Applications of Neural Networks in Structure-Activity Relationships of a Small Number of Molecules

Igor V. Tetko, Alexander I. Luik,* and Gennadiy I. Poda

Institute of Bioorganic and Oil Chemistry, Murmanskaya, 1, Kiev-094, 253094, Ukraine

Received August 27, 1992

We investigated the applications of back propagation artificial neural networks (ANN) for a small dataset analysis in the field of structure-activity relationships. The derivatives of carboquinone were used as an example. It's been found that in this case the use of the same neural network results in unambiguous classification of new molecules. Predictions can be improved with statistical analysis of independent prognosis sets. We suggest that the sign criterion be used as a classification rule. We also compared neural networks with FALS and ALS in leave-one-out prediction. ANN applied to the same dataset has shown the same predictive ability as ALS but poorer than FALS.

Artificial neural networks (ANN) is one of recently emerged directions in the field of information processing technology. A simple conceptual model of brain application has been found useful in a lot of fields, where there is a need to solve different pattern recognition and object classification problems. ANN was used not only for classification of physiologically active substances[1] but also for solving the quantitative structure-activity relationship (QSAR)[2,3] problem (for other ANN applications see also review of Zupan and Gasteiger[4]). The authors reported that the results of classification and prediction by ANN were better than the results obtained by other methods such as discriminant analysis or method of k-nearest neighbors. In the case of QSAR, neural networks have enhanced predictive capabilities relative to multiple linear regression.

Shown in Figure 1 is a typical neural network. The neurons are designated as circles or nodes. The number of layers is arbitrary and networks generally have n layers (usually $n = 3$). The data are input to A, transformed on hidden layers, and output to B. Each input layer node corresponds to a single independent variable. Similarly, each output layer node corresponds to a different dependent variable. Each neuron value (O_j) ranging from 0 to 1 is calculated by eq 1, where O'_i are neurons values

$$O_j = 1/(1 + e^{-\alpha y_j}) \equiv f(y_j),\ y_j = \sum W_{ij}O'_i - \theta_j \qquad (1)$$

at the $n - 1$ layer; W_{ij} is the weight of the bond connecting i and j neurons; θ_j is a threshold value for neuron j; α is a parameter that expresses the nonlinearity of the neuron's operation. Usually α and θ are the same for all neurons in a layer. Neural network training is achieved by minimizing an error function E with respect to the bond weights W_{ij} until it's value

$$E = E(W_{ij}) = \sum(O_j - t_j)^2 \qquad (2)$$

becomes small enough (t_j is a training pattern). Back propagation algorithm[5] has been used. In this algorithm bond weights W_{ij} starting from random values are changed by a gradient descent method during the training process. Once the training is completed, then these weights are held fixed for the testing mode of network operation.

One of the main problems arising when using networks is determining the number of nodes in hidden layers. The number of nodes determines the number of adjustable network parameters. Too many nodes cause a network to "memorize" a dataset. Networks with few nodes may be insufficient to use all information from data and classify molecules. T. Andrea and H. Kalayeh[2] investigated neural network prediction as a function of the number of adjustable network parameters P.[6] They found that parameter ρ calculated by

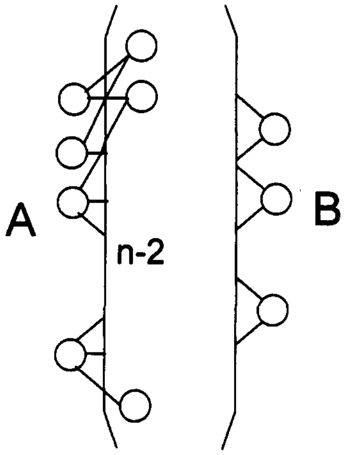


Figure 1. n-Layer neural network.

$$\rho = \frac{\text{no. of datapoints}}{P} \qquad (3)$$

has an optimal value for ANN in the range

$$1.8 < \rho < 2.2. \qquad (4)$$

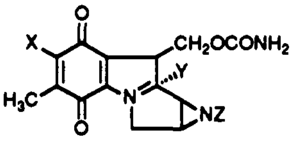
Models with $\rho > 2.2$ were unable to extract all the relevant features from a dataset and gave poor prediction. Models with $\rho < 1$ overfitted the trained set and were unable to predict accurately; $\rho > 4$ is considered optimal for linear regression models.[2] There are several cases when the number of available molecules is too small to achieve $\rho > 1.8$. This occurs when there are few molecules or many classes for classification. We investigated behavior of neural networks in this case.

We analyzed the well-suited data that had been already intensively investigated by ALS,[7] FALS,[8] and neural network[9] methods. Overfitting lowered the predictive capabilities of the models in ref 9, as shown below. The experimental data on derivatives of carboquinone and their activities were taken from literature[8,9] and the same compound numbers were used. Data are shown in Table I.

At first we examined neural network having three layers.[10] ANN parameters are shown in Table II, part A;

0022-2623/93/1836-0811$04.00/0

Table I. Structure, Descriptors, and Observed Activities in Mitomycins[a]



				part A							part B[b]		
no.	X	Y	Z	F_X	$\sigma_{m\text{-}x}$	$V_{w\text{-}x}$	Y_{OMe}	Y_{OH}	$E_{s\text{-}z}$	rank	σ^*_Y	B_1	rank
1	NH_2	OMe	H	0.02	−0.16	0.177	1	0	1.24	3+	1.81	1.00	3
2	NHEt	OMe	H	−0.11	−0.24	0.493	1	0	1.24	3+	1.81	1.00	3
3	NH_2	OMe	Me	0.02	−0.16	0.177	1	0	0	2+	1.81	1.52	3
4	NH_2	OMe	Et	0.02	−0.16	0.177	1	0	−0.07	2+	1.81	1.52	3
5	NH_2	OMe	Ac	0.02	−0.16	0.177	1	0	−0.07	2+	1.81	1.00	3
6	NH_2	OH	Me	0.02	−0.16	0.177	0	1	0	2+	1.55	1.52	3
7	NMe_2	OMe	H	0.10	−0.15	0.441	1	0	1.24	2+	1.81	1.00	3
8	NH_2	OMe	COPh-*o*-Cl	0.02	−0.16	0.177	1	0	−1.19	+	1.81	2.36	2
9	NH_2	OMe	COPh-*p*-Cl	0.02	−0.16	0.177	1	0	−1.19	+	1.81	2.36	2
10	NHPh	OMe	H	−0.02	−0.12	0.892	1	0	1.24	+	1.81	1.00	2
11	OMe	OMe	H	0.26	0.12	0.304	1	0	1.24	+	1.81	1.00	2
12	OMe	OMe	Me	0.26	0.12	0.304	1	0	0	+	1.81	1.52	2
13	OMe	OH	Me	0.26	0.12	0.304	0	1	0	±	1.55	1.52	1
14	NH_2	H	Me	0.02	−0.16	0.177	0	0	0	−	0.49	1.52	1
15	NH_2	OMe	SO_2Me	0.02	−0.16	0.177	1	0	−1.54	−	1.81	2.11	1
16	OMe	H	Me	0.26	0.12	0.304	0	0	0	−	0.49	1.52	1

[a] Me = CH_3, Et = C_2H_5, Ac = CH_3CO, Ph = C_6H_5. [b] Additional descriptors and class rank which were used only in the leave-one-out prediction.

Table II. Parameters of Neural Networks

	neurons				
layer	part A	part B	part C	α	θ
1	6	5	4		
2	2–16	3	3	2.5	0
3	5	5	3	5.0	0

Table III. Number of Misclassification in Dependence from a Number of Neurons in the Hidden Layer

	number of neurons in the hidden layer					
	1	2	3	4	5	6–16
number of misclassification	8	8	3	−	−	−

α and θ are the same as in ref 9. The parameters of molecules F_x, $\sigma_{m\text{-}x}$, $V_{w\text{-}x}$, Y_{OMe}, Y_{OH}, and $E_{s\text{-}z}$ were rescaled to take the values between 0.1 and 1 by the following equation

$$O_i = 0.9\frac{X_i - X_{i,\min}}{X_{i,\max} - X_{i,\min}} + 0.1 \qquad (5)$$

where $X_{i,\min}$ and $X_{i,\max}$ are the minimum and the maximum data. We used from 2 to 12 neurons in hidden layer. All of these ANN are overfitted, even for the smallest network parameter $\rho = 16/(6 \times 2 + 2 \times 5) < 1.8$.

Each class of five ranks of activity was represented by setting the corresponding element of output pattern to 1 and the remaining four elements to zero. For example, the rank + had the training pattern (0,0,1,0,0). Training was terminated when E became less than 0.01. We obtained an output vector as the result of classifications. Compounds were classified according to the element with the maximum value in the output vector. The output vector (0.01,0.20,0.69,0.85,0.11) shows that the analyzed molecule belongs to ± class.

The results of classification by ANN with different number of neurons in hidden layer are shown in Table III. Networks with four or more neurons in hidden layer classified all derivatives in complete accordance with observation.

We next examined predictive ability of neural network. According to ref 9, five molecules (**2**, **4**, **7**, **10**, and **16**) were removed from training patterns and networks trained on a diminished piece of data. We used networks with four neurons in hidden layer. These networks correctly classified all molecules from the reduced training set. We obtained however, when using random starting weights W^0_{ij}, different predictions for the removed compounds (see Table IV, part A). Molecules **2**, **4**, and **7** showed stable prognosis results but molecules **10** and **16** showed large dispersion of prediction. The same result we obtained using network with 12 neurons in hidden layer (see Table IV, part B).

We think if neural networks are overfitted ($\rho < 1$) it will be inadequate to use results of only one neural network prognosis to obtain correct network predictions. This occured in the quoted work,[9] where the authors drew conclusion on the result of only one prognosis. They used a network with 12 neurons in hidden layer. According to their work, molecules **2**, **4**, and **16** were predicted correctly while **7** and **10** were misclassified. Practically with the same probability they could have obtained results showing that molecule **16** was misclassified (see Table IV). To alleviate this shortcoming we propose using statistical analysis of network prognosis and the sign criterion.[11] For example, to determine molecular activity for one of two ranks of activity H_0 and H_1 and after n treatments, the molecule was predicted m times as having H_0 rank and $n - m$ ($m < n - m$) times as having H_1 rank. Then at the level p of significance

$$p < \sum_{r=0}^{m+1} \frac{m!}{(m-r)!r!} 2^{-n} \qquad (6)$$

molecule has the rank H_1.

Obtaining different ANN prognoses for the different trainings of the same network can be easily explained using an ANN with only two descriptors. While training neural network builds discriminative function that represents surface over the space of independent variable, the shape of the boundary between classes is variable when the

Table IV. Predicted Molecular Rating for Random Starting W_{ij}[a] Weight Matrix

no.	no. of starting random W_{ij} matrix										obsd rank
	1	2	3	4	5	6	7	8	9	10	
					Part A. Calculated Rank[b]						
2	3+	3+	3+	2+	3+	3+	3+	3+	3+	3+	3+
4	2+	2+	2+	2+	2+	2+	2+	2+	2+	2+	2+
7	3+	3+	2+	3+	3+	3+	3+	3+	3+	3+	2+
10	3+	2+	3+	3+	2+	3+	3+	3+	3+	2+	+
16	±	+	±	–	+	–	+	–	±	–	–
					Part B. Calculated Rank[c]						
2	3+	3+	3+	3+	3+	3+	3+	3+	3+	3+	3+
4	2+	2+	2+	2+	2+	2+	2+	2+	2+	2+	2+
7	3+	3+	3+	3+	3+	3+	2+	3+	3+	3+	2+
10	3+	2+	3+	3+	3+	3+	3+	3+	3+	2+	+
16	–	+	–	+	–	±	–	–	+	±	–

[a] Initial random weights of matrices were in the range 0–1. [b] Classification by ANN with four neurons in the hidden layer. [c] Classification by ANN with 12 neurons in the hidden layer.

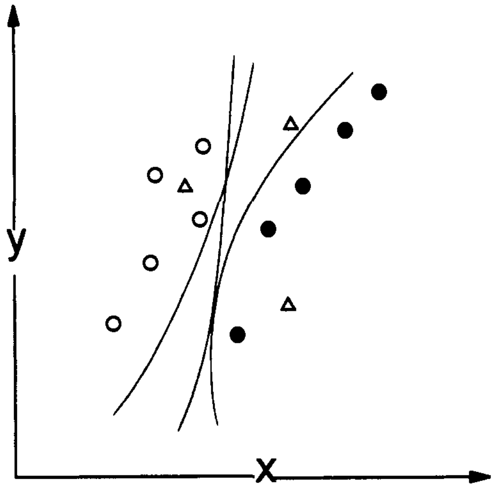


Figure 2. ANN classification of molecules, that have two activity ranks (○ and ●), and new molecules (△) rank prediction in the space of two variables (X,Y). The curves show various ANN-discriminating shapes.

number of molecules is small in comparison with adjustable parameters (see Figure 2). A molecule having values belonging to the boundary could be classified indefinitely. If the number of adjustable parameters is diminished, boundaries will be more stable and smoother. Conversely, increasing them leads to more complex discriminating shapes. Using the above criterion allows determination near which class the molecule is. This criterion can also be used to classify molecules for number of classes greater than two.

Another approach can be used to exclude uncertainty in prediction. We can exclude some parameters in input layer and diminish the number of neurons in hidden layer to obtain a network that will show stable prediction. However, some molecules from training set will not be classified correctly i.e. we will loose some piece of information. Therefore we used neural network with minimal number of neurons in hidden layer that classified correctly all the molecules from training set and evaluated the rank of new molecules by eq 6.

We used 50 random starting weight matrices to obtain a statistically significant prediction (see Table V). Molecules 2 and 4 were predicted correctly with a highly significant level of $p < 0.01$ and molecules 7 and 10 were misclassified with the same level. Molecule 16 was not classified to one of the above mentioned classes. But we can regard that molecule to belong to + or – rank with $p < 0.01$.

We tried to diminish the number of descriptors so that only the most significant of them be used for training and classification. We consequently excluded each input parameter from the datasets and analyzed how the diminished network was trained for all molecules. We used ANN with four neurons in the hidden layer. The network correctly classified all 16 molecules when we had excluded the V_{w-x} or σ_{m-x}. We could even diminish the number of neurons in hidden layer to three when we excluded V_{m-x} parameter without aggravation of training. Further excluding of descriptors or diminishing the number of neurons in hidden layer led to misclassification by ANN molecules from learning sets.

After that we examined the dataset with excluded V_{w-x} parameter because in this case we obtained the smallest network that correctly classified all molecules. The network structure and parameters are shown in Table II, part B. Table V, part B shows the results of prediction for 50 random starting weight matrices. In general the results are similar to aforesaid. Now molecule 16 has been classified as having – or ± rank of activity. Here parameter truncating and network diminishing haven't resulted in better predictive ability.

We also analyzed neural networks by leave-one-out prediction. The predictive ability was compared with ALS[7] and FALS[8] methods. To obtain clearer result of methods comparison we used the descriptor set described in refs 7 and 8. Compounds were classified only into three ranks of activity, using four descriptors: σ_{m-x}, V_{w-x}, and two new σ^*_y, B_1^z descriptors (see Table I, part B). Structure of neural network used is shown in Table II, part C. This is a network with a minimal number of neurons in hidden layer, which classified correctly all 16 molecules from training set into three ranks of activity. Here $\rho = 16/(4 \times 3 + 3 \times 3) < 1.8$. Table VI shows the result of classification by leave-one-out method. Four molecules (10, 11, 13, and 14) were misclassified. This method has the same predictive ability as the ALS method (where four molecules were erroneously predicted) but worse than FALS89 (only two molecules were misclassified). We used 10–30 calculation with random W_{ij}^0 to obtain significant results at the level of $p < 0.01$.

In summary, we showed that if parameter ρ defined by eq 3 is less than 1.8 the results of the ANN prediction must be validated by statistical method. We think that

Table V. Predicted Activities of Mitomycins Derivatives for 50 Starting Random Weight W_{ij} Matrix by ANN with four (part A) or three (part B) Neurons in the Hidden Layer[a]

	part A						part B[b]						
	calculated rank (no. of calc)						calculated rank (no. of calc)						
no.	3+	2+	+	±	–	pred	3+	2+	+	±	–	pred	obsd
2	48	2	0	0	0	3+	45	5	0	0	0	3+	3+
4	0	50	0	0	0	2+	0	50	0	0	0	2+	2+
7	43	6	0	0	1	3+	44	5	0	0	1	3+	2+
10	38	12	0	0	0	3+	40	10	0	0	0	3+	+
16	0	0	21	3	26	+ or –	0	0	9	17	24	± or –	–

[a] ANN parameters are shown in Table II, part A, and Table II, part B, correspondingly. [b] Parameter $V_{w\text{-}x}$ was excluded from input parameters set.

Table VI. Observed and Calculated Activities of Mitomycins Derivatives by Leave-One-Out Method[a]

compd	obsd	ANN	FALS	ALS
1	3	3	3	3
2	3	3	3	3
3	3	3	3	3
4	3	3	3	3
5	3	3	2	2
6	3	3	3	3
7	3	3	3	3
8	2	2	2	2
9	2	2	2	2
10	2	3	2	2
11	2	3	2	2
12	2	2	2	2
13	1	2	1	2
14	1	1	1	1
15	1	2	2	2
16	1	1	1	2

[a] Descriptors $\sigma_{m\text{-}x}$, $V_{w\text{-}x}$ (Table I, part A) and σ^*_y, B_1^z (Table I, part B) were used; ANN parameters are shown in Table II, part C.

the use of sign criterion will solve this problem. The comparison of ANN with ALS and FALS89 showed that ANN gave the same predictive ability as ALS but it's prognosis was worse in comparison with FALS89.

References

(1) Chastrette, M.; De Saint Laumer, J. Y. Structure–Odor Relationships Using Neural Networks. *Eur. J. Med. Chem.* 1991, *26*, 829–833.

(2) Andrea, T. A.; Kalayeh, H. Applications of Neural Networks in Quantitative Structure–Activity Relationships of Dihydrofolate Reductase Inhibitors. *J. Med. Chem.* **1991**, *34*, 2824–2836.

(3) Aoyama, T.; Suzuki, Y.; Ichikawa, H. Neural Networks Applied to Quantitative Structure-Activity Relationships Analysis. *J. Med. Chem.* 1990, *33*, 2583–2590.

(4) Zupan, J.; Gasteiger, J. Neural Networks: A New Method for Solving Chemical Problems or Just a Passing Phase? *Anal. Chim. Acta* 1991, *248*, 1–30.

(5) *Parallel Distributed Processing Exploration in Microstructure of Cognition*; Rumelhart, D. E., McClelland, J. L., Eds.; the MIT Press: Cambridge, MA, 1986; Vols. 1 and 2.

(6) This is a number of bonds in a neural network.

(7) Moriguchi, I. In *Structure-Activity Relationship-Quantitative Approaches*; Fujita, T., Ed.; Nankodo: Tokyo, Japan, 1986; Chapt. 9; pp 220–231.

(8) Moriguchi, I.; Hirono, S.; Liu, Q.; Matsushita, Y.; Nakagawa, T. Fuzzy Adaptive Least Squares and Its Use in Quantitative Structure–Activity Relationships. *Chem. Pharm. Bull.* **1990**, *38*, 3373–3379.

(9) Aoyama, T.; Suzuki, Y.; Ichikawa, H. Neural Networks Applied to Structure–Activity Relationships. *J. Med. Chem.* **1990**, *33*, 905–908.

(10) This method is programmed by the Borland C++ language for a personal computer (IBM PC/AT 80386).

(11) For example: Jonhos, N. N.; Leone, F. G. *Statistic and Experimental Design in Engineering and the Physical Sciences*; John Wiley & Sons: New York, 1977; Vol. 1, pp 314–364.