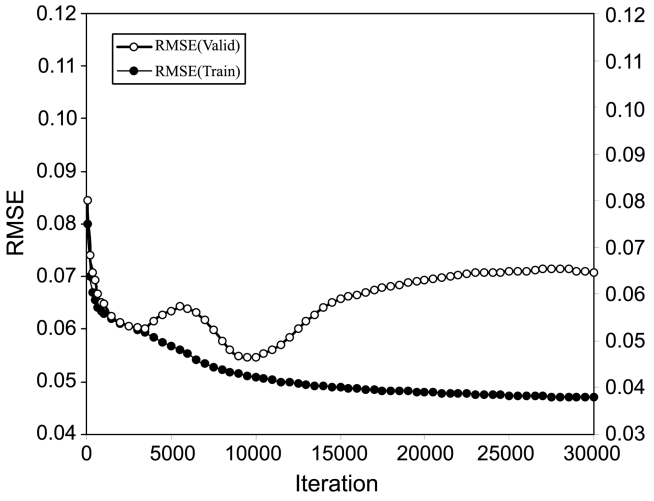

Figure 2. Plot of RMSE for calculated values of E_T^N for training and validation sets *vs.* the number of iterations.

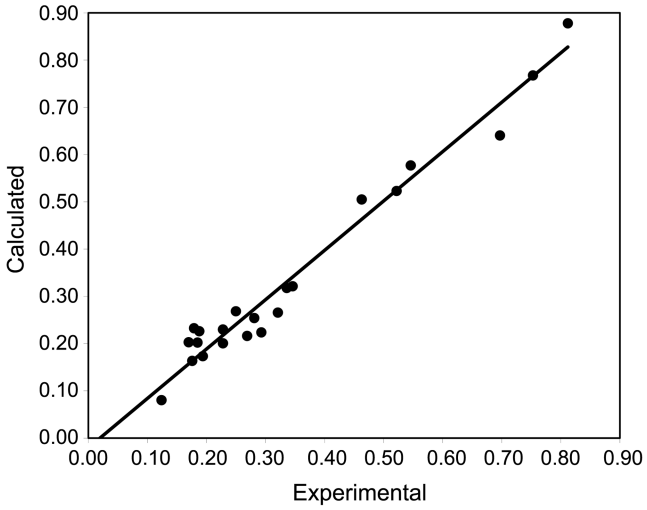

Figure 3. Plot of the calculated values of E_T^N from the ANN model *vs.* the experimental values of it for prediction set.

nodes in hidden layer is optimum value.

This network consists of five inputs (including q^+, μ, π_I, V_m and q_t), the same descriptors in the MLR model, and one output for E_T^N. Then an ANN with architecture 5-9-1 was generated. It is note worthy that training of the network was stopped when the RMSEV started to increases *i.e.* when overtraining begins. The overtraining causes the ANN to loose its prediction power.[37] Therefore, during training of the networks, it is desirable that iterations are stopped when overtraining begins. To control the overtraining of the network during the training procedure, the values of RMSET and RMSEV were calculated and recorded to monitor the extent of the learning in various iterations. Results obtained showed that after 10000 iterations the value of RMSEV started to increase and overfitting began (Figure 2).

The generated ANN was then trained using the training set for the optimization of the weights and biases. For the evaluation of the predictive power of the generated ANN, an optimized network was applied for prediction of the E_T^N values of various solvents in the prediction set, which were not used in the modeling procedure. Then calculated values of the E_T^N for various solvents in training, validation and prediction sets using the ANN model were obtained.

Figure 3 demonstrates plot of the calculated values of E_T^N for 24 solvents in prediction set *versus* the experimental values of it.

As expected, the calculated values of E_T^N are in good agreement with those of the experimental values. The correlation equation for the calculated values of E_T^N in prediction set using the ANN model and the experimental values is as follows:

$$E_T^N(\text{cal}) = 1.0444E_T^N(\text{exp}) - 0.0199 \quad (4)$$

$$(R = 0.985;\ RMSE = 0.0375;\ F_{1,23} = 741.14)$$

Plot of the residual values for E_T^N of solvents in prediction set *versus* the experimental values of it has been demonstrated in Figure 4.

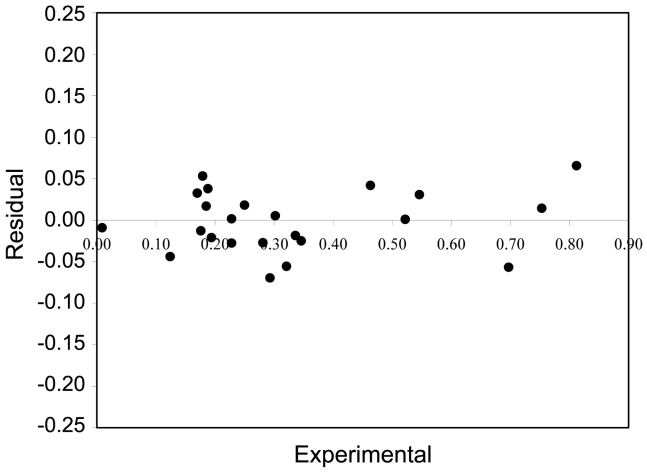

Figure 4. Plot of the Residual for calculated values of E_T^N from the ANN model *vs.* the experimental values of it for prediction set.

As can be seen the model did not show proportional and systematic error, because the slope (a = 1.0444) and intercept (b = 0.0199) of the correlation equation are not significantly different from unity and zero, respectively and the propagation of errors in both sides of zero are random shown in Figure 4.

Table 2 compares the results obtained using the MLR and ANN models. The correlation coefficient (R) and RMSE of the models for total, training, validation and prediction sets show potential of the ANN model for prediction of E_T^N values of various solvents.

As a result, it was found that properly selected and trained neural network could fairly represent the dependence of normalized polarity parameter on molecular descriptors. Then the optimized neural network could simulate the complicated nonlinear relationship between E_T^N values and the molecular descriptors. The correlation coefficient (R) and RMSE are 0.903 and 0.0887 for the prediction set by the MLR model should be compared with the values of 0.985

Table 2. Comparsion of statistical parameters obtained by the MLR and ANN models for correlation of normalized polarity parameter with molecular descriptors[a]

Model	R_{tot}	R_{train}	R_{valid}	R_{pred}	$RMSE_{tot}$	$RMSE_{train}$	$RMSE_{valid}$	$RMSE_{pred}$
MLR	0.876	0.874	0.871	0.903	0.1025	0.1043	0.1027	0.0887
ANN	0.973	0.971	0.975	0.985	0.0492	0.0510	0.0465	0.0375

[a] tot, train, valid and pred in subscript letters are referring to the total, training, validation and prediction sets.

and 0.0375, respectively, for the ANN model. It can be seen from Table 2 that although the parameters appearing in the MLR model are used as inputs for the generated ANN, the statistics is shown a large improvement. These improvements are due to the fact that E_T^N of the solvents shows nonlinear correlations with the molecular descriptors.

Conclusions

A five-descriptor nonlinear computational neural network model has been developed for prediction of normalized polarity parameter for various solvents with diverse chemical structures using quantitative-structure property relationship. Comparison of the values of RMSE and other statistical parameters in Table 2 for training, validation and prediction sets for the models show superiority of the ANN model over the regression model. Root-mean square error of 0.0887 for the prediction set by the MLR model should be compared with the value of 0.0375 for the ANN model. Since the improvement of the results obtained using nonlinear model (ANN) is considerable, it can be concluded that the nonlinear characteristics of molecular descriptors on the E_T^N values of solvents is serious and interactions between various molecular descriptors are important. Then the optimized neural network could simulate the complicated nonlinear relationship between normalized polarity parameter and the molecular structure for various solvents.

References

1. Reichardt, C. *Solvents and Solvent Effects in Organic Chemistry*, 3rd ed.; VCH: 2003; Chap. 4-7.
2. Marcus, Y. *J. Chem. Soc., Perkin Trans. 2* **1994**, 1015.
3. Cativiela, C.; Garcia, J. I.; Gil, J.; Martinez, R. M.; Mayoral, J. A.; Salvatella, L.; Urieta, J. S.; Mainer, A. M.; Abraham, M. H. *J. Chem. Soc., Perkin Trans. 2* **1997**, 653.
4. Gholami, M. R.; Habibi-Yangjeh, A. *Int. J. Chem. Kinet.* **2000**, *32*, 431.
5. Gholami, M. R.; Habibi-Yangjeh, A. *J. Phys. Org. Chem.* **2000**, *13*, 468.
6. Gholami, M. R.; Habibi-Yangjeh, A. *Int. J. Chem. Kinet.* **2001**, *33*, 118.
7. Habibi-Yangjeh, A.; Gholami, M. R.; Mostaghim, R. *J. Phys. Org. Chem.* **2001**, *14*, 884.
8. Harifi, A.-R.; Habibi-Yangjeh, A.; Gholami, M. R. *J. Phys. Chem. B* **2006**, *110*, 7073.
9. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, Germany, 2000.
10. Famini, G. R.; Wilson, L. Y. *J. Phys. Org. Chem.* **1999**, *12*, 645.
11. Famini, G. R.; Penski, C. E.; Wilson, L. Y. *J. Phys. Org. Chem.* **1992**, *5*, 395.
12. Famini, G. R. *Chemosphere* **1997**, *35*, 2417.
13. Lowrey, A. H.; Famini, G. R.; Wilson, L. Y. *J. Chem. Soc., Perkin Trans. 2* **1997**, 1381.
14. Cronce, D. T.; Famini, G. R.; Soto, J. A. D.; Wilson, L. Y. *J. Chem. Soc., Perkin Trans. 2* **1998**, 1293.
15. Engberts, J. B. F. N.; Famini, G. R.; Perjessy, A.; Wilson, L. Y. *J. Phys. Org. Chem.* **1998**, *11*, 261.
16. Famini, G. R.; Benyamin, D.; Kim, C.; Veerawat, R.; Wilson, L. Y. *Collect. Czech. Chem. Commun.* **1999**, *64*, 1727.
17. Habibi-Yangjeh, A. *Indian J. Chem. B* **2003**, *42*, 1478.
18. Habibi-Yangjeh, A. *Indian J. Chem. B* **2004**, *43*, 1504.
19. Patterson, D. W. *Artificial Neural Networks: Theory and Applications*; Simon and Schuster: New York, 1996; Part III, Chap. 6.
20. Bose, N. K.; Liang, P. *Neural Network Fundamentals*; McGraw-Hill: New York, 1996.
21. Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*; Wiley-VCH: Weinhein, 1999.
22. Agatonovic-Kustrin, S.; Beresford, R. *J. Pharm. Biomed. Anal.* **2000**, *22*, 717.
23. Xing, W. L.; He, X. W. *Anal. Chim. Acta* **1997**, *349*, 283.
24. Bunz, A. P.; Braun, B.; Janowsky, R. *Fluid Phase Equilib.* **1999**, *158*, 367.
25. Homer, J.; Generalis, S. C.; Robson, J. H. *Phys. Chem. Chem. Phys.* **1999**, *1*, 4075.
26. Goll, E. S.; Jurs, P. C. *J. Chem. Inf. Comp. Sci.* **1999**, *39*, 974.
27. Vendrame, R.; Braga, R. S.; Takahata, Y.; Galvao, D. S. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1094.
28. Gaspelin, M.; Tusar, L.; Smid-Korbar, J.; Zupan, J.; Kristl, J. *Int. J. Pharm.* **2000**, *196*, 37.
29. Gini, G.; Cracium, M. V.; Konig, C.; Benfenati, E. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1897.
30. Urata, S.; Takada, A.; Uchimaru, T.; Chandra, A. K.; Sekiya, A. *J. Fluorine Chem.* **2002**, *16*, 163.
31. Jalali-Heravi, M.; Masoum, S.; Shahbazikhah, P. *J. Magn. Reson.* **2004**, *171*, 176.
32. Koziol, J. *Internet Electron. J. Mol. Des.* **2002**, *1*, 80.
33. Wegner, J. K.; Zell, A. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1077.
34. Valkova, I.; Vracko, M.; Basak, S. C. *Anal. Chim. Acta* **2004**, *509*, 179.
35. Jalali-Heravi, M.; Masoum, S.; Shahbazikhah, P. *J. Magn. Reson.* **2004**, *171*, 176.
36. Habibi-Yangjeh, A.; Nooshyar, M. *Bull. Korean Chem. Soc.* **2005**, *6*, 139.
37. Habibi-Yangjeh, A.; Nooshyar, M. *Physics and Chemistry of Liquids* **2005**, *43*, 239.
38. Habibi-Yangjeh, A.; Danandeh-Jenagharad, M.; Nooshyar, M. *Bull. Korean Chem. Soc.* **2005**, *26*, 2007.
39. Habibi-Yangjeh, A.; Danandeh-Jenagharad, M.; Nooshyar, M. *J. Mol. Model* **2006**, *12*, 338.
40. Marcus, Y. *The Properties of Solvents*; John Wiley and Sons: New York, 1999.
41. Turner, J. V.; Maddalena, D. J.; Cutler, D. J. *Int. J. Pharm.* **2004**, *270*, 209.
42. *Matlab 6.5*; Mathworks: 1984-2002.
43. Demuth, H.; Beale, M. *Neural Network Toolbox*; Mathworks: Natick, MA, 2000.