Bioorganic & Medicinal Chemistry Letters, Vol.3, No.4, pp. 645-651, 1993
Printed in Great Britain

0960-894X/93 $6.00 + .00

THE USE OF NEURAL NETWORKS FOR VARIABLE SELECTION IN QSAR[1]

James H. Wikel*, and Ernst R. Dow
Lilly Research Laboratories, Eli Lilly and Company, Indianapolis, Indiana 46285

(*Received in USA* 11 *December* 1992)

Abstract.
The application of a back-propagation neural network has been found to be an efficient and effective tool to identify pertinent variables for QSAR studies.

Quantitative Structure Activity Relationship (QSAR) studies are attempts to derive mathematical models relating the biological activity of a series of compounds to one or more properties of the molecules. These properties, or descriptors, may be derived from numerous sources including empirically derived properties such as refractive index, octanol/water partition coefficient or spectral data. Alternatively these properties may be of theoretical derivation obtained from computational programs. Programs such as MOPAC provide quantum chemical descriptors as partial atomic charges, heats of formation, and energies of the highest occupied molecular orbital (E_{HOMO}).[2] A program is also available for the calculation of the octanol/water partition coefficient. [3] Extensive lists of substituent parameters describing electronic (s), lipophilic (p), and steric properties (mr) are also available. [4] The early stage of a QSAR study requires the collection of many of these descriptors for each structure into a dataset from which the chemist attempts to derive a model. Hansch et. al. pioneered research in this area of QSAR demonstrating the use of regression analysis to derive a model.[5] In the intervening years other methods have been developed and applied.[6] One recurrent problem in all of the studies is the fact that a dataset contains more descriptors than compounds, that is, more columns of parameters than rows of compounds. This poses two insidious problems. The first problem is that of observing a chance correlation described by Topliss and Edwards. [7] Secondly, many of the descriptors are correlated and thus redundant.

Neural networks are part of a new era of evolving computer technology in which a computer system has been designed to learn from data in a manner emulating the learning pattern in the brain.[8] Neural networks are typically used when the problem is not understood well enough to write a procedural program or expert system and there are a large number of observations. Using neural networks, the solution to the problem is sought as follows (Figure 1): 1) an answer is calculated by multiplying each input by the connection weight to each hidden unit; 2) the products are summed at each hidden unit where a non-linear transfer function is applied; 3) the output of each hidden unit is then multiplied by the connection weight from the hidden unit to the output unit where it is summed and interpreted. The neural network "learns" by repeatedly passing through the data and adjusting its connection weights to minimize the error; in this case, the predicted versus the actual biological activity. A neural network is thus a mathematical model to describe a non-linear hypersurface. The increasing interest and availability of neural network software has prompted several groups to apply this technology in QSAR studies. Aoyama reported the application of neural networks as a substitute for discriminant analysis.[9] Aoyama et. al. and subsequently Andrea et. al. applied neural networks in QSAR in a manner similar to multiple regression analysis.[10,11] Livingstone et. al. reported the use of neural networks as a tool to provide 2D visualization of n-dimensional property space.[12] Recently Livingstone described the importance of the number of hidden units in the neural network on the predictive value of the derived model.[13] We report here our initial results in the use of neural networks to identify descriptors most relevant to the biological activity. After the neural network identified the more important descriptors, a classical regression analysis was used to verify their importance.

Published data were used for this study. Dataset 1 reported by Dunn et. al. consisted of 13 compounds and 5 descriptors. [14] We added 58 additional descriptors to this dataset in order to compound the problem for the neural network (Table 1).[15] The second dataset was reported by Selwood et. al. consisting of 31 compounds and 53 descriptors (Table 1). [16] These two datasets were selected to represent a QSAR study in the early phase (13 compounds) and in a more developed phase (31 compounds) of progress. Since it is not possible to represent para- or meta-substituted compounds appearing in the dataset from Dunn as words to the network, a binary representation was used: para 0 1, meta 1 0. Missing values for individual atom descriptors due to differing number of atoms were assigned a value of zero. In a typical neural network application, the dataset is divided into two sets. One group, the larger of the two, would be used to train the network while the smaller subset would be used to evaluate the predictive power of the network. QSAR dataset are typically small in the early stages of the project and thus it becomes impractical to reduce them. We used a cross-validation technique to train the network. The datasets were sorted on activity and a single exemplar was removed for training purposes. In the Dunn dataset, every third observation was removed to provide 4 training sets each containing 12 compounds. In the Selwood dataset, every fourth observation was removed to provide 7 training sets each containing 30 compounds. In each case the observation that was removed served as the test case for the network. If the datasets were not sorted before this techniques was applied, the predictive power of the resulting neural network was compromised. Inputs from the training sets were randomized for presentation to the network.

All networks were of the back-propagation type and trained on a Sun SPARCstation 2 using the program NeuralWorks. [17] The networks were trained to predict activity. A hyperbolic tangent transfer function was used with a learning rate coefficient of 0.3 and a momentum term of 0.4. All inputs were normalized between -1 and +1. Ten hidden units and a bias unit were used for all the networks. Hidden units of 0, 2, 5, and 10-5 were included but the best generalization was obtained with 10 hidden units but relatively few training epochs. If we trained the networks to convergence (the global minima), the generalization ability was extremely poor. Therefore, one of the "local minimas" was the best solution. The networks were very consistent in picking the same descriptors as being important even though they were in different local minima. The Dunn and Selwood datasets were trained for 38 and 91 epochs respectively requiring less than 15 seconds of CPU time.

After the networks have been trained and those with the best generalization ability selected, the values of the hidden unit weights were extracted and visualized (Figure 2).[18,19] In these plots, the horizontal axis represents the 10 hidden units, the vertical axis is labeled for each of the descriptors, and the plotted value is the last output of the hidden unit with color coding. The descriptor label h_W is the value of the connection weight between the hidden unit and the output unit (Figure 1). Each vertical bar represents a different color palette reflecting sensitivity for the numerical value of the hidden unit. Positive coefficients are represented in yellow, negative coefficients are in green.

In Figure 2, the color maps indicate the magnitude of the hidden weight value with the color intensity, i.e., a black value is near zero. The color map on the right of each pair of color maps use a discontinuous cutoff with only the largest weights in color. The color map on the left of each pair reflect color in a decreasing sensitivity level for the weight values. In this manner, the 63 descriptors in the Dunn dataset (Figure 2a) was reduced to a approximately 21 more pertinent ones labeled in red. The largest weight values were identified visually with six descriptors. In the Selwood dataset (Figure 2b), the 53 descriptors were reduced to approximately 24 with nine of them associated with the largest weights labeled in red. In the Selwood dataset two descriptors, LOGP and ATCH4, stand out as important in the color map on the left side.

Neural networks have the potential to overfit, or memorize, the data. For the Dunn dataset there are (63 inputs + 1 bias) * 10 hidden units + 10 output weights for 650 adjustable parameters. However nearly all of these are close to zero and therefore do not contribute. However, if we had trained to convergence, nearly all of these parameters would have a large value and then overfit the data. It is imperative in this method that the network is not overtrained.

Multiple regression analysis was then applied to the dataset using the most important descriptors identified by the neural network.[20] For the Dunn dataset, sigma and the Swain-Lupton F descriptor were identified by the neural net and reported by Dunn. The neural net also identified, as part of the 21 most pertinet descriptiors, the electrotopological descriptor S(3) unavailable to Dunn but included in our expansion of the dataset descriptors. A plot of Activity vs. S(3) indicates a distinct difference between the meta- and para-substituted series (Plot 1). Equations 1 and 2 were developed indicating a positive correlation of the biological activity with the electronic environment of atom 3. Atom 3 is the aromatic carbon at the meta position bearing the substituent and atom 5 would be the comparable atom always bearing the hydrogen substituent. A significant regression equation could not be developed for S(5). The neural network placed more significance on the substituent bearing position suggesting the importance of the local electronic environment. Equation 3 was developed for the Selwood dataset and three of the descriptors identified by the neural network. Not surprisingly LOGP and ATCH4 were found in this model.

$$\text{Activity (meta isomers)} = 2.98(\pm 0.07) + 0.25(\pm 0.05)\ S(3) \qquad \text{(Equation 1)}$$

$$n=7 \quad F=20.74 \quad r=0.90 \quad p=0.006$$

$$\text{Activity (para isomers)} = -0.28(\pm 1.04) + 1.83(\pm 0.55)\ S(3) \qquad \text{(Equation 2)}$$

$$n=6 \quad F=11.26 \quad r=0.86 \quad p=0.028$$

$$\text{(Equation 3)}$$

$$-\log(IC50) = 4.41(\pm 1.29)\ \text{ATCH4} + 0.23(\pm 0.07)\ \text{LOGP} + 0.01(\pm 0.00)\ \text{MOFI\_X} - 1.64(\pm 0.52)$$

$$n=31 \quad F=13.50 \quad r=0.77 \quad p=0.000$$

$$F=11.69\ (\text{ATCH4}),\ 22.80\ (\text{LOGP}),\ 11.68\ (\text{MOFI\_X})$$

We have shown the application of neural networks in QSAR to aid in the identification of important parameters related to biological activity. This has been demonstrated using datasets with relatively few compounds, 13 and 31 compounds with a larger number of descriptors, 63 and 53 descriptors respectively. The important variables were quickly discerned by visualizing the hidden unit weights. These were then confirmed using standard regression analysis. The neural network was able to identify multi-variate relationships that the original authors did not reveal but were validated with standard multiple regression analysis. This is not surprising since in QSAR studies it is generally not possible to identify a single model that is omnipotent. In the final analysis the models attempt to explain the observed data and serve as a basis on which to design further experiments. This technique of determining which variables are important may be applicable in any problem in which there are non-trivial relationships between the input and the output variables. The use of neural networks in this manner has facilitated the rapid identification of important variables in a QSAR study.

Table 1. List of QSAR Descriptors

Dataset 1) Sterimol terms[a]:L1, B1, B2, B3, B4; substituent parameters[b]: s, p, F, R, MR; Topological indices [c]: S(1-19),T(1-19),$^{(1-4)}c$,$^{(1-4)}c^v$,c^{4PC}, ^1K, ^2K, ^3K, ^0K_a, ^1K_a, ^2K_a, ^3K_a; indicator variables for meta-position and para-position.

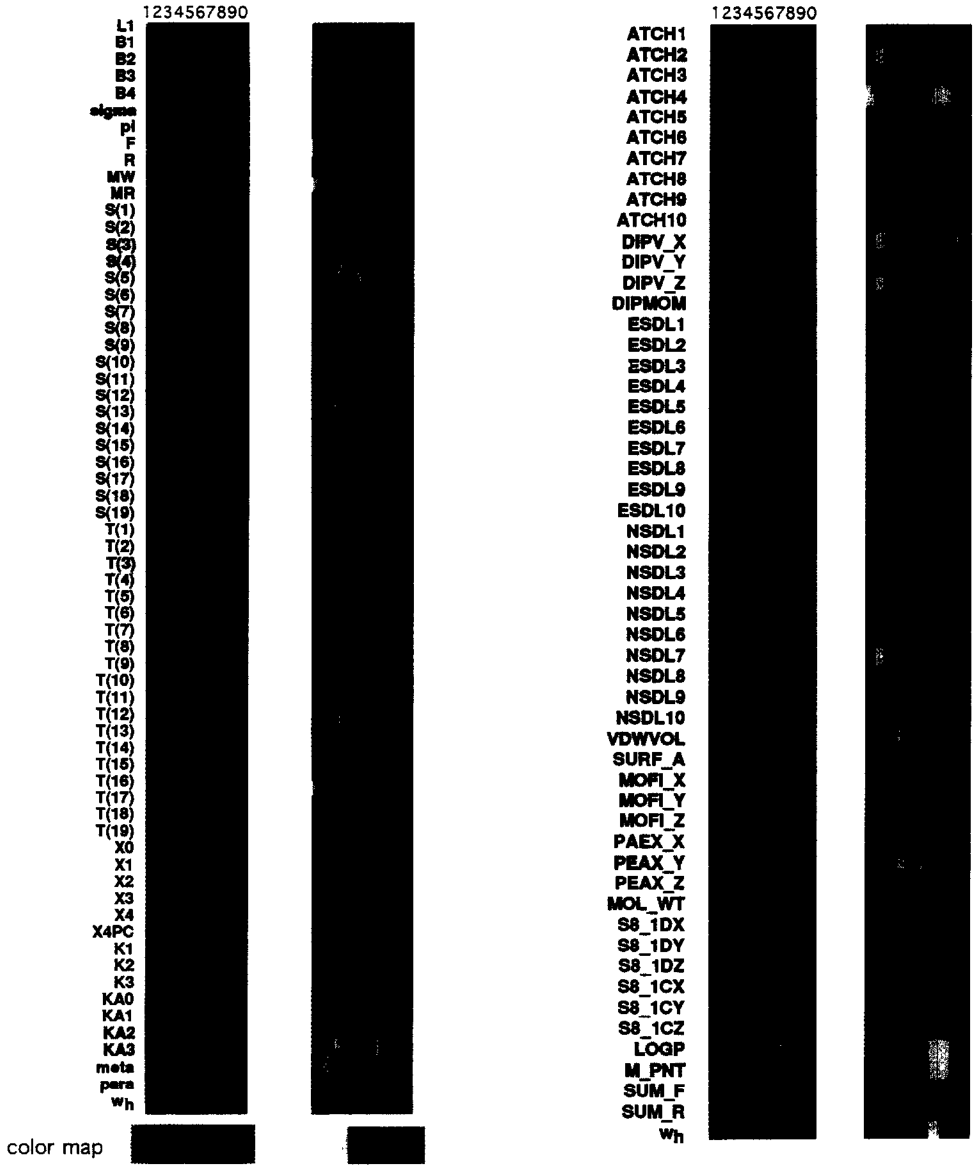
Dataset 2)[d] ATCH 1-10, DIPV_X, DIPV_Y, DIPV_Z, DIPMOM, ESDL 1-10, NSDL 1-10, VDWVOL, SURF_A, MOFI_X, MOFI_Y, MOFI_Z, PAEX_X, PAEX_Y, PAEX_Z, MOL_WT, S8_1DX, S8_1DY, S8_1DZ, S8_1CX, S8_1CY, S8_1CZ, LOGP, M_PNT, SUM_F,SUM_R

a) Verloop, A.; Hoogenstraaten, W.; Tipker, J. *Drug Design*; Ariens, E. J., Ed.; Academic Press: New York, 1976; Vol 7, pp.165-207. b) Hansch, C.; Leo, A. *Substituents Constants for Correlation Analysis in Chemistry and Biology*. John Wiley & Sons, New York, 1979. c) MOLCONN-X program, Hall Associates Consulting, 2 Davis Street, Quincy, MA 02170

References and Notes

1. Presented at the National ACS Meeting Symposium on Neural Networks in Chemistry, April 1992 San Francisco CA
2. Stewart, J. J. P., MOPAC, QCPE Bull. No 455, 1983, **3**, 43
3. Pomona College Medchem Software
4. Hansch, C.; Leo, A. Substituents Constants for Correlation Analysis in Chemistry and Biology. Wiley, New York, 1979.
5. Hansch, C. ; Fujita, T. *J. Amer. Chem. Soc.* **1964**, 86, 1616.
6. Martin, Y. C., *Medicinal Research*; Grunewald, G., Ed.; Marcel Dekker: New York, 1978; Vol 8.
7. Topliss, J. G.; Edwards, R. P. *J. Med. Chem.* **1979**, 22, 1238.
8. Rumelhart, D. B., Parallel Distributed Processing, Feldman, J. A.; Hayes, P. J.; Rumelhart, D. B., Ed.; The MIT Press: London, 1982; Vol 1, pp 318-363.
9. Aoyama, T.; Suzuki, Y.; Ichikawa, H. *J. Med. Chem.* **1990**, 33, 905.
10. Aoyama, T.; Suzuki, Y.; Ichikawa, H. *J. Med. Chem.* **1990**, 33, 2583.
11. Andrea, T. A.; Kalayeh, H. *J. Med. Chem.* **1991**, 34, 2824.
12. Livingstone, D. J.; Hesketh, G.; Clayworth, D. *J. Mol. Graph.* **1991**, 9, 115.
13. Livingstone, D. J.; Salt, D. W. *Bioorg. Med. Chem. Lett.* **1992**, 2, 213.
14. Dunn, W. J.; Greenberg, M. J.; Callejas, S. S. *J. Med. Chem.* **1976**, 19, 1299.
15. The descriptors added to the original dataset were the sterimol term, L1, B1 through B4, (Verloop, A.; Hoogenstraaten, W.; Tipker, J. *Drug Design*; Ariens, E. J., Ed.; Academic Press: New York, 1976; Vol 7, pp. 165-207.) and molecular connectivity indices (Kier, L. B.; Hall, L. H. Molecular Connectivity Analysis; Bawden, D., Ed.; Research Studies Press Ltd.: England, 1986) obtained from the program MOLCONN-X, Hall Associates Consulting, Quincy, MA 02170.
16. Selwood, D. L.; Livingstone, D. J.; Comley, J. C.; O'Dowd, A. B.; Hudson, A. T.; Jackson, P.; Jandu, K. S.; Rose, V. S.; Stables, J. N. *J. Med. Chem.* **1990**, 33, 136.
17. NeuralWorks Professional II/PLUS, NeuralWare, Inc., Pittsburgh, PA 15276.

Figure 2. Input and hidden unit weights after training.



a. Dunn dataset

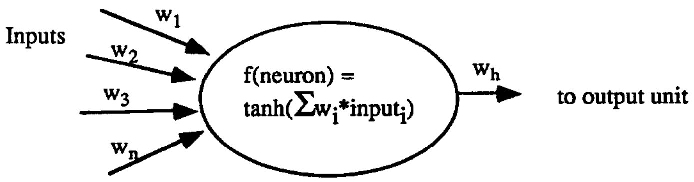
b. Selwood dataset

18. NCSA PalEdit, National Center for Supercomputer Applications, University of Illinois, Champaign-Urbana, IL 61820.

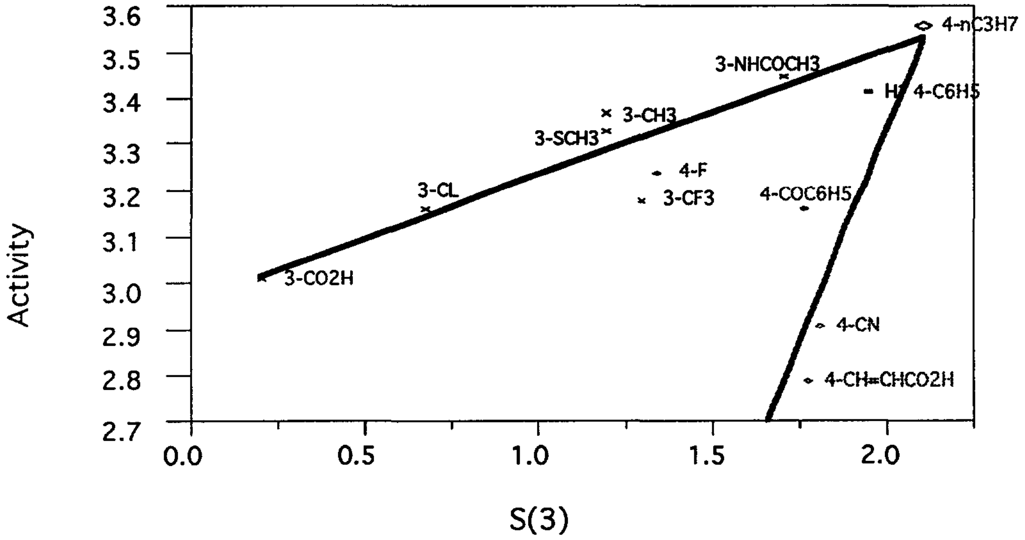
19. Spyglass Transform, Spyglass, Inc., Champaign, IL 61820

20. Regression analysis was performed using a Macintosh IIcx computer using the JMP Statistical Visualization Software from SAS Institute, Cary, NC. Definition of terms as follows: n is the number of observations, F is a measure for the significance of the equation, r is the correlation coefficient. The numbers in parenthesis are the standard errors.

Figure 1. An individual hidden layer neuron in a neural network



Plot 1. Dunn Dataset: Model using S(3) by Substituent Position.


Activity
S(3)
4-nC3H7
3-NHCOCH3
H
4-C6H5
3-CH3
3-SCH3
4-F
3-CF3
4-COC6H5
3-CL
3-CO2H
4-CN
4-CH=CHCO2H