Novel method for the display of multivariate data using neural networks

D. J. Livingstone

SmithKline Beecham Pharmaceuticals, The Frythe, Welwyn, Herts, UK

G. Hesketh and D. Clayworth

AEA Technology, Harwell Laboratory, Oxfordshire, UK

A neural network has been used to reduce the dimensionality of multivariate data sets to produce two-dimensional (2D) displays of these sets. The data consisted of physicochemical properties for sets of biologically active molecules calculated by computational chemistry methods. Previous work has demonstrated that these data contain sufficient relevant information to classify the compounds according to their biological activity. The plots produced by the neural network are compared with results from two other techniques for linear and nonlinear dimension reduction, and are shown to give comparable and, in one case, superior results. Advantages of this technique are discussed.

Keywords: nonlinear mapping, principal components analysis, unsupervised learning, pattern recognition, quantitative structure–activity relationships, artificial intelligence

INTRODUCTION

The study of quantitative structure–activity relationships (QSAR) requires physicochemical parameters to describe the biologically active molecules. The techniques of computational chemistry are being applied increasingly in the generation of such descriptors.[1–3] This often leads to the production of data sets that have more columns (parameters) than rows (compounds), and the analysis of such data sets can pose problems. Multiple linear regression, for example, is not a suitable method but there are a variety of multivariate statistical techniques that may be employed.[4,5] Among these, pattern recognition display methods can be quite informative, particularly in the early stages of an investigation when redundant information has not been eliminated. Two such techniques are plots of principal components (PCA) and nonlinear mapping (NLM), which give linear and nonlinear dimension reduction, respectively, of high-dimensional data sets.[6] These are examples of "unsupervised learning" pattern recognition in that the property of interest (in this case biological activity) is not used in the analysis, and thus the danger of chance correlations inherent in other techniques is reduced.[7] There are advantages and disadvantages to both of these methods, which we will discuss further in this report.

Neural networks are computer systems, implemented in hardware or software, that are intended to simulate some of the functions of the brain.[8,9] One feature of the human brain is the ability to recognize patterns, and thus it is not surprising that neural networks have been proposed as pattern recognition devices.[10,11] A recent report has demonstrated that a neural network may be used to classify two sets of structure–activity data successfully.[12] This was equivalent to carrying out discriminant analysis because the biological activity data was ranked according to four scores. In another application of neural networks to pattern recognition it was shown that a network may be used to identify points that are constrained to lie on an m-dimensional surface embedded in an n-dimensional data space, where $m < n$.[13] We took a different approach and used a neural network that will give a low-dimensional display (here two dimensions) of a high-dimensional data set. The utility of this technique has been investigated by the examination of three QSAR data sets, containing 23, 33 and 70 parameters, respectively, and the results compared with NLM and PCA.

EXPERIMENTAL

Experimental details of the biological testing and the calculation of physicochemical properties for the three data sets used are given in the references cited. The first data set consisted of 16 analogues of antimycin-a_1 that had killing activity towards filariae.[14] A total of 53 physicochemical parameters were calculated to describe this set of com-

Address reprint requests to Dr. Livingstone at SmithKline Beecham Pharmaceuticals, The Frythe, Welwyn, Herts, AL6 9AR, UK.
Received 15 August 1990; accepted 19 September 1990

pounds, but this set was reduced to 23 by the removal of correlated features. The second data set was derived from 13 pyrethroids with measured neurotoxicity in an isolated housefly haltere nerve preparation.[3] Seventy physicochemical properties were calculated for these compounds and although this set contained redundant information that could be removed by various means,[15] it was decided to use it as an example of the preliminary examination of a large data matrix. The third data set was composed of 13 analogues of γ-amino butyric acid (GABA), which showed varying agonist activity at the central nervous system GABA receptor.[6] A set of 33 descriptors was calculated for this data set.

PCA was carried out using the general purpose statistics package GENSTAT(NAG Ltd., Oxford, UK) and NLM using the pattern recognition package ARTHUR (Infometrix Inc., Seattle, WA 98121, USA). The neural network technique employed for dimension reduction, devised at AEA Technology, Harwell, UK, has been termed a reversible nonlinear dimensionality reduction (ReNDeR) procedure and was implemented on a Science Applications International Corporation (SAIC) Sigma neurocomputer workstation using SAIC back propagation neural network software. This workstation is an IBM PC clone with a custom accelerator board claimed to be capable of 22 Mflops performance in neural network applications. The ReNDeR network consists of an input and output layer of neurons and three hidden layers, as shown in Figure 1, with all layers fully connected. The input and output layers contain as many neurons as there are parameters in the data set; the encoding and decoding layers contain a smaller number of neurons, as described in results. The layer of two neurons is used to provide the x- and y-coordinates of compounds for the two-dimensional display. If a higher dimensional display is required (e.g., three-dimensional for use with a graphics workstation), this central layer may be expanded. Network training is clearly problem dependent but, as an example, 5 million cycles for a 23-7-2-7-23 network for data set 1 (Figure 3) took approximately 16 hours. It is almost certainly unnecessary to run a network for as long as this; in these trials the networks were left to run overnight.

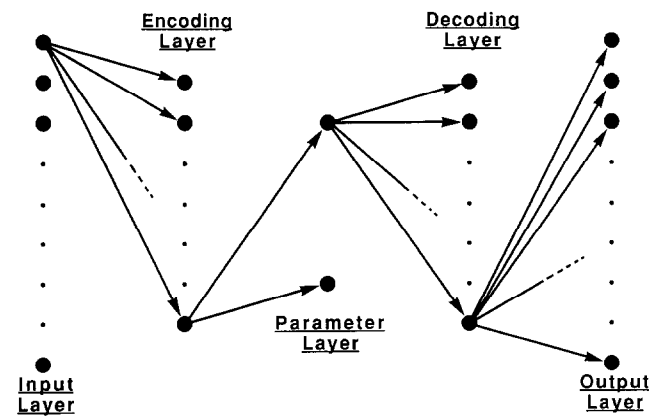

Figure 1. Diagram of the ReNDeR back-propagation neural network. The network is fully connected, and only a few connections are shown for clarity

RESULTS

A nonlinear map of the antimycin data set tends to show some clustering of the active compounds into one region of the map, as seen in Figure 2. This serves to confirm that the data set contains useful information, as was demonstrated by the generation of some reasonably successful regression equations.[14] The NLM itself does not distinguish compounds sufficiently to allow anything other than qualitative predictions of activity to be made and even this can be misleading because inactive compounds appear quite close to the actives. The display produced by ReNDeR with 7 encoding and decoding neurons, on the other hand, appears to cluster the active compounds into a tighter group, as shown in Figure 3. The inactive compounds in this plot are

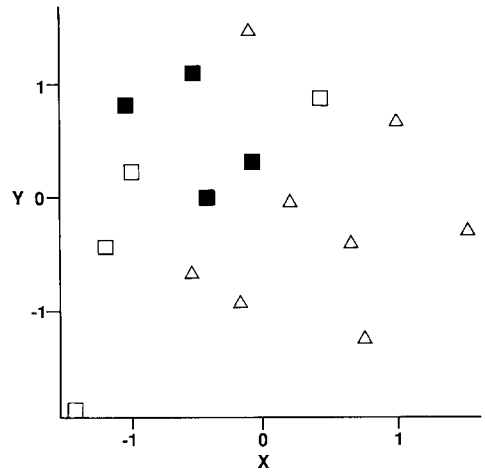

Figure 2. Nonlinear map of the antimycin data set based on 23 physicochemical parameters: open squares, inactive compounds; filled squares, active compounds; and open triangles, intermediate compounds. (Figure reprinted with permission from J. Med. Chem. *1990, 33, 136–42.)*

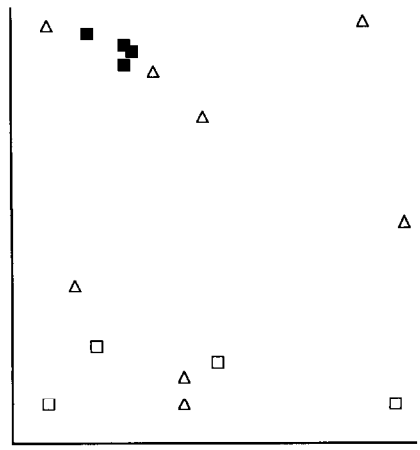

Figure 3. ReNDeR map of the antimycin data: open squares, inactive compound; filled squares, active compounds; and open triangles, intermediate compounds

also well removed from the area of active compounds and thus we might have more confidence in predictions based on this type of display.

Analysis of the pyrethroid data by PCA or NLM did not produce any plots that were useful in classifying of the compounds, and they will not be reproduced here. This is not an unusual result in the preliminary analysis of a data set: Raw data may contain sufficient "noise" to obscure any useful patterns or, indeed, may not contain any useful information at all. That these data contained a considerable amount of redundancy was shown by factor analysis, which revealed that 8 factors were sufficient to account for 99% of the variance in the set.[16] This analysis also showed that there were no simple correlations between biological activity and the factors derived from the physiocochemical data. A combination of at least three factors was required to account for any of the biological responses. A display of this data set produced by ReNDeR, using 30 neurons in each of the hidden layers, is shown in Figure 4, where it can be seen that the activity classes are mixed. This is encouraging in light of the results from the other display methods and from the factor analysis.

Treatment of the GABA dataset by NLM and PCA showed that the best display of the original 33 parameter set was obtained by PCA. This was improved by the removal of correlated descriptors and the selection of parameters according to their discriminant ability with respect to biological activity,[6] as shown in Figure 5. This plot is based on principal components derived from 4 parameters and thus represents a later stage in the data analysis procedure. Figure 6 shows the results from ReNDeR mapping, based on the starting set of 33 parameters with 20 encoding and decoding neurons. It is interesting to note that the separation of compounds is comparable to the PCA plot. Perhaps of even greater interest is the fact that this result was obtained from the starting data set and thus is unsupervised learning whereas the PCA plot in Figure 5 involved implicit supervision through the choice of the most discriminatory parameters. In this example data display using the neural network appears to give a superior result to either NLM or PCA.

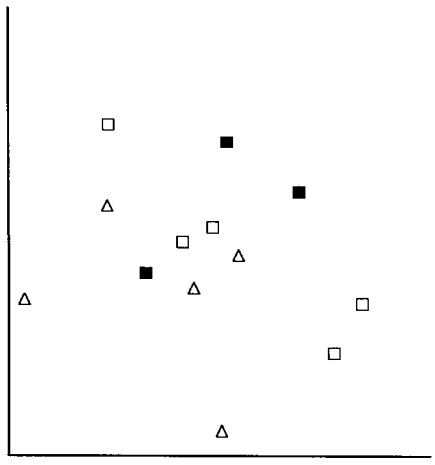

Figure 4. ReNDeR map of the pyrethroid data: filled squares, potent compounds; open triangles, weak compounds; and open squares, inactive compounds

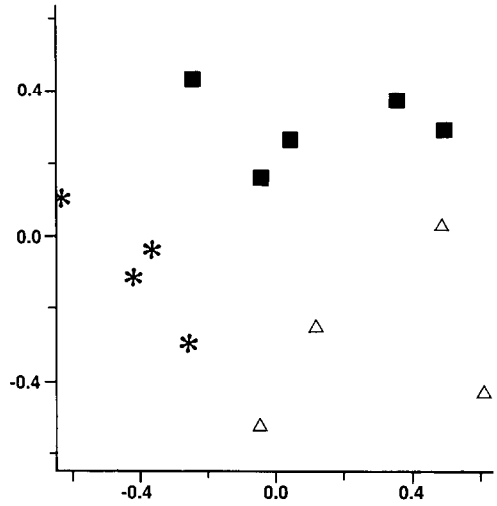

Figure 5. Principal components plot of the GABA data based on four selected parameters; filled squares, potent agonist; open triangles, weak agonist; and stars, no agonist activity. (Figure reprinted with permission from J. Comp. Aid. Mol. Design *1989, 3, 55–65.)*

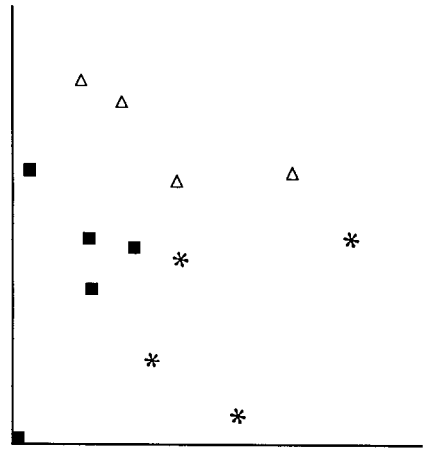

Figure 6. ReNDeR map of the GABA data based on 33 parameters: filled squares, potent agonist; open triangles, weak agonist; and stars, no agonist activity

DISCUSSION

The advantages of display techniques are clear because they produce readily understood synopses of large data matrices and may be used in an unsupervised learning sense, thus minimizing the possibility of chance correlations. There are drawbacks to the use of either NLM or PCA. For example, PCA imposes a linear structure on the variables that may obscure useful information contained in nonlinear combinations. This technique also seeks to preserve the variance of the data set in the first few principal components. If an important variable (in terms of the biological data) contains a small amount of variance this may not be included in the analysis until the later eigenvectors are calculated. It is also possible that such a variable may not be clearly associated

with any particular principal component and thus may not be seen to be of importance.

In the case of NLM, the interpoint distances in the n-dimensional space are important. Nonlinear mapping aims to preserve all interpoint distances and thus may obscure useful patterns that are represented by a particular set of distances. To produce a nonlinear map it is necessary to minimize an error function, such as

$$E = \sum_{i>j} \frac{(D_{ij}^* - d_{ij})^2}{(D_{ij}^*)^\rho} \quad (1)$$

where D_{ij}^* is the distance between a pair of points in n-space, d_{ij} is the distance between a pair of points in 2-space, and ρ is a weighting factor. When $\rho = 2$, all distances are given the same weight. The NLM shown in Figure 2, for example, used a value of 2 for ρ. If the value of this parameter is changed to -2, large interpoint distances are preserved at the expense of smaller ones. This tends to emphasize the clustering of points in a nonlinear map[6] and may serve to improve Figure 2 in terms of its predictive power. It is interesting to see that the ReNDeR mapping shown in Figure 3 appears to give this sort of clustering of the compounds.

A problem common to both of these techniques is the difficulty of moving from the 2D representation back to the n-dimensional parameter space. In other words, it is not possible to select a point on a PCA plot or a nonlinear map and translate this into the values required for the physicochemical parameters to place a compound at that point. It is also not easy to see how changes in the value of a particular property will affect the position of a point on the 2D plot. A plot involving just two parameters is easier to interpret, in this respect, but suffers from the disadvantage that only a small proportion of the information in the data set is considered. In the case of ReNDeR mapping, however, it is possible to move between 2-space and n-space because all the connection weights between neurons are known. This technique, therefore, promises the ease of interpretation of a 2D plot combined with the information content of a multivariate display method. One advantage of principal components analysis over nonlinear mapping is that the combination of variables for each principal component is known. Thus, although the combination may be complex and may suffer from the fact that it is a linear combination calculated so as to preserve variance, it may be that some interpretation is possible. Because the connection weights are known for ReNDeR, the possibility of interpretation also exists without the imposition of a linear combination on the variables. It appears that the ReNDeR technique offers some unique features compared with the linear and nonlinear display methods shown here. One potential drawback to ReNDeR is the computation time required to train the network, compared with the much lower requirements of either PCA or NLM. A neural network, of course, is well suited for distributed processing and it is anticipated that as parallel processors become more generally available this computing requirement will be seen to be less of a hurdle.

We also examined another form of dimension reduction using neural networks, known as Kohonen mapping.[17] Our preliminary results with this method have been disappointing. However, because Kohonen mapping uses a number of adjustable parameters, further experiments are needed to fully evaluate this technique. The performance of the ReNDeR approach is also subject to variability because it is necessary to choose a suitable number of encoding and decoding neurons.

It has been demonstrated that a neural network may be used as a dimension reduction device for data analysis. The results obtained by this technique are comparable to those achieved using other methods, and may even be superior. It is also possible to transfer more easily from 2-space to n-space using this technique, the penalty appearing to be a greater requirement in computer time, at least in the initial generation of the 2D display.

REFERENCES

1 Kikuchi, O. *Quant. Struct.–Act. Relat.* 1987, **6**, 179–84
2 Hyde, R.M. and Livingstone, D.J. *J. Comp. Aid. Mol. Design* 1988, **2**, 145–55
3 Livingstone, D.J., Ford, M.G. and Buckley, D.S. in *Neurotox '88: Molecular Basis of Drug and Pesticide Action* (G.G. Lunt, Ed.) Elsevier, Amsterdam (1988) 483–95
4 Salt, D.W. and Ford, M.G. in *Neurotox '88: Molecular Basis of Drug and Pesticide Action* (G.G. Lunt, Ed.) Elsevier, Amsterdam (1988), 469–482
5 Livingstone, D.J. *Pestic. Sci.* 1989, **27**, 287–304
6 Hudson, B., Livingstone, D.J. and Rahr, E. *J. Comp. Aid. Mol. Design* 1989, **3**, 55–65
7 Topliss, J.G. and Edwards, R.P. *J. Med. Chem.* 1979, **22**, 1238–44
8 Karna, K.N. and Breen, D.M. *Neural Networks* 1989, **1**, 4–23
9 Eliot, L.B. and Holliday, F. *Neural Networks* 1989, **1**, 96–106
10 Carpenter, G.A. *Neural Networks* 1989, **2**, 243–57
11 Coolen, A.C.C. and Kuijk, F.W. *Neural Networks* 1989, **2**, 495–506
12 Aoyama, T., Suzuki, Y. and Ichikawa, H. *J. Med. Chem.* 1990, **33**, 905–8
13 Saund, E. *IEEE Trans. Pattern Anal. Machine Intell.* 1989, **11**, 304–14
14 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. and Stables, J.N. *J. Med. Chem.* 1990, **33**, 136–42
15 Ford, M.G. and Livingstone, D.J. *Quant. Struct.–Act. Relat.* 1990, **9**, 107–14
16 Ford, M.G., Greenwood, R., Turner, C.H., Hudson, B. and Livingstone, D.J. *Pestic. Sci.* 1989, **27**, 305–26
17 Kohonen, T. *Self-Organization and Associative Memory*, Springer-Verlag, New York (1988)