Journal of Molecular Graphics and Modelling 27 (2008) 506–511

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

Highly correlating distance/connectivity-based topological indices 5. Accurate prediction of liquid density of organic molecules using PCR and PC-ANN

Mojtaba Shamsipur [a], Raouf Ghavami [b], Hashem Sharghi [b], Bahram Hemmateenejad [b,c,*]

[a] Department of Chemistry, Razi University, Kermanshah, Iran
[b] Department of Chemistry, Shiraz University, Shiraz, Iran
[c] Medicinal & Natural Products Chemistry Research Center, Shiraz University of Medical Sciences, Shiraz, Iran

ARTICLE INFO

Article history:
Received 15 March 2008
Received in revised form 16 August 2008
Accepted 2 September 2008
Available online 13 September 2008

Keywords:
Topological Sh indices
Liquid density
Principal component regression
Artificial neural network
Correlation ranking

ABSTRACT

The primary goal of a quantitative structure–property relationship (QSPR) is to identify a set of structurally based numerical descriptors that can be mathematically linked to a property of interest. Recently, we proposed some new topological indices (Sh indices) based on the distance sum and connectivity of a molecular graph that derived directly from two-dimensional molecular topology for use in QSAR/QSPR studies. In this study, the ability of these indices to predict the liquid densities (ρ) of a large and diverse set of organic liquid compounds (521 compounds) has been examined. Ten different Sh indices were calculated for each molecule. Both linear and non-linear modeling methods were implemented using principal component regression (PCR) and principal component-artificial neural network (PC-ANN) with back-propagation learning algorithm, respectively. Correlation ranking procedure was used to rank the principal components and entered them into the models. PCR analysis of the data showed that the proposed Sh indices could explain about 91.82% of variations in the density data, while the variations explained by the ANN modeling were more than 97.93%. The predictive ability of the models was evaluated using external test set molecules and root mean square errors of prediction of 0.0308 g ml^{-1} and 0.0248 g ml^{-1} were obtained for liquid densities of external compounds by linear and non-linear models, respectively.

© 2008 Elsevier Inc. All rights reserved.

1. Introduction

In general, development of a QSPR involves three steps: structural encoding, feature selection, and model building. Structural encoding involves the use of numerical descriptors to encode the structural features of a compound. Feature selection is then employed to determine which subset of the descriptors best relates to the property of interest. Models built from the best subset of descriptors from a direct link between descriptors and the property of interest. Finally, validation determines the level of the model's predictive capabilities for unknown compounds.

Topological indices (TIs) play an important role for the analysis of molecular diversity and lead to optimization through the well-established quantitative structure–property/activity relationships (QSPR/QSAR) [1–5]. These indices are numerical quantities generated from a graph–theoretical representation of the molecular structure through mathematical invariants [2,6]. So far, the kinds of TIs that have been reported in the literature have exceeded 100, which include Randic index [7], Balaban index [8], Wiener index [9] and Schultz index [10]. Despite the large achievements attained in this field, existing topological indices approaches to QSAR/QSPR need further improvements to obtain indices with higher correlating ability and more generalization. Recently, we proposed a set of new topological indices, named as Shamsipur indices (Sh_1–Sh_{10} indices) and used them for prediction of different physical and thermodynamic functions of alkanes and alkenes isomers and [11–13] and octanol–water partition coefficient of organic compounds [14].

In QSAR/QSPR studies, a regression model of the form $y = X\mathbf{b} + e$ may be used to describe a set of predictor variables (X) with a predicted variable (y) by means of a regression vector ($\mathbf{b}$). However, the collinearity, which often existed between independent variables, creates a severe problem in certain types of mathematical treatment such as matrix inversion [15]. A better predictive model can be obtained by ortogonalization of the

* Corresponding author at: Department of Chemistry, Shiraz University, Shiraz, Iran. Tel.: +98 711 2284822; fax: +98 711 2286008.
E-mail address: hemmatb@sums.ac.ir (B. Hemmateenejad).

1093-3263/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2008.09.005

variables by means of principal component analysis (PCA) and the consequent method is called principal component regression (PCR) [16–18]. In order to reduce the dimensionality of the independent variable space, a limited number of principal components (PCs) are used and therefore a major question will arise after the PCA is how many and which PCs constitute a good subset for predictive purposes? Hence, the selection of significant and informative PCs is the main problem in almost all PCA-based calibration methods [19–21].

Different methods have been addressed to select the significant PCs for calibration purposes. The simplest and most common one is a top-down variable selection where the factors are ranked in the order of decreasing Eigen-values and the factors are introduced into the calibration model one after the other. In the other method, called correlation ranking, the factors are ranked by their correlation coefficient with the property to be correlated (*i.e.* a dependent variable) [21]. Better results are often achieved by this method.

Because of the complexity of the relationships existed between the activity/property of the molecules and the structures, non-linear modeling methods are often used to model the structure–activity/property relationships. Artificial neural networks (ANNs) as non-parametric non-linear modeling techniques have attracted increasing interest in the recent years [22,23]. Multilayer feed-forward neural networks (MLF-ANN) trained with back-propagation learning algorithm become increasingly popular techniques [22–25]. The flexibility of ANN for discovering a more complex relationship causes that this method find wide application in QSAR/QSPR studies, which recently reviewed by Duch et al. [26]. The principal component-artificial neural network (PC-ANN), which combines the PCA with ANN, is another version of the PCR, which models the non-linear relationships between the PCs and dependent variable was proposed by Gemperline et al. to improve the training speed and decrease the overall calibration error [27].

Normal density (ρ) is one of the important physicochemical property used to characterize and identify a compound and is defined as the ratio of mass to volume at pressure of 1 atm and temperature of 20 °C. In addition, densities can be used to predict or estimate other physical properties such as critical pressure, viscosity, thermal conductivity, diffusion coefficients, and surface tension [28]. Commonly, the density of a substance is inversely proportional to the temperature and is determined by a geometric factor describing or relating to the size and shape of the electron cloud of the molecule. In addition, the magnitude of intermolecular forces determines how strongly or how loosely the molecules are held together. Since there is a wide variety in the shapes of organic molecules as well as a dynamic interplay between the molecules under thermal motion, it is expected that the experimental value of density is a complex composite of several factors. The rapid growth of combinatorial chemistry, where literally millions of new compounds are synthesized and tested without isolation, could render computational prediction methods of liquid density very useful for many applications [29,30]. A successful strategy for predicting liquid density is development of QSPR models. Recently, a QSPR model by using the CODESSA (comprehensive descriptors for structural and statistical analysis) program for the correlation and prediction of densities of 303 organic liquid compounds has been reported [31].

The goal of the present investigation is to examine the ability of the proposed Sh indices as predictor variables in QSPR-based prediction of the liquid density of organic compounds. Both linear and non-linear modeling methods were employed for estimating the liquid density of an extensive set of organic compounds including several structurally diverse groups of compounds (alkanes, alkenes, alkynes, cycloalkanes, cycloalkenes, aliphatic alcohols, ethers, esters, aldehydes, ketones, carboxylic acids, amines, aromatic hydrocarbons, nitrile and nitro).

2. Experimental

2.1. Liquid density data

The QSPR treatment started with the assembly of the data set. The observed liquid density data of diverse organic compounds were recompiled from the several literature sources [17,33–37]. The choice was based on maximum diversity of the structure of compounds and the numerical values of densities. The final set of 521 diverse organic compounds was representative for all major classes of organic compounds containing C, H, N, and O, and included saturated and unsaturated hydrocarbons, aldehyde, ketone, amino, ether, ester, carboxylic acid, hydroxyl, nitrile, nitro, aromatic and non-aromatic cyclic compounds (Table S1; Supplementary materials).

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 [11–14]. A home-made program (written in MATLAB environment) was used to calculate the Sh indices. The calculated indices were collected in a data matrix with 521 × 10 dimension. Each chemical is now a point in the 10-dimensional space, X^{10} (Table S2; supplementary materials).

2.3. Linear modeling: principal component regression

Because of some collinearity 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 [32]:

$$\mathbf{D} = \mathrm{U}\mathbf{S}\mathrm{V}^{\mathrm{T}} \quad (1)$$

where **U** and **V** are the orthonormal matrices spanned the respective row and column spaces of the data matrix (**D**) and **D** is the descriptors data matrix of the calculated Sh indices whose number of rows and columns are the number of molecules and number of Sh indices, respectively. **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 data set of 521 compounds was randomly divided to 431 calibration (or training) samples and 90 prediction samples. The PCs of the calibration samples (included in the row matrix **U**) were calculated by Eq. (1) and those of prediction samples were calculated by Eq. (2) using calculated **S** and **V** matrices of calibration data.

$$\mathbf{U}_{\mathrm{p}} = \mathbf{D}_{\mathrm{p}}\mathbf{S}^{-1}\mathbf{V} \quad (2)$$

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 liquid density 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 density data was determined first. The stepwise entrance of the PCs to the PCR model was based on their decreasing correlation with the density. Some statistical parameters such as the squared of the correlation coefficient (R^2), squared of the leave-one-out cross-validation correlation coefficient (R^2_{CV}), the standard error of estimation (SE), the root-mean-square error (RMS), relative error of prediction (REP) and the Fisher's criterion at the 95% probability level were calculated to estimate the quality of the resulted models.

2.4. Non-linear modeling: PC-ANN

To model the liquid density-Sh indices more accurately, artificial neural network was employed to process the non-linear relationships between the selected PCs in the previous section and the density data. The PC-ANN model was the same as we reported previously [23,33]. In the same manner as PCR analysis, the data sets were classified into calibration (or training) and prediction sets (see Section 2.3), however, since ANN needs a validation set through learning procedure, 51 samples of calibration set was selected randomly as validation samples. The validation 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 PCs of the validation set were used according to Eq. (2). Using the same prediction samples for PCR and PC-ANN models allows us to have a real comparison between linear and non-linear methods.

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. The summation is put through a non-linear transfer function, here a sigmoid, and then the resulting values are passed to the output layer, which contains a single neuron which represents the predicted density values.

2.5. Software

A Pentium III personal computer with Windows XP operating system was used. All the necessary programs for PCA, PCR, ANN, calculation of Sh indices and other statistical analysis were written in MATLAB (ver. 6.5, MathWork Inc.) environment.

3. Results and discussion

Table S1 (supplementary materials) lists the name of the compounds used in this study and their corresponding experimental liquid density value. In this list, the experimental density values ranged from 0.5943 (2-methylpropene) to 1.2203 (formic acid) g ml^{-1} with the respective average and standard deviation of 0.7860 and 0.109. To show the distribution of the experimental densities, their histogram-plot is shown in Fig. 1. The histogram, which not shows a normal distribution of the density data, indicates that the density of 83.11% of the compounds is located between 0.5943 and 0.8788 g ml^{-1}. Besides, it is obvious from the histogram that 7.50% of compounds have a density greater than 1.000 g ml^{-1}. Among these, four of them including

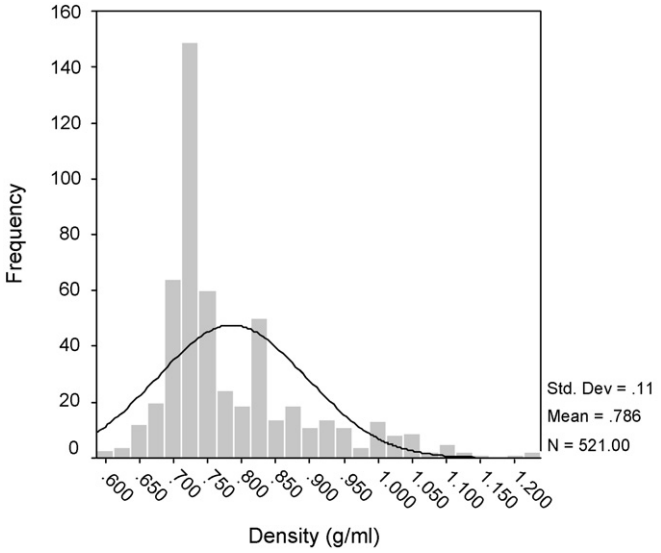

Fig. 1. Histogram of the distribution of the observed liquid density for the total data set of 521 organic compounds used in this study. The solid curve is the fitting of the density data to the normal distribution.

3-nitrotoluene, nitrobenzene, 2-naphthol and, formic acid possess high density.

The 10 Sh topological indices, which are easily calculated from the two-dimensional structure of the molecules, can be obtained from supplementary materials. Science there is some collinearity between the Sh indices; orthogonal transformation of the indices was performed by principal component analysis on the Sh data matrix. The overall goal of PCA is to reduce the dimensionality of a data set, while simultaneously retaining the information present in the data. The results showed that the eight PCs could explain 99.98% of variances in the Sh data matrix and therefore these PCs were used as the input variables of PCR and PC-ANN models.

3.1. PCR modeling

As it is shown in Table S1, a wide variety of organic molecules including saturated and unsaturated hydrocarbons, esters, aldehydes, organic acids, alcohols, ethers, amines, nitrile, nitro, and aromatic compounds have been investigated in this article. First, attempts were made to develop PCR-based QSPR models for each subset of compounds, separately. Therefore, the original set of 521 compounds was divided into subsets according to the elements represented in the compounds: (1) 336 aliphatic and aromatic compounds containing only C and H atoms (set CH); (2) 51 hydroxy containing compounds (set OH); (3) 37 nitrogen containing compounds (set N); (4) 97 different type of other compounds with wide variety of functional groups (mixed set). 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 1. As can be seen, the PCR model obtained for set *N* used 7 numbers of PCs and the number of factors used for CH and OH subsets are 8. For all subsets, the factors selected by the correlation ranking procedures are different from those of Eigen-value ranking. The models obtained for all subsets resulted in high statistical qualities, measured by the squares of correlation coefficient ($R^2 > 0.96$), root-mean-square error (RMS < 0.0254), and relative error of prediction (REP < 2.74). The higher statistical qualities obtained for the CH subset (containing only C and H atoms) can be attributed to simple structure of these compounds. It should be noted that the results obtained by the CR-PCR procedure

Table 1
Linear multivariate regression models and statistical parameters of compounds properties using PC indices

Subset	*N*	Equation	R^2	S.E.	RMS	REP	*F*	R^2_{CV}
CH	336	$\rho = 0.7313 + 0.0468\ PC_2 + 0.0270\ PC_1 + 0.0111\ PC_3 - 0.0079\ PC_8 - 0.0063\ PC_4 - 0.0043\ PC_{10} - 0.0038\ PC_5 + 0.0032\ PC_6$	0.9568	0.0121	0.0120	1.64	905	0.9532
OH	51	$\rho = 0.8628 + 0.0863\ PC_2 + 0.0255\ PC_7 + 0.0169\ PC_1 - 0.0126\ PC_3 - 0.0108\ PC_4 - 0.0082\ PC_5 - 0.0070\ PC_{10} - 0.0051\ PC_9$	0.9776	0.0155	0.0141	1.63	229	0.9355
N	37	$\rho = 0.9265 + 0.1145\ PC_2 + 0.0747\ PC_1 - 0.0497\ PC_3 - 0.0311\ PC_5 + 0.0187\ PC_7 - 0.0152\ PC_6 + 0.0128\ PC_9$	0.9719	0.0287	0.0254	2.74	143	0.9508
Total	521	$\rho = 0.7856 + 0.0884\ PC_2 - 0.0361\ PC_3 + 0.0271\ PC_6 - 0.0176\ PC_5 + 0.0167\ PC_1 + 0.0154\ PC_8 - 0.0148\ PC_9 + 0.0056\ PC_7$	0.9214	0.0308	0.0305	3.88	750	0.9131

N denotes number of structures.

were better than EV-PCR. Therefore, the results of the latter method are not included in Table 1.

The usefulness of QSPR models is not just their ability to reproduce known data, but also they should have ability to produce a good estimation for any external sample [34]. The predictive abilities of models are strongly affected by the over-fitting problem. In QSPR analyses, over-fitting problem is obtained when uninformative variables enter to the models. Another source of over-fitting is the use of exceeded number of factors in PCA-based regression methods such as PCR. There are several methods in use to estimate the quality of the models [35–38]. Cross-validation is the most frequently used validation methods [39]. Therefore, to check the prediction ability and overfitting of the resulting models, the leave-one-out cross-validation (LOO-CV) procedure was applied for each subset of models. The squared correlation coefficient for cross-validation (R^2_{CV}) was then calculated by the following equation $R^2_{CV} = 1 - (PRESS/SSD)$, where PRESS and SSD are the predicted residual sum of squares and the sum of the squared deviation from the mean, respectively. The results of LOO-CV examination for each subset of organic compounds are listed in the last column of Table 1. The cross-validation results show that all models (regression expressions) presented in the Table 1 have R^2_{CV} values greater than 0.93; therefore, all are reasonable QSPR models. This shows the success of employed Sh indices in modeling the liquid density of different subsets of organic compounds.

Although the QSPR models obtained for subsets of molecules had good performances and could predict the density of the related molecules with low error, it will be more helpful if a single model can be obtained for modeling the density of the entire set of molecules. In the last row of Table 1, the CR-PCR model obtained for the density of entire set of compounds by the correlation ranking procedure is listed. The trend of PCs, in the order of decreasing their correlation, is PC2 > PC3 > PC6 > PC5 > PC1 > PC8 > PC9 > PC7, which is not in the same direction as their decreasing Eigen-values. The resulting equation had correlation coefficients $R^2 = 0.9214$, RMS = 0.0305, REP = 3.88, $F = 750$, $R^2_{CV} = 0.9131$. The eight factors used in this equation can explain 92.14% of the variance in the density of all data set of liquid organic compounds.

Further attempts were made to examine the quality of the resulted model by randomly splitting the data set into the calibration set (431 molecules) and prediction set (90 molecules). The resulted CR-PCR model was the same as that obtained for entire set of molecules. The R^2 value, RMS error and REP for the prediction set are 0.9352, 0.0308, and 3.92, respectively. This means that the eight PCs selected by correlation ranking procedure can explain at least 93.52% variance in density values of the external data. The calculated density obtained with this method are presented in Table S1 (supplementary materials), the corresponding graph of calculated versus observed density is given in Fig. 2 and the statistical parameters for the best-fitted model are given in Table 2.

3.2. PC-ANN modeling

To increase the prediction ability of the models obtained between the PCs and density, a non-linear modeling method was employed in this study. Typically, superior models can be found using ANNs because they implement non-linear relationships and

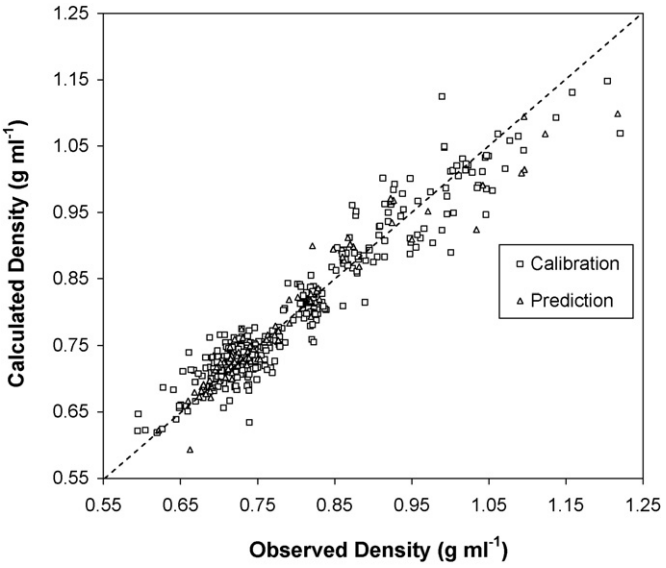

Fig. 2. Plot of the calculated liquid density by CR-PCR against the observed values. The dash line is the ideal fit to the straight line.

Table 2
Statistics of principal component regression and artificial neural network models with one hidden-layer neurons for calculating liquid density

Statistical parameters	CR-PCR		CR-PC-ANN		
	Calibration set	Prediction set	Calibration set	Prediction set	Validation set
N	431	90	380	90	51
S.E.	0.0309	0.0272	0.0156	0.0248	0.0220
RMS	0.0306	0.0308	0.0156	0.0262	0.0220
REP	3.89	3.92	1.98	3.35	2.78
R^2	0.9182	0.9352	0.9793	0.9562	0.9529
%RE range	(−14.18)–(13.70)	(−9.58)–(10.59)	(−5.79)–(7.10)	(−10.18)–(8.35)	(−9.48)–(5.54)

because they have more adjustable parameters than the linear models. Therefore, in this study we suggested the use of ANN as the non-linear model. A fully connected, three-layered feed-forward ANN model with back-propagation [40] learning algorithm was employed. In the same manner as PCR analysis, correlation ranking was used to select the most relevant set of PCs as input of ANN. The set of PCs used in PCR analysis (*i.e.* PC2, PC3, PC6, PC5, PC1, PC8, PC9, and PC7, ranked based on decreasing correlation) was selected based on linear correlation. This subset of PCs can be used as input of ANN model. However, since ANN is a non-linear method, better results may be obtained if non-linear correlation for PC ranking is used. The non-linear correlation ranking of PCs for us in ANN was previously developed in our research group by Hemmateenejad [41]. The results of the ANN-based non-linear correlation between each extracted PCs and liquid density is represented in Table 3. As it is observed, the order of PCs based on their decreasing non-linear correlation is PC2, PC3, PC6, PC5, PC1, PC8, PC9, and PC7, which is completely matched with that obtained by linear correlation ranking. Thus, this subset of PCs was used as input of ANN model.

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 was split into a calibration set, a prediction set and a validation set.

The eight PCs were test with several ANN architectures. The ANN models were confined to a single hidden layer and a sigmoid transfer function, as a more versatile transfer function, was used in this layer. 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 prediction set; the corresponding value is then calculated for the validation set for each of configuration. 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 RMS error of the prediction 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 further changed.

A response surface methodology was used to optimize network parameters. The surface plot of RMS error as a function of linear rate and momentum in three different numbers of nodes in hidden layer is shown in Fig. 3. The results indicate that an ANN with 8 PCs as input variables, 6 nodes in its hidden layer (8-6-1 structure), learning rate of 0.25, and momentum of 0.75 resulted in the optimum network performance. The network was trained using calibration data and it was evaluated by prediction samples. The best predicon results were obtained after 15000 iterations. The

Table 3
The results of ANN modeling between each extracted PC and liquid density

PC No.	N_H[a]	TF_H[b]	R_P^2	$RMSE_P$
1	6	Tan	0.2361	0.0888
2	5	Log	0.6404	0.0617
3	7	Log	0.3901	0.0787
4	5	Tan	0.0267	0.0994
5	6	Log	0.2530	0.0927
6	6	Tan	0.3193	0.0843
7	5	Log	0.0345	0.1044
8	7	Tan	0.1371	0.0991
9	7	Log	0.0920	0.0959
10	6	Log	0.0152	0.0998

Tan: tangent, Log: logarithm.
[a] Number of nodes in hidden layer.
[b] Type of transfer function used in hidden layer.

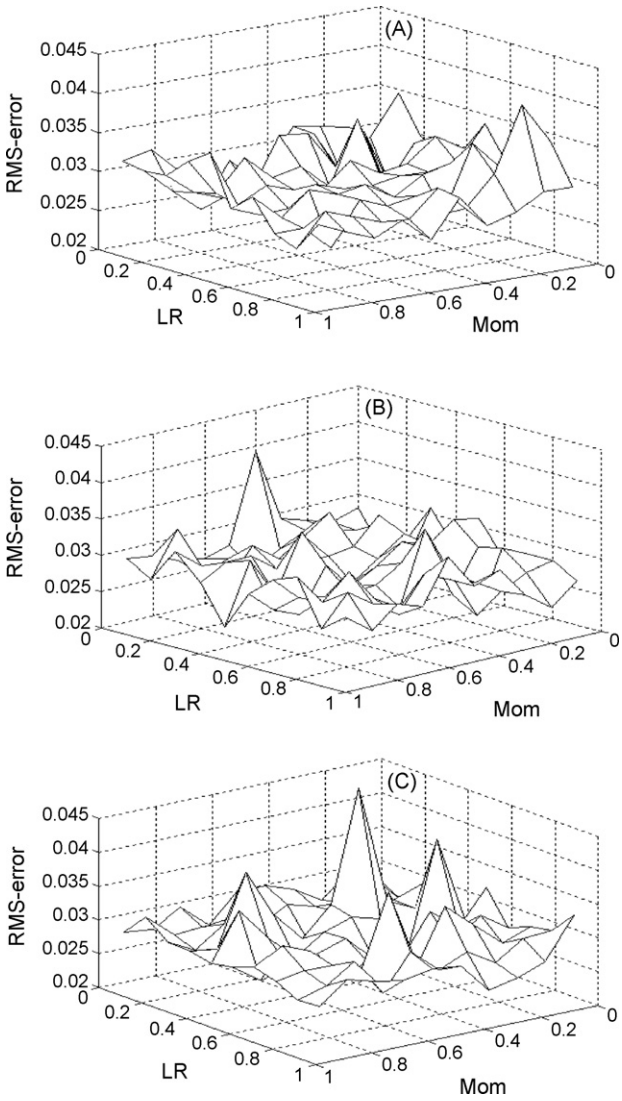

Fig. 3. Optimization of linear rate (LR), momentum (Mom) and number of hidden layer nodes (n_H) for ANN modeling; (A) $n_H = 5$; (B) $n_H = 6$ and (C) $n_H = 7$.

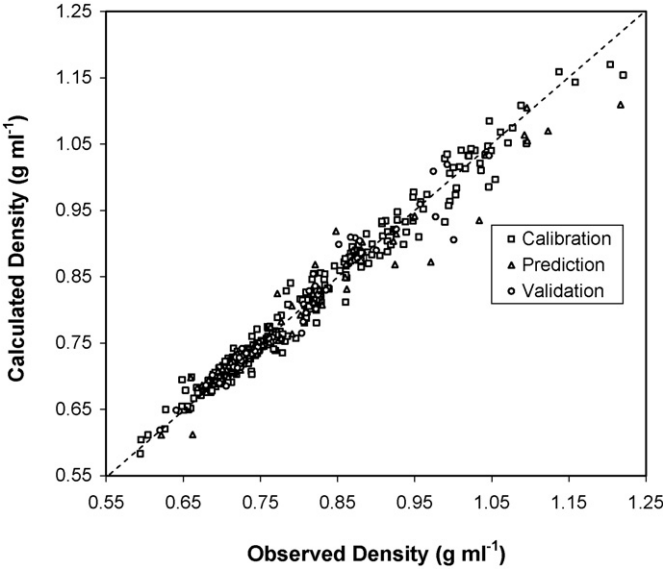

Fig. 4. Plot of the calculated density by PC-CR-ANN against the observed values. The dash line is the ideal fit to the straight line.

predicted values of liquid density resulted from the optimized PC-ANN procedures model are shown in Table S1 and are plotted in Fig. 4 against the corresponding observed values, and the statistical parameters for the best-fitted model are represented in Table 2. 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 non-linear relationships between the proposed Sh topological indices and the density 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 [41–44].

4. Conclusions

The usefulness of the some newly proposed topological indices (Sh indices) in quantitative structure–liquid density relationship analysis were examined to predict the liquid density of a wide variety of 521 organic compounds with various heteroatoms by using principal component regression and principal component-artificial neural network modeling methods. PCR analysis of the data showed that proposed Sh indices could explain about 91.82% of variations in the density data; while the variations explained by the ANN modeling were more than 97.93%. The linear and non-linear models could predict the density of molecules with the respective root mean square errors lower than 0.0305 g ml^{-1} and 0.0248 g ml^{-1}. These results demonstrated that liquid densities 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. These results confirm the suitability of the indices in QSPR analysis of the liquid density data.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmgm.2008.09.005.

References

[1] N. Trinajstić, Chemical Graph Theory, ACS, Boca Raton, FL, 1992.
[2] J. Devillers, A.T. Balaban, Topological Indices and Related Descriptors in QSAR and QSPR, Gordon & Breach, New York, 1999.
[3] M. Karelson, in: A. Jossey-Bass (Ed.), Molecular Descriptors in QSAR/QSPR, Wiley, New York, 2000.
[4] J. Devillers, New trends in (Q)SAR modeling with topological indices, Curr. Opin. Drug. Discov. Dev. 3 (2000) 275–279.
[5] L. Pogliani, From molecular connectivity indices to semiempirical connectivity terms: recent trends in graph theoretical descriptors, Chem. Rev. 100 (2000) 3827–3858.
[6] M. Randic, in: P.V.R. Schleyer, L. Allinger, T. Clark, J. Gasteiger, P.A. Kollman, H.F. Schaefer (Eds.), The Encyclopedia of Computational Chemistry, Wiley, Chichester, UK, 1998.
[7] M. Randic, Characterization of molecular branching, J. Am. Chem. Soc. 97 (1975) 6609–6615.
[8] A.T. Balaban, Applications of graph–theory in chemistry, J. Chem. Inf. Comput. Sci. 25 (1985) 334–343.
[9] H. Wiener, Correlation of heats of isomerization, and differences in heats of vaporization of isomers, among the paraffin hydrocarbons, J. Am. Chem. Soc. 69 (1947) 2636–2638.
[10] H.P. Schultz, Topological, Organic-chemistry. 1. Graph–theory and topological indexes of alkanes, J. Chem. Inf. Comput. Sci. 29 (1989) 227–228.
[11] M. Shamsipur, B. Hemmateenejad, M. Akhond, Highly correlating distance/connectivity-based topological indices. 1. QSPR studies of alkanes, Bull. Korean Chem. Soc. 25 (2004) 253–259.
[12] M. Shamsipur, R. Ghavami, B. Hemmateenejad, H. Sharghi, Highly correlating distance-connectivity-based topological indices. 2. Prediction of 15 properties of a large set of alkanes using a stepwise factor selection-based PCR analysis, QSAR Comb. Sci. 23 (2004) 734–753.
[13] M. Shamsipur, B. Hemmateenejad, R. Ghavami, H. Sharghi, Highly correlating distance-connectivity-based topological indices. 4. Stepwise factor selection-based PCR models for QSPR study of 14 properties of monoalkenes, Polish J. Chem. 81 (2007) 269–294.
[14] M. Shamsipur, R. Ghavami, B. Hemmateenejad, H. Sharghi, Highly correlating distance-connectivity based topological indices. 3. PCR and PC-ANN based prediction of the octanol–water partition coefficient of diverse organic molecules, Int. Elect. J. Mol. Des. 4 (2005) 882–910.
[15] D.C. Montgomery, E.A. Peck, Introduction to Linear Regression Analysis, Wiley, New York, 1982.
[16] I.T. Jolliffe, Principal Component Analysis, Springer, New York, 1986.
[17] J.H. Kalivas, P.M. Lang, Mathematical Analysis of Spectral Orthogonality, Marcel Dekker, New York, 1994.
[18] G. Puchwein, Selection of calibration samples for near-infrared spectrometry by factor-analysis of spectra, Anal. Chem. 60 (1988) 569–573.
[19] Y.L. Xie, J.H. Kalivas, Evaluation of principal component selection methods to form a global prediction model by principal component regression, Anal. Chim. Acta 348 (1997) 19–27.
[20] J.M. Sutter, J.H. Kalivas, P.M. Lang, Which principal components to utilize for principal component regression, J. Chemometr. 6 (1992) 217–225.
[21] J. Sun, A correlation principal component regression-analysis of NIR data, J. Chemometr. 9 (1995) 21–29.
[22] H.F. Chen, Quantitative predictions of gas chromatography retention indexes with support vector machines, radial basis neural networks and multiple linear regression, Anal. Chim. Acta 609 (2008) 24–36.
[23] B. Hemmateenejad, M. Shamsipur, R. Miri, M. Elyasi, F. Foroghinia, H. Sharghi, Linear and nonlinear quantitative structure–property relationship models for solubility of some anthraquinone, anthrone and xanthone derivatives in supercritical carbon dioxide, Anal. Chim. Acta 610 (2008) 25–34.
[24] B. Hemmateenejad, M.A. Safarpour, F. Taghavi, Application of ab initio theory for the prediction of acidity constants of some 1-hydroxy-9,10-anthraquinone derivatives using genetic neural network, J. Mol. Struct. (Theochem.) 635 (2003) 183–190.
[25] D.T. Manallack, B.G. Tehan, E. Gancia, B.D. Hudson, M.G. Ford, D.J. Livingstone, D.C. Whitley, W.R. Pitt, A consensus neural network-based technique for discriminating soluble and poorly soluble compounds, J. Chem. Inf. Comput. Sci. 43 (2003) 674–679.
[26] W. Duch, K. Swaminathan, J. Meller, Artificial intelligence approaches for rational drug design and discovery, Curr. Pharm. Des. 13 (2007) 1497–1508.
[27] P.J. Gemperline, J.R. Long, V.G. Gregoriou, Nonlinear multivariate calibration using principal components regression and artificial neural networks, Anal. Chem. 63 (1991) 2313–2323.
[28] Z.Y. Liu, Z.C. Chen, Estimation of critical pressures of pure substances from data of density and vaporization heat of liquids, Chem. Eng. Biochem. Eng. J. 59 (1995) 127–132.
[29] J.J. Llano, R. Garcia, R. Rosal, H. Sastre, F. Diez, Program to estimate the density, viscosity and thermal-conductivity in pure fluids and mixtures, Afinidad 50 (1993) 243–249.
[30] G.M. Kontogeorgis, A. Fredenslund, D.P. Tassios, Chain-length dependence of the critical density of organic homologous series, Fluid Phase Equil. 108 (1995) 47–58.
[31] M. Karelson, A. Perkson, QSPR prediction of densities of organic liquids, Comput. Chem. 23 (1999) 49–59.
[32] E.R. Malinowski, Factor Analysis in Chemistry, Wiley-Interscience, New York, 2002.
[33] M. Shamsipur, B. Hemmateenejad, M. Akhond, Multicomponent acid–base titration by principal component-artificial neural network calibration, Anal. Chim. Acta 461 (2002) 147–153.
[34] P. Gramatica, E. Papa, QSAR modeling of bioconcentration factor by theoretical molecular descriptors, QSAR Comb. Sci. 22 (2003) 374–385.
[35] S. Wold, Validation of QSARs, Quant. Struct., Act. Relat. 10 (1991) 191–193.
[36] R.D. Cramer, J.D. Bunce, D.E. Patterson, I.E. Frank, Cross-validation, bootstrapping, and partial least-squares compared with multiple-regression in conventional QSAR studies, Quant. Struct., Act. Relat. 7 (1988) 18–25.
[37] A. Golbraikh, A. Tropsha, Beware of q2! J. Mol. Graph. Model. 20 (2002) 269–276.
[38] P. Gramatica, E. Giani, E. Papa, Statistical external validation and consensus modeling: a QSPR case study for Koc prediction, J. Mol. Graph. Model. 25 (2007) 755–766.
[39] W. Zhang, A. Tropsha, Novel variable selection quantitative structure–property relationship approach based on the *k*-nearest-neighbor principle, J. Chem. Inf. Comput. Sci. 40 (2000) 185–194.
[40] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, Nature 323 (1986) 533–536.
[41] B. Hemmateenejad, Correlation ranking procedure for factor selection in PC-ANN modeling and application to ADMETox evaluation, Chemom. Intel. Lab. Syst. 75 (2005) 231–245.
[42] B. Hemmateenejad, M. Shamsipur, Quantitative structure–electrochemistry relationship study of some organic compounds using PC–ANN and PCR, Int. Elect. J. Mol. Des. 3 (2004) 316–334.
[43] B. Hemmateenejad, M.A. Safarpour, R. Miri, N. Nesari, Toward an optimal procedure for PC-ANN model building: prediction of the carcinogenic activity of a large set of drugs, J. Chem. Inf. Model. 45 (2005) 190–199.
[44] B. Hemmateenejad, Optimal QSAR analysis of the carcinogenic activity of drugs by correlation ranking and genetic algorithm-based PCR, J. Chemometr. 18 (2004) 475–485.