Journal of Computer-Aided Molecular Design, 3 (1989) 55–65
ESCOM

J-CAMD 056

Pattern recognition display methods for the analysis of computed molecular properties

Brian Hudson, David J. Livingstone and Elizabeth Rahr

Department of Physical Sciences, Wellcome Research Laboratories, Langley Court, Beckenham, Kent BR3 3BS, U.K.

Received 5 May 1988
Accepted 24 August 1988

Key words: Quantitative structure activity relationships; Computer-aided molecular design; Multivariate analysis; Pattern recognition; Unsupervised learning; Display methods; Principal components analysis; Non-linear mapping

SUMMARY

Pattern recognition methods, particularly the 'unsupervised learning' techniques, are well suited for the preliminary analysis of the large data sets produced by computer chemistry. The use of linear and non-linear display methods for such exploratory analysis are exemplified with the aid of two data sets of biologically active molecules. Advantages and disadvantages of these techniques are discussed.

INTRODUCTION

Current approaches to computer-aided molecular design increasingly involve the use of computed structural and quantum mechanical parameters to produce a set of molecular property descriptors for a series of compounds [1]. This is because it has been recognised that neither purely structural methods nor substituent constant based QSAR can provide the necessary description of the physicochemical properties of a molecule in isolation. The analysis of the resulting data requires the use of statistical techniques which may either be thought of as the use of QSAR techniques in Computer Chemistry (CC) or the use of molecular modelling to produce 'new' parameters for QSAR. Whatever the terminology the result is the same.

In the case where biological activity data are quantitative and the number of descriptor variables small, a single or multiple linear regression approach are most often adopted. In many circumstances this is the analytical method of choice although a drawback of these techniques is that non-linear or discontinuous relationships may be obscured. Where the biological data are not quantitative or where more than one mechanism is involved or where there are a large number of molecular descriptors, regression analysis is not appropriate and other techniques should be used. For the former, where the data may consist of scores or a classification, discriminant analysis [2] may be used in place of regression. In the situation where the number of physicochemical parameters is large compared to the number of compounds in the data set, it has been recognised that there are dangers of chance correlation if a regression approach is used [3].

The usual case in CC/QSAR is for both the above problems to apply. In particular 'wide' data sets are produced; often wider than they are long. These are not amenable to treatment by regression analysis. Such data sets may be contracted by statistical means [4], removing redundancy and 'ill-conditioned' parameters, so as to allow their analysis by regression or other techniques. The danger, however, is that in the process of removing such data useful information may be lost.

There is a need therefore for methods to analyse these data sets without the risk of chance correlations and without the need for data reduction with its potential for information loss. One answer to this problem is the use of 'unsupervised learning' pattern recognition techniques.

Pattern recognition is a general term applied to methods of data analysis which seek to find a 'pattern' in a data set. A successful regression equation, for example, can be thought of as a pattern which has been found in a data set. Pattern recognition methods can use either supervised or unsupervised learning. The expression 'learning' derives from the origins of some pattern recognition methods in artificial intelligence research [5]; algorithms were constructed so as to 'learn' from a set of data in order to be able to generate a classification or discrimination rule. This is known as supervised learning where the algorithm learns from the classification of objects in the data set; the data set may also be referred to as a training set where, from the opposite viewpoint, the classification of objects is used to 'train' the algorithm in order to produce the decision rules.

Unsupervised learning, on the other hand, does not require knowledge of the class membership of the objects (compounds) in the training set and so may be used both prior to, and following, any biological testing. Since the biological data is not used in the analysis there *should* be little possibility of a seemingly useful pattern emerging by chance.

There are a number of unsupervised learning pattern recognition methods but we shall just consider here the 'display' techniques. A multivariate data set may be thought of as a collection of objects in N-dimensional space where each dimension corresponds to a descriptor variable and each object corresponds to a compound. The N-space coordinates of each compound correspond to the values of the physicochemical descriptors for that compound. Display methods are means by which such N-dimensional spaces may be reduced to two or three dimensions so that a human, arguably the best pattern recogniser, can evaluate them. Three different display philosophies, two linear and one non-linear, will be discussed using example data sets.

METHODS

Electronic properties were calculated using the molecular orbital programs CNDO/2 [6] and MOPAC [7]. These properties included partial atomic charges, electrophilic and nucleophilic superdelocalisabilities [8], dipole moment and the X, Y and Z components of the dipole moment, where the frame of reference corresponds to the modelling coordinate system. Other structure-related properties were calculated using PROFILES [8] and included measures of molecular shape and bulk, distances between heteroatoms and substituent dimensions. Capture and collation of all calculated parameters and output to the data analysis software was performed using PROFILES. Data handling and analysis was performed using RS/1 [9] and the pattern recognition program ARTHUR [10]. All calculations were performed on a VAX 11/750. Plots were generated on a Hewlett–Packard HP7221 plotter.

Two example data sets were used.

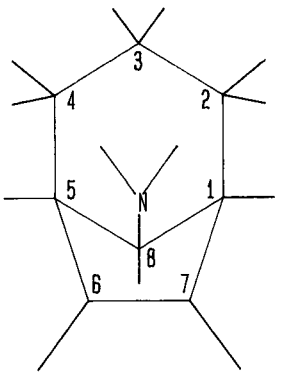

Fig. 1. Structure of antiviral bicyclic amines.

(i) GABA dataset

The compounds in this dataset have been classified as either potent, weak or inactive agonists of the gamma-amino butyric acid (GABA) CNS receptor. Structures of these and the conformational fitting strategy are given in Ref. 8. A total of 33 parameters were generated for this set of 13 compounds i.e., all possible parameters within the limits of our present software, rather than parameters chosen *a priori.*

(ii) Bicyclic amine dataset

This consists of 53 compounds of the general structure shown in Fig. 1. The molecules were overlayed by fitting the carbon skeletons using SYBYL [11]. After calculation of all properties and removal of those highly correlated with other properties ($r > 0.75$) using an in-house RS/1 procedure CORCHOP [4], the following 9 parameters remained.

1 – partial atomic charge (MNDO) of atom 7 (Fig. 1),
2 – X component of the dipole moment (MNDO),
3 – Y component of the dipole moment (MNDO),
4 – total dipole moment (MNDO),
5 – the nucleophilic superdelocalisability (MNDO) at atom 7 (Fig. 1),
6 – the second principal ellipsoid axis (Ref. 8),
7 – the third principal ellipsoid axis (Ref. 8),
8 – the molecular weight,
9 – the calculated log P (Ref. 12).

All compounds have been tested for antiviral activity against Influenza A. Each compound was classified as active, intermediate or inactive on the basis of a plaque reduction assay.

VARIABLE/VARIABLE PLOTS

This is the simplest of the linear display methods and needs little comment other than that the information content of such plots is necessarily low since only two variables are considered at a

time. This can be increased by adding a third variable, but one of the major drawbacks of variable/variable plots is the large number of plots [$\frac{1}{2}$ N(N − 1) for N variables] which can be generated even for data sets of moderate size.

There are two main uses of a variable/variable plot. The first is obvious in that inter-parameter correlations are easily spotted if a straight line is observed. The second is that clustering of active compounds in a region of variable/variable plot may indicate the requirements for biological (or any other) activity. Unfortunately, biological activity can only rarely be described in terms of two parameters. The advantage of the multivariate display methods here described is that a multi-parameter molecular description can be reduced to a 2-dimensional representation which is almost as easy to interpret as a variable/variable plot.

PRINCIPAL COMPONENTS PLOTS

Principal components analysis is a well established statistical technique for dimension reduction [13,14]. Principal components are new variables created from linear combinations of the starting variables:

$$PC1 = A_1V_1 + A_2V_2 + \ldots\ldots A_nV_n$$
$$PC2 = B_1V_1 + B_2V_2 + \ldots\ldots B_nV_n$$

for variables V_1 to V_n with weighting coefficients A_1 to A_n for the first principal component and B_1 to B_n for the second component, etc. The properties of these principal components are such that:

(1) Each principal component is orthogonal (uncorrelated) with all the other principal components.
(2) The first principal component contains the largest part of the variance of the data set (information content) with subsequent principal components containing correspondingly smaller amounts of variance.

Thus, a plot of the samples on the first two principal components gives the 'best' representation, in terms of information content, of the data in two dimensions.

The GABA dataset is an example of the use of principal components plots and also illustrates the use of various parameter reduction techniques. Initially 33 parameters were calculated for each member of the dataset. The first two principal components of these parameters are plotted in Fig. 2. The clustering of the compounds can be seen on this plot but the divisions between the activity categories are not sharp enough for this plot to be useful in a predictive sense.

Of the initial 33 parameters, 9 were removed because of poor statistical distribution (i.e., skewed or kurtosed parameters). A further 12 parameters were removed using the ARTHUR routine SELECT. This identifies the parameters most important for activity classification and is related to regression analysis. The remaining 12 parameters were input to the ARTHUR routine LESLT (least squares linear transform) which optimises category pair separation in as few variables as possible.

It should be recognised that this selection process is a supervised learning method in that the biological activity is used to provide the selection criteria; the ARTHUR routine SELECT [15] is effectively a forward stepping regression procedure with a decorrelation step applied to the

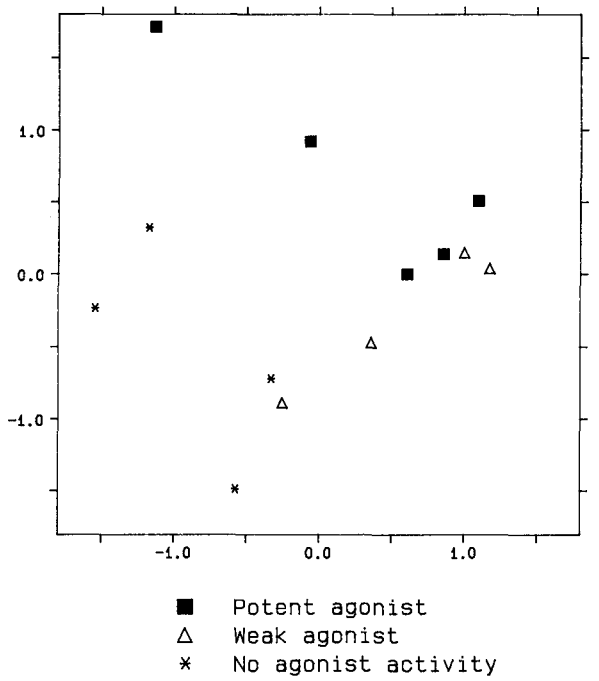

Fig. 2. Plot of the GABA data set on the first two principal components extracted from 33 parameters.

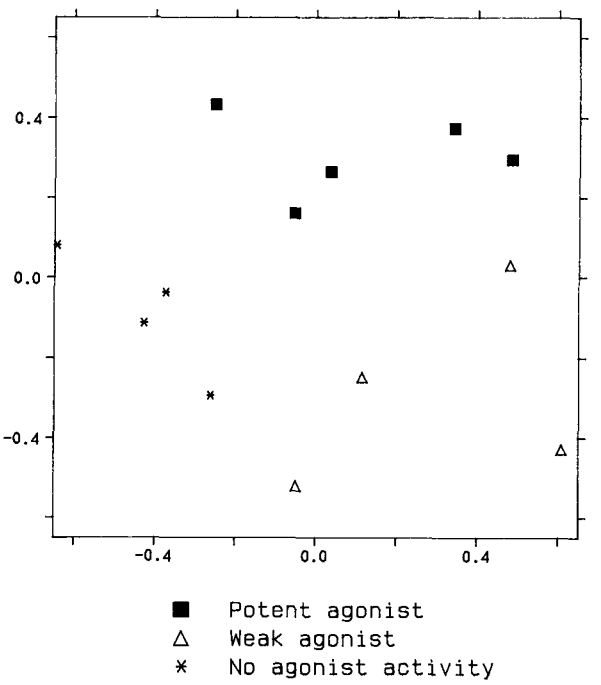

Fig. 3. The GABA data set: Principal components from 4 parameters.

remaining parameters after each parameter is selected. The routine LESLT is also obviously a supervised learning technique. There is a danger, therefore, that parameters may be selected by 'chance' [3] even though the eventual use of the parameters is unsupervised.

On the other hand, it may be argued that if there is 'real' information in a data set then any selection process should serve to improve the already observed pattern. Extra parameters chosen by chance will merely serve to retain 'noise' contained in the starting set. Indeed, inspection of Fig. 2 shows that the original data does contain useful information and any successful parameter selection should 'improve' this (cf. Fig. 3). If improvement is not observed this may point to the selection of parameters by chance. Since the unsupervised methods do not initially seek a 'fit' between biological data and the physicochemical descriptors the selection of parameters by chance *may* not be as misleading as in a multiple regression analysis.

Using this selection procedure it was found that 4 of the original 33 parameters (the charge on atom 4, two of the principal ellipsoid axes and the dipole moment) were adequate to describe the variation in biological data. Figure 3 shows the first 2 principal components of these 4 parameters in a PC plot. The 'information content' of this plot is the same as that for the 33 parameter PC plot. However, the removal of the noise (parameters unrelated to biological activity) results in greater discrimination between the various activity categories. The advantages are twofold. The better separation of the categories makes this plot much more useful for predictive purposes. Secondly, some idea of which parameters can be used to describe the biological activity may give some clues as to the nature of the biological recognition process.

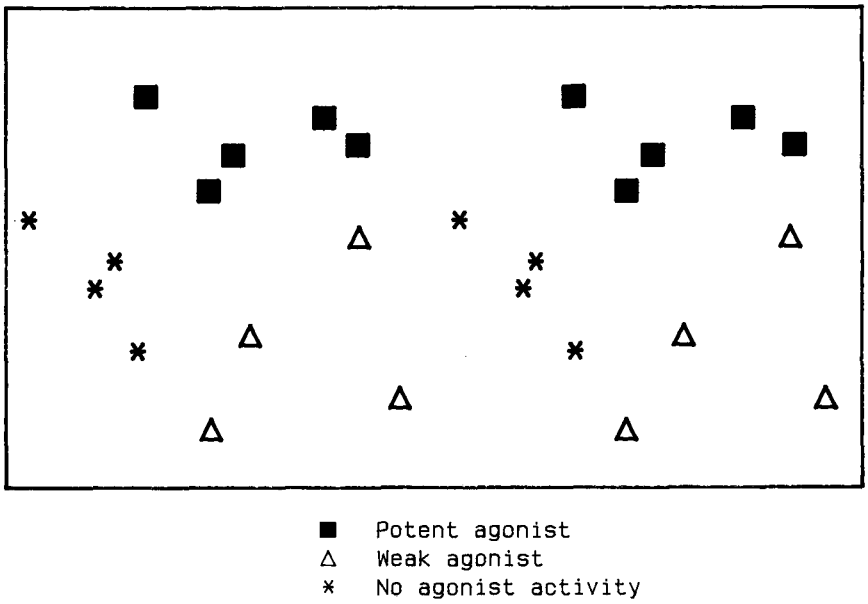

Fig. 4. Stereoscopic plot of the GABA data set on the first three principal components extracted from 4 parameters: X axis = PC1, Y axis = PC2, out-of-plane axis = PC3.

In this example, the descriptor set was easily reduced to just four variables and thus the principal components analysis does not effect a large reduction in the dimensions of the data. Figure 3, however, does exemplify the usefulness of the technique since no plot of any 2 of the 4 variables gives such a good classification. Furthermore, it may be seen by comparison of Fig. 3 with Fig. 2 that much of the 'noise' has been filtered out of the original data set, giving a clear classification of the compounds.

It is usually the case that a 2-dimensional plot of the first 2 principal components is sufficient to describe the dataset. Additional principal components do however give extra information and a useful technique is to produce a 3-dimensional plot which plots the first 3 principal components. Figure 4 shows such a plot for the GABA dataset. In some cases it may be that this will give a better description than a 2-dimensional plot of the first two PCs since the reduction in dimensionality necessarily involves some distortion of the data. Reduction to 3 dimensions introduces less distortion than reduction to 2 dimensions and so the 3-D PC1/PC2/PC3 plot is another useful representation. Whilst the 3rd PC can add some fine detail to the representation provided by the 1st and 2nd PCs, the 4th and higher PCs do not generally add anything to the description of the data.

NON-LINEAR MAPPING

This technique was introduced to the chemical literature by Kowalski and Bender [16] based on the method published by Sammon [17]. Successful applications have included the classification of geological samples [18], discrimination between analgesics and antispasmodics [19], the analysis of uptake in filarial worms [20] and the classification of antibiotics [21]. As in principal components

analysis, the dimensionality of the problem is reduced to 2 or 3 so as to enable analysis of the data by visual means. However, the linear combination of parameters forced on the data by a principal components analysis is not a feature of the non-linear map.

After autoscaling of the data the Euclidean distance in N dimensions between each pair of data points is calculated according to:

$$D_{ij}^* = \sum_{k=1}^{N} \left((P_{ik} - P_{jk})^2 \right)^{\frac{1}{2}}$$

where D_{ij}^* = the interpoint distance in N dimensional space,
P_{ik} = the value of the *k*th parameter for the *i*th data point, and
k = over all parameters P.

Note that when the number of dimensions is three and the parameters are the X, Y and Z coordinates of two points this reduces to the familiar equation for the distance between the 2 points.

A set of random positions for each data point is generated in the lower dimensional space (usually 2) of the non-linear map. The interpoint distances between these random points d_{ij} are calculated and a conjugate gradient minimiser is used to iteratively change the coordinates of each point on the map in order to minimise the error function

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

where $(D_{ij}^*)^{\rho}$ is a weighting factor. The result is a set of 2-dimensional coordinates for the data points where the distances between the points (the d_{ij}) closely reflect the distance between the points in the N-dimensional property space (the D_{ij}^*).

Figure 5 shows the non-linear map produced from the bicyclic amine data set using the ARTHUR default value of 2 for the error exponent ρ. A close clustering of the most active compounds is seen in the top left hand corner of this map. This shows that variation in biological activity can be described using a 2-dimensional representation of the original 9-dimensional property space and, furthermore, that at least some of the original 9 parameters are related to biological activity.

In the original Kowalski paper [18], an alternative method of calculating non-linear maps is discussed. This is to use an error exponent $\rho = -2$. This has the effect of preserving the larger interpoint distances at the expense of the smaller ones (when $\rho = 2$ all distances are given the same weighting). The non-linear map in Fig. 6 shows the effect of this change on the bicyclic amine data set and shows a significant improvement in the clustering of the active compounds.

In many cases, the second method is what is required in a CC/QSAR analysis. The user is generally interested in finding an area of the map associated with high biological activity. Compounds which differ greatly from the established 'recipe' of molecular properties need to be quickly recognised as such and given a lower priority as synthetic targets. This procedure is easier to do with the more focussed clustering obtained with $\rho = -2$.

The other major use of display methods is to identify areas of property space which have been 'unexplored' by a particular set of compounds. This is important when searching for novel struc-

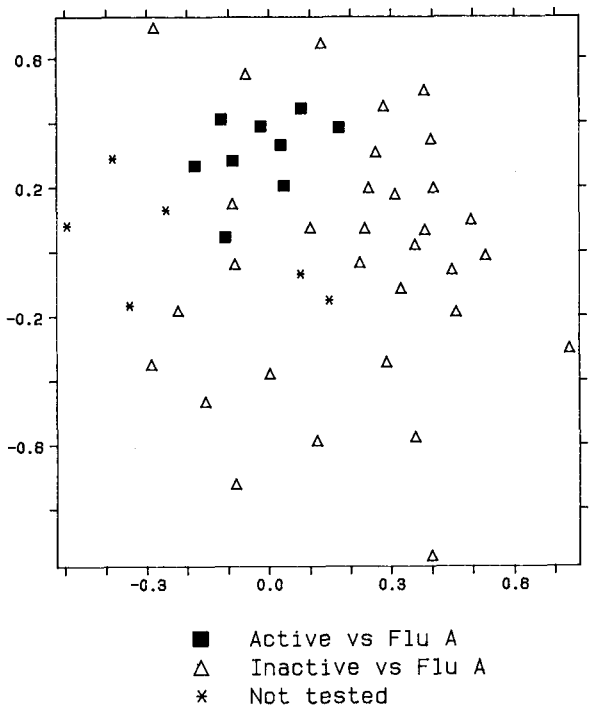

Fig. 5. Non-linear map of the bicyclic amine data set derived from 9 parameters: Error exponent = 2.

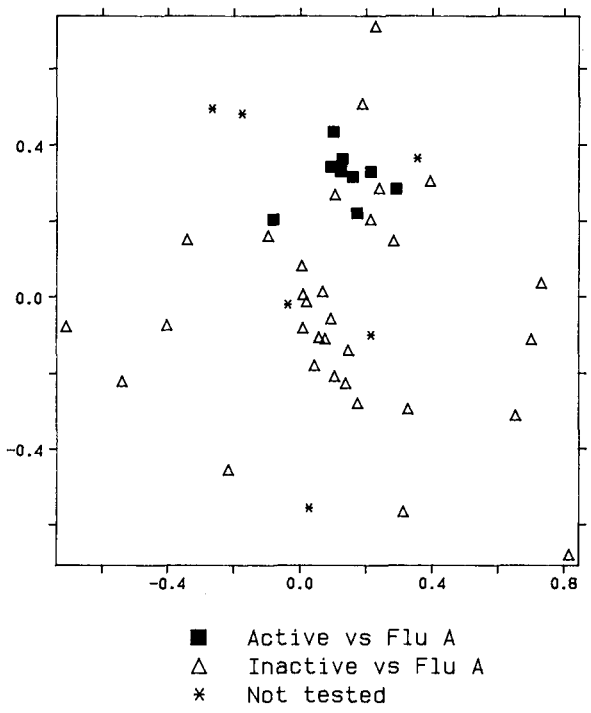

Fig. 6. Non-linear map of the bicyclic amine set, 9 parameters: Error exponent = −2.

tures. In this case $\rho = 2$ is the method of choice for the non-linear map due to the equal weighting of all interpoint distances. $\rho = 2$ is also the method of choice in the initial stages of an analysis whilst the relationships between different parameters are being explored.

DISCUSSION

The multivariate pattern recognition methods clearly constitute a very powerful tool for the analysis of molecules in terms of their physicochemical parameters. The display methods described here are particularly useful since they are easy to interpret and, with a reasonable set of parameters, give a good representation of a diverse set of molecular properties.

This, of course, raises the question of what is a 'reasonable' set of parameters. The approach taken here has been to generate as many parameters as possible for the data sets in question, the argument being that any a priori choice of descriptors may serve to introduce bias. Indeed it is possible that an injudicious choice of starting parameters may result in a useful pattern being 'missed'.

The generation of such a large set of descriptors may be criticised in terms of: (1) the effort involved; (2) redundancy in the data; (3) an increase in the possibility of chance correlations; (4) the fact that, apparently, any independent knowledge of the compounds or biological systems is not employed.

The effort involved in the production of these data sets is however slight in comparison with both the man-hours and computer-hours required to model the compounds and carry out molecular orbital calculations. Such modelling studies would routinely be undertaken as part of a CC/QSAR analysis.

The fact that there is redundancy in the data set is not a problem since the first step in the analysis is to remove redundant parameters based on a correlation matrix [4]. This method has been found to produce reasonably concise data sets in a number of studies carried out in-house.

The increased probability of chance correlations [3] is a matter that clearly requires further investigation. This problem is likely to be particularly serious if regression analysis is the only method available for the analysis of a data set since regression seeks a 'fit' between the biological and physicochemical data. Of course, with data sets of the size described here, regression analysis is only possible if a data reduction protocol, such as CORCHOP [4], results in a data set with sufficient degrees of freedom, i.e. fewer descriptors than compounds. One of the advantages of the unsupervised learning pattern recognition methods is that since the biological activity data is not used in the analysis there should be a low probability of chance effects occurring.

The final criticism, that particular knowledge of the compounds or biological systems is not used, is, of course, not true since such information may be added to the data set at any stage. It may be instructive to add any such parameters, e.g. pharmacokinetic data, at the beginning of the analysis so that correlations between the descriptors may be identified. Indeed it is one of the main advantages of principal components and non-linear map plots that they are both unsupervised methods and the biological data is not used in the analysis. This is of great importance in a CC/QSAR analysis since there is usually only a small number of observations (the biological activities of the compounds) to be described by a large number of possible parameters. Supervised methods are beset with the problem of chance correlations.

Although there are advantages in the use of these unsupervised display methods, there are also disadvantages. One particular problem with principal component plots is that, unlike a simple variable/variable plot, it is not possible to predict values of the constituent variables for a particular point on the plot. Since the coordinates of any point on the map are made up of linear combinations of the starting variables, there are an infinite number of solutions to the values of the principal component scores. Another criticism of the principal component approach is that this technique imposes a linear structure on the descriptor data. This linear combination of variables may obscure information which could be derived from non-linear combinations. Finally, the fact that principal components are calculated so as to preserve the maximum variance in the data set may, in fact, obscure useful patterns in the data if these are contained in variables which only make a small contribution to the variance (and hence have low loadings on the principal components).

There are two major drawbacks to the non-linear mapping method. Firstly, as for principal components analysis, it is not possible to specify for a particular point on the non-linear map what the values of each parameter are. It is therefore not possible to specify the parameter values necessary to place a compound in a particular region of the map. However, methods for parameter estimation are under investigation. Meanwhile, new compounds may be assessed by simply re-calculating the map with the candidates included. Secondly, the calculated non-linear map depends on the initial guess since the minimiser will find the nearest local minimum rather than the global minimum. This has the effect that the order in which the compounds are input and the inclusion

of new compounds can alter the map significantly although in most cases the relationships between the points are preserved. One approach to this problem, available as an option in ARTHUR, is to initialise the map from the principal component scores of the compounds. Several other approaches are currently being examined.

It is as predictive tools that these techniques are most useful. This is very simple since once a set of parameters which can separate the different classes of biological activity can be found a novel compound can easily be included in the analysis. The position of the novel compound with respect to the training set of compounds can then be used to assess its suitability as a synthetic target.

REFERENCES

1 Buckley, S., Ford, M.G., Leake, L.D., Salt, D.W., Burt, P.E., Moss, M.D.V., Brealey, C.J. and Livingstone, D.J., In Hadzi, D. and Jerman-Blazic, B. (Eds.) QSAR in Drug Design and Toxicology (Pharmacochemistry Library, Vol. 10), Elsevier, Amsterdam, 1987, pp. 336–339.
2 Martin, Y.C., Holland, J.B., Jarboe, C.M. and Plotnikoff, N., J. Med. Chem., 17 (1974) 409–13.
3 Topliss, J.G. and Edwards, R.P., J. Med. Chem., 22 (1979) 1238–44.
4 Livingstone, D.J. and Rahr, E., Quant. Struct.-Act. Relatsh., (1989) in press.
5 Sharaf, M.A., Illman, D.A. and Kowalski, B.R., Chemometrics, Wiley, New York, 1986, pp. 179–295.
6 CNDO, program No. 91, QCPE, Bloomington, IN.
7 MOPAC, program No. 455, QCPE, Bloomington, IN.
8 Glen, R.C. and Rose, V.S., J. Mol. Graph., 5 (1987) 79–86.
9 RS/1, Data Handling Software, BBN Software Products UK Ltd., Staines, Middlesex.
10 Infometrix Inc., Seattle, WA.
11 SYBYL, Tripos Associates, St. Louis, MO.
12 CLOGP, Medchem Software V. 3.51, April 1987, Pomona College Medicinal Chemistry Project, Pomona College, Claremont, CA.
13 Seal, H., Multivariate Statistical Analysis for Biologists, Methuen, London, 1968, pp. 101–122.
14 Chatfield, C. and Collins, A.J., Introduction to Multivariate Analysis, Chapman & Hall, London, 1980, pp. 57–79.
15 Kowalski, B.R. and Bender, C.F., Pattern Recognition, 8 (1976) 1–4.
16 Kowalski, B.R. and Bender, C.F., J. Am. Chem. Soc., 94 (1972) 5632–5639.
17 Sammon, J.W., IEEE Trans. Comput., C-18, (1969) 401–409.
18 Kowalski, B.R. and Bender, C.F., J. Am. Chem. Soc., 95 (1973) 686–693.
19 Abe, H., Kumazawa, S., Taji, T. and Sasaki, S., Biomed. Mass Spectrometry, 3 (1976) 151–154.
20 Court, J.P., Murgatroyd, R.C., Livingstone, D.J. and Rahr, E., Mol. Biochem. Parasitol., 27 (1988) 101–108.
21 Hyde, R.M. and Livingstone, D.J., J. Comput.-Aided Mol. Design, 2 (1988) 145–155.

APPENDIX

Loadings of input variables for the first 3 principal components (total explained variance = 70%)

Variable	Loading		
	PC_1	PC_2	PC_3
CMR	−0.1537	−0.2746	0.1070
1-ATCH	−0.1955	0.0964	0.2976
2-ATCH	0.0147	−0.2025	−0.3477
3-ATCH	0.1864	0.0030	0.2148
4-ATCH	−0.1827	0.0809	−0.1968
5-ATCH	−0.2229	0.0610	−0.0267
6-ATCH	0.2715	−0.0295	0.0487
X-DIPV	0.1524	−0.0847	−0.0586
Y-DIPV	0.0785	−0.0771	−0.2780
Z-DIPV	0.0728	−0.1171	0.0189
DIPMOM	−0.0187	0.1732	−0.0057
T-ENER	0.1366	0.1460	−0.2419
1-ESDL	0.2528	−0.1575	0.1198
2-ESDL	0.2210	−0.0718	−0.0203
3-ESDL	−0.2171	0.1076	−0.2481
4-ESDL	−0.1665	−0.1145	−0.2447
5-ESDL	−0.1047	0.1968	0.1576
6-ESDL	0.1282	0.0720	0.3372
1-NSDL	0.1828	−0.2529	0.1475
2-NSDL	0.1855	−0.2357	0.0254
3-NSDL	−0.0209	−0.1357	−0.3652
4-NSDL	−0.2258	0.1950	−0.0456
5-NSDL	−0.1108	0.2274	0.1410
6-NSDL	−0.0990	0.2571	0.1250
VDW_VOL	−0.2279	−0.2293	0.0314
X-MOFI	−0.1861	−0.2383	0.1362
Y-MOFI	−0.1861	−0.2664	0.0927
Z-MOFI	−0.2091	−0.2384	0.0898
X-PEAX	−0.2183	−0.1782	−0.0199
Y-PEAX	−0.2663	−0.0502	0.0839
Z-PEAX	−0.0505	−0.2172	0.0349
MOL_WT	−0.1264	−0.2627	0.1893
IHET_1	0.1854	−0.0709	−0.0520