ISSN : 0974-7419 Volume 10 Issue 5

Trade Science Inc.

Analytical CHEMISTRY

An Indian Journal

Full Paper

ACAIJ, 10(5) 2011 [330-335]

Application of multiple linear regression and artificial neural networks to predict LC50 in fish

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

A quantitative structure–activity relationship (QSAR) study has been carried out on 31 diverse organic pollutants by using molecular structural descriptors. Modeling of the logarithm values LC50 (lethal concentration required to kill 50% of a population) in fish (after 96 h) of these compounds as a function of the theoretically derived descriptors was established by multiple linear regression (MLR) and artificial neural networks (ANN). The Stepwise SPSS was used for the selection of the variables (descriptors) that resulted in the best-fitted models. For prediction logarithm values LC50 of compounds three descriptors were used to develop a quantitative relationship between the logarithm values LC50 and structural activity. Appropriate models with low standard errors and high correlation coefficients were obtained. After variables selection, compounds randomly were divided into two training and test sets and MLR and ANN used for building the best models. The predictive quality of the QSAR models were tested for an external prediction set of 8 compounds randomly chosen from 31 compounds. The regression coefficients of prediction for training and test sets for ANN model were 0.9953 and 0.9938 srespectively. Result obtained showed that ANN model can simulate the relationship between structural descriptors and the Log LC50 of the molecules in data sets accurately and Theoretical predictions coincide very well with experimental results.

© 2011 Trade Science Inc. - INDIA

KEYWORDS

Log LC50;
QSAR;
QSPR;
ANN;
MLR.

INTRODUCTION

The environmental risk assessment of organic chemicals requires information on both their physicochemical properties and toxicity. Experimental investigations are often carried out to collect this information. However, the data collection procedure, especially that conducted for toxicity determination, is extremely time consuming. One practical alternative would be to predict these properties, or toxic effects, by utilizing quantitative structure–activity relationships (QSARs). Many such models have been developed and applied in the field of aquatic toxicology[1]. In recent years, there has been an evolution in the development and application of quantitative structural activity relationships (QSAR) within the field of aquatic toxicology[2].

Quantitative structure–activity relationships (QSARs) are the fundamental basis of developed ap-

TABLE 1 : Data set and corresponding observed and ANN and MLR predicted values of Log LC50[a]

No.	Name	Log LC50 (EXP)	Log LC50 (ANN)	Log LC50 (MLR)
	Training set			
1	(2,4,5-Trichlorophenoxy)acetic acid	1.7176	1.7519	1.7364
2	(4-Chloro-2-methylphenoxy)acetic acid	2.1528	2.1399	1.9891
3	2-(2,4-Dichlorophenoxy)propionic acid	1.8493	1.8491	1.8297
4	2-(3-Chlorophenoxy)propionic acid	2.2477	2.2194	2.0605
5	2-(4-Chlorophenoxy)-2-methylpropionic acid	1.8755	2.0189	1.9351
6	2-(4-Chlorophenoxy)propionic acid	2.2477	2.2158	2.0593
7	2,4-Dichlorophenoxyacetic acid	2.1219	2.1087	1.9656
8	3,6-Dichloro-2-methoxybenzoic acid	2.4723	2.4400	2.2440
9	4-(2-Methyl-4-chlorophenoxy)butyric acid	1.4942	1.4044	1.6435
10	Benzene	1.1300	1.2524	1.1413
11	Biphenyl	0.1332	0.1823	-0.0293
12	Buprofezin	0.0358	0.0350	-0.1100
13	Buturon	1.1224	1.0092	0.5238
14	Chlorbromuron	0.8582	0.7879	1.0660
15	Chloroxuron	0.1752	0.1742	0.3586
16	Chlortoluron	1.1343	1.1393	1.3031
17	Diuron	1.1084	1.0139	1.2054
18	Fenuron	1.8980	1.8510	1.9432
19	Isoproturon	0.9313	1.2040	1.3766
20	Metoxuron	1.5089	1.5419	1.8529
21	Monuron	1.5062	1.3921	1.4791
22	Propargite	-0.8327	-0.7855	-0.7848
23	Tetradifon	-0.5406	-0.5965	-0.4416
	Test set			
24	2-(2,4,5-Trichlorophenoxy)propionic acid	1.4342	1.3477	1.5701
25	2-Phenoxypropionic acid	2.6331	2.6048	2.4812
26	Anthracene	-0.2343	-0.0798	-0.4363
27	Bromobenzene	0.7835	0.8465	0.7305
28	Fluometuron	1.3404	1.1741	1.1758
29	Monolinuron	1.3719	1.4020	1.5727
30	Neburon	0.1000	-0.0356	0.5299
31	Triclopyr	2.2521	2.1029	1.9710

[a]Log LC50 in fish after 96 h

proaches for estimating the toxicity of chemicals from their molecular structure and/or physicochemical properties[3,4]. QSARs are mathematical models that can be used to predict the physicochemical and biological properties of molecules considering that the biological activity of a new or untested chemical can be inferred from the molecular structure or other properties of similar compounds whose activities have already been assessed. The two main objectives of QSARs are to allow prediction of the biological properties of chemically characterized compounds that have not been biologically tested and to obtain information on the molecular characteristics of a compound that are important for the biological properties[3].

Artificial neural networks (ANNs) 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 (ANNs), inspired by scientist's interpretation of the architecture and functioning of the human brain[5,6]. mean, however, a methodology related to nonlinear regression techniques[7,8]. Reviews have been published concerning applications of ANN in different fields[9,10]. Recently, artificial neural networks (ANNs) have been used to a wide variety of chemical problems such as spectral analysis[11], prediction of dielectric constant[12]. and mass spectral search[13]. 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[14–18].

The main aim of the present work is development of a QSAR models by using ANN as nonlinear method to predict the Log LC50 (lethal concentration required to kill 50% of a population) in fish (after 96h) of various organic pollutants and comparison with MLR as linear method.

In the present work, a QSAR study has been carried out on the Log LC50 in fish for 31 diverse organic pollutants by using structural molecular descriptors. Linear method, multiple linear regressions (MLR) and nonlinear method, feed forward neural network with back-propagation training along with Stepwise SPSS as variable selection software were used to model the Log LC50 with the structural descriptors.

MATERIALS AND METHODS

Experimental data

The experimental data of the Log LC50, for 31 chemical compounds including various organic pollutants were taken from[19], that shown in TABLE 1. The

data set randomly was divided into two subsets in ANN: training and test sets including 23 and 8 compounds respectively.

MLR analys

The multiple linear regression (MLR) is an extension of the classical regression method to more than one dimension[20]. MLR calculates QSAR equation by performing standard multivariable regression calculations using multiple variables in a single equation. The stepwise multiple linear regressions 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[21].

Artificial neural networks (ANN)

Principles, functioning and applications of artificial neural networks have been adequately described elsewhere[22,23]. The relevant principle of supervised learning in an ANN is that it takes numerical inputs (the training data) and transfers them into desired outputs. The input and output nodes may be connected to any other nodes within the network. The way in which each node transforms its input depends on the so-called 'connection weights' or 'connection strength' and bias of the node, which are modifiable. The output values of each node depend on both the weight strength and bias values. Training of the ANN can be performed by using the backpropagation algorithm. In order to train the network using the back propagation algorithm, the differences between the ANN output and its desired value are calculated after each training iteration and the values of weights and biases modified by using these error terms.

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 backpropagation 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 net-

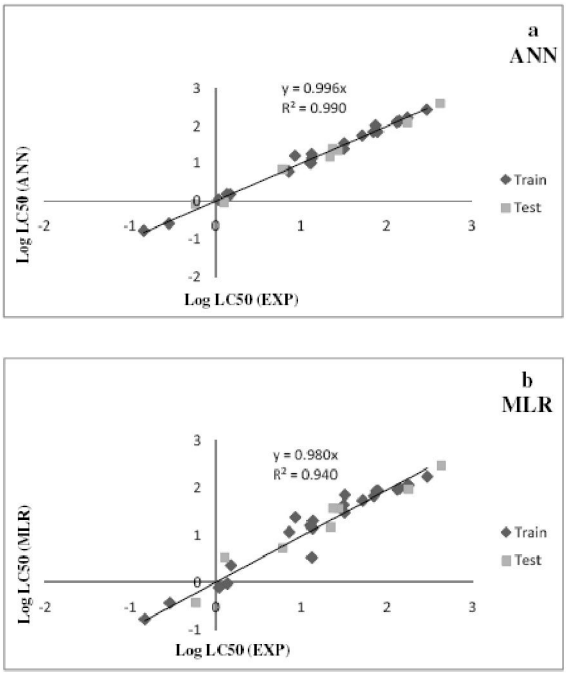

Figure 1 : Plots of predicted Log LC50 estimated by ANN (a) and MLR (b) modeling versus experimental Log LC50 compounds

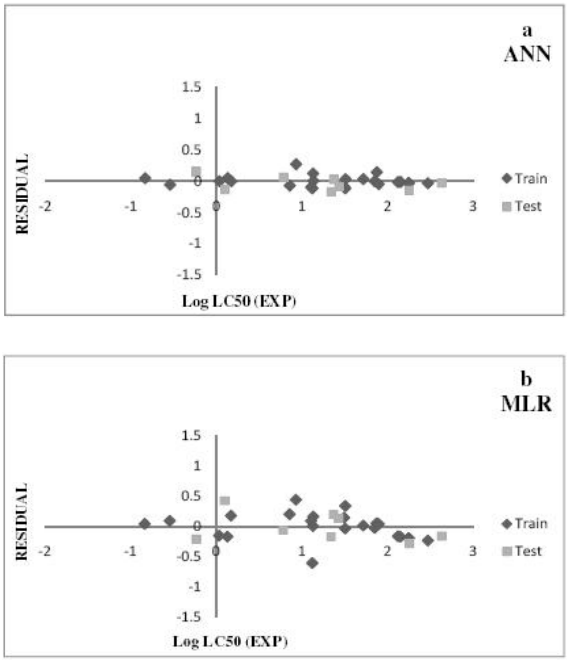

Figure 2 : Plots of residual versus experimental Log LC50 in ANN (a) and MLR (b) models

TABLE 2 : Molecular descriptors employed for the proposed ANN and MLR models

No.	Descriptor	Notation	class	Coefficien
1	Eigenvalue 10 form edge adj matrix weighted by edge degrees	EEig10x	Edge adjacency indices	-0.2982
2	Heat of formation	HF	Thermodynamic	-0.0018
3	Partition coefficient (octanol/water)	CLogP	Thermodynamic	-0.4451
	Constant			2.2393

TABLE 3 : Correlation matrix of the three descriptors and Log LC50 used in this work[a]

	EEig10x	HF	CLogP	Log LC50
EEig10x	1	0.4656	0.8411	-0.8548
HF		1	0.2632	-0.7334
CLogP			1	-0.7996
Log LC50				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	3
No. of nodes in the hidden layer	7
No. of nodes in the output layer	1
learning rate	0.3
Momentum	0.1
Epoch	1000
Transfer function	Sigmoid

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

Ft	Fc	R^2t	R^2c	Rt	Rc	SEt	SEc	Model
2207.6451	476.4930	0.9906	0.9876	0.9953	0.9938	0.0902	0.1128	ANN
345.4659	92.6994	0.9427	0.9392	0.9709	0.9691	0.2174	0.2436	MLR

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

work 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.

The QSAR models for the estimation of the Log LC50 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-Log LC50 models generation with the multivariate methods and statistical analysis.

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 of structures, 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). 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

Descriptors selection

Generally the first step in variables selection is the calculation of the correlation between variables and with seeking activity. 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 LC50 of the molecules were examined, and descriptors which showed high interrelation (i.e., r>0.9) with Log LC50 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 QSAR model and the rest were deleted. In second step, Stepwise SPSS was used for variables selection. After these process three descriptors were remained, that keeps most interpretive information for Log

LC50. TABLE 2 shows descriptors that used in ANN method. A correlation analysis was carried out to evaluate correlations between selected descriptors with each other and with Log LC50 (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 LC50 of organic pollutants. Thus, networks have three neurons in input layer, and one neuron in output layer.The networks 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 test 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 test error reaches a minimum value. TABLE 4 shows the architecture and specification of the optimized networks.

Results of ANN analysis and comparison with MLR

The QSAR models provided by the optimal ANN and MLR are presented in figure 1a and 1b where computed or predicted Log LC50 values are plotted against the corresponding experimental data. Figure 2a and 2b shows a plot of residuals versus the observed Log LC50 values. The substantial random pattern of this plot indicates that most of the data variance is explained by the proposed models.

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 LC50 values are (0.9953, 0.9938) for training and test set respectively. For comparison, a linear QSAR model relating Log LC50 to the selected descriptors were obtained by mean of MLR method. With the purpose MLR model built on the same subsets that used in ANN analysis. Multiple linear regressions (MLR) are one of the most used modeling methods in QSAR. For the best MLR model contained three selected descriptors correlation coefficient (R) between calculated and observed Log LC50 values are (0.9709, 0.9691) for training and test set respectively.

Comparison between statistical parameters in TABLE 5 reveals that nonlinear ANN model produced better results with good predictive ability than linear model.

CONCLUSIONS

QSAR analysis was performed on a series of organic pollutants using ANN method that correlate Log LC50 values of these compound to the their structural descriptors. According to obtained results it is concluded that the (EEig10x , HF, CLogP) can be used successfully for modeling Log LC50 of the under study compounds. The statistical parameters of the built ANN model 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 QSAR models proposed with the simply calculated molecular descriptors can be used to estimate the Log LC50 for new compounds even in the absence of the standard candidates.

REFERENCES

[1] S.Tao, X.Xi, F.Xu, B.Li, J.Cao, R.Dawson; Environ.Pollut., **116**, 57 **(2002)**.

[2] S.Tao, X.Xi, F.Xu, R.Dawson; Water Res., **36**, 2926 **(2002)**.

[3] S.Ekins; 'Computational Toxicology', Risk Assessment for Pharmaceutical and Environmental Chemicals, Wiley-Interscience, **(2007)**.

[4] T.W.Schultz, M.T.D.Cronin, T.I.Netzeva; J.Mol.Struct.Theochem., **622**, 23 **(2003)**.

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

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

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

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

[9] A.Guez, I.Nevo; Clin.Chim.Acta, **248**, 73 **(1996)**.
[10] V.Jakus; Chem.Listy., **87**, 262 **(1993)**.
[11] J.M.Vegas, P.J.Zufiria; 'Generalized Neural Network for Spectral Analysis', Dynamics and Liapunov Functions, Neural Networks, 17 **(2004)**.
[12] R.C.Schweitzer, J.B.Morris; Anal.Chem.Acta, **384**, 285 **(1999)**.
[13] C.S.Tong, K.C.Cheng; Chemometr.Intell.Lab.Syst., **49**, 135 **(1999)**.
[14] F.Lui, Y.Liang, C.Cao; Chemometr.Intell.Lab.Syst., **81**, 120 **(2006)**.
[15] H.Golmohammadi, M.H.Fatemi; Electrophoresis, **26**, 3438 **(2005)**.
[16] E.Baher, M.H.Fatemi, E.Konoz, H.Golmohammadi; Microchim.Acta, **158**, 117 **(2007)**.
[17] M.H.Fatemi; J.Chromatogr.A, **1038**, 231 **(2004)**.
[18] M.H.Fatemi; J.Chromatogr.A, **955**, 273 **(2002)**.
[19] L.Escuder-Gilabert, Y.Mart´ýn-Biosca, S.Sagrado, R.M.Villanueva-amanas, M.J.Medina-Hernandez; Anal.Chim.Acta, 173 **(2001)**.
[20] R.H.Myers; 'Classical And Modern Regression With Application', Pws-Kent Publishing Company, Boston, **(1990)**.
[21] J.Ghasemi, Sh,Ahmadi; Ann.Chim.(Rome)., **97(1-2)**, 69 **(2007)**.
[22] J.Zupan, J.Gasteiger; 'Neural Networks In Chemistry And Drug Design', Wiley-Vch Verlag, Weinheim, **(1999)**.
[23] L.Fausett; 'Fundamentals of Neural Networks', Prentice Hall, New York, **(1994)**.