Revue Roumaine de Chimie, **1998**, *43* (3), 255–260

ARTIFICIAL NEURAL NETWORKS APPLICATIONS. PART 4[1]

QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIPS FOR THE ESTIMATION OF THE RELATIVE TOXICITY OF PHENOLS FOR *TETRAHYMENA*

Ovidiu IVANCIUC

University "Politehnica" of Bucharest, Faculty of Chemical Technology, Department of Organic Chemistry, Splaiul Independenţei 313, 77206 Bucharest, Roumania
E-mail: o_ivanciuc@chim.upb.ro

Received March 20, 1996

The toxicity of 30 *para*-substituted phenols for *Tetrahymena pyriformis* has been investigated using Artificial Neural Networks (ANN), using as structural descriptors lg K_{ow} (the 1-octanol/water partition coefficient) and pK_a. The Quantitative Structure-Activity Relationship (QSAR) formulated with ANN is compared with a multiple linear regression QSAR, and the predictive ability of the neural model is tested with the leave-one-out method. The chance correlation of ANN models is investigated using random input patterns and a random assignment of experimental output patterns.

INTRODUCTION

Originally, research into neural networks was primarily motivated by a desire to model the working of the human brain. An Artificial Neural Network (ANN) is, basically, a computer program that can detect patterns and correlations in data. The success of this methodology in the recognition and classification of patterns has attracted much interest in recent years.[2–3]

A great number of problems from diverse branches of chemistry have already been investigated by applying neural networks;[4–7] among them, the Quantitative Structure-Property Relationships (QSPR) and Quantitative Structure-Activity Relationships (QSAR).[8–14] Because the use of ANN in QSAR studies is new, there is a need to compare the capabilities of ANN with those of more established models. Such comparisons will bring out the advantages and/or disadvantages of neural networks.

The scope of the present paper is to study the applicability of ANN to obtain reliable QSAR models when the number of data in the training set is small, and to compare the results with the estimations of the Multiple Linear Regression (MLR) model for the same set of compounds. The toxicity of 30 *para*-substituted phenols for *Tetrahymena pyriformis* is investigated, using as structural descriptors the physico-chemical parameters of the compounds. The chance correlation of ANN models is estimated using random input patterns and a random assignment of experimental output patterns. The predictive ability of the neural model is examined with the leave-one-out cross-validation method.

RESULTS AND DISCUSSION

The toxicity of a set of 30 *para*-substituted phenols for *Tetrahymena pyriformis* was recently reported.[15] The model contained as structural descriptors lg K_{ow} (the 1-octanol/water partition coefficient) and pK_a. The structure of the 30 phenols, structural descriptors and experimental data (lg BR, the

logarithm of the inverse of the IGC value in mmol/L) are presented in Table 1 and were used in ref. 15 to develop an MLR QSAR:

$$\lg BR = 1.9860\ (\pm 0.4845) + 0.6577(\pm 0.1604)\ \lg K_{ow} - 0.3171(\pm 0.0774)\ pK_a \qquad (1)$$

$$n = 30 \qquad r = 0.955 \qquad s = 0.221$$

where n is the number of compounds used in the correlation, r is the correlation coefficient, and s is the standard deviation. The standard error of estimation of each coefficient at the 95% confidence level is given in parentheses. The partial correlation coefficients are $r(\lg K_{ow}) = 0.870$ and $r(pK_a) = -0.113$, leading to the conclusion that both independent parameters must be used in order to obtain a good QSAR model.

Table 1

Structure, molecular descriptors and toxicity data for phenols

No	X	$\lg K_{ow}$	pK_a	$\lg BR_{exp}$	Res^a_1	Res^b_2
1	$CONH_2$	0.00	9.23	–0.7802	–0.0102	0.1606
2	$NHCOCH_3$	0.32	9.99	–0.8198	–0.0592	0.1515
3	CH_2CH_2OH	0.72	10.12	–0.8275	–0.1516	–0.0780
4	CH_2CN	0.90	9.97	–0.3840	0.2007	0.1996
5	OCH_3	1.34	10.20	–0.1425	0.2317	0.2246
6	CHO	1.35	7.62	0.2661	–0.0298	–0.1915
7	$COCH_3$	1.35	8.05	–0.0932	–0.2428	–0.4144
8	H	1.49	9.99	–0.4310	–0.2151	–0.2291
9	COC_2H_5	1.55	8.85	0.0557	–0.0368	–0.1434
10	CN	1.60	7.95	0.5161	0.1199	–0.0013
11	F	1.77	9.89	0.0169	0.0106	0.0029
12	OC_2H_5	1.87	10.52	0.0130	0.0846	0.1330
13	NO_2	1.91	7.15	1.4257	0.1800	0.4508
14	CH_3	1.94	10.26	–0.1920	–0.2263	–0.2005
15	Cl	2.39	9.43	0.5447	0.0936	–0.0229
16	C_2H_5	2.58	10.00	0.2058	–0.1791	–0.3060
17	Br	2.59	9.34	0.6806	–0.0976	–0.0471
18	I	2.91	9.20	0.8544	–0.3763	–0.1281
19	OC_4H_9	3.04	10.70	0.7016	0.2478	0.1092
20	$CH(CH_3)_2$	3.05	10.32	0.4732	–0.2007	–0.2462
21	COC_6H_5	3.07	8.89	1.0237	–0.2786	–0.1624
22	C_3H_7	3.18	10.28	0.6350	–0.2738	–0.1826
23	$N{=}NC_6H_5$	3.18	8.56	1.6547	0.3402	0.2917
24	C_6H_5	3.20	9.55	1.3828	0.1129	0.3205
25	$C(CH_3)_3$	3.31	10.23	0.9126	–0.2046	–0.0064
26	OC_6H_5	3.56	10.70	1.3550	0.2440	0.4206
27	$CH_2CH(CH_3)_2$	3.60	10.30	0.9797	–0.2911	–0.1078
28	cyclopentyl	3.63	9.92	1.2916	–0.0154	0.0639
29	$CH_2C_6H_5$	3.69	10.19	1.1946	–0.1049	0.0130
30	$CH_2C(CH_3)_3$	4.03	10.50	1.2326	–0.0809	–0.0743

[a] $Res_1 = \lg BR_{exp} - \lg BR_{NN}$
[b] $Res_2 = \lg BR_{exp} - \lg BR_{eq.(1)}$

The ANN used in the present study are three-layer Multi-Layer Feedforward (MLF) networks, with two input units representing the two independent parameters (lg K_{ow} and pK_a), and one output unit (representing lg BR); for training we have used the backpropagation algorithm and the activation function for the hidden layer was the hyperbolic tangent. Other important specifications for the ANN used in simulations are: input scaling between –0.9 and 0.9, initial weights scaling between –0.1 and 0.1, learning rate 0.01, and the momentum 0.8.

To optimize the ANN with respect to the QSAR model quality, network architecture, and training regime, we have used several statistical indices: the standard deviation s and the correlation coefficient r of the linear correlation between lg BR_{exp} and lg BR_{ANN} computed with the neural network: lg $BR_{exp} = a + b$ lg BR_{ANN}.

The number of neurons in the hidden layer was selected on the basis of empirical trials, in which ANN with different number of hidden neurons are trained to predict the lg BR_{exp} values. The training was done by randomly presenting to the network the patterns from the set of 30 *para*-substituted phenols, using a hyperbolic tangent as output activation function, and output scaling between –0.9 and 0.9. Ten ANN were generated, with the number of hidden neurons between 1 and 10. The training was terminated after 1000 complete cycles, and the results obtained in the evaluation of lg BR_{exp} show that a network with only one hidden neuron provides a better estimation of lgBR values than the MLR model ($r = 0.960$, $s = 0.206$). The network performances increase for three hidden neurons ($r = 0.971$, $s = 0.172$), and are almost constant for the ANN with larger number of hidden neurons. Because in certain cases a linear output function provides a better QSAR model, a set of ten networks with a linear output activation function were trained for 1000 epochs. The output scaling was between –0.9 and 0.9, and the number of hidden neurons varied between 1 and 10. The results obtained in the evaluation of lg BR show a constant improvement of the correlation coefficient when compared with the results obtained with a tanh output function. The correlation coefficient reaches an almost constant value for the size of the hidden layer between three and ten neurons, and is always greater than r in the MLR model.

The performance of a neural network with a linear output activation function is strongly dependent on the range of the output scaling. In order to test the influence of this parameter, a network with five hidden neurons and linear output activation was trained for 1000 epochs. The output scaling was modified between ±1.5 and ±10. The results for the estimation of lg BR show that the correlation coefficient has a maximum for a scaling between –6 and 6. Considering the results obtained, ANN with a linear output activation function will be considered, with an output scaling between –6 and 6.

In QSAR studies it is very important to take into consideration that MLF ANN are universal approximators: they are capable of arbitrarily accurate approximation to arbitrary mappings, provided sufficiently many hidden units are available. Recently, it was proved by Cybenko[16] and Funahashi[17] that any continuous function can be approximated on a compact set with the uniform topology, by a MLF network with one hidden layer. Hornik, Stinchcombe, and White[18] have shown that any measurable function can be approximated with a multilayer feedforward neural network. The properties of the intermediate layer activation function are not as crucial, and the sigmoid activation function is not necessary for universal approximation. This great modelling power of MLF ANN represents a property which must be considered with caution when using neural networks for quantitative property estimation. The most important problem with the use of ANN in QSAR is that any physical, chemical, or biological property of a set of chemical compounds can be approximated using as input data random numbers if the network contains a sufficiently large number of hidden neurons.

In a recent paper, Andrea and Kalayeh[8] characterized a network by a parameter, ρ, which is the ratio of the number of patterns in the training set to the number of connections. The optimum ρ values for the MLF networks lay in the range $1.8 < \rho < 2.2$ and it was reported that networks having ρ values in excess of 2.2 failed to extract relevant features and gave poor predictions. The potential of chance correlations in ANN was investigated for classification and for continuous output networks using random data as input and output patterns.[19,20] One must consider that the results for the optimum ρ value obtained

with simulations using random patterns do not adequately represent the type of data normally encountered in a QSAR study, and the chance correlations must be studied for each particular case.

The set of chemical compounds investigated in the present paper presents some problems for a neural network model: a low number of structural parameters and a low number of experimental data. The usual cross-validation procedure gives too pessimistic values for the cross-validation correlation coefficient r_{cv}, because the calculation of the output values for patterns which are extremal in one or more variables involves extrapolation, and neural networks give poor results in these cases.

An important goal of this work is to determine how to distinguish between true and random correlations in QSAR models using neural networks, due to the fact that ANN are universal approximators. In order to establish the role of randomness in developing ANN models when the number of compounds in the training set is low, we simulated a learning set of patterns by retaining the values of the output patterns assigned to the corresponding compounds, and randomly generating the input data. A neural network with 5 hidden neurons is trained for 1000 epochs using this simulated training set. If the fit obtained with the real data is consistently better than that with the random input data, the correlation obtained with the real data can confidently be said to be not due to chance factors. The procedure of generating simulated training sets was repeated 16 times, and the statistical indices of estimating lg BR are much lower than the values obtained with real data: $s_{min} = 0.323$, $s_{max} = 0.730$, $s_{mean} = 0.543$, $r_{min} = 0.087$, $r_{max} = 0.898$, $r_{mean} = 0.606$. Even networks with only one hidden neuron trained with real data outperform networks with five hidden neurons trained with random input data.

The simulations using random input data give no idea of the effects of correlation between independent variables and inhomogeneous distribution of the values of parameters. A second set of simulations was done using a training set obtained from the real one by randomly assigning a real lg BR value to a compound not necessarily possessing this activity, and retaining the values of the input patterns. A neural network with five hidden neurons was trained for 1000 cycles with 16 such random training sets; the statistical results obtained in the runs using randomly reassigned activity data are considerably lower than those obtained when actual activity data are used: $s_{min} = 0.361$, $s_{max} = 0.725$, $s_{mean} = 0.547$, $r_{min} = 0.144$, $r_{max} = 0.869$, $r_{mean} = 0.612$. Considering the results obtained, we are confident that a valid relationship has indeed been established between the structural descriptors and the observed biological properties. Even networks with only one hidden neuron give estimations of lg BR with better statistical indices than in the tests with random input patterns or randomly reassingned target data.

Table 2

Statistical results for the calibration of the neural networks with linear output activation function. H is the number of hidden neurons, N is the number of epochs, a and b represent the linear regression coefficients, r is the correlation coefficient, and s represents the standard deviation of the equation $\lg BR_{exp} = a + b \lg BR_{ANN}$

H	N	a	b	r	s
1	200	–0.1062	0.9546	0.931	0.268
2	500	–0.0271	0.9736	0.964	0.194
3	1400	0.0006	0.9875	0.976	0.161
4	2600	–0.0221	0.9914	0.983	0.133
5	2600	–0.0221	0.9913	0.983	0.133

A set of neural networks with a variable number of hidden neurons with a linear output activation function and an output scaling between –6 and 6 were trained until convergence was achieved. The learning process was finished when the correlation coefficient of the estimate increased with less than 10^{-4} in 100 epochs. The statistical results are presented in Table 2. If we compare the predictions of the MLR model represented by eq. (1) with the predictions of the ANN model presented in Table 2, it is clear that neural network outperforms regression analysis and provides superior mapping of physicochemical parameters to biological activities.

In order to estimate the prediction power of the neural model we have used the leave-one-out cross-validation. The results are presented in Table 3, showing that the best predictions are offered by a network with two hidden neurons. The calibration residuals for the network with two hidden neurons are presented in Table 1, column 6. The ANN residuals are lower than the MLR residuals, presented also in Table 1, column 7.

Table 3

Statistical results for the leave-one-out cross-validation of the neural networks with linear output activation function. The notations are the same as those from Table 2

H	*a*	*b*	*r*	*s*
1	-0.1044	0.9518	0.910	0.303
2	0.0275	0.9363	0.912	0.301
3	0.0522	0.9417	0.884	0.343
4	0.0869	0.9183	0.893	0.329
5	0.0917	0.8189	0.861	0.373

CONCLUSIONS

The neural network approach gives better predictions than the usual MLR model for the estimation of the relative toxicity of phenols for *Tetrahymena*. Because neural networks are universal approximators, when the ρ index (the ratio of the number of patterns in the training set to the number of connections of the network) has a low value, there is a danger to obtain chance correlations or to obtain a network that memorizes the data. A low value of the ρ index is a usual characteristic of the networks used in QSAR studies, and each case must be investigated in order to avoid the risk of chance correlations or memorizing the data.

In order to establish the role of randomness in developing ANN models with a certain topology of the neural network, we propose two methods. The first one uses artificial learning sets obtained by retaining the values of the output patterns assigned to the corresponding compounds, and randomly generating the input data. Then, neural networks with the same topology as the real one are trained with learning sets generated in this way, and if the fit obtained with the real training set is consistently better than that with the random input data, the correlation obtained with the real data can confidently be said to be not due to chance factors.

Because the simulations using random input data give no idea of the effects of correlation between independent variables and inhomogeneous distribution of the values of parameters, the second method uses a training set obtained from the real one by randomly assigning a real activity value to a compound not necessarily possessing this activity, and retaining the values of the input patterns. Then a neural network with the same structure as the one used in the QSAR study is trained. This procedure is repeated a number of times, and if constantly the fit obtained with the real learning set is better than the one obtained with randomly assigned activity data, one can be confident that a valid relationship has indeed been established between the structural descriptors and the observed biological properties. Otherwise, the number of hidden neurons must be lowered and/or the learning set must be expanded.

ACKNOWLEDGEMENT. Financial support was obtained from the Ministry of Research and Technology under Grant 310 TA10 and from the Ministry of the National Education under Grant 7001 T34.

REFERENCES

1. Previous Part: O. Ivanciuc, *Rev. Roum. Chim.*, **1997**, *42*, 325–332.
2. D. E. Rumelhart and J. L. McClelland, *Parallel Distributed Processing*, MIT Press, Cambridge, MA, 1986.
3. D. E. Rumelhart, G. E. Hinton and R. J. Williams, *Nature*, **1986**, *323*, 533–536.
4. P. D. Wasserman, *Neural Computing*, Van Nostrand Reinhold, New York, 1989.
5. B. J. Wythoff, *Chemom. Intell. Lab. Syst.*, **1993**, *18*, 115–155.
6. J. Gasteiger and J. Zupan, *Angew. Chem. Int. Ed. Engl.*, **1993**, *32*, 503–527.
7. J. R. M. Smits, W. J. Melssen, L. M. C. Buydens and G. Kateman, *Chemom. Intell. Lab. Syst.* **1994**, *22*, 165–189.
8. T. A. Andrea and H. Kalayeh, *J. Med. Chem.*, **1991**, *34*, 2824–2836.
9. D. Zakarya, L. Farhaoui and S. Fkih-Tetouani, *Tetrahedron Lett.*, **1994**, *35*, 4985–4988.
10. O. Ivanciuc, *Rev. Roum. Chim.*, **1995**, *40*, 567–574.
11. O. Ivanciuc, *Rev. Roum. Chim.*, **1995**, *40*, 1093–1101.
12. O. Ivanciuc, *Rev. Roum. Chim.*, **1996**, *41*, 645–652.
13. O. Ivanciuc, J.-P. Rabine, D. Cabrol-Bass, A. Panaye and J.P. Doucet, *J. Chem. Inf. Comput. Sci.*, **1996**, *36*, 644–653.
14. O. Ivanciuc, J.-P. Rabine, D. Cabrol-Bass, A. Panaye and J.P. Doucet, *J. Chem. Inf. Comput. Sci.*, **1997**, *36*, 587–598.
15. T. W. Schultz, *Bull. Environ. Contam. Toxicol.*, **1987**, *38*, 994–999.
16. G. Cybenko, *Math. Control Signals Syst.*, **1989**, *2*, 303–314.
17. K. Funahashi, *Neural Networks*, **1989**, *2*, 183–192.
18. K. Hornik, M. Stinchcombe and H. White, *Neural Networks*, **1989**, *2*, 359–366.
19. D. T. Manallack and D. J. Livingstone, *Med. Chem. Res.*, **1992**, *2*, 181–190.
20. D. J. Livingstone and D. T. Manallack, *J. Med. Chem.*, **1993**, *36*, 1295–1297.