Revue Roumaine de Chimie, 1996, *41* (7–8), 645–652

ARTIFICIAL NEURAL NETWORKS APPLICATIONS. PART 2[1]

USING THEORETICAL DESCRIPTORS OF MOLECULAR STRUCTURE IN QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIPS: ANALYSIS OF THE INHIBITION OF DIHYDROFOLATE REDUCTASE

OVIDIU IVANCIUC

Polytechnic University of Bucharest, Faculty of Chemistry, Department of Organic Chemistry, Splaiul Independenței 313, 77206 Bucharest, Roumania

Received February 11, 1994

Quantitative structure-activity relationships (QSAR), which relate biological and toxicological activities to structural features, have been employed to correlate structure to activity mainly by the use of multiple linear regression (MLR). A general problem of MLR models, namely the lack of nonlinear mapping, is resolved by the use of a new approach in computational chemistry: artificial neural networks (ANN). A comparison is made between the ability of MLR and ANN to predict the inhibitory potencies of substituted *s*-triazine derivatives on chicken liver dihydrofolate reductase; as chemical structure characteristics, three theoretical descriptors were used. Comparing the statistics of the two models, ANN performs better than MLR, providing accurate predictions of the biological activities of the *s*-triazine derivatives.

INTRODUCTION

In the study of quantitative structure-activity relationships (QSAR), mathematical models are constructed to describe the correlation between the biological activity of a molecule and its physicochemical characteristics. The purpose of QSAR is to understand the forces governing the activity of a particular class of compounds, and to assist drug design.

The Hansch approach to QSAR[2] describes the biological activity (A) in terms of a linear combination:

$$A = c_0 + \Sigma c_i P_i$$

where P_i are physicochemical parameters used to describe chemical structure, and the coefficients c_i are normally estimated by multiple linear regression (MLR) which minimises the variance between the data and the model.

In an extensive QSAR study, Hansch and coworkers[3] predicted the inhibition of chicken liver dihydrofolate reductase (DHFR) by a large set of 4,6-diamino-1,2-dihydro-2,2-dimethyl-1-(substituted-phenyl)-*s*-triazines (see Fig. 1 for the general structure). The QSAR developed was a MLR model using as structural descriptors

fects. Apart from the classical substituent parameters, there are many QSAR studies involving theoretical descriptors[4] (geometric and quantum indices); their major advantage is their easy computation using standard programs. Also, substituent parameters do not reflect the effect of the particular chemical structure under study, while quantum indices offer more sensitive and reliable molecular descriptors.

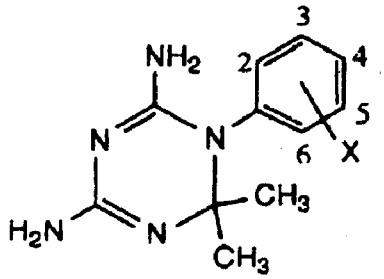

Fig. 1. – Structure of 4,6-diamino-1,2-dihydro-2,2-dimethyl-1-(substituted-phenyl)-*s*-triazines.

Recently, the QSAR reported in ref. 3 was examined with theoretical descriptors as independent variables:[5] the average molecular electrostatic field, F, the total molecular surface, S, and the saturated apolar molecular surface, S_{sa}. A good description of the pK_i was obtained, expressed in equation (1). The set of data contained 94 substituted *s*-triazine derivatives.

$$pK_i = 8.27 - 0.1222\ F_s - 0.0057\ S_{sa} + 0.0035\ S \qquad (1)$$

$$n = 94 \quad r = 0.872 \quad s = 0.35$$

The scope of the present study is to investigate the applicability of artificial neural networks (ANN) in estimating inhibitory potencies of substituted *s*-triazine derivatives on chicken liver dihydrofolate reductase, and to compare the results with the predictions of the MLR model of the same set of compounds, represented by equation (1).

Artificial neural networks (ANN)[6-8] are computer-modeled or algorithmic systems derived from a simplified concept of the brain. In a neural network, a number of nodes, called neurons, are interconnected into a net-like structure. A multi-layer perceptron (MLP) network is constructed with three or more layers of neurons: input neurons, output neurons, and one or more layers of intermediate elements called the hidden neurons.

The structure of the ANN, represented by the number of layers and the number of neurons in each layer, is adjusted so that it is appropriate for the problem under study. The network is then put through a training process in which the weights of the connections are modified recursively by a learning algorithm, based on a training set of known data, until the weights converge to fixed values. When the training process is finished successfully, the network can be used to solve new problems in a predictive process. Unlike an expert system, knowledge is represented in a neural networks in a parallel fashion in terms of weights of the connections between the neurons. The advantages of neural networks over conventional algorithms include the ability to generalize and their self-learning feature.

Recent reports have demonstrated the utility of artificial neural networks in the estimation of boiling points of organic compound,[9,10] in the correlation of the regiochemistry with structural parameters,[11,12] in the prediction of ^{13}C-NMR chemical

shifts,[13-15] IR spectra interpretation and identification of functional groups,[16,17] and QSAR.[18-21]

The major advantage in using ANN in QSAR is that with the presence of hidden layers, neural networks are able to perform nonlinear mapping of the physico-chemical parameters to the corresponding biological activity.

RESULTS AND DISCUSSION

In the present investigation we have used the set of 94 substituted *s*-triazine derivatives reported in ref. 5 to train an ANN to predict their inhibitory potencies on DHFR, with the scope to compare the ANN results with the predictions of the corresponding MLR model represented by eq. (1).

The ANN used in the present study are three-layer MLP networks, with three input units representing the three independent parameters in eq. (1), and one output unit (representing pK_i); for training we have used the backpropagation algorithm and the transfer function was the hyperbolic tangent. Other specifications for the networks used in simulations are presented in Table 1.

Table 1

Artificial Neural Network Specifications

Type of neural network	Multi-layer perceptron
Learning algorithm	Backpropagation
Bias neuron	Yes
Learning set presentation	Random
Learning rate	0.001 (NN1) / 0.005 (NN2)
Momentum	0.8
Transfer function	Tanh
No. input neurons	3
No. hidden neurons	variable
No. output neurons	1
Input scaling (min/max)	-0.9/0.9
Output scaling (min/max)	-0.9/0.9
Initial weights scaling (min/max)	-0.1/0.1

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 $pK_{i\,exp}$ and $pK_{i\,ANN}$ of the type $pK_{i\,exp} = a + b\,pK_{i\,ANN}$. MSE is defined by the following expression:

$$MSE = \frac{\Sigma_i \Sigma_j \left(Output_{ij} - Target_{ij}\right)^2}{P}$$

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. High-quality ANN predictions should have MSE and s close to zero, and r close to unity. Another statistical index used in the evaluation of ANN performances is ΔpK_{iav}, defined as $\Delta pK_{iav} = \Sigma \left| pK_{iANN} - pK_{i\,exp} \right| / P$.

While the numbers of neurons in the input and output layers are predetermined by the nature of experimental data, 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 pK_i of s-triazine derivatives.

The training was done by presenting to the network the set of 94 s-triazine derivatives, using as structural descriptors the three theoretical parameters involved in equation (1): the average molecular electrostatic field, F, the total molecular surface, S, and the saturated apolar molecular surface, S_{sa}. The target was the corresponding pK_i reported in ref. 3. The training set was presented randomly, each example being presented the same number of times. Eight ANN were generated, with the number of hidden neurons between 1 and 8. The training was terminated after 5000 complete cycles, and the results obtained in the evaluation of pK_i are presented in Table 2.

Table 2

Statistical Results in Estimating ρK_i with a Neural Network Trained for 5000 Cycles

No. of hidden neurons	MSE · 10^2	s	r	$\Delta\rho K_{i\,av}$
1	4.878	0.429	0.806	0.318
2	3.029	0.336	0.886	0.265
3	2.968	0.334	0.888	0.258
4	2.971	0.332	0.889	0.254
5	3.008	0.334	0.888	0.257
6	2.314	0.292	0.915	0.228
7	2.535	0.308	0.905	0.233
8	2.558	0.305	0.907	0.239

It can be seen from Table 2 that the networks give almost identical results in terms of MSE, r and s for all hidden layer sizes between 2 and 5 neurons; the performance of the ANN improves as the hidden layer size is equal to 6. A greater number of hidden neurons does not improve the performance of the ANN.

Following the statistical results in Table 2, an ANN with 2 hidden neurons was chosen as the one that gave a good balance between fitting the training data and a small number of adjustable parameters. The network (denoted by NN1) was trained for a period of 40000 cycles, when there was no further decrease in overall error. The MSE decreased to 0.029 and the correlation between the experimental pK_i and NN1 pK_i is given by the equation:

$$pK_{i\,exp} = 0.0443 + 0.9929\, pK_{i\,NN1} \qquad (2)$$
$$n = 94 \quad s = 0.332 \quad r = 0.889$$

which shows the good prediction obtained with the network NN1. The average difference ΔpK_{iav} is 0.26.

If we compare the predictions of the MLR model in equation (1) with the predictions of the ANN model in equation (2), it is clear that the neural network outperforms regression analysis and provides superior mapping of physicochemical parameters to biological activities.

For predictive purpose, an ANN with 6 hidden neurons was generated. This ANN, denoted NN2, was trained for 30000 cycles, when there was no further improvement in the prediction of pK_i. The final MSE was 0.016, and the correlation between $pK_{i\,exp}$ and the inhibitor potencies estimated by NN2, $pK_{i\,NN2}$ is given by the equation:

$$pK_{i\,exp} = 0.0699 + 0.9858\, pK_{i\,NN2} \qquad (3)$$
$$n = 94 \quad s = 0.243 \quad r = 0.942$$

The structure, physical parameters (F, S_{sa} and S), $pK_{i\,exp}$ and $pK_{i\,NN2}$ inhibitor potencies of 4,6-diamino-1,2-dihydro-2,2-dimethyl-1-(substituted-phenyl)-*s*-triazines inhibitors are presented in Table 3. The average difference between experimental and estimated pK_i is $\Delta pK_{iav} = 0.18$, while the maximum difference ΔpK_{imax} equals -0.69, for compound number 2

Table 3

Structure, physical parameters, experimental and calculated inhibitor potencies of 4,6-diamino-1,2-dihydro-2,2-dimethyl-1-(substituted-phenyl)-*s*-triazines inhibitors

No.	Substituent	F v/nm	S_{sa} A²	S A²	pK_i exp[a]	pK_i calc[b]
1	3–SO_2NH_2	27.9	104.9	275	5.00	4.95
2	4–SO_2NH_2	27.9	102.0	280	4.70	4.96
3	4–SO_2CH_3	23.0	132.3	296	5.25	5.09
4	3–$CONH_2$	26.4	98.0	254	5.07	4.99
5	4–$CONH_2$	26.4	102.1	262	4.95	4.96
6	3–$COCH_3$	19.6	136.1	268	5.56	5.74
7	4–$COCH_3$	19.6	133.6	267	5.69	5.75
8	3–OH	24.5	104.7	238	5.57	5.60
9	4–OH	24.5	105.0	238	5.70	5.61
10	3–CF_3	16.9	150.2	268	7.01	6.60
11	4–CF_3	16.9	157.2	263	6.77	6.70
12	4–NH_2	23.0	103.7	241	5.67	5.76
13	3–F	16.9	124.3	236	6.79	6.94
14	4–F	16.9	123.7	234	6.89	6.93
15	3–Cl	11.3	136.6	247	7.36	7.15
16	4–Cl	11.3	136.5	243	6.95	7.16
17	3,5–Cl_2	11.3	158.4	261	7.03	7.12
18	3–CH_3	11.6	138.0	247	7.08	7.10
19	4–CH_3	11.6	141.3	249	7.09	7.10
20	3–CH_2CH_3	11.6	156.4	274	7.00	7.06
21	4–$(CH_2)_3CH_3$	11.6	195.6	309	7.38	7.17
22	3,4–$(CH_2)_4$	11.6	179.2	283	7.72	7.03
23	3–$(CH_2)_5CH_3$	11.6	242.9	361	7.12	6.88
24	3–$(CH_2)_8CH_3$	11.6	310.6	429	6.53	6.51
25	4–$(CH_2)_8CH_3$	11.6	266.4	388	6.41	6.69

[a] The data were taken from ref.[3]

[b] Estimated by the network NN2

Table 2 (continued)

26	3-$(CH_2)_{11}CH_3$	11.6	335.7	453	6.38	6.34
27	3-$C(CH_3)_3$	11.6	177.3	286	6.75	7.04
28	4-$C(CH_3)_3$	11.6	181.3	292	6.71	7.05
29	3-OCH_3	17.2	139.5	265	6.41	6.59
30	4-OCH_3	17.2	140.6	267	6.48	6.52
31	3-OCH_3, 4-OCH_3	17.2	169.9	286	6.01	6.00
32	3-OCH_2CH_3	15.6	159.3	283	6.47	6.22
33	3-$O(CH_2)_2CH_3$	14.7	177.8	302	5.92	5.93
34	4-$O(CH_2)_2CH_3$	14.7	179.8	297	5.90	5.97
35	3-$O(CH_2)_3CH_3$	14.1	215.2	340	6.20	6.54
36	3-$O(CH_2)_4CH_3$	13.8	229.2	345	6.28	6.40
37	3-$O(CH_2)_5CH_3$	13.5	240.4	366	6.30	6.73
38	4-$O(CH_2)_5CH_3$	13.5	225.8	350	6.46	6.64
39	3-$O(CH_2)_8CH_3$	12.9	300.2	425	6.55	6.68
40	3-$O(CH_2)_{10}CH_3$	12.7	337.5	462	6.56	6.51
41	4-$O(CH_2)_{10}CH_3$	12.7	307.3	432	6.03	6.64
42	3-$O(CH_2)_{11}CH_3$	12.6	362.1	487	6.38	6.18
43	4-$O(CH_2)_{11}CH_3$	12.6	328.1	450	6.50	6.51
44	3-$O(CH_2)_{12}CH_3$	12.6	380.0	504	5.48	5.82
45	3-$O(CH_2)_{13}CH_3$	12.5	407.0	532	6.50	6.57
46	4-$O(CH_2)_2OC_6H_4$-4′-NH_2	13.0	166.3	367	6.76	7.40
47	3-$OCH_2C_6H_5$	10.1	159.0	340	6.93	7.23
48	4-$OCH_2C_6H_5$	10.1	145.1	340	7.53	7.28
49	3-$O(CH_2)_2OC_6H_5$	11.3	177.8	369	7.15	7.32
50	3-$O(CH_2)_2OC_6H_4$-3′-CH_3	10.7	210.8	394	7.02	7.16
51	3-$O(CH_2)_4OC_6H_5$	10.3	219.7	417	7.29	7.30
52	3-$O(CH_2)_4OC_6H_4$-3′-CF_3	10.8	260.6	445	7.54	7.23
53	3-$OCH_2C_6H_3$-3′,4′-Cl_2	9.0	198.5	372	6.78	7.13
54	4-$OCH_2C_6H_3$-3′,4′-Cl_2	9.0	190.9	372	7.14	7.13
55	3-$OCH_2C_6H_4$-4′-$CONH_2$	14.9	150.3	377	7.05	7.29
56	4-$OCH_2C_6H_4$-4′-$CONH_2$	14.9	136.9	373	7.30	7.29
57	4-$OCH_2C_6H_4$-4′-SO_2NH_2	16.5	140.1	398	7.49	7.24
58	4-$OCH_2C_6H_4$-4′-CH_2OH	12.1	169.4	367	7.35	7.41
59	3-CH_2O-*c*-C_6H_{11}	13.5	238.5	356	7.19	6.57
60	3-$CH_2NHC_6H_3$-3′,5′-$(CONH_2)_2$	17.4	137.8	401	6.98	7.23
61	3-$CH_2NHC_6H_4$-4′-SO_2NH_2	16.5	147.3	402	7.18	7.24
62	3-$CH_2OC_6H_5$	10.1	157.3	344	7.28	7.22
63	3-$CH_2OC_6H_4$-3′-Cl	9.5	180.2	359	7.18	7.15
64	3-$CH_2OC_6H_4$-3′-CN	13.1	161.3	366	7.59	7.40
65	3-$CH_2OC_6H_4$-3′-OCH_3	11.3	191.3	375	7.29	7.23
66	3-$CH_2OC_6H_4$-3′-CH_2OH	12.1	183.0	385	7.10	7.42
67	3-$CH_2OC_6H_4$-3′-CH_3	9.5	189.9	365	7.14	7.11
68	3-$CH_2OC_6H_4$-3′-CH_2CH_3	9.2	205.9	379	7.27	7.08
69	3-$CH_2OC_6H_4$-3′-$CH(CH_3)_2$	8.9	221.7	393	7.47	7.06
70	3-$CH_2OC_6H_4$-3′-$C(CH_3)_3$	8.6	236.9	404	7.24	7.06
71	3-$CH_2OC_6H_4$-3′-C_6H_5	9.2	177.0	386	6.79	7.30
72	3-$CH_2OC_6H_4$-3′-$NHCOCH_3$	13.4	183.6	376	7.64	7.36

Table 2 (continued)

73	3–$CH_2OC_6H_4$–3′–$NHCOCH_3$	15.7	143.3	364	7.46	7.27
74	3–$CH_2OC_6H_4$–3′–$NHCONH_2$	13.5	143.7	375	7.22	7.36
75	3–$CH_2OC_6H_4$–3′–$NHCSNH_2$	8.5	255.0	422	6.71	7.02
76	3–$CH_2OC_6H_4$–4′–$(CH_2)_4CH_3$	9.5	161.5	393	7.50	7.56
77	3–CH_2O–2–naphtyl	9.5	163.8	363	7.15	7.23
78	3–CH_2O–1–naphtyl	6.5	186.2	343	7.47	7.92
79	3–$CH_2SC_6H_5$	6.5	174.0	343	8.17	7.80
80	4–$CH_2SC_6H_5$	6.6	193.8	358	7.70	7.59
81	3–$CH_2SC_6H_4$–3′–CH_3	6.6	193.0	372	7.40	7.37
82	4–$CH_2SC_6H_4$–3′–CH_3	6.6	205.7	351	7.37	7.89
83	4–$CH_2SC_6H_4$–2′–CH_3	6.5	164.7	354	7.52	7.47
84	3–$SCH_2C_6H_5$	6.5	174.4	354	7.71	7.54
85	4–$SCH_2C_6H_5$	6.5	185.1	368	7.55	7.39
86	3–$SCH_2C_6H_4$–4′–Cl	6.5	189.2	373	7.13	7.36
87	4–$SCH_2C_6H_4$–4′–Cl	13.3	230.6	430	7.01	7.32
88	3–Cl, 4–$OCH_2C_6H_4CON(CH_3)_2$	22.9	113.8	302	5.66	5.67
89	3–SO_2NH_2, 4–Cl	15.9	142.2	276	6.50	6.35
90	3–NH_2, 4–CH_2CH_3	6.5	185.4	269	7.58	7.47
91	3–$CH_2SC_6H_5$, 4–Cl	16.0	188.5	351	7.40	7.23
92	3–Cl, 4–$SCH_2C_6H_5$	16.0	181.9	346	7.33	7.22
93	3–Cl, 4–$CH_2SC_6H_5$	30.0	313.9	432	6.46	6.47
94	3–Cl, 4–$O(CH_2)_8CH_3$	27.0	240.4	459	7.55	7.55
	3–Cl, 4–$(CH_2)_4C_6H_3$–2′–Cl, 4′–SO_2F					

The pK_i for compound 22 estimated by equation (1) is 6.83, instead of the experimental value of 7.72; a possible explanation for this situation is that the theoretical parameters used in correlation do not properly reflect the structure of the compounds with a bridge between positions 3 and 4.

The standard deviation of equation (2) is greater than in equation (3). This fact is a consequence of the fact that NN1 has only 11 connections (adjustable parameters), while NN2 has 31 connections. As a consequence of the good statistical indices of the equation (3), the network NN2 could be used to predict the pK_i for other *s*-triazines derivatives.

The L20%O (Leave-20%-Out of data) cross-validation was applied in order to test the prediction ability of the MLR and ANN models. The MLR model has the lowest predictive power, with r=0.727 and s=0.498, while for the network with two hidden neurons the predictive power increases (r=0.790 and s=0.445). The best L20%O predictions are offered by a network with six hidden neurons: r=0.821 and s=0.414.

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 ρK_i.

CONCLUSIONS

We can conclude from our results that ANN are able to train to predict the inhibitory potencies of substituted *s*-triazine derivatives on chicken liver dihydrofolate

reductase, when sufficient neurons are present in the hidden layer. When judged in statistical terms, the ANN results are superior to regression analysis, and they also provide accurate predictions of activities of the compounds when one uses as input data those molecular parameters which play an important role in determining biological activity.

ACKNOWLEDGEMENT. We thank the Ministry of Research and Technology for financial support of this research under Grant 469.

REFERENCES

1 Part 1: O. Ivanciuc, *Rev. Roum. Chim.*, **1995**, *40*, 567-574.
2 C. Hansch, P. P. Maloney, T. Fujita and R. M. Muir, *Nature*, **1962**, *194*, 178-180.
3 C. Hansch, B. A. Hathaway, Z. R. Guo, C. D. Selassie, S. W. Dietrich, J. M. Blaney, R. Langridge, K. W. Volz and B. T. Kaufman, *J. Med. Chem.*, **1984**, *27*, 129-143.
4 L. Y. Wilson and G. R. Famini, *J. Med. Chem.*, **1991**, *34*, 1668-1674.
5 T. Balogh and G. Náray-Szabó, *Croat. Chem. Acta*, **1993**, *66*, 129-140.
6 D. E. Rumelhart, G. E. Hinton and R. J. Williams, *Nature*. **1986**, *323*, 533-536.
7 P. D. Wasserman, "Neural Computing", Van Nostrand Reinhold, New York, 1989.
8 R. Hecht-Nielsen, "Neurocomputing", Addison-Wesley Publishing Company, Redwood City, CA, 1990.
9 L. M. Egolf and P.C. Jurs, *J. Chem. Inf. Comput. Sci.*, **1993**, *33*, 616-625.
10 H. Lohninger, *J. Chem. Inf. Comput. Sci.*, **1993**, *33*, 736-744.
11 D. W. Elrod, G. M. Maggiora and R. G. Trenary, *J. Chem. Inf. Comput. Sci.*, **1990**, *30*, 477-484.
12 D. W. Elrod, G. M. Maggiora and R. G. Trenary, *Tetrahedron Comput. Methodol.*, **1990**, *3*, 163-174.
13 V. Kvasnička, *J. Math. Chem.*, **1991**, *6*, 63-76.
14 V. Kvasnička, Š. Sklenák and J. Pospichal, *J. Chem. Inf. Comput. Sci.*, **1992**, *32*, 742-747.
15 V. Kvasnička, Š. Sklenák and J. Pospichal, *J. Mol. Struct. Theochem.*, **1992**, *277*, 87-107.
16 M. E. Munk, M. S. Madison and E. W. Robb, *Mikrochim. Acta*, **1991**, *2*, 505-514.
17 K. Tanabe, T. Tamura and H. Uesaka, *Appl. Spectrosc.*, **1992**, *46*, 807-810.
18 T. A. Andrea and H. Kalayeh, *J. Med. Chem.*, **1991**, *34*, 2824-2836.
19 T. Aoyama, Y. Suzuki and H. Ichikawa, *J. Med. Chem.*, **1990**, *33*, 905-908.
20 V. S. Rose, I. F. Croall and H. J. H. MacFie, *Quant. Struct.-Act. Relat.*, **1991**, *10*, 6-15.
21 Q. Liu, S. Hirano and I. Moriguchi, *Quant. Struct.-Act. Relat.*, **1992**, **11**, 135-141.