Indian Journal of Science and Technology Vol. 5 No. 2 (Feb 2012) ISSN: 0974- 6846

Application of artificial neural network for prediction of retention time for some pesticides in liquid chromatography

Saeid Khodadoust
Young Researchers Club, Branch Dehdasht, Islamic Azad University, Dehdasht, Iran
saeid.kh64@gmail.com

Abstract

The quantitative structure-retention relationships (QSRR) method is employed to predict the retention times (RTs) of pesticides by molecular descriptors which were calculated by Dragon software. After the calculation molecular descriptors for all molecules, a suitable set of molecular descriptors were selected by using genetic algorithm (GA) and then the data set was randomly divided into training and prediction set. The selected five descriptors were used to build QSRR models with multi-linear regression (MLR) and generalized regression neural network (GRNN) which were built and optimized with intelligent problem solver (IPS) in Statistica 7.1software. Both linear and nonlinear models show good predictive ability, of which GRNN model demonstrated a better performance than that of the MLR model. The root mean square error of cross validation (RMSECV) of the training and the prediction set for the GRNN model was 1.345 and 2.810, and the correlation coefficients (R) were 0.955 and 0.927 respectively, while the square correlation coefficient of the cross validation (Q^2_{loo}) on the GRNN model was 0.951, revealing the reliability of this model. The resulting data indicated that GRNN could be used as a powerful modeling tool for the QSRR studies.

Keywords: Pesticides, Quantitative structure-retention relationship, Genetic algorithm, Multiple linear regression, Retention time, Artificial neural networks.

Introduction

Pesticides are substances with high toxic effects and persistence in the environment have been widely used in agriculture throughout the world. They widely utilized at various stages of cultivation and during post harvest storage to protect fruit and vegetables against a range of pests and fungi and/or to provide quality preservation. The pollution of environmental compartments involves a serious risk to the environment and to human health as well, due to either direct exposure or residues in food and drinking water (Kuster *et al.,* 2006; Lambropoulou & Albanis, 2007; Rodrigues *et al.*, 2007). High-performance liquid chromatography (HPLC) (Santalad *et al*., 2009; Wang *et al.,* 2009; Khodadoust & Hadjmohammadi, 2011) and gas chromatography (GC) (Huertas-Pérez & García-Campaña, 2008; Saraji & Esteki, 2008) are the most appropriate analytical techniques for multi-residue monitoring of pesticides in natural ecosystems or water and foodstuffs for human consumption (van der Hoft & van Zoonen, 1999; Hogendoorn & van Zoonen, 2000).

Quantitative structure-retention relationship (QSRR) (Kaliszan, 1997) studies are widely investigated in GC and HPLC. The number of reports on the application of QSRR in comparative studies of retention properties of stationary phase materials for reverse phase- high performance liquid chromatography (RP-HPLC) was published in the past few years (Jalali-Heravi & Garkani-Nejad, 1993; Katritzky *et al.*, 2000; Fatemi, 2002; Fragkaki *et al.,* 2004; Luan *et al*., 2005; Flieger *et al.,* 2007). In these QSRR models multiple linear regression (MLR) (Riahi *et al.*, 2008), partial least squares (PLS) regression (Riahi *et al.,* 2009; Bodzioch *et al.,* 2010), artificial neural networks (ANN) (Fausett, 1994; Zupan & Gasteiger, 1999) and support vector machine (SVM) (Fatemi *et al.,* 2009) were used to quantitatively correlate the solute retention to the molecular descriptors. Multiple linear regression (MLR) is the method most frequently used for the statistical treatment of QSRR multivariate data consisting of a set of observed retention values and descriptors for a given set of test molecules (Neter *et al.*, 1995). In recent years ANNs have gained popularity as a powerful chemometric tool that can be used to solve chemical problems such as the optimization of chromatographic analysis (Metting & Coenegracht, 1996; Booth *et al.*, 1997; Marengo *et al.,* 1998; Guo *et al.*, 2000).

Chromatographic retention is a physical phenomenon that is primarily dependent on the interactions between the solute and the stationary phase. QSRR methodology is aimed at describing chromatographic behavior of solutes in terms of their structure and has been extensively applied for over two decades to several chromatographic systems and a large variety of solutes with different objectives of chromatographic phases and retention prediction (Héberger, 2007; Kaliszan, 2007).

QSRR provides a promising method for the estimation of the retention based on the descriptors calculated from the molecular structure (Luan *et al.,* 2005; Xia *et al*., 2007). The main steps involved in QSRR include the following: data collection, molecular descriptors obtaining and selection, correlation model development and finally model evaluation. The advantage of QSRR over other methods lies in the fact that the descriptors used to build the models can be calculated from the structure alone, and once a reliable model was built, the calculation of retention time (RT) of other compounds is not dependent on any experimental properties. The scope of this work was to establish a new QSRR model for predicting the RTs of the some pesticides, using the generalized regression neural

Indian Journal of Science and Technology Vol. 5 No. 2 (Feb 2012) ISSN: 0974- 6846

Table1. Experimental retention times of 43 pesticides.

No	pesticide	Mor07p	Mor28m	H6m	MLOGP	C005	RT(exp)	RT (MLR)	RT(GRNN)
1*	Aminocarb[a]	1.29	0.02	0.02	2.38	3.00	2.39	13.49	13.34
2	Butoxycarboxim	0.46	-0.01	0.02	0.26	2.00	3.78	8.69	9.11
3	Oxamyl	0.63	0.11	0.05	0.35	4.00	4.00	7.37	6.51
4*	Methomyl[b]	0.21	0.13	0.03	0.87	2.00	4.79	11.08	11.25
5	Vamidothion	0.45	-0.09	0.08	0.74	3.00	6.53	8.54	8.73
6	Ethiofencarbsulfon	0.77	-0.19	0.05	0.31	2.00	7.87	8.05	9.29
7	Pirimicarb	1.18	0.12	0.02	1.91	4.00	8.32	11.04	8.67
8	Dimethoate	0.23	-0.03	0.07	-0.76	3.00	9.74	5.17	7.98
9	Thiofanoxsulfone	1.12	-0.05	0.07	0.91	2.00	10.03	11.32	10.90
10	Butocarboxim	0.61	-0.06	0.01	1.60	2.00	12.40	11.37	11.92
11	Triacloprid	1.93	-0.01	0.05	1.37	0.00	13.06	16.33	17.09
12	Aldicarb	0.41	-0.20	0.08	1.60	2.00	13.52	11.18	11.32
13*	Spiroxamine[a]	2.21	-0.11	0.02	3.29	0.00	14.68	19.80	19.24
14	Fenpropimorph	2.86	0.03	0.06	3.83	0.00	14.95	23.63	19.80
15	Demeton-s-methy	0.29	0.16	0.00	1.35	2.00	16.00	12.13	13.01
16	Propoxur	2.28	0.04	0.01	2.38	1.00	17.23	17.24	18.55
17	Bendiocarb	3.07	0.16	0.01	1.88	1.00	17.53	17.78	18.35
18	Dioxacarb	2.61	0.22	0.02	1.34	1.00	17.54	16.76	17.83
19	Carbofuran	3.09	0.16	0.01	2.27	1.00	17.56	18.79	18.66
20	Carbaryl	2.12	0.09	0.03	3.03	1.00	18.57	19.41	19.26
21	Atrazine	1.31	-0.14	0.08	1.77	0.00	18.95	16.25	17.16
22*	Ethiofencarb[a]	1.56	0.07	0.02	2.92	1.00	19.21	18.16	18.95
23*	Isoproturon[b]	2.12	0.13	0.04	2.39	2.00	19.29	16.84	18.16
24	Metalaxyl	2.82	-0.01	0.08	1.91	2.00	19.30	16.00	17.91
25	Pyrimethanil	2.36	0.08	0.00	2.63	0.00	19.38	19.62	19.09
26	Diuron	1.07	0.23	0.33	2.65	2.00	19.44	23.05	21.24
27*	3,4,5-Trimethacarb[b]	1.68	-0.02	0.06	2.92	1.00	20.09	18.37	18.93
28	Isoprocarb	2.52	-0.07	0.06	2.92	1.00	20.10	18.73	19.27
29	Methiocarb	1.40	0.15	0.08	3.19	2.00	21.96	18.95	19.48
30	Linuron	1.03	0.43	0.30	2.65	2.00	22.31	24.02	22.44
31	Promecarb	1.95	-0.00	0.02	3.20	1.00	22.63	18.60	19.25
32	Iprovalicarb	3.29	0.03	0.12	3.18	0.00	22.71	23.80	21.04
33	Azoxystrobin	4.86	0.12	0.25	2.07	2.00	22.85	22.78	23.88
34	Cyprodinil	2.46	0.13	0.02	3.16	0.00	22.98	21.80	19.58
35	Fenoxycarb	3.91	0.11	0.12	3.18	0.00	24.60	25.01	21.83
36	Metolachlor	2.99	0.12	0.20	3.03	1.00	24.71	23.79	23.63
37*	Tebufenozide[a]	3.98	-0.01	0.09	3.95	0.00	25.48	25.40	20.82
38	Haloxyfopmethy	3.24	0.27	0.29	2.86	1.00	28.29	26.81	27.38
39	Indoxacarb	4.94	0.39	0.33	3.17	2.00	28.49	29.49	28.24
40*	Quizalofop-ethyl[b]	3.72	0.23	0.08	2.81	0.00	29.06	24.26	21.17
41	Haloxyfop-2-ethoxyethyl	3.01	0.39	0.23	2.76	0.00	29.58	27.71	28.27
42	Furathiocarb	3.12	0.37	0.22	3.42	2.00	30.27	26.00	28.17
43*	Fluazifop-butyl[a]	4.25	0.25	0.23	3.32	0.00	30.76	29.12	26.55

** Prediction set, a: Test set, b: Validation set*

Indian Journal of Science and Technology Vol. 5 No. 2 (Feb 2012) ISSN: 0974- 6846

network (GRNN) technique. The performance of this model was compared with those obtained by the MLR and GRNN.

Table 2. The correlation coefficient matrix for the selected descriptors by GA.

	Mor07p	MLOGP	H6m	C005	Mor28m
Mor07p	1.000				
MLOGP	0.657	1.000			
H6m	0.433	0.306	1.000		
C005	-0.587	-0.584	-0.018	1.000	
Mor28m	0.430	0.351	0.306	-0.032	1.000

Theory and methods

Equipment and software

A Pentium(R) Dual personal computer (CPU E2180 2.00GHz) with the Windows XP operating system was used. Dragon software (Ver. 3.0) (http://www.disat. Unimib.it/chm.) was used for calculation of the molecular descriptors from molecular geometries which had been previously generated and optimized by means of the Hyperchem program (Ver. 7.0). Statistica 7.1 software (StatSoft, 2006) and GA toolbox in MATLAB 7 were used for the development of models.

Data set and descriptor generation

The data set for this investigation was taken from the literature (Pang *et al.,* 2006). A complete list of the compounds' names and their corresponding experimental RTs are summarized in Table 1. Chromatographic separation was performed at 40°C on an Atlantis dC18 column, 150 mm×2.1 mm, 3µm particle. Detection and quantification were performed with an AB API3000 LC-MS-MS equipped with an ESI Turbo Ion Spray source. The chemical structures of the 43 molecules studied were drawn with Hyperchem software. Then obtained structures were preoptimized by using MM+ molecular mechanics force field, and then a further precise optimization was done with the AM1 semi-empirical method. The molecular structures were optimized using the Polak-Ribiere algorithm till the root mean square gradient was 0.01. The Dragon software that include Constitutional, Topological, Geometrical, Charge, GETAWAY (Geometry, Topology and Atoms-Weighted Assembly), WHIM (Weighted Holistic Invariant Molecular descriptors), 3D-MoRSE (3D-Molecular Representation of Structure based on Electron diffraction), Molecular Walk Counts, BCUT descriptors, 2D-Autocorrelations, Aromaticity Indices, Randic Molecular Profiles, Radial Distribution Functions, Functional Groups, Atom-Centred Fragments, Empirical and Properties was used to calculate the descriptors and a total of 1,243 molecular descriptors, from 18 different types of theoretical descriptor, were calculated for each molecule. In this case, to reduce redundancy in the descriptor data matrix, correlation of the descriptors with each other and with the RTs of the molecules was examined and collinear descriptors (i.e. r > 0.9) were detected. Among the collinear descriptors, that with the highest correlation with RTs was retained and the others were removed from the data matrix. The remaining descriptors were collected in an n×m data matrix (C), where n=43 and m=443 are the number of compounds and descriptors, respectively. In order to obtain practical QSRR models, the significant descriptors should be selected from of these molecular descriptors.

Genetic algorithm for variable selection

As it is impossible to generate a large number of molecular descriptors for each compound in the data set, the problem becomes how efficient in selecting the set of molecular descriptors that yields an accurate relationship. Genetic algorithm (GA) (Goldberg, 1989; Leardi *et al.*, 1992) is a stochastic optimization method inspired by evolution theory, here used to select within the 443 molecular descriptors the most significant for developing a reliable predictive model. To select the most relevant descriptors, the evolution of the population was simulated (Massart *et al.,* 1997; Waller & Bradley, 1999; Aires-de-Sousa *et al.,* 2002; Ahmad & Gromiha, 2003). Each individual of the population, defined by a chromosome of binary values, represented a subset of descriptors. The number of genes on each chromosome was equal to the number of the descriptors. The number of the genes with a value of unity was kept relatively low to maintain a small subset of descriptors (Siripatrawan & Harte, 2007). As a result, the probability of generating zero for a gene was set at least 70% greater than the probability of generating unity. The operators used here were crossover and mutation. The probability of application of these operators was varied linearly with generation renewal (0-0.1% for mutation and 70-90% for crossover). A population size of typically 200 individuals was chosen, and evolution was allowed over, typically, 50 generations. For a typical run, evolution of the generations was stopped when 90% of the generations took the same fitness. The five most significant descriptors (Table 2) which were selected by GA for building QSSR models are: moriguchi octanol water partition coefficient (MLOGP), H autocorrelation of lag 6/weighted by atomic masses (H6m), 3D-MoRSE signal 07/weighted by atomic polarizability (Mor07p), 3D-MoRSE signal 28/weighted by atomic masses (Mor28m) and CH_3X (C005). As was shown in Table 2, there is not any significant correlation between these descriptors. Because of the retention behavior of pesticide compounds is complex, which involves in several kinds of inter and intra molecular interactions, besides the MLR, nonlinear ANN was adopted to explore the relationship between those descriptors and RTs.

Multiple Linear Regressions (MLR)

The main advantages of MLR over other methods are computational simple and the capability of deriving coefficients which directly relate to the original data. MLR,

Indian Journal of Science and Technology Vol. 5 No. 2 (Feb 2012) ISSN: 0974- 6846

however, can only be used in the situation that the number of descriptors less than the number of samples. In addition, MLR are sometime over fitting the data, dimensionality of data, poor prediction and inability to work on ill conditional data.

The QSRR equations are obtained by a stepwise MLR following the multi-linear form:

$$RT = b_0 + b_1D_1 + b_2D_2 + \cdots b_nD_n \quad (1)$$

where RT is the retention times, D_1, D_2 and D_n are the descriptors, b_0 is the intercept, and n is the number of the descriptors. The regression coefficients of the descriptors (b_1, b_2. . . b_n) are determined by using the least squares method. About 80% of the data set was randomly classified into the training set to select descriptors subset and build models; the remaining 20% was used as prediction set in multi-linear regression. This 20% data set was separated into selection, prediction set for ANN modeling, and illustrated in Table 1.

Artificial neural network (ANN)

ANNs are inspired from the information-processing pattern of the biological nervous system (Qin *et al.*, 2009). Input, hidden and output layers are the main components of most neural networks. The input layer takes information directly from input files, and the output layer sends information directly to the outside world through computer or any other mechanical control system. There may be many hidden layers between input and output layers. We processed our data with different ANNs looking for a better model (Acevedo-Martýnez *et al.*, 2006). The advantage of ANN is the inclusion of nonlinear relations in the model. In this study, ANN calculations were performed with Statistica 7.1 by intelligent problem solver (IPS) and by customizing the number of neurons (from 5 to 15) with a single hidden layer. IPS is a toolbox capable of doing the first two tasks, creating and testing neural networks for data analysis and prediction tasks. This program can search automatically for the optimal type/architecture of ANN such as MLP and GRNN (belonging to the group of Bayesian Neural Networks). IPS can automatically design a number of neural networks and recommend a list of ANN architectures. The optimization of the architecture and of the input vector by the IPS was performed on the basis of validation error minimization. The number of compounds in the training set, validation and test sets were 34, 4, and 5, respectively, while the compound for each set was randomly selected. The neural networks were trained using the training subset only. The validation subset was used to keep an independent check on the performance of the networks during training, with deterioration in the selection error indicating over-learning. If over-learning occurs, the network will stop training the network and restore it to the state with minimum validation error. The test set was purely used to check that the select error was not artificial. The network model will generalize if the validation and test errors are close together. The optimal network architecture was determined experimentally with ISP, which built and selected the best models from linear (LIN), multilayer perceptron with linear output neuron (MLP) as well as generalized regression neural networks (GRNN).

Model validation

Model validation is crucial of QSRR modeling. The calibration and predictive capability of a QSRR model should be tested through model validation. The most widely used squared correlation coefficient (R^2) can provide a reliable indication of the fitness of the model, thus, it was employed to validate the calibration capability of a QSRR model. As for the validation of predictive capability of a QSRR model, two basic principles (internal validation and external validation) were existed. The cross validation (CV) is one of the most often used methods for internal validation. A good CV result (Q^2) often indicates a good robustness and high internal predictive ability of a QSRR model. The statistical external validation can be applied at the model development step, in order to determine both the generalizability of QSRR models for the true predictive power of model, by properly employing a prediction set for validation (Héberger, 2007; Kaliszan, 2007; Xia *et al*., 2007). The internal predictive capability of a model was evaluated by cross validation coefficient (Q^2 or R^2) using the following equation:

$$Q^2 = 1 - \frac{\sum(y_i - y_0)^2}{\sum(y_i - y_m)^2} \quad (2)$$

Another useful parameter was the root mean square error of cross validation (RMSECV), also employed to evaluate the performance of developed models. Therefore the overall performance of MLR, MLP and GRNN was evaluated in terms of RMSECV, which was calculated from the following equation:

$$\text{RMSECV(f)} = \sqrt{\frac{\sum_{i-1}^{n}(y_i - y_0)^2}{n}} \quad (3)$$

where in the above expressions, y_i is the experimental value of dependent variable for the ith object, y_0 is the predicted values for the ith object based on the model built with f factor, y_m is the mean of observed values and n is the number of molecules in data set (Lang, 2005; Acevedo-Martýnez *et al.*, 2006).

Table 3. Molecular descriptors employed for the proposed MLR model.

No	Descript	Group	coefficient	Std.	t-value
1	Mor07p	3D-MoRSE descriptors	0.969	0.604	1.604
2	MLOGP	Molecular properties	2.389	0.740	3.229
3	H6m	GETAWAY descriptors	19.913	6.901	2.885
4	C005	atom-centred fragments	-1.568	0.654	-2.399
5	Mor28m	3D-MoRSE descriptors	8.462	4.655	1.818

Indian Journal of Science and Technology Vol. 5 No. 2 (Feb 2012) ISSN: 0974- 6846

Fig. 1. Plot of experimental vs. predicted RTs by MLR

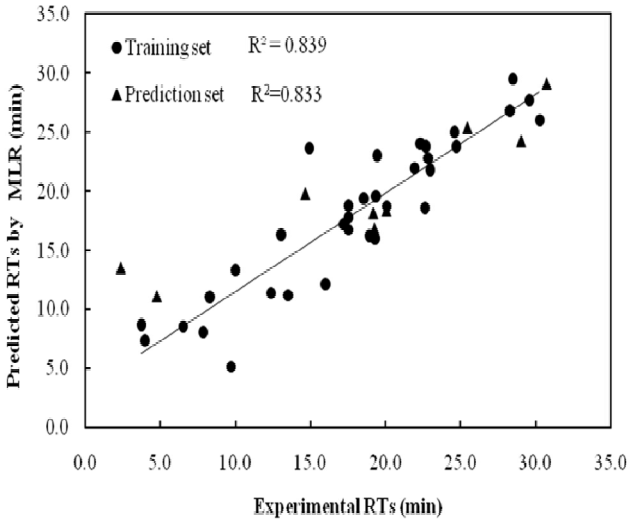


Fig.2. Neural networks architectures used in the regression analysis. Profile of GRNN 5:5-34-2-1:1.

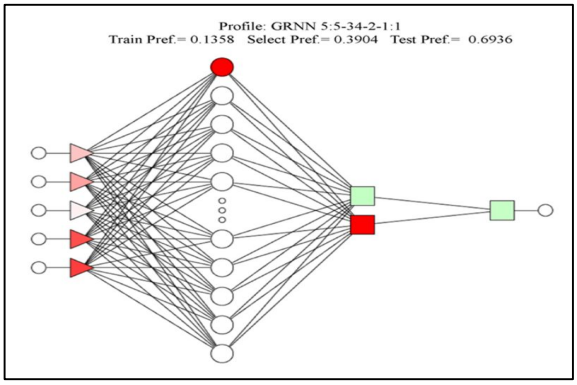
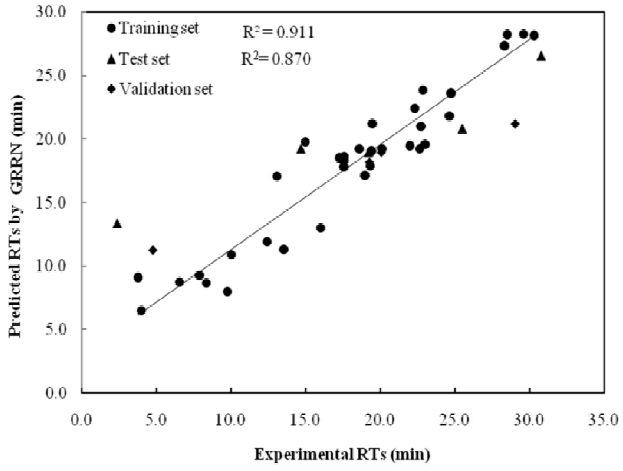


Fig. 3. Plot of experimental vs. predicted RTs by GRNN.



Results and discussions

Multiple Linear regressions (MLR)

For comparison purpose, the MLR models were built through a step-wise regression by using following descriptor subsets for training set: MLOGP, H6m, Mor07p, Mor28m and C005. The built models were used to predict the external prediction set. The statistical characteristics of MLR model using five descriptors were listed in Table 3 and the predicted values for all the pesticides were given in Table 1. According to the criteria for a good model mentioned above, the MLR model using five descriptor chosen by GA method had satisfactory predictive ability. The resulting equation including the best five descriptors is as follows:

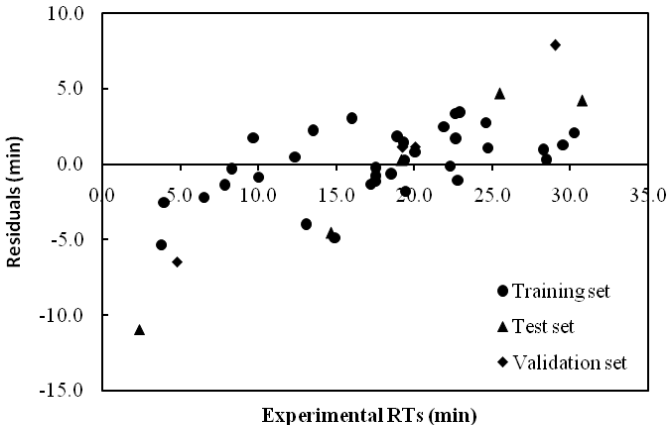
RT= 10.327(±4.655) + 2.389(±0.740) MLOGP +19.913(±6.901) H6m - 1.568(±0.654) C005+ 8.462(±4.655) Mor28m + 0.969(±0.604) Mor07p (4)

The plot of experimental vs. predicted RTs by MLR were shown in Fig. 1.

Non-linear GRNN network

The process of building the model in the GRNN network is divided into two stages (Xu, 1994; Todeschini & Consonni *et al.,* 2002). In the first step, in the space of the input signals groups of similar cases are localized. This stage is realized using the radial layer of the GRNN network. In the second stage, the regression approximation of the searched relationship is formed. Based on the earlier input space division by radial layer and the degree of similarity of the considered input signal to particular class, the decision is made and the result is obtained. The quality of the work of the GRNN 5:5-34-2-1:1 network is shown in Table 4 and the predicted values were given in Table 1. It can be seen that in this respect the quality of the best GRNN network is better than the quality of the MLR. The scatter plot of experimental vs. predicted RTs calculated by this model was shown in Fig. 3. It was evident that the predicted values agreed well with experimental values while Fig. 2 shows the architecture of this neural network.

Fig. 4. Plot of residuals vs. experimental RTs for The GRNN model as the best model



The statistical result of GRNN was listed in Table 4, and all the results were in accord with the criteria for a good predictive model. In order to compare the MLR

Indian Journal of Science and Technology Vol. 5 No. 2 (Feb 2012) ISSN: 0974- 6846

model with GRNN model, validate and test set in GRNN model were evaluated together. The rather better result of GRNN model than MLR model as shown in Table 4 demonstrated the complexity of chromatography retention process. Obtained results reveal that the reliability and good predicatively of the GRNN model for prediction of RTs for understudy pesticides. Among of MLR and GRNN models, GRNN have better statistical parameters including R= 0.955, F= 329.924, RMSECV= 1.345 for training set and R= 0.924, F= 42.614, RMSECV= 2.810 for prediction set, respectively. Fig. 4 showed a plot of the residuals vs. experimental RTs for GRNN model. The residuals were equally distributed on both sides of zero line which indicates that no symmetric error exists in the development of our GRNN as the best model.

Table 4. Statistical results of the MLR and ANN models

Model	Data set	Q_{LOO}^2	RMSECV	R	F
MLR	Training	0.800	2.835	0.916	167.043
	Predictio		2.615	0.913	35.037
GRNN	Training	0.951	1.345	0.955	329.924
	Validatio		1.863	0.986	
	Test		2.384	0.950	
	Predictio		2.810	0.927	42.614

Molecular descriptors

The statistical parameters of MLR model constructed by chosen (with GA) five descriptors are shown in Table 2. Among them, the lipophilicity parameter MLOGP represents the extent of hydrophilic/hydrophobic interactions. The positive coefficient of MLOGP indicating that an increase in MLOGP, result in an increase in RTs values. Another affected descriptor is H6m, which was weighted by atomic mass and is belong to the GETAWAY descriptors (Xu et al.,1994). GETAWAY descriptors are based on the representation of molecular geometry in terms of an influence matrix (H-GETAWAY) or influence-distance matrix (R-GETAWAY). The Molecular Influence Matrix (H) is defined as:

$$H = M.(M^T.M)^{-1}.M^T \quad (5)$$

where M is the molecular matrix constituted by the centered Cartesian coordinates and the superscript T refers to the transposed matrix. The diagonal elements h_{ij} of the H matrix, called leverage, encode atomic information and are considered to represent the influence of each atom in determining the whole shape of the molecule. For example mantle atoms always have higher h_{ij} values than atoms near the molecule center. Moreover, the magnitude of the maximum leverage in the molecule depends on the size and shape of the molecule itself. The Influence-distance matrix (R) involves a combination of the elements of H matrix with those of the Geometric Matrix.

The H6m mean effect has a positive sign (Table 3), which reveals that the RT is directly related to this descriptor. Hence, it was concluded that by increasing the molecular mass the value of this descriptor increased, caused to RTs of pesticides in LC increased.

Mor07p and Mor28m are the other descriptors, appearing in these models and are belong to the 3D-MoRSE descriptor (Schuur *et al*., 1996; Schuur & Gasteiger, 1997). The 3D-MoRSE descriptor is calculated using following expression:

$$I = \sum_{i=2}^{N}\sum_{j=1}^{i-1} W_i.W_j \frac{sin(s.r_{ij})}{s.r_{ij}} \quad (6)$$

Where *s* is scattering angle, r_{ij} is interatomic distance between i th and *j*th atom, W_i and W_j are atomic properties of i th and *j*th atom, respectively, including atomic number, masses, van der Waals volumes, Sanderson electronegativities, and polarizabilities. Mor07p and Mor28m display a positive sign, which indicate that the RTs is directly related to these descriptors.

Final descriptor C005 is one of the Ghos-Crippen atom-centred fragments related to the methyl group attached to any electronegative atom (O, N, S, P, Se, halogens) fragment. It gives information about the number of predefined structural features in the molecule, while it has shown negative influence on the prediction RTs. For this reason, RT of understudy pesticides is inversely related to this descriptor.

Conclusion

In conclusions this study, an accurate QSRR models for estimating the RT of some pesticides were developed for a series of 43 pesticides by employing the MLR and GRNN modeling approaches. Starting from the same set of descriptors included in the best MLR model, more robust models were obtained by the nonlinear methods of ANNs. The obtained results by GRNN model were compared with the results obtained by MLR model. The results demonstrated that GRNN model was more powerful in the RTs prediction of the pesticide compounds than MLR. A suitable model with high statistical quality and low prediction errors was eventually derived.

References:

1. Acevedo-Martýnez J, Escalona-Arranz JC, Villar-Rojas A, Tellez-Palmero F, Perez-Roses R, Gonzalez L and Carrasco-Velar R (2006) Quantitative study of the structure-retention index relationship in the imine family. *J. Chromatogr. A* 1102, 238-251.
2. Ahmad S and Gromiha M M (2003) Design and training of a neural network for predicting the solvent accessibility of proteins. *J. Comput. Chem.* 24, 1313-1320.
3. Aires-de-Sousa J, Hemmer MC and Gasteiger J (2002) Prediction of ^{1}H NMR Chemical shifts using neural networks. *Anal. Chem.* 74, 80-90.

Indian Journal of Science and Technology Vol. 5 No. 2 (Feb 2012) ISSN: 0974- 6846

4. Bodzioch K, Durand A, Kaliszan R, Baczek T and Vander Heyden Y (2010) Advanced QSRR modeling of peptides behavior in RPLC. *Talanta.* 81, 1711-1718.
5. Booth TD, Azzaoui K and Wainer IW (1997) Prediction of chiral chromatographic separations using combined multivariate Regression and Neural Network. *Anal. Chem.* 69, 3879-3883.
6. Consonni V, Todeschini R and Pavan M (2002) Structure/Response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors. *J. Chem. Inf. Comput. Sci.* 42, 682.
7. Fatemi MH (2002) Simultaneous modeling of the Kovats retention indices on OV-1 and SE-54 stationary phases using artificial neural networks. *J. Chromatogr. A* 955, 273-280.
8. Fatemi MH, Baher E and Ghorbanzade'h M (2009) Predictions of chromatographic retention indices of alkylphenols with support vector machines and multiple linear regression. *J. Sep. Sci.* 32, 4133-4142.
9. Fausett L (1994) Fundamentals of neural networks, Prentice Hall, NY.
10. Flieger J, Swieboda R and Tatarczak M (2007) Chemometric analysis of retention data from salting-out thin-layer chromatography in relation to structural parameters and biological activity of chosen sulphonamides. *J. Chromatogr. B* 846, 334-340.
11. Fragkaki AG, Koupparis MA and Georgakopoulos CG (2004) Quantitative structure-retention relationship study of α-,β_1-, and β_2-agonists using multiple linear regression and partial least-squares procedures. *Anal. Chim. Acta.* 512, 165-171.
12. Goldberg DE (1989) Genetic algorithms in search, Optimisation and machine learning. Addison-Wesley, Massachusetts, MA.
13. Guo W, Lu Y and Zheng XM (2000) The predicting study for chromatographic retention index of saturated alcohols by MLR and ANN. *Talanta* 51, 479-488.
14. Héberger K (2007) Quantitative structure-(chromatographic) retention relationships. *J. Chromatogr. A*1158, 273-305.
15. Hogendoorn E and Van Zoonen P (2000) Recent and future developments of liquid chromatography in pesticide trace analysis. *J. Chromatogr. A* 892, 435-453.
16. Huertas-Pérez JF and García-Campaña AM (2008) Determination of N-methylcarbamate pesticides in water and vegetable samples by HPLC with post-column chemiluminescence detection using the luminol reaction. *Anal. Chim. Acta.* 630, 194-204.
17. Jalali-Heravi M and Garkani-Nejad Z (1993) Prediction of gas chromatographic retention indices of some benzene derivatives. *J. Chromatogr. A* 648, 389-393.
18. Kaliszan R (1997) Structure and Retention in Chromatography. A chemometric approach, Harwood Academic Publishers, Amsterdam.
19. Kaliszan R (2007) QSRR: Quantitative structure-(Chromatographic) retention relationships. *Chem. Rev.* 107, 3212-3246.
20. Katritzky AR, Chen K, Maran U and Carlson DA (2000) QSPR Correlation and Predictions of GC Retention Indexes for Methyl-Branched Hydrocarbons Produced by Insects. *Anal. Chem.* 72, 101-109.
21. Khodadoust S and Hadjmohammadi MR (2011) Determination of N-methylcarbamate insecticides in water samples using dispersive liquid-liquid microextraction and HPLC with the aid of experimental design and desirability function. *Anal. Chim. Acta.* 699, 113-119.
22. Kuster M, Alda ML and Barceló D (2006) Analysis of pesticides in water by liquid chromatography-tandem mass spectrometric techniques. *Mass Spectrom. Ver.* 25, 900-916.
23. Lambropoulou DA and Albanis TA (2007) Liquid-phase micro-extraction techniques in pesticide residue analysis. *J. Biochem. Biophys. Methods.* 70, 195-228.
24. Lang B (2005) Monotonic multi layer perceptron networks as universal approximators. In: W. (Eds.), Formal Models and Their Applications. *Intl. Conf. Artificial Neural Networks, 2005, Lecture Notes in Comput. Sci.,* 3697, Springer, Berlin. pp. 31.
25. Leardi R, Boggia R and Terrible M (1992) Genetic algorithms as a strategy for feature selection. *J. Chemom.* 6, 267-281.
26. Luan F, Xue C X, Zhang R S, Zhao C Y, Liu M C, Hu Z D and Fan B T (2005) Prediction of retention time of a variety of volatile organic compounds based on the heuristic method and support vector machine. *Anal. Chim. Acta* 537, 101-110.
27. Marengo E, Gennaro MC and Angelino SJ (1998) Neural network and experimental design to investigate the effect of five factors in ion-interaction high-performance liquid chromatography. *J. Chromatogr. A* 789, 47-55.
28. Massart DL, Vandeginste BGM, Buydens LMC, Jong SDE, Leui PJ, Smeyers-Verbeke J (1997) Handbook of chemometrics and qualimetrics: Part A, Elsevier, Netherlands.
29. Metting HJ and Coenegracht PMJ (1996) Neural networks in high-performance liquid chromatography optimization: response surface modeling. *J. Chromatogr. A* 728, 47-53.
30. Neter J, Wasserman W, Kutner M (1995) Applied linear statistical models, 3rd edn, Irwin, Homewood.
31. Pang GF, Liu YM, Fan CL, Zhang JJ, Cao YZ, Li XM, Li ZY, Wu YP and Guo TT (2006) Simultaneous determination of 405 pesticide residues in grain by accelerated solvent extraction then gas chromatography-mass spectrometry or liquid chromatography-tandem mass spectrometry. *Anal. Bioanal. Chem.* 384, 1366-1408.
32. Qin LT, Liu SS, Liu HL and Tong J (2009) Comparative multiple quantitative structure-retention

Indian Journal of Science and Technology Vol. 5 No. 2 (Feb 2012) ISSN: 0974- 6846

relationships modeling of gas chromatographic retention time of essential oils using multiple linear regression, principal component regression, and partial least squares techniques. *J. Chromatogr. A* 1216, 5302-5312.

33. Riahi S, Ganjali MR, Pourbasheer E and Norouzi P (2008) QSRR Study of GC Retention Indices of essential-oil compounds by multiple linear regression with a genetic algorithm. *Chromatographia.* 67, 917-922.
34. Riahi S, Pourbasheer E, Ganjali MR and Norouzi P (2009) Investigation of different linear and nonlinear chemometric methods for modeling of retention index of essential oil components: Concerns to support vector machine. *J. Hazard. Mater.* 166, 853-859.
35. Rodrigues AM, Ferreira V, Cardoso VV, Ferreira E and Benoliel MJ (2007) Determination of several pesticides in water by solid-phase extraction, liquid chromatography and electrospray tandem mass spectrometry. *J. Chromatogr. A* 1150, 267-278.
36. Santalad A, Srijaranai S, Burakham R, Glennon JD and Deming RL (2009) Cloud-point extraction and reversed-phase high-performance liquid chromatography for the determination of carbamate insecticide residues in fruits. *Anal. Bioanal. Chem.* 394, 1307-1317.
37. Saraji M and Esteki N (2008) Analysis of carbamate pesticides in water samples using single-drop microextraction and gas chromatography-mass spectrometry. *Anal. Bioanal. Chem.* 391, 1091-1100.
38. Schuur JH, Selzer P and Gasteiger J (1996) The Coding of the three-dimensional structure of molecules by molecular transforms and Its application to structure-spectra correlations and studies of biological activity. *J. Chem. Inf. Comput. Sci.* 36, 334-344.
39. Schuur JH and Gasteiger J (1997) Infrared spectra simulation of substituted benzene derivatives on the basis of a 3D Structure Representation. *Anal. Chem.* 69, 2398-2405.
40. Siripatrawan U and Harte BR (2007) Solid phase microextraction/gas chromatography/mass spectrometry integrated with chemometrics for detection of Salmonella typhimurium contamination in a packaged fresh vegetable. *Anal. Chim. Acta.* 581, 63-70.
41. StatSoft (2006) Inc. STATISTICA (data analysis software system), version 7.1. http:// www.statsoft.com.
42. Todeschini R and Consonni V (2000) Handbook of Molecular Descriptors, Wiley-VCH, Weinheim.
43. Van der Hoft GR and van Zoonen P (1999) Trace analysis of pesticides by gas chromatography. *J. Chromatogr. A* 843, 301-322.
44. Waller CL and Bradley MP (1999) Development and validation of a novel variable selection technique with application to multidimensional quantitative structure−Activity relationship studies. *J. Chem. Inf. Comput. Sci.* 39, 345-355.
45. Wang S, Mu H, Bai Y, Zhang Y and Liu H (2009) Multiresidue determination of fluoroquinolones, organophosphorus and N-methyl carbamates simultaneously in porcine tissue using MSPD and HPLC-DAD. *J. Chromatogr. B.* 877, 2961-2966.
46. Xia BB, Ma WP, Zhang XY and Fan BT (2007) Quantitative structure-retention relationships for organic pollutants in biopartitioning micellar chromatography. *Anal. Chim. Acta.* 598, 12-18.
47. Xu L, Krzyzak A and Yuille AL (1994) *Neural Networks.* 7, 609-628.
48. Zupan J and Gasteiger J (1999) Neural networks in chemistry and Drug design, Wiley-VCH Verlag, Weinheim.