Bioorganic & Medicinal Chemistry Letters, Vol.3, No.2, pp. 329-332, 1993
Printed in Great Britain

0960-894X/93 $6.00 + .00

Quantitative Structure - Activity Relationship Studies of Aromatic and Heteroaromatic Nitro Compounds Using Neural Network

Nanda Ghoshal*, Sudhindra N. Mukhopadhyay+,

Tapan K. Ghoshal+ and Basudeb Achari

Medicinal Chemistry Division, Indian Institute of Chemical Biology, Calcutta - 700032, India.

+ Electrical Engineering Department, Jadavpur University, Calcutta - 700032, India.

(*Received in USA* 30 *September* 1992)

Abstract: A back propagation type neural net was applied for correlating the mutagenic activity of a dataset of 197 compounds with energy of the LUMO and hydrophobicity. The network system with 4-4-1 configuration produced good duplication of the observed activities (SD=0.789, r=0.919) in the training cycle excluding nine outliers and also showed good prediction ability (SD=1.049, r=0.853).

Recently considerable interest has been generated in applying artificial neural network (ANN) approach for quantitative structure-activity relationship (QSAR) studies[1-6]. Here we report the results of a QSAR study for correlating the mutagenic activity of diverse classes of aromatic and heteroaromatic nitro compounds with energy of lowest unoccupied molecular orbital ($\mathcal{E}_{LUMO}$) and hydrophobicity by applying back propagation (BP) type ANN.

A dataset of 197 compounds (their observed activity and physicochemical parameters) has been taken from published results[7]. The four input variables are $\mathcal{E}_{LUMO}$, hydrophobicity and two indicator variables I_1 and I_a. The value of I_1 is unity for compounds containing three or more fused rings and zero for compounds with two or less rings. The value of I_a is unity for five examples of acenthrylenes and zero for the rest. A simple three layer NN configuration (4-4-1) was used. For all four input cases four processing elements (neurons) have been used in the hidden layer since use of 1 or 2 neurons led to inferior fit. More complex configurations, viz. 4-10-1 and 4-8-4-1 (with two hidden layers) have also been tried earlier[8]. Although these required lesser number of iterations for the same quality of fit, they were better avoided bearing in mind the consequences of overfitting with large numbers of neurons[6].

The BP program of McClelland and Rumelhart[9] was used as basic module for exploration. A learning rate of 0.05 and a momentum term of 0.09 were used. The programs were run on a PC/AT 486 machine. A total of 20,000 iterations were needed for reaching the results reported; increasing the number of iterations to 40,000 resulted in marginal improvement only.

Several sets of training were carried out for a systematic comparison with regression analysis[7]. The results are summarised in Table 1. In the first set 188 compounds were used as in regression leaving out 9 outliers (compounds 189-197). It appears that NN provides an improved fit with better correlation coefficient and marked reduction in SD. When the outliers were used in training (set II) the fitment became comparable but marginally poorer (the results of these 9 outliers are given in Table 2). However, by excluding those outliers in the prediction cycle contributions of the first 188 compounds to SD and r were found to be 0.819 and 0.912 respectively. This demonstrates the outlier filtering ability of ANN configuration used.

In set III only 2 inputs ,viz., ΣLUMO and hydrophobicity were used for training. The fitment becomes markedly poorer underlining the significance of indicator variables[8,10]. So at this stage it would be premature to treat the ANN approach as a substitute[6] for the one using genuine indicator variables.

For testing the prediction ability of ANN a number of test sets were chosen by random selection of 50 data and using the remaining 147 data for training. A typical case is shown in set IV (table 1) in which 2 outliers are included in the test set. While the training results are comparable to those of set II, values of 1.049 for SD and 0.853 for r in the test set showed the good prediction ability of ANN.

In the original regression fit by Hansch et al[7] the 9-nitroanthracene type structures (Table II) and some heterocyclic compounds (Table III) were poorly predicted. When trained neural nets (sets I and II) were used for the analysis, the results shown in Table 2 were obtained. It may be observed that both the neural networks gave consistent results. However, these results are mostly far out from experimental value though they are better in

some cases. The poor prediction ability suggests that either these should be treated as separate groups or, if these groups are to be considered together, additional descriptor indicator variables must be used.

References:

1. Aoyama, T.; Suzuki, Y.; Ichikawa, H. *J. Med. Chem.* **1990**, *33*, 905-908.
2. Aoyama, T.; Suzuki, Y.; Ichikawa, H. *J. Med. Chem.* **1990**, *33*, 2583-2590.
3. Aoyama, T.; Ichikawa, H. *Chem. Pharm. Bull.* **1991**, *39*, 358-366.
4. Aoyama, T.; Ichikawa, H. *Chem. Pharm. Bull.* **1991**, *39*, 372-378.
5. Aoyama, T.; Ichikawa, H. *Chem. Pharm. Bull.* **1991**, *39*, 1222-1228.
6. Andrea, T. A.; Kalayeh, H. *J. Med. Chem.* **1991**, *34*, 2824-2836.
7. Debnath, A.K.; Lopez de Compadre, R.L.; Debnath, G.; Shusterman, A.J.; Hansch, C. *J. Med. Chem.* **1991**, *34*, 786-797.
8. Ghoshal, N.; Mukhopadhyay, S.N.; Ghoshal, T.K. *Technical Report from Centre for Knowledge Based Systems*, E.E. Department, Jadavpur University, Calcutta, 1991.
9. McClelland, J.L.; Rumelhart, D.E. *Explorations in Parallel Distributed Processing*, MIT Press, Cambridge, 1988.
10. Silipo, C.; Hansch, C. *J. Am. Chem. Soc.* 1975, 97, 6849-6861.

Table 1. Results of Regression Analysis and Training by ANN.

Set No.	I	II	III	IV	Regression[a]
No. of data	188	197	197	147	188
NN config.	4-4-1	4-4-1	2-3-1	4-4-1	-
SD	0.789	0.959	1.210	0.958	0.886
r	0.919	0.877	0.795	0.877	0.900

SD is defined as $\sqrt{\text{Sum of squared error/data points.}}$

(a) Results from equation 2 of reference 7.

Table 2. Prediction of Activity by ANN and Regression Analysis

Compd. No.[b]	Obsd.[c]	ANN predicted By setI	ANN predicted By setII	Predicted by regression[d]
I-189	-0.70	1.7471	0.9868	1.50
I-190	0.57	-1.3122	-1.1876	-1.68
I-191	0.77	3.7478	3.7975	3.40
I-192	-0.22	2.7009	2.0959	2.41
I-193	-0.22	2.8234	2.2236	2.48
I-194	0.63	-1.4403	-1.1436	-2.40
I-195	-2.94	-0.1149	-0.3917	0.18
I-196	-2.00	0.8884	0.5206	1.49
I 197	2.54	-1.2843	-0.9969	-1.35
II-1	1.64	2.6206	2.1665	2.39
II-2	0.30	0.7681	0.5821	1.35
II-15	-0.52	3.2710	2.7333	2.82
II-17	-0.95	3.4573	3.2736	3.01
II-19	a	0.8548	0.8874	0.54
II-20	a	0.9129	0.9442	0.62
II-21	a	0.3929	0.7418	1.07
II-22	a	1.3735	1.1862	1.63
II-23	a	1.5642	1.3998	1.87
II-25	a	3.4469	3.2628	3.00
II-26	a	3.5736	3.2249	3.10
III-1	0.64	-0.8821	-0.7221	-1.04
III-2	1.97	-0.7685	-0.6797	-0.97
III-3	1.02	-1.2808	-1.0111	-2.05
III-4	1.03	-1.0095	-0.9989	-2.18
III-5	2.59	-1.1912	-0.8997	-1.43
III-6	2.24	-1.1989	-0.9983	-2.11

(b) The Roman numerals I,II,III stand for table nos. of Reference 7 & the Arabic numerals for compound nos. in respective Table.

(c) As reported in reference 7, a = inactive.

(d) Results from reference 7.