Revue Roumaine de Chimie, **1997**, *42* (4), 325–332

ARTIFICIAL NEURAL NETWORKS APPLICATIONS. Part 3[1]

A QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIP FOR THE ACTINIDIN HYDROLYSIS OF SUBSTITUTED-PHENYL HIPPURATES

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 7, 1995

The hydrolysis of 29 phenyl hippurates ($XC_6H_4OCOCH_2NHCOC_6H_5$) by the cystein protease actinidin has been investigated with artificial neural networks (ANN), using as structural descriptors σ (the Hammett constant), π_3' (the hydrophobic constant for the more hydrophobic of the two meta substituents), and MR_4 (the molar refractivity of *para* substituents). The quantitative structure activity relationship (QSAR) formulated with ANN is compared with a multiple linear regression QSAR. The chance correlation of ANN models is investigated using a random assignment of experimental output patterns, and the predictive ability of the neural networks is tested with the leave-one-out method.

INTRODUCTION

Artificial Neural Networks (ANN) are a promising new method for solving hard problems by virtue of their ability to construct an internal representation of the problem which allows accurate predictions to be made for similar problems.[2–4] In many chemical areas there is a wealth of experimental data but a scarcity of rules to define the chemical process. In such cases ANN, which employ learning procedures to develop internal representations from examples, may be able to discern patterns in the data which would allow the development of very accurate models of the chemical phenomena.[5–7]

Recently there has been growing interest in the application of neural networks in the field of quantitative structure-property relationships[8–10] and Quantitative Structure-Activity Relationships (QSAR).[11–12] It has been demonstrated that this new technique is often superior to the traditional MultiLinear Regression (MLR) analysis. The commonest type of ANN used in QSAR is the Multi-Layer Feedforward (MLF) network, consisting of three types of layers of units (or artificial neurons): a layer of input units is connected to one or more layer(s) of hidden units, which is connected to a layer of output units. The activity of the input units represents the raw information that is fed into the network. The activity of each hidden unit is determined by the activities of the input units and the weights on the connections between the input and hidden units. Similarly, the behaviour of the output units depends on the activity of the hidden units and the weights between the hidden and output units.

The scope of the present paper is to study the applicability of ANN to investigate the hydrolysis of phenyl hippurates by the cysteine protease actinidin, and to compare the results with the predictions of the MLR model for the same set of compounds. The chance correlation of ANN models is investigated using a random assignment of experimental output patterns. The predictive ability of the neural model is tested with the leave-one-out cross-validation method.

RESULTS AND DISCUSSION

Actinidin is a cysteine protease isolated from the fruit of *Actinidia chinensis* (Kiwi fruit); the cystein protease actinidin hydrolysis of 29 phenyl hippurates, with the formula $XC_6H_4OCOCH_2NHCOC_6H_5$, has been investigated using MLR QSAR.[13] The model contained as structural descriptors three empirical substituent parameters: σ, the sum of the Hammett constant for the substituents in the positions 3, 4, and 5; π_3', the hydrophobic constant of the more hydrophobic of the two *meta* substituents; MR_4, the molar refractivity of the substituent in the 4-position of phenyl hippurates. The structures of the 29 phenyl hippurates, the substituent constants, and the experimental $\log 1/K_m$ (where K_m is the Michaelis constant) are presented in Table 1 and were used in ref. 13 to develop a MLR QSAR. With all 29 experimental data, the following equation is obtained:

Table 1

Structure, physical parameters, experimental and calculated $\log 1/K_m$ for the actinidin-catalyzed hydrolysis of $XC_6H_4OCOCH_2NHCOC_6H_5$ at 25°C, pH 6.0

No.	Substituent	σ	π_3'	MR_4	$\log 1/K_m$		
					exp[a]	calc[b]	residual[c]
1	H	0.00	0.00	0.10	2.77	2.85	−0.08
2	3-F	0.34	0.14	0.10	3.01	3.08	−0.07
3	3-Cl	0.37	0.71	0.10	3.63	3.61	0.02
4	3-Br	0.39	0.86	0.10	3.64	3.68	−0.04
5	3-I	0.35	1.12	0.10	3.93	3.72	0.21
6	3-CH_3	−0.07	0.56	0.10	3.26	3.31	−0.05
7	3-t-Bu	−0.10	1.98	0.10	3.66	3.69	−0.03
8	3-CF_3	0.43	0.88	0.10	3.47	3.71	−0.24
9	3-CN	0.56	0.00	0.10	3.08	3.36	−0.28
10	3-NO_2	0.71	0.00	0.10	3.53	3.51	0.02
11	3-$NHCOCH_3$	0.21	0.00	0.10	3.18	3.02	0.16
12	3-$CONH_2$	0.28	0.00	0.10	3.15	3.08	0.07
13	3-SO_2NH_2	0.46	0.00	0.10	3.42	3.26	0.16
14	4-F	0.06	0.00	0.09	2.72	2.90	−0.18
15	4-Cl	0.23	0.00	0.60	3.04	3.11	−0.07
16	4-CH_3	−0.17	0.00	0.56	2.95	2.79	0.16
17	4-$COCH_3$	0.50	0.00	1.12	3.47	3.48	−0.01
18	4-CN	0.66	0.00	0.63	3.62	3.64	−0.02
19	4-NO_2	0.78	0.00	0.74	3.80	3.79	0.01
20	4-OCH_3	−0.27	0.00	0.79	2.87	2.87	0.00
21	4-NH_2	−0.66	0.00	0.54	2.55	2.65	−0.10
22	4-$CONH_2$	0.36	0.00	0.98	3.47	3.47	0.00
23	4-SO_2NH_2	0.57	0.00	1.23	3.22	3.21	0.01
24	3,5-$(CH_3)_2$	−0.14	0.56	0.10	3.37	3.30	0.07
25	3-CH_3-5-Et	−0.14	1.02	0.10	3.52	3.55	−0.03
26	3,5-$(OCH_3)_2$	0.24	0.02	0.10	3.60	3.06	0.54
27	3,5-Cl_2	0.74	0.71	0.10	3.84	3.79	0.05
28	3,5-$(NO_2)_2$	1.42	−0.28	0.10	3.90	3.90	0.00
29	3,4,5-Cl_3	0.97	0.71	0.60	4.01	4.01	0.00

[a] The data were taken from ref. 3

[b] Estimated by the network NN2

[c] Residual = $\log 1/K_{m\,exp} - \log 1/K_{m\,NN2}$

$$\log 1/K_m = 2.976(\pm 0.012) + 0.798(\pm 0.454)\,\sigma + 0.456(\pm 0.292)\,\pi_3' + 0.090(\pm 0.058)MR_4 \quad (1)$$

$$n = 29 \quad r = 0.883 \quad s = 0.191$$

where n is the number of compounds used in the correlation, r is the correlation coefficient, and s represents 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 low, and the most important descriptor is the Hammett parameter σ: $r(\sigma) = 0.676$, $r(\pi_3') = 0.438$, $r(MR_4) = -0.077$. The collinearity between variables is quite low as indicated by the intercorrelation matrix:

	σ	π_3'	MR_4
σ	1	–0.172	0.057
π_3'		1	–0.342

The low values of the intercorrelation coefficients indicate that all the three structural parameters can be used in a MLR equation. Equation (1) has a strong outlier, namely the compound **26**, with 3,5-$(OCH_3)_2$ substituents. Therefore, in the derivation of the equation (2) this compound was not used.

$$\log 1/K_m = 2.927(\pm 0.010) + 0.720(\pm 0.399)\sigma + 0.492(\pm 0.273)\pi_3' + 0.140(\pm 0.077)MR_4 \quad (2)$$

$$n = 28 \quad r = 0.912 \quad s = 0.169$$

A significant increase of r and a decrease of s is a strong indication that compound **26** is an outlier. In ref. 13 the compound **23**, with the 4-SO_2NH_2 substituent, was considered also an outlier, although in eq. (2) its residual was comparable with the residuals of other compounds in the set. After the deletion of compound **23**, the following MLR equation was computed:

$$\log 1/K_m = 2.896(\pm 0.009) + 0.738(\pm 0.376)\sigma + 0.504(\pm 0.256)\pi_3' + 0.245(\pm 0.125)MR_4 \quad (3)$$

$$n = 27 \quad r = 0.927 \quad s = 0.157$$

The increase of r is not so great and we consider that compound **23** is not an outlier, as it was considered in ref. 13.

The neural networks used in the present study are three-layer MLF networks, with three input units representing the three independent parameters in eq. (1–3), (σ, π_3', and MR_4, respectively), and one output unit (representing $\log 1/K_m$); for training we have used the backpropagation algorithm, and the transfer function was the hyperbolic tangent. Other important specifications for the ANN used in simulations are: input scaling between –0.9 and 0.9, output 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.

The quality of ANN output was assessed by a set of statistical variables: the mean square error (MSE) in scaled units, and the standard deviation s and the correlation coefficient r of the linear correlation between $\log 1/K_{m\,exp}$ and $\log 1/K_{m\,ANN}$ of the type:

$$\log 1/K_{m\,exp} = a + b \log 1/K_{m\,ANN}$$

MSE is defined by the following expression:

$$MSE = \frac{\sum_i \sum_j (Output_{ij} - Target_{ij})^2}{P \cdot O}$$

where $Output_{ij}$ represents the output of the neuron j in the output layer for the pattern i, and $Target_{ij}$ represents the desired output in scaled units for the same neuron and pattern. The first summation goes over all P patterns in the training set, while the second summation goes over all O output neurons. The two statistical indicators, r and s, are used to compare the results of the ANN model with the MLR model from eq. (1–3). Another statistical index used in the evaluation of the ANN performances is the mean residual *mres*, defined as:

$$mres = \frac{\sum |\log 1/K_{m\,exp} - \log 1/K_{m\,ANN}|}{P}$$

While the numbers of neurons in the input and output layers are predetermined by the number of structural descriptors used in the model, 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 $\log 1/K_m$ values for the actinidin hydrolysis of phenyl hippurates.

The training was done by randomly presenting to the network the set of 29 phenyl hippurates. 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 $\log 1/K_m$ are presented in Table 2.

Table 2

Statistical results in estimating $\log 1/K_m$ with neural networks trained for 1000 cycles, with a variable number of hidden neurons

No. of hidden neurons	s	r	MSE·10²
1	0.189	0.877	5.212
2	0.169	0.903	4.218
3	0.169	0.903	4.189
4	0.168	0.903	4.169
5	0.172	0.898	4.385
6	0.169	0.903	4.194
7	0.168	0.903	4.174
8	0.168	0.904	4.169
9	0.169	0.903	4.192
10	0.178	0.891	4.631

It can be seen from Table 2 that a network with 4 neurons in the hidden layer gives the best performances. Because a greater number of hidden neurons does not improve the performance of the ANN, all neural networks used in this study are provided with 4 hidden neurons. There is an important advantage in adopting a network with a small number of hidden units: the network can generalize the input patterns better, and this results in superior predictive power. However, caution is needed because a network with insufficient hidden units will not be able to extract all the relevant correlation between physicochemical parameters and biological activity, and no reliable predictions may be obtained.

Feedforward neural networks are universal approximators which are capable of arbitrarily accurate approximation to arbitrary mappings, provided sufficiently many hidden units are available.[14–16] It is proved that any continuous function can be approximated on a compact set with the uniform topology, by a layered network with one hidden layer. However, this great modeling power of MLF ANN leads to some pitfalls that have to be considered when using neural networks. The most important problem with the use of ANN in QSAR is that if the hidden layer contains a great number of neurons, any physical, chemical, or biological property of a finite set of chemical compounds can be approximated using as input patterns random numbers and not significant structural descriptors.

Also, with a large number of hidden neurons there is a danger of overfitting, when the neural network simply memorizes the entire data set, and it will not be able to extract relevant correlation of the input patterns and give meaningful interpretation of other unknown examples.

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, we have retained the input patterns assigned to the corresponding compounds, and the set of real activity values in Table 1 were transformed in output patterns by randomly assigning an activity data to a compound not necessarily possessing this activity. A neural network with four hidden neurons was trained for 1000 cycles with ten such random training sets, and the statistical results are presented in Table 3. We found that the statistical indices in the runs using randomized activity data are considerably lower than those obtained when actual activity data are used (i.e., $r = 0.903$, $s = 0.168$, and MSE = 0.042). Even when only one hidden neuron is used, the statistical indices are better than in the tests with randomized data. Thus, we are confident that a valid relationship has indeed been established between the structural descriptors and the observed biological property.

Table 3

Statistical results in estimating log $1/K_m$ with neural networks trained for 1000 cycles, with randomly assigned output values (activity data)

s	r	MSE
0.299	0.647	0.129
0.353	0.438	0.176
0.392	0.038	0.217
0.338	0.507	0.161
0.305	0.629	0.131
0.357	0.415	0.180
0.376	0.283	0.202
0.325	0.560	0.150
0.351	0.446	0.174
0.353	0.439	0.176

Following the statistical results in Table 2, an ANN with 4 hidden neurons was chosen as the one that gave a good balance between fitting the training data and a small number of weights. The network (denoted by NN1) was trained for a period of 20000 cycles, when there was no further decrease in overall error. The MSE decreased to 0.029 and the correlation between the experimental $\log 1/K_{m\,exp}$ and $\log t1/K_{m\,NN1}$ is given by the equation:

$$\log 1/K_{m\,exp} = -0.070 + 1.033\ \log 1/K_{m\,NN1} \qquad (4)$$

$$n = 29 \quad r = 0.937 \quad s = 0.136 \quad mres = 0.100$$

The largest error is presented by the compound **26**, with $\log 1/K_{m\,NN1} = 3.20$. If we compare the predictions of the MLR model represented by eq. (1) with the predictions of the ANN model in eq. (4), it is clear that the neural network outperforms regression analysis and provides superior mapping of physicochemical parameters to biological activities.

In order to investigate the predictive character of the ANN model, we have used the leave-one-out (LOO) cross-validation method. In the LOO technique, an untrained network is first created. Then one pattern is taken out of the training set of patterns and the network is trained with the remaining patterns. When the learning process is finished, the network predicts the output value for the pattern which was eliminated from the learning set. The pattern is then put back in the set and the next one is taken out to

repeat the process, starting with the untrained network. In this manner, each pattern serves as an unknown once and as a training pattern all the other times. The LOO predictions of the network NN1 are not well correlated with the experimental values:

$$\log 1/K_{m\,exp} = 0.567 + 0.836\ \log 1/K_{m\,NN1\,LOO}$$
$$n = 29 \quad r = 0.768 \quad s = 0.251 \quad mres = 0.197 \tag{5}$$

The largest prediction error is obtained for the compound **26**, with log1/K = 2.99, and a residual equal to 0.61. The low statistical indices of the LOO prediction is due to the presence of an outlier, namely compound **26**, and to the known fact that neural networks give poor extrapolations. The extrapolation appears in the LOO procedure when the eliminated pattern is extremal in one or more variables.

Because compound **26** is an outlier in the ANN model, we have developed another neural network, denoted NN2, with a training set not containing this compound. After 20000 cycles there was no further improvement in the prediction of $\log 1/K_m$. The final MSE was 0.018, and the correlation between the experimental $\log 1/K_{m\,exp}$ and $\log 1/K_{m\,NN2}$ is given by the equation:

$$\log 1/K_{m\,exp} = -0.050 + 1.012\ \log 1/K_{m\,NN2}$$
$$n = 28 \quad r = 0.958 \quad s = 0.114 \quad mres = 0.080 \tag{6}$$

The values of $\log 1/K_{m\,NN2}$ estimated by the NN2 neural network are presented in Table 1. The largest difference between experimental and estimated log1/K equals 0.28 for compound **9**. The ANN model NN2 outperforms the MLR model with the same number of compounds presented in eq. (2). While in ref. 13 the compound **23** was considered an outlier, in the ANN model its activity is estimated with precision. Even if we compare the NN2 model with the MLR developed with **27** compounds, presented in eq. (3), it is clear that the neural network provides a better estimation of the experimental activity of actinidin in the hydrolysis of phenyl hippurates.

The $\log 1/K_{m\,NN2}$ for compound **26** predicted by the NN2 network is 3.06, instead of the experimental value of 3.60; there are two possible explanations for this situation: (a) the small number of 3,5-disubstituted phenyl hippurates in the training set which hinders the ANN to give a proper representation for this class of compounds; (b) the theoretical parameters used in correlation do not properly reflect the structural influence on the actinidin hydrolysis of this compound.

An intersting behaviour is the one of compound **23**: while in the MLR model this is an outlier,[13] the ANN model estimates very well its activity, with a low residual. A possible explanation is the nonlinear mapping and the interactions between structural descriptors considered in the neural network, which is not considered in the MLR model.

The network NN2 provides better predictions than the network NN1, as indicated by the results of the LOO cross-validation:

$$\log 1/K_{m\,exp} = 0.235 + 0.930\ \log 1/K_{m\,NN2\,LOO}$$
$$n = 28 \quad r = 0.879 \quad s = 0.189 \quad mres = 0.151 \tag{7}$$

The largest prediction error is obtained for the compound **23**, with $\log 1/K_{m\,NN2\,LOO} = 3.57$, and a residual equal to –0.35. The good predictive power of the network NN2 may be used to obtain an estimation of the activity for new, untested compounds.

A recent paper by Andrea and Kalayeh[11] cautions against the use of too many hidden units and describes the effect of too many units as a tendency for the network to "memorize" the data. The networks described in this report were characterised by a parameter, ρ, which is the ratio of the number of patterns in a set to the number of connections. The optimum ρ values for the MLF networks[11] 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 was investigated for classification ANN[17] and for continuous output MLF ANN[18] using random data as input and output patterns. As a general guideline it was proposed that ρ should exceed 3 to keep the chance correlation coefficient below 0.700.

For both networks NN1 and NN2, with the structure 3:4:1, the value of ρ is equal to 1.38, below the optimum ρ values.[11,18] However, correlation coefficients in the runs using randomized activity data presented in Table 3 are lower than 0.700, indicating that there is no danger for random correlations. 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.

CONCLUSIONS

In conclusion, it seems that the neural network approach gives both a useful and simple QSAR model for the computation and prediction of the biological activity, in our case the actinidin hydrolysis of substituted phenyl hippurates. If one considers the statistical parameters of the models, the neural network QSAR surpasses the usual multilinear regression analysis QSAR, providing accurate predictions of the activities of the compounds if the input patterns are represented by those molecular parameters which play an important role in determining biological activity.

The major difference between the two models, MLR and ANN, is that the first one uses a well defined function to fit the data, while the neural network performs a model-free mapping of the molecular structure descriptors to predict the molecular property under investigation.

An important goal of this investigation was 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, we propose to compare the statistical indices of the neural network model with the statistics of a number of ANN trained with a modified learning set: the input patterns (structural descriptors) are assigned to the corresponding compounds and input neurons, while the set of real property values is transformed in output patterns by randomly assigning a property data to a compound not necessarily possessing this property. If the statistical indices of the ANN model are better than in the tests with randomized data, one can be confident that a valid relationship has indeed been established between the structural descriptors and the experimental property. Otherwise, the number of hidden neurons must be lowered and/or the learning set must be expanded.

ACKNOWLEDGEMENT. We thank the Ministry of Research and Technology for financial support of this research under Grants 572B TA4 and TB11.

REFERENCES

[1] Previous Part: O. Ivanciuc, *Rev. Roum. Chim.*, **1996**, *41*, 645–652.
[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] O. Ivanciuc, *Rev. Roum. Chim.*, **1995**, *40*, 567–574.
[9] O. Ivanciuc, *Rev. Roum. Chim.*, **1995**, *40*, 1093–1101.
[10] O. Ivanciuc, J.-P. Rabine, D. Cabrol-Bass, A. Panaye and J. P. Doucet, *J. Chem. Inf. Comput. Sci.*, **1996**, *36*, 644-653.
[11] T. A. Andrea and H. Kalayeh, *J. Med. Chem.*, **1991**, *34*, 2824–2836.
[12] D. Zakarya, L. Farhaoui and S. Fkih-Tetouani, *Tetrahedron Lett.*, **1994**, *35*, 4985–4988.
[13] A. Carotti, C. Hansch, M.M. Mueller and J.M. Blaney, *J. Med. Chem.*, **1984**, *27*, 1401–1405.
[14] G. Cybenko, *Math. Control Signals Syst.*, **1989**, *2*, 303–314.
[15] K. Funahashi, *Neural Network,* **1989**, *2*, 183–192.
[16] K. Hornik, M. Stinchcombe and H. White, *Neural Networks,* **1989**, *2*, 359–366.
[17] D. T. Manallack and D.J. Livingstone, *Med. Chem. Res.*, **1992**, *2*, 181–190.
[18] D. J. Livingstone and D. T. Manallack, *J. Med. Chem.*, **1993**, *36*, 1295–1297.