ISSN : 0974-7419 Volume 10 Issue 5

Trade Science Inc.

Analytical CHEMISTRY
An Indian Journal

Full Paper

ACAIJ, 10(5) 2011 [319-324]

Modeling quantitative structure-property relationships (QSPR) for a set of pesticides or toxicants

Mehdi Alizadeh
Department of Chemistry, Gachsaran Islamic Azad University, Gachsaran, (IRAN)
E-mail : Mehdi.Alizadeh85@ yahoo.com

Received: 24th September, 2010 ; Accepted: 4th October, 2010

ABSTRACT

In this paper, a quantitative structure–property relationships (QSPR) study based on feed-forward artificial neural network (ANN) with back-propagation learning rule and multiple linear regression (MLR) methods has been carried out to predict the Solubility behavior of pesticides. Accurate description of the water Solubility of 38 compounds including commonly used insecticides, herbicides and fungicides and some metabolites is successfully achieved. The Stepwise SPSS was used for the selection of the variables that resulted in the best-fitted models.The regression coefficients of prediction for training and test sets for ANN model were 0.997 and 0.992 respectively. The proposed nonlinear QSPR model (ANN) exhibits a high degree of correlation between observed and computed water Solubility and a good predictive performance that supports its application for the prediction of the Solubility behavior of unknown pesticides. A multiple linear regression (MLR) based on the same selected descriptors shows a significantly worse predictive capability.

KEYWORDS

Pesticides;
QSPR;
QSAR;
MLR;
ANN.

INTRODUCTION

Solubility in water plays one of the most important roles among many physicochemical parameters that characterize a chemical pollutant. It influences behavior of the chemical compound in many physical and biological processes, involving information on the ability of the compound to take part in metabolic processes as well as assessing its environmental persistence, transport and fate[1].

Pesticides, as a consequence of massive use in agriculture and other human activities, are widely diffuse environmental contaminants subjected in Europe and USA to restrictive legislation aimed at the protection of natural ecosystems and health safeguard. Rather than a well identifiable chemical class, the term “pesticide” identifies a large spectrum of structurally different compounds. A wide structural variability also characterizes the pesticide subfamilies (insecticides, herbicides and fungicides) that group together molecules according to the target of biocide activity[2,3]. Activity traditional experimental determination methods usually need special equipment and samples, as well as large amounts of money and manpower. Despite methods development and automation, it is unlikely that laboratory determinations can cope with the pace that new pollutants are identified. Hence there is an increasing need for using the technology of quantitative structure–property/activity relationships (QSPR/QSAR), which correlates and predicts property data of pollutants from their structural descriptors, and can be used to study the physicochemical properties and generate predicted data efficiently[4]. The main advantage of quantitative structure property relationships (QSPR), like quantitative structure activity relationships (QSAR), 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[5,6].

Artificial neural network (ANN)[7] modelling represents the most common non-linear approach to the investigation of structure–property relationships. ANN statistical treatment, which does not require the preliminary knowledge of the mathematical form of the relationship between the descriptors and the model response, allows overcoming possible inaccuracy of MLR related with the existence of non-linear effects or colinearities among the descriptor variables. In the context of QSRR studies, the better predictive capability of the ANN-based models compared with the performance of the related MLR models has been largely documented[8-13].

Artificial neural networks are among the best available tools to generate nonlinear models. Artificial neural networks are parallel computational devices consisting of groups of highly interconnected processing elements called neurons. Artificial neural networks (ANN), inspired by scientist's interpretation of the architecture and functioning of the human brain[14,15] mean, however, a methodology related to nonlinear regression techniques[16,17]. Reviews have been published concerning applications of ANN in different fields[18,19]. ANNs have been applied to QSPR analysis since the late 1980s due to its flexibility in modeling of nonlinear problems, mainly in response to increase accuracy demands; they have been widely used to predict many physicochemical properties[20–24].

In the present work, a QSPR study has been carried out on the logarithm water solubility (Log S_w) for 38 diverse pesticides or toxicants by using structural molecular descriptors. The linear method MLR and nonlinear method feed forward neural network with backpropagation training along with Stepwise SPSS as variable selection software were used to model the Log S_w with the structural descriptors.

TABLE 1 : Data set and corresponding observed and (ANN,MLR,) predicted values of Log S_w (mg/l) (25°C)[a]

No.	Name	log S_w (EXP)	log S_w (ANN)	log S_w (MLR)
	Training set			
1	Ethoprophos	2.87	2.864	3.324
2	Phorate	1.70	1.690	1.743
3	Trietazine	1.30	1.312	1.108
4	Lindan	0.86	0.911	-0.173
5	Dichlorofention	-0.61	-0.654	0.080
6	PCB31	-0.84	-0.971	-0.837
7	Aldrin	-1.77	-1.641	-1.357
8	Paraoxon-ethyl	2.86	2.882	2.374
9	Fenitrothion	1.58	1.557	1.597
10	Thiobencrab	1.45	1.404	1.955
11	Isodrin	-1.85	-1.620	-1.385
12	Allethrin	0.66	0.596	0.908
13	Isocarbophos	1.85	1.953	1.502
14	Flumetralin	-1.15	-1.148	-1.193
15	Procymidone	0.65	0.643	1.104
16	Chinomethionate	0.00	0.007	0.633
17	2,4'-DDD	-1.00	-0.821	-0.990
18	Endrin	-0.60	-0.875	-1.463
19	2,4'-DDT	-1.07	-1.497	-1.424
20	Ethion	0.30	0.310	0.201
21	4,4'-DDT	-2.26	-1.984	-1.618
22	Benalaxyl	1.57	1.554	1.587
23	Famphur	2.04	2.054	1.556
24	PCB202	-3.83	-3.840	-3.679
25	Dicofol	-0.10	-0.118	-0.227
26	Phosalone	0.48	0.527	0.446
27	Fenarimol	1.15	1.146	0.424
		Test set		
28	Terbufos	0.70	0.439	1.083
29	Carbofuran	2.50	2.428	2.103
30	Benfuresate	2.42	2.183	1.913
31	PCB52	-1.81	-1.398	-1.330
32	PCB70	-1.39	-1.454	-1.382
33	2,4'-DDE	-0.85	-0.959	-1.248
34	4,4'DDE	-1.40	-1.261	-1.438
35	PCB149	-2.37	-2.321	-2.325
36	4,4'DDD	-1.05	-1.196	-1.171
37	Fenpropathrin	-0.48	-0.626	-0.066
38	Tetramethrin	0.26	-0.226	0.375

[a] Logarithm water solubility (mg/l) (25°C)

TABLE 2 : Molecular descriptors employed for the proposed QSPR models

No.	Descriptor	Notation	Type	Coefficient
1	Mean atomic van der waals volume (scaled on carbon atom)	Mv	Constitutional	-11.82915(±3.15366)
2	Volume	Volume	Molecular properties	-0.00712(±3.28211)
3	Maximal electrotopological positive variation	MAXDP	Topological	0.49396(±0.00283)
4	Superpendentic index	SPI	Topological	-0.00007(±0.18904)
5	Kier flexibility index	PHI	Topological	0.17031(±0.00003)
	Constant			11.56081(±0.15071)

METHODS

Stepwise multiple linear regression

The multiple linear regression (MLR) is an extension of the classical regression method to more than one dimension[25]. MLR calculates QSPR equation by performing standard multivariable regression calculations using multiple variables in a single equation. The stepwise multiple linear regression is a commonly used variant of MLR. In this case, also a multiple-term linear equation is produced, but not all independent variables are used. Each variable is added to the equation at a time and a new regression is performed. The new term is retained only if equation passes a test for significance. This regression method is especially useful when the number of variables is large and when the key descriptors are not known[26].

Artificial neural networks

Principles, functioning and applications of artificial neural networks have been adequately described elsewhere[27,28]. A three-layer feed-forward network formed by one input layer consisting of a number of neurons equal to the number of descriptors, one output neuron and a number of hidden units fully connected to both input and output neurons, were adopted in this study. The most used learning procedure is based on the back-propagation algorithm, in which the network reads inputs and corresponding outputs from a proper data set (training set) and iteratively adjusts weights and biases in order to minimize the error in prediction. To avoid overtraining and consequent deterioration of its generalization ability, the predictive performance of the network after each weight adjustment is checked on unseen data (validation set).

In this work, training gradient descent with momentum is applied and the performance function was the mean square error (MSE), the average squared error between the network outputs and the actual output.

Computer hardware and software

All calculations were run on a Pentium IV personal computer with windows XP as operating system. The molecular 3D structures of data set were sketched using hyperchem (ver. 7.1), then each molecule was “cleaned up” and energy minimization was performed using geometry. Optimization was done using semiempirical AM1 (Austin Model) Hamiltonian method. After optimization, 3D structures with lower energy conformers obtained by the aforementioned procedure were fed into dragon (ver. 5.2-2005) and ChemOffice 2005 molecular modeling software ver. 9, supplied by Cambridge Software Company, for calculation of the structural molecular descriptors (constitutional, topological, connectivity, geometrical, getaway, thermodynamic and charge descriptors) also hyperchem can calculate several descriptors. Through these descriptors which have values further than 90% zero or have equal values further than 90% are not useful and cut. Then Descriptor selection was accomplished by using Stepwise SPSS (SPSS Ver. 11.5, SPSS Inc.). other calculations were performed in the MATLAB (version 7.0, MathWorks, Inc.) environment.

RESULTS AND DISCUSSION

Experimental data

Water solubility (mg/l)(25°C) of 38 compounds including pesticides or toxicants were taken from the literature[29] that shown in TABLE 1. The QSPR models for the estimation of the Log S_w of various compounds are established in the following five steps: 1) molecular structure input and generation of the files containing the chemical structures stored in a computer–readable for-

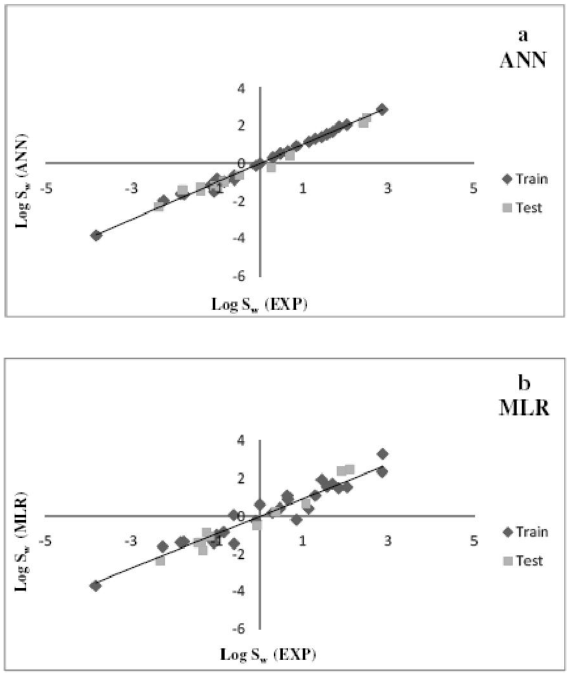

Figure 1 : Plots of predicted Log S_w estimated by ANN (a) and MLR (b) modeling versus experimental Log S_w compounds

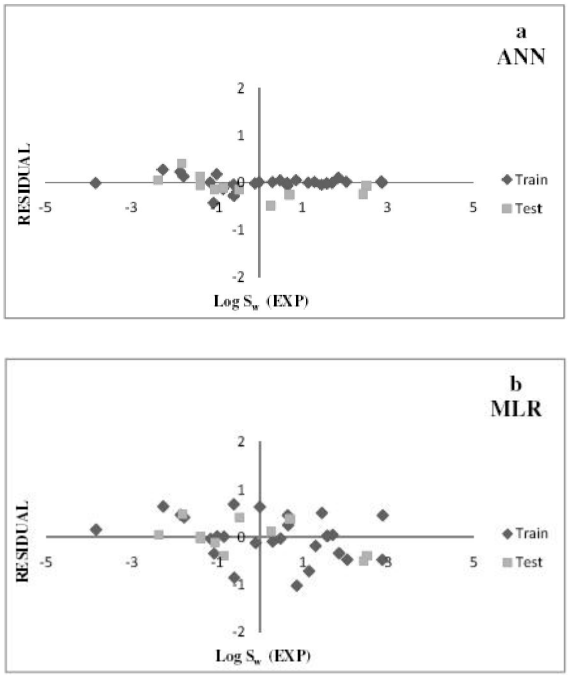

Figure 2 : Plots of residual versus experimental Log S_w in ANN (a) and MLR (b) models

mat; 2) quantum mechanics geometry optimization with a semi–empirical (AM1) method; 3) structural descriptors computation; 4) structural descriptors selection; 5) structure–Solubility models generation with the multivariate methods(ANN,MLR) and statistical analysis.

The data set was divided into two subsets in ANN and MLR: a training set of 27 compounds and a test set of 11 compounds.

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 Log S_w of the molecules were examined, and descriptors which showed high interrelation (i.e., r>0.9) with Log S_w and low interrelation (i.e., r<0.9) with each other were detected. For each class of the descriptor just one of them was kept for construction the final QSPR model and the rest were deleted. In second step, Stepwise SPSS was used for variables selection. After these processing five descriptors were remained, that keeps most interpretive information for Log S_w. TABLE 2 shows five descriptors and their coefficients (± confidence interval) that used in MLR method. A correlation analysis was carried out to evaluate correlations between selected descriptors with each other and with Log S_w (TABLE 3).

ANN optimization

A three-layer neural network was used and starting network weights and biases were randomly generated. Descriptors selected by stepwise method were used as inputs of network and the signal of the output node represent the Log S_w of pestisides. Thus, this network has five neurons in input layer and one neuron in output layer. The network performance was optimized for the number of neurons in the hidden layer (hnn), the learning rate (*lr*) of back-propagation, momentum and the epoch.As weights and biased are optimized by the back-propagation iterative procedure, training error typically decreases, but validation error first decreases and subsequently begins to rise again, revealing a progressive worsening of generalization ability of the network. Thus training was stopped when the validation error reaches a minimum value. TABLE 4 shows the architecture and specification of the optimized network.

Results of ANN analysis and comparison with MLR

The QSPR models provided by the optimal ANN

TABLE 3 : Correlation matrix of the five descriptors and Log S_w used in this work[a]

	Mv	VOLUME	MAXDP	SPI	PHI	Log S_w
Mv	1	-0.376	-0.505	-0.095	-0.458	-0.841
VOLUME		1	0.448	0.270	0.482	0.007
MAXDP			1	0.338	-0.040	0.523
SPI				1	0.126	-0.127
PHI					1	0.237
Log S_w						1

[a]The definitions of the descriptors are given in TABLE 2

TABLE 4 : Architecture and specification of the generated ANNs

No. of nodes in the input layer	5
No. of nodes in the hidden layer	7
No. of nodes in the output layer	1
learning rate	0.3
Momentum	0.1
Epoch	4000
Transfer function	Sigmoid

TABLE 5 : Statistical parameters obtained using the ANN and MLR models[a]

Ft	Fc	Rt	Rc	SEt	SEc	Model
525.194	3561.449	0.992	0.997	0.207	0.138	ANN
215.189	284.339	0.979	0.959	0.316	0.452	MLR

[a]c refers to the calibration (training) set; *R* is the correlation coefficient; *R*^{2} is the correlation coefficient; and *F* is the statistical *F* value

and MLR are presented in figure 1a and 1b where computed or predicted Log S_w values are plotted against the corresponding experimental data. Figure 2a and 2b shows a plot of residuals versus the observed Log S_w values. The substantial random pattern of this plot indicates that most of the data variance is explained by the proposed model.

The agreement between computed and observed values in ANN training and test sets are shown in TABLE 1. The statistical parameters calculated for the ANN model are presented in TABLE 5. Goodness of the ANN-based model is further demonstrated by the high value of the correlation coefficient *R* between calculated and observed Log S_w values 0.997 and 0.992 for training and test set, respectively.

For comparison, a linear QSPR model relating Log S_w values to the selected descriptors were obtained by means of MLR method. With the purpose MLR model built on the same subsets that used in ANN analysis. Multiple linear regression (MLR) is one of the most used modeling methods in QSPR. Comparison between statistical parameters in TABLE 5 reveals that nonlinear ANN model produced better results with good predictive ability than linear model.

CONCLUSIONS

QSPR analysis was performed on a series of pesticides or toxicants using ANN and MLR methods that correlate Log S_w values of these compound to the their structural descriptors. According to obtained results it is concluded that the ANN can be used successfully for modeling Log S_w property of the under study compounds. The statistical parameters of the built QSPR models were satisfactory which showed the high quality of the chose descriptors. High correlation coefficients and low prediction errors obtained confirm good predictive ability of ANN model. The QSPR models proposed with the simply calculated molecular descriptors can be used to estimate the water solubility values for new compounds even in the absence of the standard candidates.A non-linear modeling approach based on artificial neural networks allows to significantly improve the performance of the QSPR model.

REFERENCES

[1] S.N.Bhattachar, L.A.Deschenes, J.A.Wesley; Drug Discovery Today, **11**, 1012 **(2006)**.

[2] E.Hogendoorn, P.Van Zoonen; J.Chromatogr.A, **892**, 435 **(2000)**.

[3] G.R.VanDerHoft, P.Van Zoonen; J.Chromatogr.A, **843**, 301 **(1999)**.

[4] Sh.D.Chen, Xi.L.Zeng, Z.Y.Wang , H.Xi.Liu; Sci.Total Environ., **382**, 59 **(2007)**.

[5] S.Shahmirani, Ev.Farahani, J.Ghasemi; Ann.Chem., **96**, 327 **(2006)**.

[6] J.Ghasemi, S.Asadpour, A.Abdolmaleki; Anal.Chim.Acta, **588**, 200 **(2007)**.

[7] J.Zupan, J.Gasteiger; 'Neural Networks in Chemistry and Drug Design', Wiley-VCH Verlag, Weinheim, **(1999)**.

[8] M.H.Fatemi; J.Chromatogr.A, **955**, 273 **(2002)**.

[9] H.Li, Y.X.Zhang, L.Xu; Talanta, **67**, 741 **(2005)**.

[10] K.L.Peterson; Anal.Chem., **64**, 379 **(1992)**.

[11] F.Ruggieri, A.A.D'Archivio, G.Carlucci, P.Mazzeo; J.Chromatogr.A, **1076**, 163 **(2005)**.

[12] Y.L.Loukas; J.Chromatogr.A, **904**, 119 **(2000)**.

[13] T.Suzuki, S.Timofei, B.E.Iuoras, G.Uray, P.Verdino, W.M.F.Fabian; J.Chromatogr.A, **922**, 13 **(2001)**.

[14] W.S.Mculloch, W.Pitts; Bull.Math.Bioph., **5**,115 **(1943)**.

[15] D.E.Rumelhart; 'Parallel Distributed Processing', London, Mit Press, **(1982)**.

[16] J.Zupan, J.Gasteiger; Anal.Chim.Acta, **248**, 1 **(1991)**.

[17] D.T.Manallack, D.D.Ellis, D.J.Livingstone; J.Med.Chem., **37**, 3758 **(1994)**.

[18] A.Guez, I.Nevo; Clin.Chim.Acta, **248**, 73 **(1996)**.

[19] V.Jakus; Chem.Listy., **87**, 262 **(1993)**.

[20] F.Lui, Y.Liang, C.Cao; Chemometr.Intell.Lab.Syst., **81**, 120 **(2006)**.

[21] H.Golmohammadi, M.H.Fatemi; Electrophoresis, **26**, 3438 **(2005)**.

[22] E.Baher, M.H.Fatemi, E.Konoz, H.Golmohammadi; Microchim.Acta, **158**, 117 **(2007)**.

[23] M.H.Fatemi; J.Chromatogr.A, **1038**, 231 **(2004)**.

[24] M.H.Fatemi; J.Chromatogr.A, **955**, 273 **(2002)**.

[25] R.H.Myers; 'Classical and Modern Regression with Application', Pws-Kent Publishing Company, Boston, **(1990)**.

[26] J.Ghasemi, Sh,Ahmadi; Ann.Chim.(Rome)., **97(1-2)**, 69 **(2007)**.

[27] J.Zupan, J.Gasteiger; 'Neural Networks In Chemistry And Drug Design', Wiley-Vch Verlag, Weinheim, **(1999)**.

[28] L.Fausett; 'Fundamentals Of Neural Networks', Prentice Hall, New York, **(1994)**.

[29] [SRC] Syracuse Research Corporation; Physical/Chemical Property Database (PHYSPROP): SRC Environmental Science Center, Syracuse, **(2009)**.