December 2008 Volume 7 Issue 10

Trade Science Inc.

Analytical CHEMISTRY
An Indian Journal

Full Paper

ACAIJ, 7(10) 2008 [751-756]

Prediction of retention times for a large set of pesticides or toxicants based on quantitative structure-retention relationships

A.H.M.Sarrafi, M.Alizadeh, F.Ahmadi*
Chemistry Department, Gachsaran Azad University, Gachsaran, (IRAN)
Tel : (098)-742-3333923; Fax : (098)-742-3333533
E-mail : Ahmadi@iaug.ac.ir
Received: 2nd December, 2008 ; Accepted: 7th December, 2008

ABSTRACT

In this paper, three different multivariate calibration methods feed-forward artificial neural networks (ANN) with back-propagation learning rule, Partial Least Squares (PLS) and Multiple Linear Regression (MLR) were applied to predict the retention time of 103 diverse pesticides or toxicants in gas chromatography-mass spectrometry (GC-MS) by using molecular structural descriptors. Five descriptors are considered to account for the effect of solute structure on the retention time. These are (solvation connectivity index chi-1, 3D -Balaban index, H autocorrelation of lag 3/weighted by atomic sanderson electronegativities, relative negative charge, Wiener-type index from Z weighted distance matrix (Barysz matrix). The Stepwise SPSS was used for the selection of the variables that resulted in the best-fitted models. After variables selection, 103 compounds randomly are divided into three training, validation and test sets. The mean square error (MSE) of training, test and validation sets for the ANN model are 0.0008676, 0.0014 and 0.0013, respectively. Result obtained showed that nonlinear model can simulate the relationship between structural descriptors and the retention times of the molecules in data set accurately.

KEYWORDS

Retention times;
QSRR;
PLS;
ANN;
Pesticides.

1. INTRODUCTION

High-performance liquid chromatography (HPLC) and gas chromatography (GC) are the most appropriate analytical techniques for multi-residue monitoring of pesticides in natural ecosystems or water and foodstuffs for human consumption, which, as a consequence of persistency and toxicological effects of these icrocontaminants, has become in the last decades an essential aspect of environmental protection and human health, safeguard policy[1]. As a potential alternative to expensive and time-consuming experimental trial-and-error approach traditionally adopted to optimize chromatographic separations, retention predictive models have received considerable attention in recent years[2]. An important property that has been extensively studied in QSRR is the chromatographic retention time. A QSRR study involves the prediction of chromatographic retention parameters using molecular structure. QSRR studies are widely investigated in gas chromatography (GC) and high-performance liquid chromatography (HPLC)[3]. The chromatographic parameters are expected to be proportional to a free energy change that is related to the solute distribution on the column. Chromatographic retention is a physical phenomenon that is primarily dependent on the interactions between the

solute and the stationary phase. Molecular group contribution methods are widely employed to estimate gas chromatographic retention parameters[4]. 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. Chromatography is the most suitable analytical tool for pesticide determination[5,6].

Quantitative structure-activity relationship (QSAR) methods represent an attempt to correlate structural and/or property descriptors of compounds with biological activities. These descriptors characterizing topological, connectivity, geometrical and getaway properties of a series of molecules have been traditionally determined mainly empirically, and only more recently by computational methods.

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[7,8] mean, however, a methodology related to nonlinear regression techniques[9,10]. Reviews have been published concerning applications of ANN in different fields[11,12]. Partial Least Squares (PLS) was introduced by Wold and Krishnaiah[13] and is commonly used in chemometrics as a modeling alternative to Ordinary Least Squares (OLS) when the predictor matrix is poorly conditioned. PLS regression is one of the standard calibration methods used in many chemical applications[14].

In the present work, a QSRR study, has been carried out on the GC retention times (t_R) for 103 diverse pesticides or toxicants by using structural molecular descriptors. The two linear methods MLR and PLS and nonlinear method feed forward neural network with back-propagation training along with Stepwise SPSS as variable selection software were used to model the retention times with the structural descriptors.

2. METHODS

2.2. Stepwise multiple linear regression

The multiple linear regressions (MLR) are an extension of the classical regression method to more than one dimension[15]. MLR calculates QSAR equation by performing standard multivariable regression calculations using multiple variables in a single equation. The stepwise multiple linear regressions are 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[16].

2.3. Partial least squares (PLS)

The PLS model will try to find a few PLS factors (also known as components or latent variables) that explain most of the variation in both predictors and responses. Factors that explain response variation well provide good predictive models for new responses, and factors that explain predictor variation well are well represented by the observed values of the predictors. The Partial Least Squares (PLS) regression method is well suited for problems with multicollinear predictor and response variables. PLS is explained in detail in literature[17,18]. To obtain the PLS model with the best predictive performance, the number of PLS components that optimize the predictive ability of the model should be determined. This is typically done by cross-validation, a procedure in which the available data within the training set are split into several subgroups called validation sets. The prediction residual sum of squares (PRESS) for the test samples is determined as a function of the number PLS components retained in the regression model that was formed with the training data. The procedure is usually repeated several times, with each subset in the training set being part of the test samples at least once[19].

2.4. Artificial neural networks

Principles, functioning and applications of artificial neural networks have been adequately described elsewhere[20,21]. 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.

2.5. 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) for calculation of the structural molecular descriptors (constitutional, topological, connectivity, geometrical, getaway 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.). PLS regression (PLS_Toolbox, version 2.1, Eigenvector Company) and other calculations were performed in the MATLAB (version 7.0, MathWorks, Inc.) environment.

3. RESULTS AND DISCUSSION

3.1. Datasets

Retention times (t_R) of 103 compounds including pesticides or toxicants were taken from the literature[22] that shown in TABLE 1.

TABLE 1: Data set and corresponding observed and (ANN, MLR, PLS) predicted values of Retention time (TR)

No.	Name	t_R	t_R	t_R	t_R
	Training set	(EXP)	(ANN)	(MLR)	(PLS)
1	Ethoprophos	1.195	1.199	1.274	1.310
2	Demton-s-methyl	1.217	1.295	1.297	1.307
3	Omethoate	1.239	1.260	1.249	1.268
4	Terbufos	1.303	1.318	1.349	1.351
5	Chlorbufan	1.313	1.320	1.336	1.366
6	Atrazine	1.323	1.308	1.318	1.329
7	Trietazine	1.326	1.344	1.338	1.357
8	Lindan	1.34	1.325	1.345	1.350
9	PCB15	1.344	1.360	1.410	1.396
10	Disulfoton	1.35	1.334	1.346	1.370
11	Dimetoate	1.356	1.284	1.279	1.289
12	Carbofuran	1.36	1.367	1.352	1.384
13	4,4'-DDM	1.39	1.395	1.427	1.422
14	PCB31	1.399	1.420	1.465	1.446
15	Benoxactor	1.401	1.433	1.424	1.437
16	Fenchlorphos	1.429	1.412	1.482	1.516
17	Phosphamidon	1.43	1.414	1.399	1.324
18	Aldrin	1.434	1.460	1.551	1.581
19	PCB52	1.443	1.492	1.520	1.496
20	Paration-methyl	1.449	1.441	1.441	1.455
21	Metalaxyl	1.449	1.489	1.464	1.444
22	Pentanochlor	1.451	1.436	1.345	1.371
23	Pirimiphos	1.452	1.451	1.536	1.525
24	Paraoxon-ethyl	1.463	1.491	1.438	1.482
25	Metolachlor	1.464	1.488	1.456	1.436
26	Chlorpyriphos	1.476	1.480	1.522	1.574
27	Fenitrothion	1.477	1.482	1.476	1.477
28	Malathion	1.478	1.477	1.477	1.446
29	Thiobencrab	1.478	1.434	1.396	1.430
30	Isodrin	1.485	1.513	1.569	1.580
31	Fenthion	1.501	1.517	1.497	1.491
32	Allethrin	1.503	1.523	1.547	1.536
33	Pendimethalin	1.516	1.494	1.469	1.446
34	Isocarbophos	1.52	1.500	1.511	1.493
35	PCB70	1.522	1.494	1.521	1.495
36	Isofenphos	1.523	1.529	1.540	1.580
37	Tridimenol	1.533	1.556	1.537	1.517
38	Bromophos-ethyl	1.539	1.532	1.548	1.589
39	Chlorfenvinphos	1.54	1.587	1.580	1.611
40	PCB101	1.545	1.567	1.573	1.545
41	2,4'-DDE	1.545	1.609	1.599	1.560
42	Alpha-endosulfan	1.549	1.630	1.599	1.650
43	Phenthoate	1.551	1.597	1.571	1.572
44	Chlorbenside	1.557	1.459	1.483	1.491
45	Prothiofos	1.572	1.555	1.553	1.600
46	Tetrachlorvinphos	1.576	1.560	1.592	1.603
47	Chinomethionate	1.577	1.562	1.473	1.466
48	PCB87	1.581	1.562	1.570	1.547
49	4,4'DDE	1.582	1.616	1.604	1.557
50	Iodofenphos	1.589	1.582	1.550	1.558
51	Fenamiphos	1.59	1.559	1.516	1.528

Countinue next page

Full Paper

No.	Name	t_R (EXP)	t_R (ANN)	t_R (MLR)	t_R (PLS)
	Training set				
52	2,4'-DDD	1.6	1.599	1.592	1.549
53	Binapacryl	1.604	1.572	1.552	1.541
54	PCB149	1.608	1.626	1.616	1.597
55	Endrin	1.612	1.628	1.625	1.626
56	PCB118	1.614	1.566	1.574	1.545
57	PCB153	1.627	1.627	1.625	1.594
58	Beta-endosulfan	1.637	1.576	1.579	1.650
59	4,4'DDD	1.637	1.601	1.591	1.547
60	PCB141	1.639	1.627	1.625	1.596
61	Sulprophos	1.647	1.629	1.592	1.598
62	4,4'-DDT	1.652	1.640	1.626	1.579
63	Benalaxyl	1.652	1.647	1.674	1.611
64	PCB187	1.653	1.666	1.663	1.646
65	Haloxyfop-2-ethoxyle	1.656	1.657	1.644	1.704
66	PCB185	1.661	1.669	1.662	1.647
67	PCB167	1.661	1.627	1.625	1.594
68	Edifenphos	1.664	1.658	1.622	1.648
69	PCB202	1.665	1.672	1.721	1.695
70	PCB128	1.666	1.625	1.623	1.597
71	Brompropylate	1.67	1.671	1.644	1.648
72	Fenpropathrin	1.673	1.688	1.731	1.705
73	Dicofol	1.678	1.652	1.598	1.601
74	Tetramethrin	1.679	1.665	1.630	1.639
75	Leptophos	1.687	1.691	1.714	1.700
76	Tertradifon	1.688	1.677	1.621	1.634
77	Phosalone	1.689	1.701	1.690	1.707
78	Pyrazophos ethyl	1.695	1.698	1.713	1.734
79	Fenarimol	1.697	1.688	1.663	1.653
80	Permethrin	1.7	1.688	1.742	1.757
81	Azinphos-methyl	1.7	1.690	1.642	1.638
	Test set				
1	Fonofos	1.337	1.338	1.370	1.424
2	Benfuresate	1.432	1.451	1.440	1.469
3	Methiocrab	1.483	1.428	1.387	1.376
4	Bromophos-methyl	1.504	1.465	1.515	1.537
5	Procymidone	1.544	1.537	1.522	1.504
6	Crotoxyphos	1.561	1.561	1.573	1.569
7	Buprofezin	1.59	1.569	1.567	1.558
8	Ethion	1.636	1.544	1.658	1.585
9	Famphur	1.664	1.632	1.648	1.581
10	Pyridaphention	1.68	1.679	1.618	1.675
11	PCB194	1.701	1.674	1.726	1.695
	Validation set				
1	Phorate	1.247	1.268	1.314	1.340
2	Dichlorofention	1.387	1.417	1.465	1.524
3	Trichoronate	1.472	1.451	1.483	1.530
4	Paration	1.495	1.494	1.469	1.508
5	Flumetralin	1.533	1.636	1.619	1.628
6	Quinalophos	1.549	1.531	1.538	1.592
7	Methidathion	1.587	1.607	1.554	1.531
8	2,4'-DDT	1.627	1.644	1.632	1.581
9	PCB-138	1.65	1.627	1.624	1.595
10	PCB180	1.674	1.649	1.674	1.645
11	Imidan	1.687	1.682	1.640	1.631

TABLE 2: Molecular descriptors employed for the proposed QSRR models

No.	Descriptor	Notation	Type	Coefficient
1	solvation connectivity index chi-1	X1sol	Connectivity	0.096 (±0.014)
2	3D -Balaban index	J3D	Geometrical	-0.026 (±0.009)
3	H autocorrelation of lag 3 / weighted by atomic sanderson electronegativities	H3e	Getaway	-0.047 (±0.029)
4	relative negative charge	RNCG	Charge	-0568 (±0.249)
5	Wiener-type index from Z weighted distance matrix (Barysz matrix)	WhetZ	Topological	-0.0000989 (±0.0000989)

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

	X1soL	J3D	H3e	RNCG	WhetZ	t_R
X1soL	1	-0.186	0.605	0.019	0.653	0.810
J3D		1	0.039	0.207	-0.144	-0.513
H3e			1	-0.123	0.557	0.306
RNCG				1	-0.432	-0.180
WhetZ					1	0.474
t_R						1

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

The QSRR models for the estimation of the retention times 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 format; (2) quantum mechanics geometry optimization with a semi-empirical (AM1) method; (3) structural descriptors computation; (4) structural descriptors selection; (5) structure-retention models generation with the multivariate methods(ANN,MLR,PLS) and statistical analysis.

The data set was divided into three subsets in ANN, MLR and PLS: a training set of 81 compounds, a test and a validation sets both of 11 compounds.

3.2. Descriptors selection

Generally the first step in variables selection is the calculation of the correlation between variables and with seeking property. In the present case, to decrease the redundancy existed in the descriptors data matrix, the correlations of descriptors with each other and with the t_R of the molecules were examined, and descriptors which showed high interrelation (i.e., $r>0.9$) with t_R 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 QSRR 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 retention time. 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 retention time (TABLE 3).

3.3. 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 retention time 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.

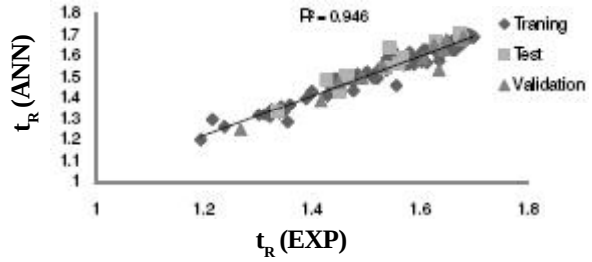

Figure 1 : Plots of predicted t_R estimated by ANN modeling versus experimental t_R compounds

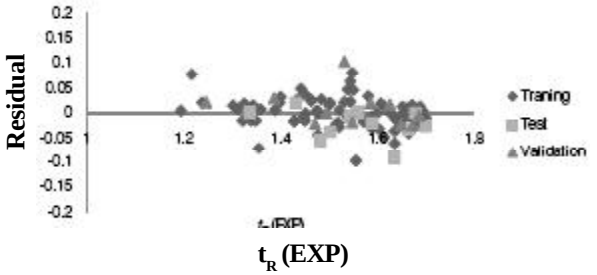

Figure 2 : Plots of residual versus experimental t_R in ANN model

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	0.6
Momentum	0.1
Epoch	5300
Transfer function	Sigmoid

3.4. Results of ANN analysis and comparison with MLR and PLS

The nonlinear QSRR model provided by the optimal neural network is presented in figure 1 where computed or predicted retention time values are plotted against the corresponding experimental data. Figure 2 shows a plot of residuals versus the observed retention time 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, validation and test sets are shown in TABLE 1 and TABLE 5. The statistical parameters calculated for the ANN, MLR and PLS models 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 t_R values 0.9728, 0.9604 and 0.9605 for training, validation and test set, respectively.

For comparison, a linear QSRR model relating retention times to the selected descriptors were obtained by means of MLR and PLS methods.

With the purpose MLR and PLS models built on

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

Ft	Fv	Fc	Rt	Rv	Rc	SEt	SEv	SEc	Model
107.357	106.792	1394.641	0.960	0.960	0.973	0.031	0.037	0.029	ANN
66.553	78.729	511.820	0.939	0.947	0.931	0.042	0.036	0.044	MLR
27.359	30.040	349.153	0.867	0.877	0.903	0.051	0.043	0.041	PLS

[a]c refers to the calibration (training) set; v refers to validation set; t refers to test set; R is the correlation coefficient; and F is the statistical F value.

Full Paper

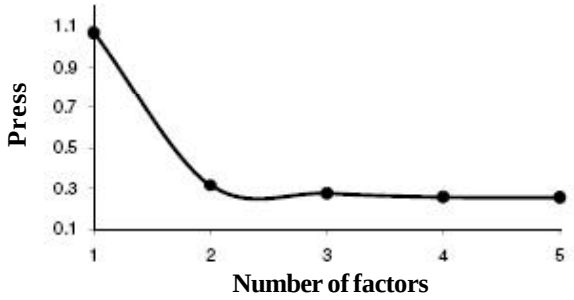

Figure 3: Plot of press versus PC number for the PLS model

the same subsets that used in ANN analysis.

Multiple linear regressions (MLR) are one of the most used modeling methods in QSRR. The colinearity problem of the MLR method has been overcome through the development of the partial least-squares projections to latent structures (PLS) method, which has been shown to be an efficient approach in monitoring many complex processes, reducing the high dimensional strongly cross-correlated data to a much smaller and interpretable set of principal components or latent variables. The number of significant factors for the PLS algorithm was determined using the cross-validation method. The optimum number of factors was concluded as the first local minimum in the PRESS versus number of factors plot. Figure 3 shows the plot of PRESS versus number of factors for the PLS model. The best PLS model contained five selected descriptors in two latent variables space.

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

4. CONCLUSIONS

QSRR analysis was performed on a series of pesticides or toxicants using ANN, MLR and PLS methods that correlate t_R values of these compound to the structural descriptors.

According to obtained results it is concluded that the X1sol, J3D, H3e, RNCG and Whetz can be used successfully for modeling t_R property of the under study compounds. The statistical parameters of the built QSRR 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 QSRR models proposed with the simply calculated molecular descriptors can be used to estimate the chromatographic retention times for new compounds even in the absence of the standard candidates.

REFERENCES

[1] M.Aschi, A.A.D.Archivio, M.A.Maggi, P.Mazzeo, F.Ruggieri; Anal.Chim.Acta, **582**, 235 **(2007)**.
[2] R.kaliszan; 'Quantitative Structure-Chromatographic Retention Relationships', John Wiley and Sons, New York, **(1987)**.
[3] J.Ghasemi, S.Asadpour, A.Abdolmaleki; Anal. Chim.Acta, **588**, 200 **(2007)**.
[4] I.Teodora, I.Ovidiu;, Internet electronic journal of molecular design **1**, 94 **(2002)**.
[5] E.Hogendoorn, P.Van Zoonen; J.Chromatogr.A, **892**, 435 **(2000)**.
[6] G.R.Van Der Hoft, P.Van Zoonen; J.Chromatogr.A, **843**, 301 **(1999)**.
[7] W.S.Mculloch, W.Pitts; Bull.Math.Bioph., **5**, 115 **(1943)**.
[8] D.E.Rumelhart; Parallel Distributed Processing. London: Mit press, **1982**.
[9] J.Zupan, J.Gasteiger; Anal.Chim.Acta, **248**, 1 **(1991)**.
[10] D.T.Manallack, D.D.Ellis, D.J Livingstone; J.Med. Chem., **37**, 3758 **(1994)**.
[11] A.Guez, I.Nevo; Clin.Chim.Acta, **248**, 73 **(1996)**.
[12] V.Jakus; Chem.Listy., **87**, 262 **(1993)**.
[13] H.Wold, P.R.Krishnaiah; Academic Press, New York, **36**, 391**(1966)**.
[14] P.Geladi, J.Chemom., **2**, 231 **(1988)**.
[15] R.H.Myers; Classical and Modern Regression With ApplicationPws-Kent Publishing Company: Boston, **(1990)**.
[16] J.Ghaseni, Sh, Ahmadi; Ann.Chim.(Rome)., **97 (1-2)**, 69 **(2007)**.
[17] A.Lorber, L.E.Wangen, B.R.Kowalski, J.Chemom, 1, 19 **(1987)**.
[18] A.Hoskuldsson, Journal of Chemometrics, **2**, 211 **(1988)**.
[19] H.Swierenga, A.P.D.Weijer, R.J.V.Wijk, L.M.C. Buydens, Chemom.Intell.Lab.Syst, **49**, 1 **(1999)**.
[20] J. Zupan, J. Gasteiger, Neural Networks In Chemistry And Drug Design, Wiley- Vch Verlag, Weinheim, **(1999)**.
[21] L.Fausett, Fundamentals Of Neural Networks, Prentice Hall, New York, **(1994)**.
[22] L.I.Xiuyong, Feng Luana, S.I.Hongzong, H.U. Zhide, L.I.U.Mancang, Toxicol.Lett., **175**, 136 **(2007)**.