Self-Organization of Neural Networks with Active Neurons for Bioctivity of Chemical Compounds Forecasting by Analogues Complexing GMDH algorithm

Ivakhnenko, A.G., Kovalishyn V.V., Tetko I.V., Luik A.I.,
Ivakhnenko G.A., Ivakhnenko N.A.

Gai@gmdh.kiev.ua http://come.to/GMDH

Poster for the ICANN'99 Conference.

INTRODUCTION

The rational search of the compounds with desired biological activity includes 1) description of molecular structure and 2) determination of relationships between calculated parameters and the investigated activity. In the last years interest to 3D methods of investigation of connection between a structure and activity (CoMFA[1], Electron-topological method ETM, etc.) has essentially increased.

Artificial neural networks (ANNs) have became one of the leading methods in this field.[2] It was shown that ANN allow to construct more adequate relations between a structure and activity of compounds in comparison with traditional methods, such as the multiple regression analysis, linear discriminant analysis, etc. However, there are some difficulties (such as limitation in speed, local minima, overfitting/overtraining problems, etc.) with an application of these methods for analysis of data sets with a large number of input parameters and, particularly, three-dimensional electronic parameters of compounds generated by 3D QSAR approaches, such as CoMFA.

The current study analyses a new method, i.e. neural networks with active neurons,[3,4] that could be used in such studies.

Analogues Complexing GMDH algorithm

Analogues Complexing (AC) algorithm is one of the spectrum of Group Method of Data Handling algorithms[5]. It is used for forecasting, extrapolation and pattern recognition of ill-defined objects and time series analysis.

It was shown that in the case of insufficient a priori information, not very accurate measurements, noisy and short data sample, this algorithm can derive non-physical models of the analyzed process that are superior to other analyzed methods.

Almost all objects for recognition and control in economics, ecology, biology and medicine are undeterministic or fuzzy. Thus this algorithm is potentially powerful method that can be used in such studies.

The algorithm consists of the next two steps:

1. Selection of the analogues.

The input information is presented as sample of initial data, that describes analyzed chemical compounds. Each molecule A_0 corresponds to one characteristic point in the space of variables. As measure of proximity the Euclidean distance between the characteristic points is used. The point A_1 nearest to the analyzed point is called the first analog. The second nearest point is called the second analog A_2 and so on up to the last analogue A_N.

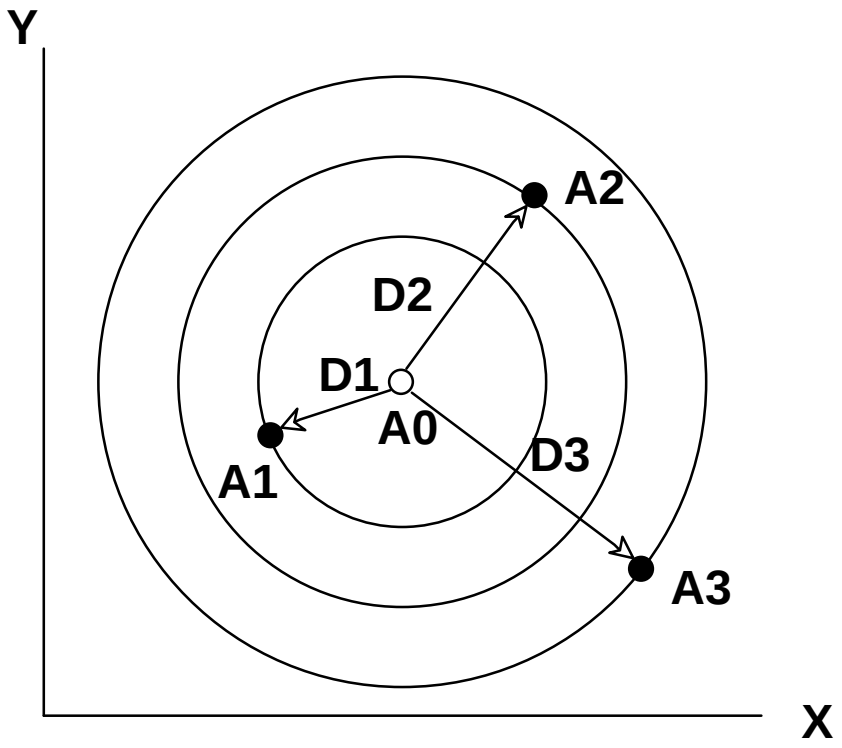

Figure 1. Calculation of three analogues in two-dimensional space.

2. Combining of forecasts

The forecast A_{0F} of a molecule A_0 is be defined as

$$A_{0F} = \sum_{i=1}^{M} w_i * A_i \qquad (1)$$

where w_i is weight of analogue i given by $w_i = d_i/D$, d_i is the distance from a characteristic point of investigated molecule to the i analogue, and D is sum of d_i distances; M is a number of complexing analogues. The M = 1-3 were analyzed in our studies.

Neural Network with Active Neurons

The neural network with active neurons were used. The active neurons are defined as algorithms that are able during the learning (it is also known as self-organizing of active neuron) to select inputs necessary to minimize the given objective function of the neuron.

Number of active neurons in each layer is equal to number of variables given in initial data sampling. Each active neuron predicts its own variable (either activity of molecules or input variable).

The output variables of previous layers are used as additional inputs for the neurons of next layer. Each layer of active neurons acts similar to Kalman filter: output set of variables repeated the input set but with filtration of noise.

The computing modules are united in a multi layered structure with the purpose to increase the algorithm accuracy by a more complete processing of the input information from lower to higher order lever. The optimal layer (“the best layer”) is determined for each analyzed variables. The variables with higher level of noise have the higher numbers of the corresponding “best layers”.

Architecture of a neural network with active neurons

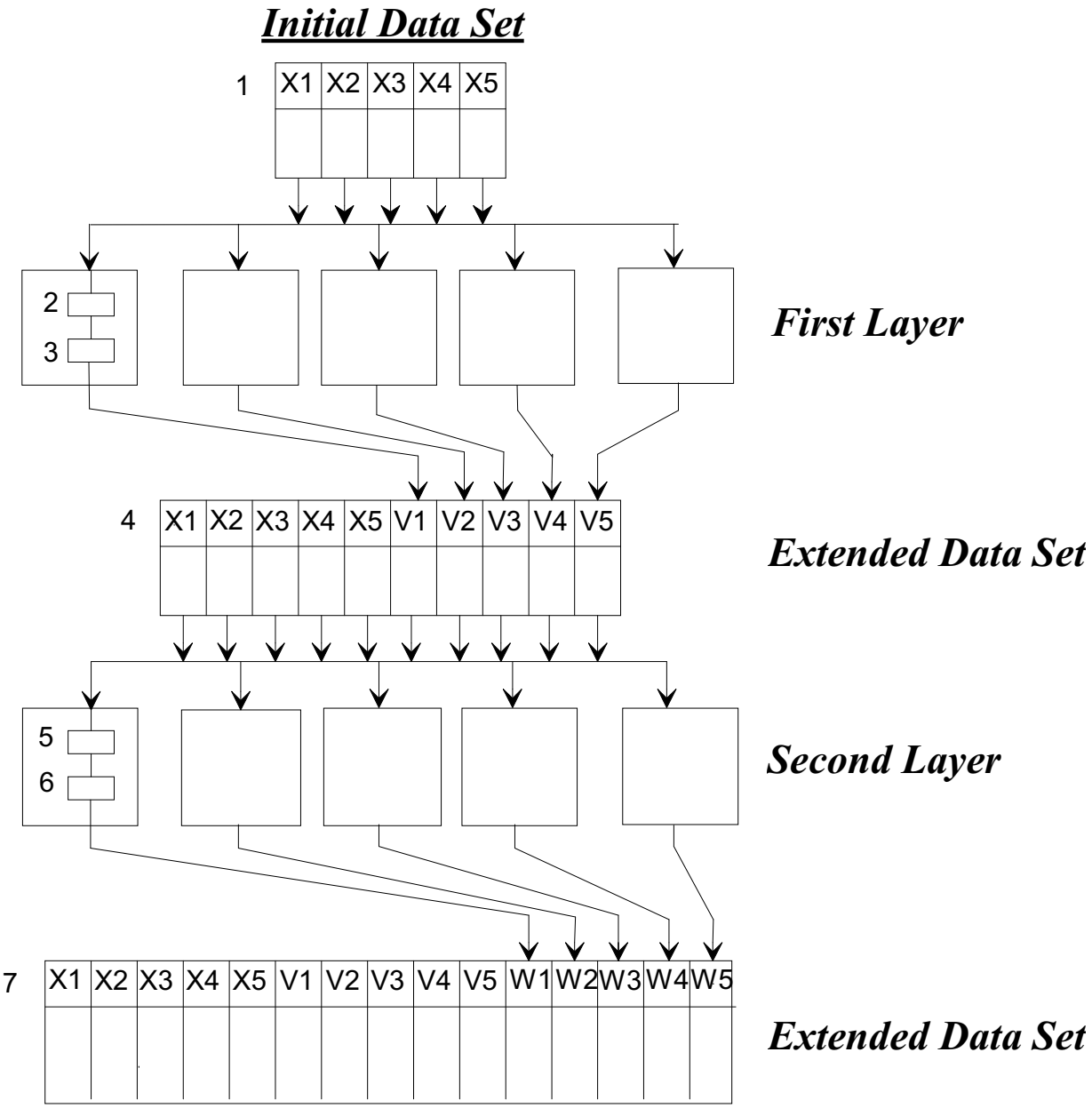

Figure 2. 1 - initial experimental data. **2+3** the active neuron of the first layer. It consists of block of the convergence and selection of an optimal set of variables (**2**) and the forecasting block (**3**). 4- the input for the second layer extended by outputs of the first layer. **5+6** the active neuron of the second layer. 7 the input for the third layer extended by outputs of the first two layer.

Divergence and convergence of input data variables

The number of input variables increases from layer to layer. The variables, which are not effective, should be excluded from the set of input variables analyzed at the given layer. This could significantly speed-up the algorithm. All variables - candidates are checked for their forecasting ability of analyzed activity (e.g., by analogs complexing) and only a certain number of the most effective variables is included into set of variables selected for each layer of the neuronet.

Let us note that in back-propagation neural networks the set of input variables is optimized only once for the first layer. In ANN with active neurons the set of variables is optimized at each layer. This could improve prediction accuracy of this method compared to other approaches.

The regularity criterion was used for selection of an optimal set of variables:

$$CR = \sqrt{\frac{\sum_1^N \left(A_i - A_{if}\right)^2}{\sum_1^N \left(A_i - \overline{A}\right)^2}} \rightarrow \min, \qquad (2)$$

where A_i - activity of molecule i, A_{if} - predictions of molecules activity, $\overline{A}$ -average value of activity of all molecules, N - number of investigated molecules.

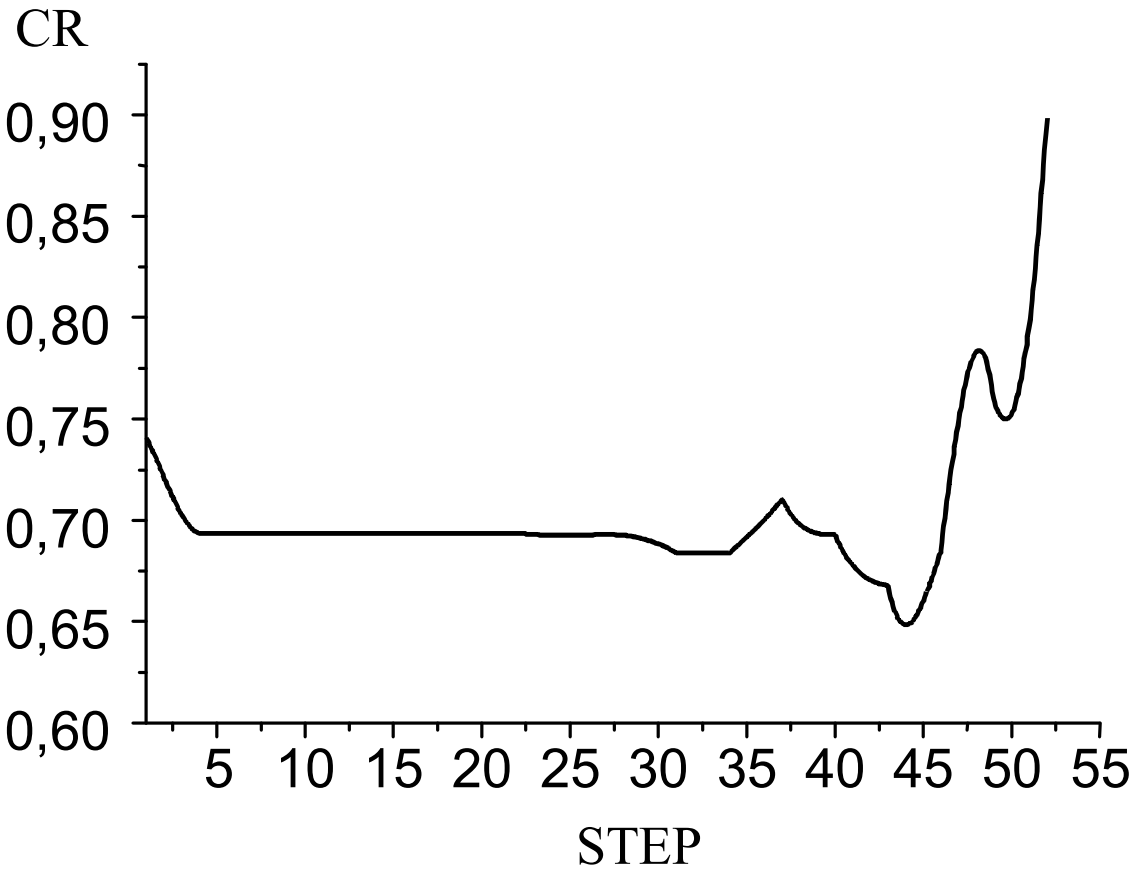

Figure 3. Procedure of an optimization of a set of input variables for antimycin analogues for the first layer of neural networks with active neurons. Criterion regularity (CR) displayed as a function of the optimization step. On each step one variable was excluded from the input set and the CR was analyzed. Nine variables providing the minimum CR were selected for further training of the first layer of the neural network. Let us note that number of active neurons in the network is 10 (9 variables preselected at this step + analyzed activity).

Detection of the optimal set of variables

The most comprehensive method for detection of the optimal set of variables simply analyze all possible combinations of the N variables and the subset that minimizes the CR is selected. However, such method requires analysis of $N!$ combinations of variables and cannot be applied for N > 10-12. As an alternative way the step-wise algorithms can be used.

The first method (*growing algorithm*) selects the variables by successive addition to the optimized set (that is empty at the beginning of optimization) of a new variable (selected from the initial set) decreasing the criterion. Only one variable is added per step. The optimization is terminated if addition of any new variable does not decrease the criterion.

The second method (*decreasing algorithm*) starts with initial set of variables and eliminates one inefficient variable at each step of optimization. The inefficient variable is determined as variable which elimination minimizes the CR. The optimization is terminated if excluding of any new variable increases CR (Fig. 3).

The set with minimum CR detected by any one of the two algorithms was used.

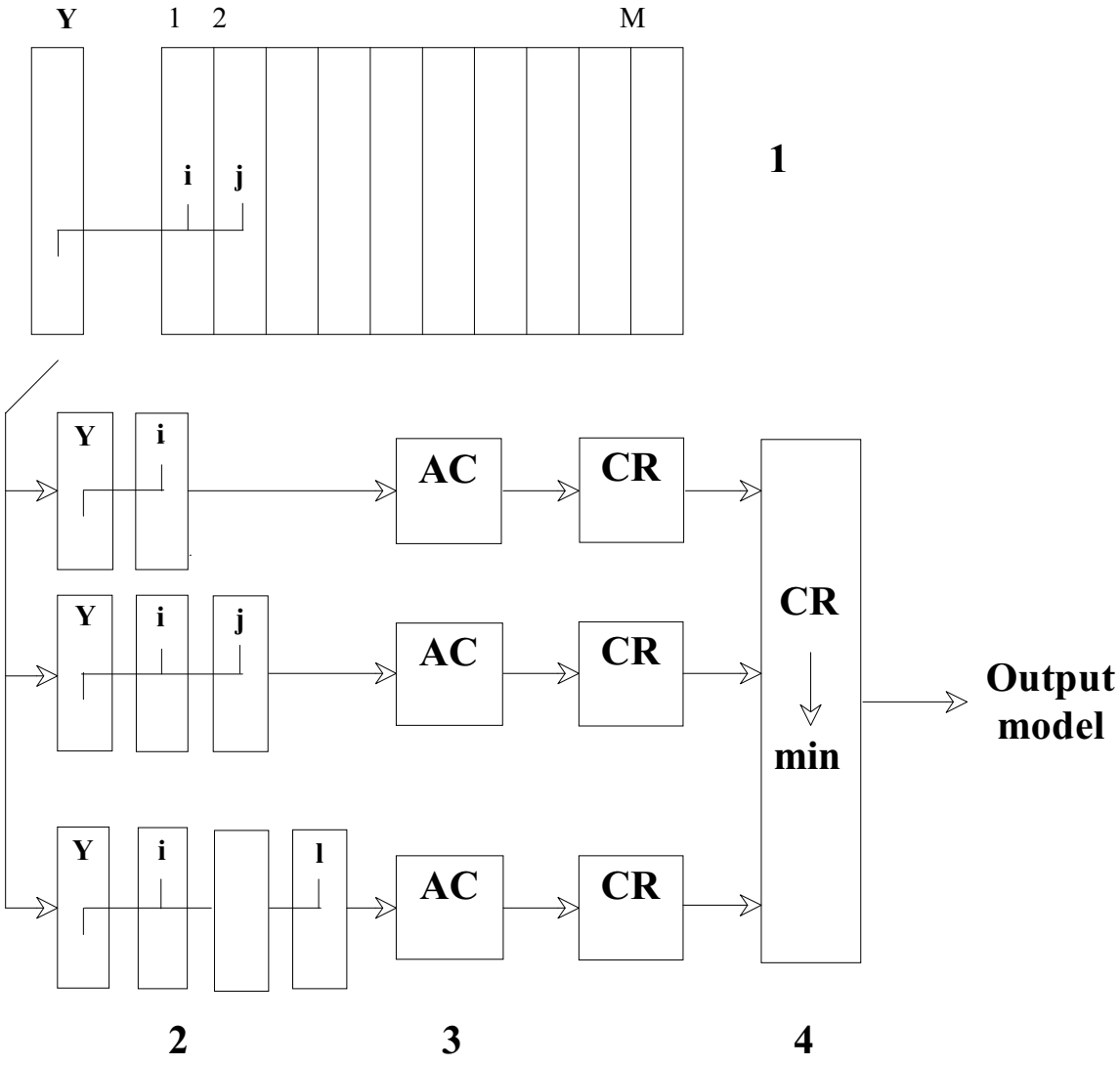

Figure 4. Procedure of an optimization of the initial variable set for the increasing algorithm. 1 - initial data; 2 - variable sorting block; 3 - analogues complexing algorithm; 4 - selection of the optimal subset.

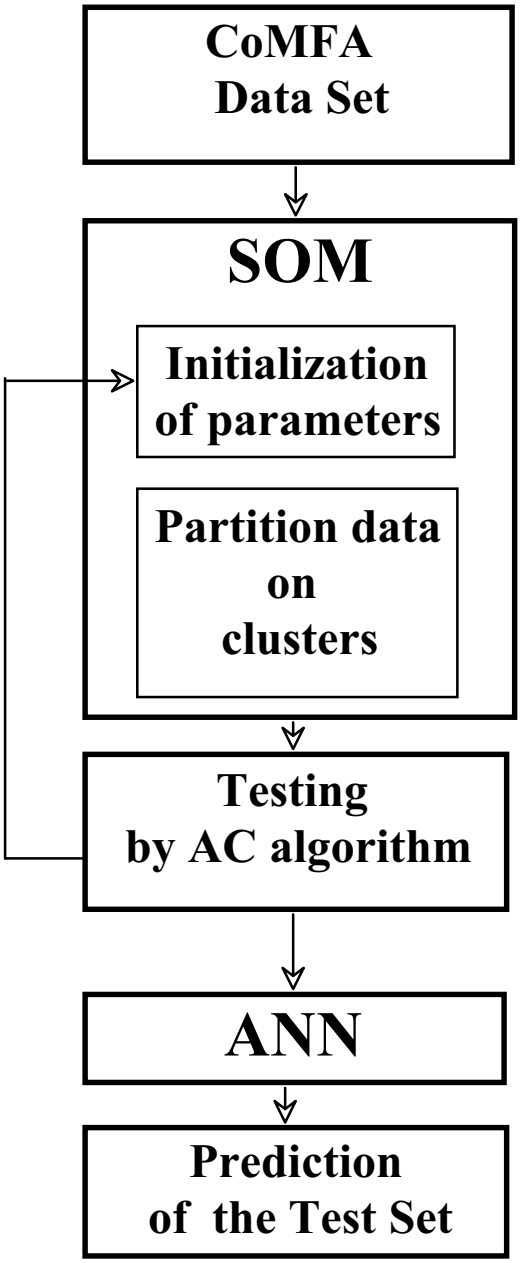

Figure 5. The process of investigation of the CoMFA data set included three stages. **1**) The Kohonen's Self-Organizing Map (SOM) were used to find clusters in the input data. **2**) The centers of clusters were used as inputs for the AC algorithm and an optimal clusterization of input space was detected. **3**) The optimized centers of clusters were fitted to neural networks.

QSAR DATASETS

Three data sets were analyzed:

1) 53 antimycin analogues with antifilarial activity;[6]

2) 35 monosubstituted benzenes with charge-transfer properties;[7]

3) the CoMFA dataset including 82 benzylpiperidine derivatives with AchE inhibitory activity[8].

Results

Table 1. The leave-one-out results calculated for QSAR examples.

	total	ANN with active neurons			back-propagation ANN	
data set	params	no[1]	First layer	best layer	all params	pruned params
Antimycin analogues	53	6	0.74[2] (0.5)[3]	0.91 (0.81)	0.66 (0.43)	0.91 (0.67)
Benzenes	31	2	0.74 (0.51)	0.95 (0.89)	0.89 (0.78)	0.97 (0.95)
Benzylpipe-ridines	188 (28224)[4]	4	0.86 (0.71)	0.89 (0.78)	0.56 (0.55)	0.72 (0.73)

[1]cardinal number of the layer (the best layer) with the lowest error of the network; [2]correlation coefficient R; [3]cross-validated q^2; [4]number of CoMFA parameters before preprocessing with SOM.

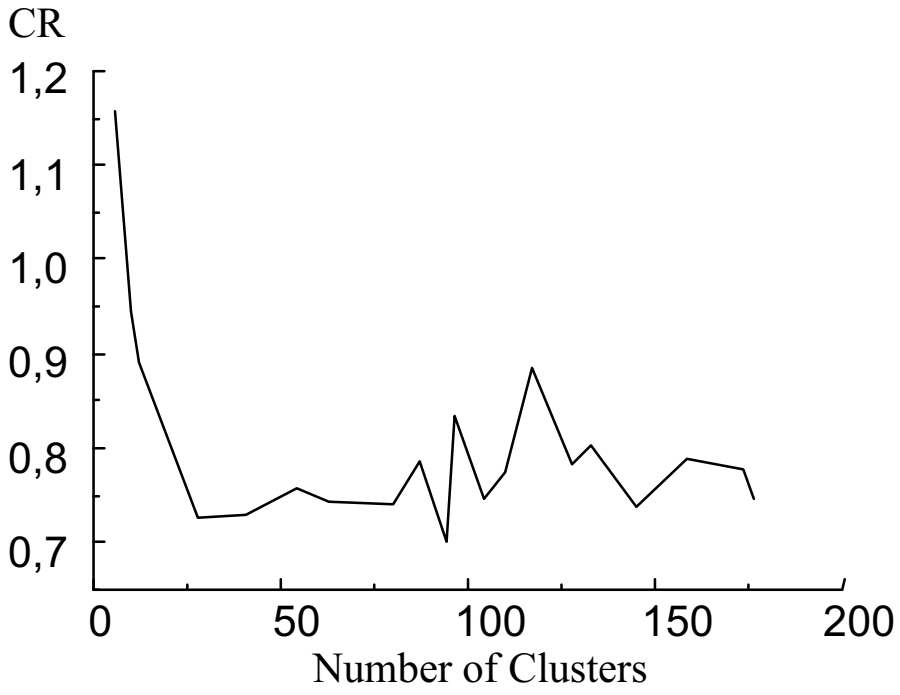

Figure 6. Procedure of detection optimal amount of a clusters for CoMFA data set. Criterion regularity (CR) displayed as a function of the amount of clusters. The number of clusters provided the minimum CR was selected for training of neural networks.

Table 2. The number of clusters determined with SOM and value of criterion regularity determined by AC algorithm.

Size	Clusters	CR	Size	Clusters	CR
6 (2x3)	6	1.16	99 (9x11)	96	0.83
10 (2x5)	10	0.94	108(8x12)	104	0.75
12 (2x6)	12	0.89	117(9x13)	110	0.78
28 (2x14)	28	0.73	126(9x14)	117	0.88
42 (3x14)	41	0.73	132(11x12)	128	0.78
56 (4x14)	54	0.76	143(11x13)	133	0.80
65 (5x13)	63	0.74	156(12x13)	145	0.74
84 (6x14)	80	0.74	168(12x14)	158	0.79
90 (9x10)	87	0.79	182(13x14)	173	0.78
96 (8x12)	94	**0.70**	195(13x15)	176	0.75

Table 3. The CR values calculated for different layers of the neural network trained with the AC algorithm

	Variable									
no	4	6	12	13	19	20	31	33	41	act.
1	0.66	0.30	0.63	1.00	0.68	0.69	0.27	0.93	0.93	0.63
2	0.42	0.25	0.68	0.96	0.59	0.48	0.21	0.90	0.93	0.60
3	**0.29**[1]	0.25	0.52	0.91	0.50	0.48	0.21	0.90	0.91	0.56
4	0.32	0.21	0.50	0.95	0.57	0.60	0.21	0.90	0.91	0.56
5	0.29	0.26	0.36	0.87	0.47	0.56	0.21	0.90	0.90	0.53
6	0.29	0.25	0.35	0.87	0.45	**0.45**	0.21	**0.89**	0.89	**0.41**[2]
7	0.27	0.24	0.35	0.87	0.45	0.45	0.21	0.89	**0.79**	0.47
8	0.31	0.30	**0.34**	0.87	0.45	0.46	0.21	0.89	0.78	0.52
9	0.31	**0.19**	0.34	**0.75**	0.42	0.46	**0.20**	0.89	0.78	0.48
10	0.31	0.23	0.47	0.87	**0.42**	0.45	0.21	0.89	0.80	0.48

[1]in bold are shown the minimum CR (minimum error) calculated for the corresponding variable;
[2]the layer with minimum CR for activity corresponds to “the best” layer detected for the analyzed data set.

Table 4. Variables optimized using neural networks trained with AC algorithm and back-propagation algorithm

data set	neural network with active neurons	BNN
Antimycines	4,6,12,13,19,20,31,33,41	4,6,11,13,14,35,50,52
Benzenes	2-5,11,13,18,20,21,26,30	3-6,11,14
Benzylpiperidines	17,36,41,55,61,89,117,118121,124,139,145,155,159,167-170,172,177,181,183	22,28,29,32,40,57,75, 81-83,91,106,108,120,125, 159,182,183

Table 5. The leave-one-out results for the optimized sets

	active neurons		R, (q^2)	
dataset	no	R, (q^2)	BNN	MLR
Antimycines I[a]	6	0.91 (0.81)	0.62(0.37)	0.45 (-0.5)
Antimycines II	7	0.87(0.72)	0.91 (0.67)	0.82 (0.66)
Benzenes I	5	0.93(0.87)	0.97 (0.95)	0.96 (0.92)
Benzenes II	5	0.93(0.87)	0.97 (0.95)	0.96 (0.92)
Benzylpiperidines I	4	0.89 (0.78)	0.64(0.41)	0.60 (0.1)
Benzylpiperidines II	2	0.76(0.54)	0.76 (0.76)	0.41 (-3.3)

[1]I, II refer to sets optimized by neural networks with active neurons (I) and by BNN (II).

CONCLUSION

The calculated result show that the ANN with active neurons can be successfully used in QSAR studies for prediction activity of new compound. The results for CoMFA data set demonstrate that number of parameters to be fitted to neural network can be reduced by considering the values in the centers of clusters estimated with SOM algorithm.

The results for antimycins and benzylpiperidines by neural networks with active neurons were better than results of back-propagation neural networks, while the opposite was true for benzenes. It is interesting that the statistical coefficients calculated for data set of benzenes were the highest amid analyzed data. The AC algorithm was mainly developed for fuzzy dataset with a high level of noise and, probably its application for benzenez is not optimal. The further studies are required to provide a more objective comparison of the methods.

Acknowledgments

This study was partially supported by INTAS-Ukraine grant 95-0060. The authors thank Prof. Jacques R. Chretien and Dr. Philippe Bernard (University of Orleans) for providing us the CoMFA data.

References

1. R. D. III. Cramer, D. E. Patterson, J. D. J. Bunce, Comparative Molecular Field Analysis (COMFA).1. Effect of shape on binding of steroids to carrier proteins *J. Am. Chem. Soc.* 110: 5959- 5967 (1998)

2. D.J. Livingstone, D. T. Manallack, and I. V. Tetko, Data Modelling with Neural Networks - Advantages and limitations, *J. Comp. Aid. Mol. Design*. 11:135-142 (1997).

3. A.G. Ivakhnenko, G.A. Ivakhnenko, and J.-A. Muller, Self-organization of neural networks with active neurons, *Pattern Recognition and Image Analysis* 2:185-196 (1994).

4. A.G. Ivakhnenko, V.V. Kovalishyn, I.V. Tetko, A.I. Luik, G.A. Ivakhnenko, and N.A. Ivakhnenko, Application of self-organizing neural networks with active neurons for prediction of bioactivity of chemical compounds by the analogues search algorithm, *Problems of control and information* in press.

5. H.R. Madala, A.G. Ivakhnenko. *Inductive Learning Algorithms for Complex Systems Modeling*, CRC Press Inc., Boca Raton (1994).

6. V.V. Kovalishyn, I.V. Tetko, A.I. Luik, V.V. Kholodovych, A.E.P. Villa, and D.J. Livingstone, Neural network studies. 3. Variable selection in the cascade-correlation learning architecture, *J. Chem. Inf. Comput. Sci*. 38:651-659 (1998).

7. Selwood, D.L.; Livingstone, D.J.; Comley, J.C.W.; O'Dowd, A.B.; Hudson, A.T.; Jackson, P.; Jandu, K. S.; Rose, V. S.; Stables, J. N. Structure-Activity Relationships of Antifilarial Antimycin Analogues: A Multivariate Pattern Recognition Study. *J. Med. Chem*. *33*:136-142 (1990).

8. P. Bernard, D.B. Kireev, J.R. Cretien, P-L. Fortier, and L. Coppet, Automated docking of 82 N-benzylpiperidine derivatives to mouse acetylcholinesterase and comparative molecular field analysis with ìnaturalî aligment, *J. Comp. Aid. Mol. Design*. in press (1998).