ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

Application of MLR, PLS and Artificial Neural Networks for Prediction of GC/ECD Retention Times of Chlorinated Pesticides, Herbicides, and Organohalides

Saeid Asadpour[1*], Mahmood Chamsaz[1], Md Jelas Haron[2]

[1] Department of Chemistry, Faculty of Sciences, Ferdowsi University of Mashhad, 91775,Mashhad, Iran
[2] Department of Chemistry, Faculty of Science, University Putra Malaysia, 43400, Selangor, Malaysia.

ABSTRACT

Quantitative structure–retention relationship (QSRR) models correlating the retention times of diverse chlorinated pesticides, herbicides, and organohalides in gas chromatography/electron capture detector (GC/ECD) system and their structures were developed based on different multivariate regression techniques by using molecular structural descriptors.Modeling of retention times of these compounds as a function of the theoretically derived descriptors was established by multiple linear regression (MLR), partial least squares (PLS) regression and Artificial Neural Networkwith batch back propagation algorithm (BBP-ANN).The stepwise regression using SPSS was used for the selection of the variables that resulted in the best-fitted models. The aim of this paper was to compare the performances of different linear and nonlinear multivariate calibration techniques. The predictive quality of the QSRR models were tested for an external prediction set of 12 compounds randomly chosen from 38 compounds. The best model obtained from the training set based onhighest external predictive R^2 value and lowest RMSEP value also showed good internal predictive power. The ANN method with Batch Back Propagation (BBP) algorithm was used to model the structure-retention relationships, more accurately. The squared regression coefficients of prediction for the MLR, PLS and ANN models were 0.951, 0.948 and 0.968, respectively

Keywords: Molecular descriptors, QSRR, MLR, PLS, ANN, Chlorinated Pesticides, Herbicides, Organohalides

Corresponding author
Email:s.asadpour@gmail.com

ISSN: 0975-8585

INTRODUCTION

High-performance liquid chromatography (HPLC) and gas chromatography (GC) are the most appropriate analytical techniques for multi-residue monitoring of pesticides in natural ecosystems or water and foodstuffs for human consumption. As a potential alternative to expensive and time-consuming experimental trial-and-error approach traditionally adopted to optimize chromatographic separations, retention predictive models have received considerable attention in recent years [1].

An important property that has been extensively studied in quantitative structure property relationship(QSPR) is the chromatographic retention time. QSRR study involves the prediction of chromatographic retention parameters using molecular structure. These studies are widely investigated in gas chromatography (GC) and high-performance liquid chromatography (HPLC). The chromatographic parameters are expected to be proportional to a free energy change that is related to the solute distribution on the column. Chromatographic retention is a physical phenomenon that is primarily dependent on the interactions between the solute and the stationary phase. Molecular group contribution methods are widely employed to estimate gas chromatographic retention parameters [2]. The difficulty of this approach is represented by the definition of a consistent set of groups and by the necessity to compute the contribution of each group from a statistically significant number of molecules where the respective group is present. This method is limited to molecules containing only the groups presented in the calibration set of molecules. In addition, some group contribution schemes are not comprehensive enough to cover multiple substitutions of functional groups. With the aid of QSRR the interactions associated with this phenomenon can be related to the constitutional, molecular graph (topological), geometrical, electrostatic, and quantum descriptors of the molecules. Gas chromatographic QSRR models have been successfully developed for a large number of compound classes: alkanes, alkenes, alkylbenzenes [3], polycyclic aromatic hydrocarbons, various hydrocarbons from naphthas, various aromatic compounds [4], alkanes, alkenes, alcohols, esters, ketones, monoterpenes, di– and tricyclic methyl esters and alcohols, and monocyclic ketones and alcohols, chlorinated alkanes, chlorinated benzenes [5], chlorinated dibenzodioxins [6], polychlorinated biphenyls, polyhalogenated biphenyls , polychlorinated dibenzofurans, pyrazines, diverse compounds [7], odor–active aliphatic compounds with oxygen–containing functional groups [8], stimulants and narcotics, anabolic steroids, sulfur vesicants [9], diverse organic compounds [10]. The main advantage of QSPR, like QSRR, lies in the fact that once such a relationship is ascertained with an adequate statistical degree of confidence, it can be of valuable assistance in the prognosis of the behavior of new molecules, even before they are actually synthesized [11].

Chemical systems are typically multivariate, i.e. multiple measurements are madesimultaneously. Therefore, most chemometrics methods fall under the class of statistical techniques known as multivariate analysis. The measurement and analysis of dependence between variables is fundamental to multivariate analysis[12]. Multivariate calibration is the collective term used for the development of aquantitative model for the reliable prediction of properties of interest (y1, y2, . . . , yp) froma number of predictor variables (x1, x2, . . . , xp).

ISSN: 0975-8585

However, multivariate calibration is ageneral selectivity and reliability enhancement tool[13]. It is applicable todetermination of major constituents as well as micro-component and other qualities and fora very wide range of instrument types. The advantage of multicomponent analysis using multivariate calibration methodsis the speed of the method of determination for the components of interest in a mixture, asa separation step can be avoided[14], [15].

In the present work, a QSRR study, has been carried out on the GC/ECD retention times (t_R) for 38 diverse chlorinated pesticides, herbicides, and organohalides by using structural molecular descriptors. The two linear methods MLR and PLS and nonlinear method Artificial Neural Networks (ANN) with Batch Back Propagating algorithm along with Stepwise SPSS as variable selection software were used to model the retention times with the structural descriptors.

MATERIALS AND METHODS

Retention times (t_R) of 38 compounds including chlorinated pesticides, herbicides, andorganohalideswere taken from the literature [16], and are presented in Table 5. The analyts are extracted from the water sample, and then sample components are separated, identified, and measured by a high-resolution fused silica capillary column of a gas chromatograph/electron capture detector (GC/ECD) system. The 38 molecules divided into two subgroups with 26 and 12 members for calibration and predication sets respectively.

Table 5. Experimental retention times of 38 compounds, [P]indicates test set.

No.	*Compound*	*t_R(min)*	*No.*	*Compound*	*t_R (min)*
1	**Hexachlorocyclopentadiene**	**9.64**	20	**Heptachlor Epoxide**	**27.20**
2	**Etridiazole**	**11.41**	21	**Chlordane-gamma [P]**	**28.65**
3	**Chloroneb**	**12.39**	22	**Endosulfan I**	**29.36**
4	**Propachlor[P]**	**14.69**	23	**Chlordane-alpha[P]**	**29.58**
5	**Trifluralin**	**16.29**	24	**Dieldrin**	**30.95**
6	**HCH-alpha**	**17.01**	25	**4,4'-DDE[P]**	**31.97**
7	**Hexachlorobenzene**	**17.44**	26	**Endrin**	**32.24**
8	**Simazine[P]**	**17.86**	27	**Butachlor**	**32.65**
9	**Atrazine**	**18.23**	28	**Endosulfan II**	**32.81**
10	**HCH-beta**	**18.33**	29	**Chlorbenzilate**	**32.98**
11	**HCH-gamma**	**18.71**	30	**4,4'-DDD**	**33.49**
12	**HCH-delta [P]**	**19.21**	31	**Endrin Aldehyde [P]**	**33.96**
13	**Chlorthalonil**	**20.27**	32	**Endosulfan Sulfate**	**35.43**
14	**Metribuzin**	**21.88**	33	**4,4'-DDT[P]**	**35.80**
15	**Heptachlor [P]**	**22.78**	34	**Methoxychlor[P]**	**39.38**
16	**Alachlor**	**22.86**	35	**cis-Permethrin**	**44.98**
17	**Aldrin**	**24.81**	36	**trans-Permethrin[P]**	**45.42**
18	**Metolachlor**	**25.02**	37	**Pentachlorointrobenzene**	**19.02**
19	**Cyanazine[P]**	**25.21**	38	**4,4-Dibromobiphenyl**	**25.6**

The QSRR models for the estimation of the retention times of various compounds are established in the following six steps: molecular structure input and generation of the files containing the chemical structures stored in a computer–readable format; quantum mechanics geometry optimization with a semi–empirical method; structural descriptors computation; structural descriptors selection; structure–retention models generation with the multivariate methods and statistical analysis.

Computer hardware and software

All calculations were run on aPentium IV personal computer with windows XP as operating system. The molecular structures of data set were sketched using ChemDraw Ultra module of CS ChemOffice 2005 molecular modeling software ver. 9, supplied by Cambridge Software Company. The sketched structures were exported to Chem3D module in order to create their 3D structures. Each molecule was "cleaned up" and energy minimization was performed using Allinger's MM2 force filed by fixing Root Mean Square (RMS) gradient at 0.1 kcal/mol. Further geometry optimization was done using semiempirical AM1 (Austin Model) Hamiltonian method and closed shell restricted wave function available in the MOPAC module until the RMS value becomes smaller than 0.001 kcal/mol. The lower energy conformers obtained by the aforementioned procedure were fed into Excel spreadsheet for calculation of the structural molecular descriptors by add-in ChemSAR. The ChemSAR generate descriptors include physicochemical, thermodynamic, electronic and spatial descriptors available in the 'Analyze' option of the Chem3D packing. The descriptors calculated accounts four important properties of the molecules: physicochemical, thermodynamic, electronic and steric, as they represent the possible molecular interactions which determined the retention times of the studied molecules.Descriptor selection was accomplished by using stepwise regression using SPSS. PLS regression (PLS_Toolbox, version 2.1, Eigenvector Company) and other calculations were performed in the MATLAB (version 7.0, Mathworks, Inc.) environment.

The commercial ANN software NeuralPower version 2.5 (CPC-X Software) was used throughoutthe study. This software allows the user to select the network type, the number of hidden layers and hidden layer neurons, the iterations used during the model training and the transfer functions. The networkarchitecture consisted of an input layer with six neurons (six variables), an output layer with oneneuron (one response), and a hidden layer.

MLR and PLS Analysis

MLR is one of the most used modeling methods in QSRR. The colinearity problem of the MLR method has been overcome through the development of the PLSprojections to latent structures. This method, which has been shown to be an efficient approach in monitoring many complex processes, reducing the high dimensional strongly cross-correlated data to a much smaller and interpretable set of principal components or latent variables. The program used for MLR analysis was written in SPSS. In MLR analysis, however, the number of compounds in samples should be at least five times greater than the number of descriptors and of course the descriptors should be orthogonal. In order to minimize the information overlap in descriptors

and to reduce the number of descriptors required in regression equation, the concept of non-redundant descriptors was used in our study. The best equation is selected on the basis of the highest multiple correlation coefficient (r^2). MLR method provides equation linking the structural features to the t_R of the compounds:

$$\mathbf{t_R} = a_0 + a_1\mathbf{d_1} + \cdots + a_n\mathbf{d_n} \quad (1)$$

Where the intercept (a_0) and the regression coefficients of the descriptors (a_i) are determined by using the least-squares method.$\mathbf{d_i}$ has the common definition, variable or descriptor in this case, the elements of this vector are equivalent numerical values of a 3D structures of the molecules or structural descriptors.

The modeling by PLS method was performed in the MATLAB 7.0 and using PLS_Toolbox 2. PLS is a linear modeling technique where information in the descriptor matrix **X** is projected onto a small number of underlying ("latent") variables called PLS components, referred to as latent variables. The matrix **Y** is simultaneously used in estimating the "latent" variables in **X** that will be most relevant for predicting the **Y** variables. All descriptor variables were preprocessed by auto scaling. The number of significant factors for the PLS algorithm was determined using the cross-validation method. With cross-validation, one sample was kept out (leave one out) of the calibration and used for prediction. The process was repeated so that each of the samples was kept out once. The predicted values of left-out samples were then compared to the observed values using prediction error sum of squares (PRESS). The PRESS obtained in the cross validation was calculated each time that a new PC was added to the model. The optimum number of factors was concluded as the first local minimum in the PRESS versus number of factors plot. PRESS is defined as

$$PRESS = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \quad (2)$$

Where $\hat{y}_i$ is the estimated value of the *i*th object and $y_i$is the corresponding reference value of this object.

Artificial Neural Network

Artificial neural networks(ANNs) are computer programs designed to model the relationships between independent and dependent variables and are capable of modeling complex, non-linear relationships directly from the raw data. Unlike classical statistical techniques, such as response surface methodology, ANNs do not require the prior assumption of the nature of the relationships between input and output parameters, nor do they require the raw data to be transformed prior to model generation.ANNs are typically organized in layers where these layers are made up of a numberof interconnected nodes which contain an activation function. Input vectors are presented to the network via the inputlayer which communicates to one or more 'hidden layers'where the actual processing is done via a system ofweighted 'connections'.Most ANNs contain some form of 'learning rule' whichmodifies the weights of the connections according to inputpatterns that it is presented with. There are many

ISSN: 0975-8585

differentkinds of learning rules used by neural networks, in thiswork, Batch Back Propagation Neural Networks ((BBP-ANN) wasused. In BP-ANN, 'learning' is a supervised process thatoccurs with each cycle of 'Iterations' (i.e., each time thenetwork is presented with a new input pattern) through aforward activation flow of inputs and the backwards errorpropagation of weight adjustment.

RESULTS AND DISCUSSION

The main aim of the present work was developing a QSRR model to prediction of the retention times of Chlorinated Pesticides, Herbicides, and Organohalides. Chromatographic retention is based on interactionsbetween the solute and the stationary phase and the aim of thepresent work is to find which of the available topological, geometrical, constituteonal, and physical descriptors that we computedare related to the retention of the compounds present in this study.Therefore, the development of a robust and interpretable QSRRmodel, which is able to accurately predict the Rt, is necessary.

Descriptors Selection

Generally the first step in variables selection is the calculation of the correlation between variables and with seeking property. In the present case, to decrease the redundancy existed in the descriptors data matrix, the correlations of descriptors with each other and with the t_R of the molecules were examined, and descriptors which showed high interrelation (i.e., r>0.95) were detected. For each cluster of the descriptors which have close correlation coefficients just one of them was kept for construction the final QSRR model and the rest were removed. In second step, descriptors were analyzed for the existence of a constant or near constant pattern, and those were also removed. The remaining descriptors were gathered in an $n \times m$ data matrix **X**, where n and m are the number of molecules and descriptors, respectively. A column vector (***y***) was made by the t_R data. In order to obtain practical model, the number of descriptors should be decreased. Stepwise regression using SPSS, was used for variables selection using training. After these processing 6 descriptors subset were remained, which keep most interpretive information for t_R. A total of 6 descriptors were calculated for each compound in the data set. These descriptors deemed as important in their correlation with experimental retention time are presented in Table 1.

Table 1.Molecular descriptors employed for the proposed QSRR models.

No.	Descriptor	Notation	Group
1	Balaban Index	Bindx	Steric
2	Critical Volume	Vc	Thermodynamic
3	Lumo Energy	Lumo	Electronic
4	Radius	Rad	Steric
5	Repulsion Energy	NRE	Electronic
6	Vapor Pressure	VP	Thermodynamic

ISSN: 0975-8585

Results of ANN analysis and comparison with MLR and PLS

At first we constructed two different linear models MLR and PLS that both of them are mostly used modeling methods in QSRR. Table 2shows the parameters of the MLR model corresponding to the 6 independent variables and standardized coefficients (also named beta coefficients) allows comparing the relative weight of the variables in the MLR model. The greater the absolute value of a coefficient, the greater the weight of the variable in the model.

Table 2. Model parameters value and standardized coefficients for MLR model.

	Model parameters		Standardized coefficients
Source	Value	Standard error	Value
Intercept	-30.547	5.190	-
NRE	0.001	1.45-04	0.043
Vc	0.015	0.003	0.046
VP	-4.967	2.031	-1.127
Bindx	-3.8E-05	7.69E-06	0.043
Rad	7.326	1.131	1.768
Lumo	-3.465	1.079	-0.773

The colinearity problem of the MLR method has been overcome through the development of the partial least-squares projections to latent structures (PLS) method. Fig. 1 shows the plot of PRESS vs. number of factors for the PLS model.The best PLS model contained 6 selected descriptors in4 latent variables space. For this in general, the number of components (latent variables) is less than the number of independent variables in PLS analysis.

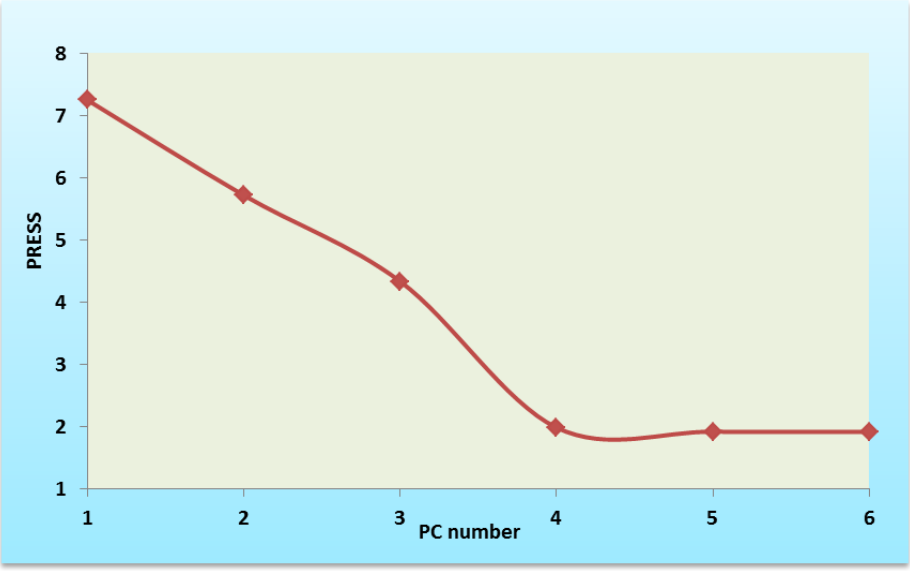


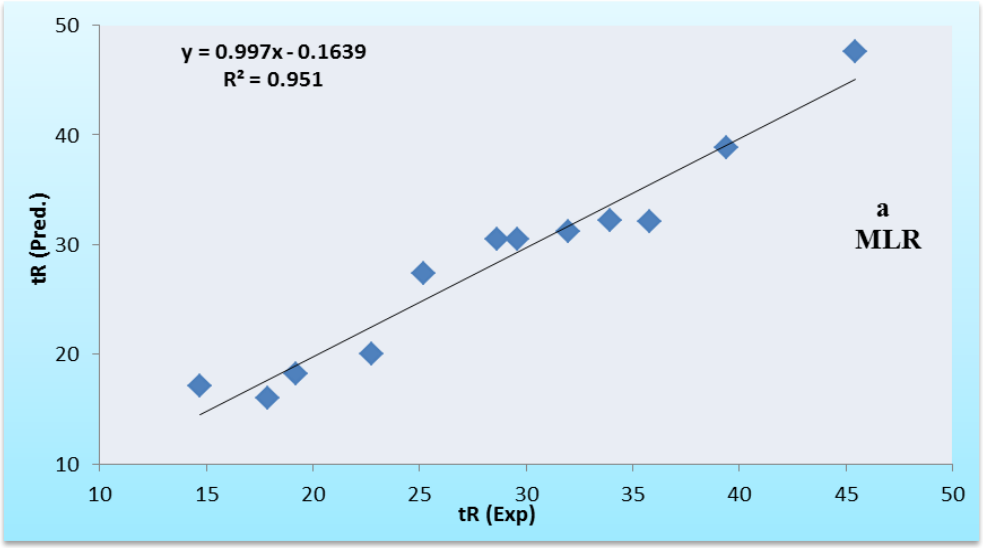
Fig. 1.PRESS versus number of factors for the PLS model

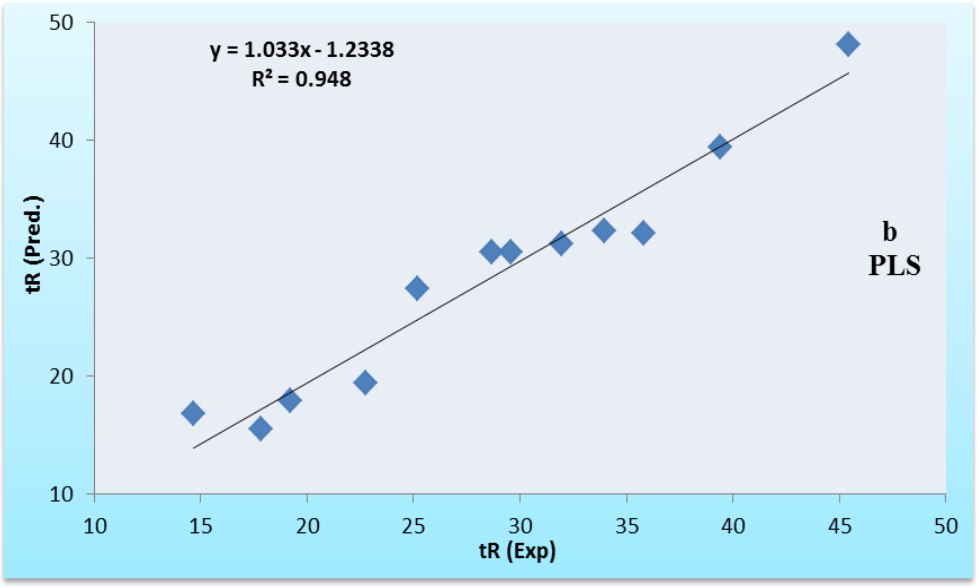
After that for construction of artificial neural network model (ANN) and to determine the optimal network topology, the number ofneurons in the hidden layer was iteratively determined by developing several networks that vary only with the size of hidden layer and

ISSN: 0975-8585

simultaneously observing the change in the root mean squared errors(RMSE). The transfer function was chosen sigmoid and other parameters for network were chosen asthe default values of the used software. The experimental data of central composite design were used as the training and the test data of the artificial neural network with Batch Back Propagation (BBP) algorithm. At the start of the training, weights were initialized with random values. The neural network was trained with the data obtained from 26experimental points.

R^2 is probably the most popular measure of how well a regression model fits the data. A value of R^2 near zero indicates no linear relationship, while a value near one indicates a perfect linear fit.





ISSN: 0975-8585

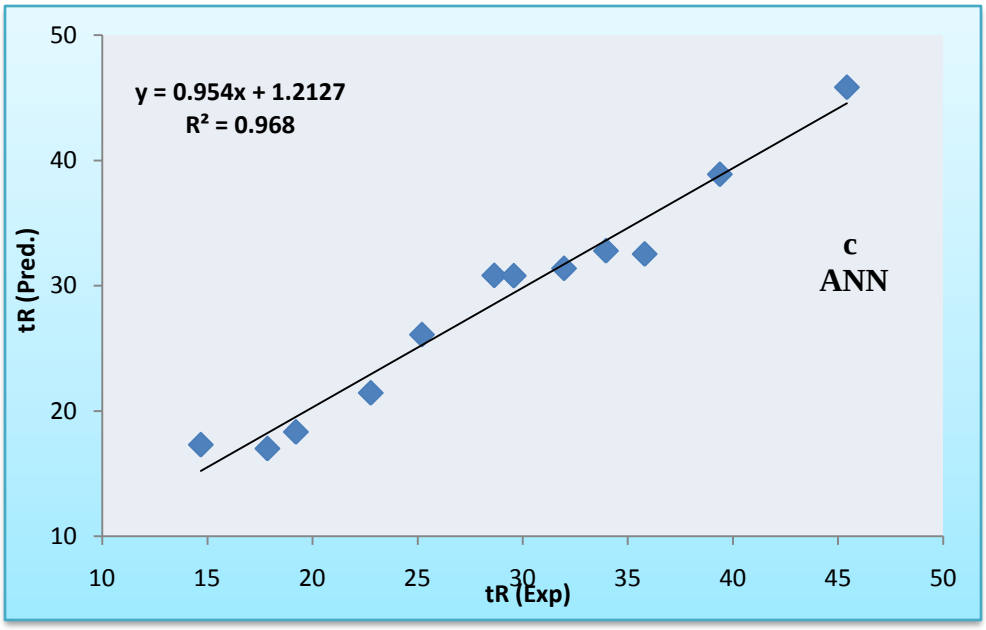


Fig. 2.Predicted t_R values by MLR (a), PLS (b) and ANN (c) modeling versus experimental t_R values.

Plots of predicted t_R versus experimental t_R, obtained by the MLR, PLS and ANN modeling, are shown inFigs. 2a, 2b and 2c, respectively. The agreement observed between the predicted experimental values in Fig. 2and the random distribution of residuals about zero mean confirms the good predictive ability of these modeling. In Table 3, the predicted values of t_R obtained by the MLR, PLS and ANN methods and the percent relative errors of prediction are presented.

Table 3. Comparison of MLR, PLS and ANN models for experimental and predicted values of t_R for test set.

		MLR model		PLS model		ANN model	
No.	**Exp. t_R**	**Pred. t_R**	**RE (%)**	**Pred. t_R**	**RE (%)**	**Pred. t_R**	**RE (%)**
4	14.69	17.108	16.451	16.773	14.170	17.306	17.809
8	17.86	15.975	-10.551	15.49	-13.270	17.013	-4.745
12	19.21	18.232	-5.091	17.862	-7.017	18.326	-4.601
15	22.78	20.080	-11.852	19.403	-14.824	21.443	-5.867
19	25.21	27.342	8.458	27.418	8.758	26.105	3.551
21	28.65	30.443	6.259	30.465	6.335	30.81	7.538
23	29.58	30.449	2.937	30.479	3.039	30.784	4.072
25	31.97	31.180	-2.472	31.168	-2.509	31.386	-1.827
31	33.96	32.198	-5.189	32.304	-4.876	32.786	-3.458
33	35.8	32.085	-10.375	32.157	-10.176	32.524	-9.151
34	39.38	38.853	-1.338	39.406	0.066	38.894	-1.235
36	45.42	47.573	4.740	48.135	5.977	45.831	0.905

For the constructed models, four general statistical parameters were selected to evaluate the prediction ability of the model for retention times. The statistical parameters root mean

squares error of prediction (RMSEP) relative error of prediction (REP), standard error of prediction (SEP) and squared regression coefficient (R^2) are summarized in Table 4.

Table 4.Statistical parameters obtained by applying the MLR, PLS and ANN models to the test set.

Parameter	MLR	PLS	ANN
RMSEP	2.012	2.164	0.755
REP (%)	7.0711	7.613	2.64
SEP	2.102	2.260	0.789
R^2	0.951	0.948	0.968
No. PCs[a]	-	4	
No. DSs[b]	6	6	6

[a]Number of factors
[b]Number of Descriptors

CONCLUSIONS

QSRR analysis was performed on a series of chlorinated pesticides, herbicides, and organohalides using molecular modeling program ChemOffice 2005 which generate molecular descriptors, has allowed establishing a new numerical model to correlate t_R values of these compound to the their structural descriptors. The statistical parameters of the built QSRR models were satisfactory which showed the high quality of the chose descriptors. High correlation coefficients (0.951, 0.948 and 0.968for MLR, PLS and ANN respectively) and low prediction errors obtained confirm good predictive ability of these models. Comparison of the values of statisticalparameters obtained using models of Artificial Neural Network with Batch Back Propagationlearning rules (BBP-ANN), PLS and MLR for predicting of retention time shows superiority of theBBP-ANN overthose of non-linear and especially linear models

The QSRR models proposed with the simply calculated molecular descriptors can be used to estimate the chromatographic retention times for new compounds even in the absence of the standard candidates.

REFERENCES

[1] Kaliszan R. Quantitative structure–chromatographic retention relationships. John Wiley & Sons, New York, 1987.
[2] Teodora I, Ovidiu I. Internet Electronic Journal of Molecular Design 2002;1: 94-107.
[3] IvanciucO, Ivanciuc T, Klein DJ, Seitz WA, BalabanAT.SAR QSAR Environ Res2001; 11: 419-52.
[4] Gautzsch R, Zinn P. Chromatoghraphia1996; 43: 163-176.
[5] SabljicA. J Chromatogr1985; 319: 1-8.
[6] Sekusak S, SabljicA.JChromatogr1993; 628: 69-79.
[7] RohrbaughRH,JursPC.Anal Chem1988; 60: 2249-2253.
[8] Anker LS, JursPC, EdvardsPA.Anal Chem1990; 62: 2676-2684.

[9] WoloszynTF, JursPC. AnalChem1992; 64: 3059-3063.
[10] Katritzky AR, Ignatchenko ES, BarcockRA, Lobanov VS, Karelson M. Anal Chem1994; 66: 1799-1807.
[11] Ghasemi J, AsadpourS, AbdolmalekiA. AnalChimActa2007; 588: 200–206.
[12] Anderson TW. An Introduction to Multivariate Statistical Analysis, Wiley, New York, 1984.
[13] BeebeKR, Kowalski BR.AnalChem1987; 59: 1007A.
[14] MerasID,PenaAM, MansillaAE, Salinas F.Analyst1993; 118: 807-813.
[15] GhasemiJ, AhmadiSh.Ann di Chim2007;97: 69-83.
[16] JW Munch, JW. US Environmental Protection Agency (EPA), Eichelberger, Method 508.1.