ISSN : 0974-7419 Volume 10 Issue 3

Trade Science Inc.

Analytical CHEMISTRY

An Indian Journal

Full Paper

ACAIJ, 10(3) 2011 [183-187]

# Prediction chemical toxicity of organic pollutants using artificial neural networks (ANN)

**Mehdi Alizadeh**

**Department of Chemistry, Gachsaran Islamic Azad University, Gachsaran, (IRAN)**

**E-mail : mehdi.alizadeh85@yahoo.com**

*Received: 19th August, 2010 ; Accepted: 29th August, 2010*

## ABSTRACT

An artificial neural networks (ANN) study, has been carried out on 38 diverse organic pollutants for prediction chemical toxicity by using molecular structural descriptors. Modeling of logarithm values of chronic toxicity in fish (Log Chv) of these compounds as a function of the theoretically derived descriptors was established by 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 Log Chv of compounds, three descriptors were used to develop a quantitative relationship between the Log Chv and structural properties. 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 ANN used for building the best models. The predictive quality of the ANN models were tested for an external prediction set of 11 compounds randomly chosen from 38 compounds. The regression coefficients of prediction for the ANN model were 0.9940, 0.9955 for training and test sets respectively. Result obtained showed that ANN can simulate the relationship between structural descriptors and the Log Chv of the molecules in data sets accurately. 

## KEYWORDS

Molecular descriptors;
Log Chv;
QSAR;
ANN.

## INTRODUCTION

Quantitative structure-activity relationships (QSARs) are the fundamental basis of developed approaches for estimating the toxicity of chemicals from their molecular structure and physicochemical properties[1,2]. 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[1].

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 process-

ing elements called neurons. Artificial neural networks (ANNs), inspired by scientist's interpretation of the architecture and functioning of the human brain[3,4]. mean, however, a methodology related to nonlinear regression techniques[5,6]. Reviews have been published concerning applications of ANN in different fields[7,8]. Recently, artificial neural networks (ANNs) have been used to a wide variety of chemical problems such as spectral analysis[9], prediction of dielectric constant[10] and mass spectral search[11]. ANNs have been applied to QSAR 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[12-16]. The main aim of the present work is development of a QSAR models by using ANN as nonlinear method to predict the logarithm values of chronic toxicity in fish of various organic pollutants.

In the present work, a QSAR study, has been carried out on the logarithm values of chronic toxicity in fish (Log ChV) for 38 diverse organic pollutants by using structural molecular descriptors. Nonlinear method, feed forward neural network with back-propagation training along with Stepwise SPSS as variable selection software were used to model the Log ChV with the structural descriptors.

## MATERIALS AND METHODS

### Experimental data

The experimental data of the logarithm values of chronic toxicity in fish (Log Chv), for 38 chemical compounds including various organic pollutants were taken from literature[17], that shown in TABLE 1. The data set randomly was divided into two subsets in ANN: training and test sets including 27 and 11 compounds respectively.

### Artificial neural networks (ANN)

Principles, functioning and applications of artificial neural networks have been adequately described elsewhere[18,19]. 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,

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

| No. | Name | log Chv (EXP) | log Chv (ANN) | Residual |
|---|---|---|---|---|
| | **Training set** | | | |
| 1 | (2,4,5-trichlorophenoxy)acetic acid | 1.2912 | 1.1685 | 0.0591 |
| 2 | 2-(2,4,5-trichlorophenoxy)propionic acid | 0.9490 | 0.9470 | -0.0020 |
| 3 | 2-(2,4-dichlorophenoxy)propionic acid | 1.4551 | 1.4186 | -0.0365 |
| 4 | 2-(4-chlorophenoxy)-2-methylpropionic acid | 1.4939 | 1.6230 | 0.1291 |
| 5 | 2-(4-chlorophenoxy)propionic acid | 1.9431 | 1.9224 | -0.0207 |
| 6 | 2,4-dichlorophenoxyacetic acid | 1.7851 | 1.7472 | -0.0379 |
| 7 | 2,4-dimethylphenol | 0.0546 | 0.0953 | 0.0407 |
| 8 | 2-chlorophenol | 0.3581 | 0.3157 | -0.0424 |
| 9 | 2-phenoxypropionic acid | 2.4181 | 2.4769 | 0.0588 |
| 10 | 2-phenyl phenol | -0.2197 | 0.0674 | 0.2871 |
| 11 | 3,6-dichloro-2-methoxybenzoic acid | 2.2027 | 2.1707 | -0.0320 |
| 12 | 4-(2-methyl-4-chlorophenoxy)butyric acid | 1.0342 | 1.1031 | 0.0689 |
| 13 | Anthracene | -0.8125 | -0.7893 | 0.0232 |
| 14 | Biphenyl | -0.3635 | -0.4137 | -0.0502 |
| 15 | Bromobenzene | 0.4103 | 0.4018 | -0.0085 |
| 16 | Buturon | 0.7800 | 0.5552 | -0.2248 |
| 17 | Chlorbromuron | 0.4472 | 0.2542 | -0.1930 |
| 18 | Chloroxuron | -0.3665 | -0.2431 | 0.1234 |
| 19 | Chlortoluron | 0.8031 | 0.8603 | 0.0572 |
| 20 | Fenuron | 1.7348 | 1.7923 | 0.0575 |
| 21 | Fluometuron | 1.0414 | 1.1409 | 0.0995 |
| 22 | Isoproturon | 0.5637 | 0.4297 | -0.1340 |
| 23 | Monolinuron | 1.0855 | 1.0343 | -0.0512 |
| 24 | Monuron | 1.2520 | 1.1843 | -0.0677 |
| 25 | Pentachorophenol | -0.9393 | -0.9198 | 0.0195 |
| 26 | Phenol | 0.6289 | 0.6997 | 0.0708 |
| 27 | Tetradifon | -1.2366 | -1.2524 | -0.0158 |
| | **Test set** | | | |
| 28 | (4-chloro-2-methylphenoxy)acetic acid | 1.8300 | 1.7709 | -0.0591 |
| 29 | 2-(3-chlorophenoxy)propionic acid | 1.9431 | 1.8643 | -0.0788 |
| 30 | 2,4-dichlorophenol | 0.0611 | -0.0612 | -0.1223 |
| 31 | 3-methyl-4-chlorophenol | 0.0656 | -0.0742 | -0.1398 |
| 32 | Benzene | 0.8814 | 0.6747 | -0.2067 |
| 33 | Buprofezin | -0.5361 | -0.4936 | 0.0425 |
| 34 | Diuron | 0.7646 | 0.7627 | -0.0019 |
| 35 | Metoxuron | 1.2435 | 1.4082 | 0.1647 |
| 36 | Neburon | -0.4510 | -0.3866 | 0.0644 |
| 37 | Propargite | -1.5850 | -1.4886 | 0.0964 |
| 38 | Triclopyr | 1.9279 | 1.9179 | -0.0100 |

[a]Log Chv in fish after 30 days

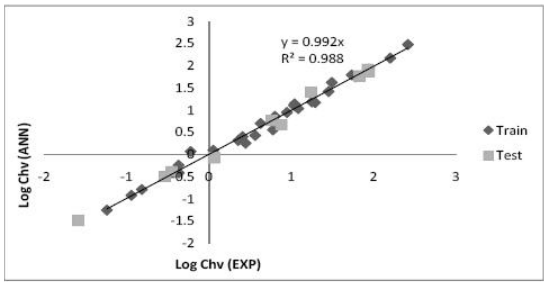


**Figure 1 : Plots of predicted Log Chv estimated by ANN modeling versus experimental Log Chv compounds**

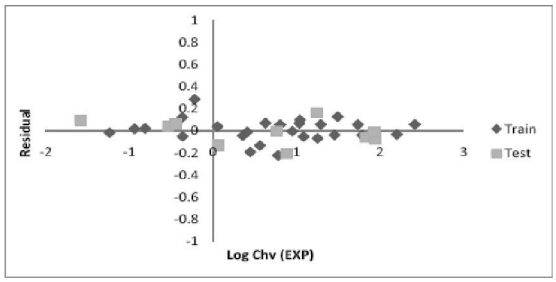


**Figure 2 : Plots of residual versus experimental Log Chv in ANN model**

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 back propagation 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 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.

The QSAR models for the estimation of the Log Chv 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 Chv models generation with the ANN method 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 Chv of the molecules were examined, and descriptors which showed high interrelation (i.e., $r>0.9$) with Log Chv and low interrelation (i.e., $r<0.9$) with each

Full Paper

**TABLE 2 : Molecular descriptors employed for the proposed ANN model**

| No. | Descriptor | Notation | class |
|---|---|---|---|
| 1 | Partition coefficient (octanol/water) | CLogP | Thermodynamic |
| 2 | Heat of formation | HF | Thermodynamic |
| 3 | Ievrege-weighted autocorrelation of lag 7/weighted by atomic sanderson electronegativities | HATS7e | Getaway |

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

| | CLogP | HF | HATS7e | Log chV |
|---|---|---|---|---|
| CLogP | 1 | 0.2135 | -0.0471 | -0.7526 |
| HF | | 1 | -0.5247 | -0.6586 |
| HATS7e | | | 1 | 0.5243 |
| Log chV | | | | 1 |

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

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 this process three descriptors were remained, that keeps most interpretive information for Log Chv. 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 Chv (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 Chv 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.

**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 | 0.6 |
| Momentum | 0.8 |
| Epoch | 2200 |
| Transfer function | Sigmoid |

**TABLE 5 : Statistical parameters obtained using the ANN model[a]**

| Ft | Fc | $R^2t$ | $R^2c$ | Rt | Rc | SEt | SEc | Model |
|---|---|---|---|---|---|---|---|---|
| 2076.6734 | 990.4636 | 0.9881 | 0.9910 | 0.9940 | 0.9955 | 0.1058 | 0.1130 | ANN |

[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

## Results of ANN analysis

The nonlinear QSAR model provided by the optimal neural networks is presented in figure1 where computed or predicted Log Chv values are plotted against the corresponding experimental data. Figure 2 shows a plot of residuals versus the observed Log Chv 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 TABLE1. 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 Chv values are (0.9940, 0.9955) for training and test set respectively.

## CONCLUSIONS

QSAR analysis was performed on a series of organic pollutants using ANN method that correlate Log Chv values of these compound to their structural descriptors. According to obtained results it is concluded that the (CLogP , HF, HATS7e) can be used successfully for modeling Log Chv 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 ANN model proposed with the simply calculated molecular descriptors can be used to estimate the logarithm values of chronic toxicity for new compounds even in the absence of the standard candidates.

## REFERENCES


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

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

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

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

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

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

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

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

**[9]** J.M.Vegas, P.J.Zufiria; 'Generalized Neural Network for Spectral Analysis', Dynamics and Liapunov Functions, Neural Networks, 17 **(2004)**.

**[10]** R.C.Schweitzer, J.B.Morris; Anal.Chem.Acta, **384**, 285 **(1999)**.

**[11]** C.S.Tong, K.C.Cheng; Chemometr.Intell.Lab.Syst., **49**, 135 **(1999)**.

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

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

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

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

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

**[17]** L.Escuder-Gilabert, Y.Mart´ýn-Biosca, S.Sagrado, R.M.Villanueva-amañas, M.J.Medina-Hernández; Anal.Chim.Acta, 173 **(2001)**.

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

**[19]** L.Fausett; Fundamentals of Neural Networks, Prentice Hall, New York, **(1994)**.