Journal of Computer-Aided Molecular Design, 2 (1988) 145–155
ESCOM

J-CAMD 033

Perspectives in QSAR: Computer chemistry and pattern recognition

R.M. Hyde and D.J. Livingstone

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

Received 15 February 1988
Accepted 6 April 1988

Key words: Quantitative structure activity relationships; Molecular mechanics; Quantum mechanics; Pattern recognition; Multivariate analysis; Supervised and unsupervised learning

SUMMARY

Computer chemistry allows a detailed description of properties for a wide range of molecular environments. In these respects it offers substantial benefits to the QSAR (Quantitative Structure Activity Relationship) analyst. Problems associated with the resulting wide data matrices are, it is proposed, amenable to solution through multivariate 'pattern recognition' techniques.

INTRODUCTION

QSAR (Quantitative Structure Activity Relationship) analysis is a well established tool for optimising series of bio-active molecules. The technique is statistically-based and aimed at extracting the maximum information from biological test data on compounds of known structure and physicochemical properties. If the physicochemical properties can be calculated or predicted then relationships identified in the analysis can be used predictively. Hence the technique allows the search for optimal biological activity to be focussed on promising areas of chemistry.

QSAR analysis is potentially limited by the availability of physicochemical properties which are routinely and reliably predictable, and by the range of molecular environments within which predictions can be made. In both these respects, it would appear that computational chemistry, traditionally regarded as complementary to QSAR, has a direct contribution to make by providing a comprehensive and widely applicable source of calculable properties. A brief summary of the development of QSAR will, hopefully, put this claim in context and allow an assessment of its validity and potential. With the anticipated benefits come problems which threaten the statistical foundations of QSAR. However, it is proposed that these problems are not insuperable, and a working system, involving pattern recognition techniques, which allows advantage to be taken of computational chemistry as a source of QSAR property data is described.

BACKGROUND

Few would argue with the historical assertion that QSAR started with partition coefficients. The earliest examples of QSAR (in all but name) were reports of relationships between the potency of local anaesthetics and the oil/water partition coefficients [1,2]. In these and subsequent studies, the partition system was serving as a chemical model for the distribution of compounds between aqueous and non-aqueous biological phases. By the mid-1960s, QSAR as such had become established, and although recognisably descended from these earlier studies, was fast becoming an accepted tool for the optimisation of bio-active compounds. Hansch et al. established what has become the industry standard [3], water/octanol partition, and introduced the substituent constant π. The partition coefficient was now, within the right molecular environment, a predictable property. Thus if a correlation were observed with biological activity, it was possible to use the observation predictively.

The typical QSAR was still a lead optimisation exercise aimed at enhancing bioavailability and based on the tacit assumption that each compound in the series was equi-potent at the biological target. The parameters used were generally based on chemical model systems and, specifically, the linear free energy model.

Having established this basic model in which goodness-of-fit was assessed and variables selected by multiple linear regression analysis, attention was focussed on the determination and evaluation of improved parameters to describe steric [4], electronic [5] and hydrophobic effects. The description of electronic effect has been particularly prolific with 27 different substituent constants listed by Franke as 'the most important' [6, p. 100].

Whilst it cannot be denied that the advances in our knowledge of physicochemical descriptors has allowed a much more successful characterisation of molecular properties, the very wealth of this information can prove an embarrassment. It is well known that a regression equation should not be judged just on the number of terms included. Of equal importance, if chance correlations are to be avoided, is the number of terms *considered* in the analysis [7]. As will be seen later, this is important when parameters generated from computer chemistry are used.

A further extension of QSAR came when use was made of substituent constants such as π to predict the affinity of a compound for an enzyme. Here there was no bioavailability step but a model for a single interaction between ligand and macromolecule. In such cases, π no longer modelled partition into a membrane, but rather the hydrophobic effect which is an important component of ligand–macromolecule interactions. This duality is easy to live with since in both cases, partition from water to membrane or a hydrophobic interaction with a macromolecule, a high positive π value simply means that the group in question is not enthusiastic about hydrogen bonding to water molecules.

Similarly while the role of σ in predicting bioavailability is clear, its occurrence as a correlate for ligand–macromolecule interactions is less direct. Quite simply, those electronic properties which influence ionisation also influence electronically-based ligand–macromolecule interactions. The relevance of steric parameters to both membrane permeability and ligand–macromolecular interactions needs no explanation.

QSAR at this stage was working on a 'fixed menu' of properties which could be subdivided into steric, hydrophobic and electronic. These three classes could be assigned predictively within a limited range of molecular environments. In some cases, e.g. log P and MR, this could be done for

a diversity of chemical environments with the advent of CLOGP and CMR [8]. This allowed parameter values to be calculated through a fragment addition approach working from a database of measured values. However, for other parameters, particularly the electronic ones, the situation was more restricted.

Consequently, when variation occurred at more than one part of the molecule, equations involving indicator variables were sometimes found. (The authors believe that the use of such a variable rarely contributes to understanding, often makes the statistics look better than they really are and very often should be replaced or augmented by a weighting factor.)

Another development of QSAR was the appearance of square terms, generally in log P or π. These were adopted to account for non-linear dependency. Reservations about the use of the square term led workers to use relationships based on mathematical models to interpret the relationship between biological activity and physicochemical properties [9–11].

The most popular analytical tool has been, and remains to this day, Multiple Regression Analysis (MRA) although a number of workers have reported the use of other techniques such as cluster [12], factor [13] and discriminant [14] analysis, pattern recognition [15, 16] and principal components [17]. Chen and co-workers reported an interesting comparison of several of these methods [18].

Despite the advances therefore, QSAR was still to some extent limited (a) in the scope of molecular environments for which physicochemical properties could be assigned, and (b) in the detail with which differences in properties could be described. Working alongside computational chemists it became clear that there was nothing in principle against the idea of using the computed properties as a direct input into the QSAR property database. These properties would differ in two respects from the substituent constants they would augment. Firstly, they would not be based directly on chemical model systems, and secondly they would be specific to the molecules in question in their selected conformation. By adopting this approach, QSAR analysis could be possible in hitherto inaccessible molecular environments and with a 'resolution' of property description far greater than possible with the traditional parameters. If this approach offered so much, why was it not being pursued? The answer is that it was, but not in the comprehensive way now advocated.

The use of parameters derived from molecular orbital calculations might be considered a phenomenon of recent times but in fact dates back to the same period as the birth of 'traditional' QSAR. In 1964 Yoneda and Nitta [19] showed a correlation between the antibacterial activity of nitro furan derivatives and a quantity related to the nucleophilic reactivity of the molecule. In 1965 Snyder and Merril [20] reported a relationship between hallucinogenic potency and the energy of the highest occupied molecular orbital (E_{HOMO}) while Neely and co-workers [21] described analgesic potency in terms of E_{HOMO} and log P. However, despite these early examples there have been few, if any, reports of the *systematic* use of computer chemistry-generated molecular descriptors in QSAR.

It is clear that, as for all good things, there is a drawback. In this case, it is the analytical problems which the wide property data matrices would present. Serious though this problem may seem in the light of traditional QSAR, a solution can be found, and will be described after the next section, in which the construction of the data matrix is described.

THE PROPERTY DATA MATRIX

In our investigations, molecular modelling, quantum chemical calculations and data analysis are all carried out on a DEC VAX 11/750 minicomputer with 5 MB of physical memory, approx, 1 GB of disk space and a floating point processor. Molecular mechanics software includes the commercial program suite SYBYL (Tripos Associates Inc., St. Louis, MO, U.S.A.) and an in-house software package WMM (Wellcome Molecular Modelling). Results from SYBYL are visualised on an Evans & Sutherland PS 300 high-resolution calligraphic display (Evans & Sutherland, Salt Lake City, UT, U.S.A.) and the results from WMM are displayed on a SIGMA 5688 medium resolution raster graphics device (Sigmex Ltd., Horsham, Sussex, U.K.)

Quantum mechanical calculations may be initiated from either modelling package, routines currently in use include CNDO, MOPAC and GAUSSIAN 80 (Qantum Chemistry Program Exchange, University of Indiana, Bloomington, IN, U.S.A.).

A more comprehensive solution to data handling and analysis is still under development but Fig. 1 shows the current status of the system along with an indication of intended links and additions.

As can be seen from this figure, the system is based around a general purpose data handling package, RS/1 (BBN Software Products UK Ltd., Staines, Middlesex, U.K.), which performs a variety of manipulations of data in tables along with graph plotting, curve fitting and some statistical functions. This package also includes a high-level programming language (RPL – research programming language) which allows the easy development of QSAR procedures, e.g. compound selection and parameter deletion.

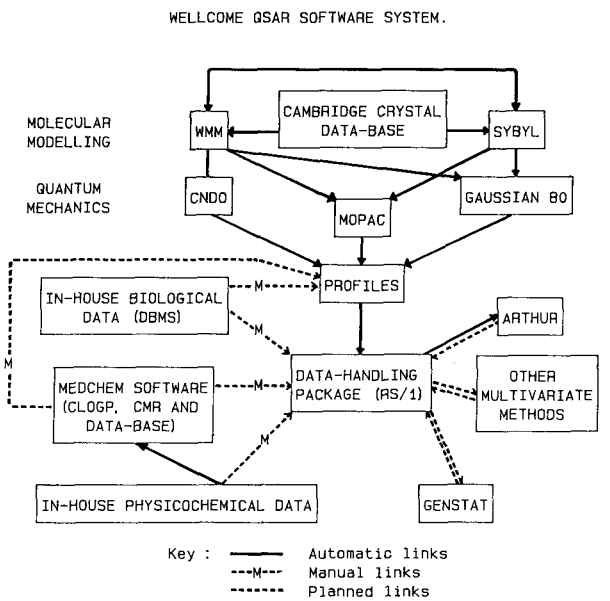

Fig. 1. Diagram of the relationship between the software packages and databases which comprise the Wellcome QSAR system.

The generation of a starting data set involves the construction of a set of molecules using the modelling package(s). Structural information from the Cambridge crystallographic database (Crystallographic Data Centre, University Chemical Laboratory, Cambridge, U.K.) may be used as an aid to the assignment of conformation. Having arrived at a suitable representation of the structures of the starting compounds (a necessary prerequisite) they are then submitted to one of the quantum mechanics programs for the calculation of atom charges, molecular orbital energy levels, dipole moment, etc. The files produced by these programs are then processed by an in-house package (PROFILES) [22] to produce a data set, in the form of an RS/1 data table, for use in subsequent analyses. Biological data and physicochemical property data, from the literature, calculated [8] or measured in-house, may be added at this stage or may be included in the RS/1 data table at a later time.

The data table thus produced resembles a typical QSAR data table with one important exception and that is the number of parameters it contains. These data tables are often wider than they are long, (or 'over-square'). They contain more parameters than they do compounds, in some cases many more parameters.

THE OVER-SQUARE MATRIX

Table 1 shows a list of the parameters generated, using this system, in a typical study. It may be seen that the set of parameters derived from computer chemistry is augmented by familiar QSAR descriptors such as MR and log P. The dimensions of the property database, 79 compounds and 81 parameters, are typical and contrast with classical QSAR studies based on substituent constants where the number of compounds comfortably exceeds the number of parameters tested.

Multiple regression analysis with simultaneous inclusion of all variables is clearly not possible for an over-square matrix in which degrees of freedom $\leqslant$ zero. Does this mean, therefore, that QSAR must reject the potential benefits of comprehensive property description through computer chemistry? A positive approach is required; the data are available, and it is up to QSAR to rise to the challenge.

The first move in solving the problem is to establish the information content of the data matrix. Whilst it is *expected* that a parameter set such as this will contain more information than is coded for by a set of substituent constants, this is not necessarily the case. What is certainly true is that the increase in the number of parameters and the information content are unlikely to be linearly related. Thus it should be possible to reduce the width of the matrix significantly with only marginal loss in information content. This can be achieved through the identification and elimination of redundant column vectors. In practice this means the removal of those descriptors which show a 'high' correlation with others in the set. This may, at first sight, seem to be an unproductive step since considerable effort was required to generate the set in the first place. However, if the correlations which are identified are genuine, then little or no information is lost from the set.

Other criteria may be used as an aid to parameter reduction, e.g. the spread of values for a given parameter, expressed as skewness and kurtosis. This will be discussed in a further publication.

If, after reducing the width of the data set, the property matrix is now taller than it is wide, regression analysis is at least an option. However, there are other constraints to be considered, of which the most important is the need to protect against chance correlations. In other words, merely having a positive number of degrees of freedom may not be enough. Topliss [7], for example,

TABLE 1
A LIST OF THE MOLECULAR DESCRIPTORS USED IN A TYPICAL STUDY (81 PARAMETERS, 79 COMPOUNDS)

	Whole molecule properties
'Bulk' descriptors:	Molecular weight, van der Waals' volume, dead space volume, collision diameter, approach diameter, surface area, molar refraction.
'Shape' descriptors:	Moment of inertia in the X, Y and Z axes, principal ellipsoid axes in the X, Y and Z directions.
Electronic and energy descriptors:	Dipole moment, X, Y and Z vectors of the dipole moment, energies (total, core-core repulsion and electronic)
Hydrophobicity descriptiors:	Log P
	Substituent properties
For 2 substituents:	Coordinates (X, Y and Z) of the centre, ellipsoid axes (X, Y and Z) of the substituent[a]
	Atom-centred properties
Electronic:	Atom charges, nucleophilic superdelocalizability, electrophilic superdelocalizability, for atom Nos. 1–14
Shape:	Inter-atomic distances between 6 pairs of heteroatoms.

[a] These are similar parameters to the STERIMOL substituent constants [23].

recommends that for a set of 20 compounds no more than 10 variables should be *considered* in a regression analysis if the probability of obtaining a chance correlation with $r^2 \geq 0.8$ is to be kept at the 1% level, or less. Note that this refers to the number of parameters considered, not included, in the regression analysis. As Franke pointed out [6, pp. 167–168] many workers have misinterpreted these results and have incorrectly assumed that a meaningful QSAR should be based on at least five observations per variable included in the regression equation. Goodness-of-fit of a regression eqation can be judged simply on the basis of its statistics and the confidence intervals of the coefficients.

If, after reduction of the data there are still more descriptors than compounds, it might appear that the set is not amenable to analysis at all. This is not the case. Two techniques which are claimed to make quantitative correlations between activity data and multiple descriptor data for such sets are SIMCA [24] and PLS [25] and, indeed, the reported applications appear to support this [26–29].

In addition to these two methods there are a number of other ways in which these data sets may be examined. These methods may be collectively called 'unsupervised learning' pattern recognition. Pattern recognition means exactly what the name implies and that is any means by which patterns in data are perceived by the analyst. Thus, multiple regression, cluster analysis, π/σ plots,

principal components, discriminant analysis, PLS, etc., are all examples of 'pattern recognition' methods. In the analysis of QSAR data generated from computer chemistry, pattern recognition methods to be found in ARTHUR (Infometrix Inc., Seattle, WA, USA) along with data pre-processing procedures written in RPL have been used. In addition to the techniques within the ARTHUR package other multivariate methods are under evaluation and will be incorporated into the system in the future. This may be carried out using the general statistical programming package GENSTAT (Numerical Algorithms Group, Oxford, U.K.) or by linking specific packages.

'Supervised' and 'unsupervised' learning derives from artificial intelligence research in which attempts have been made to devise computer programs which 'learn', e.g. the linear learning machine [30]. The objective of the learning process is to produce a classification rule, or rules, which may be used to distinguish between two or more classes of object. In the context of QSAR, the objects are compounds and the classes are the biological classification, e.g. toxic/non-toxic, active/intermediate/inactive, etc., although, of course, this classification may be some quantitative measure of activity. The term 'supervised' means that knowledge of class membership is used to supervise or train the method, i.e. the biological data is used in the analysis.

A number of supervised and unsupervised learning methods are listed in Table 2 with references to some applications. In addition, the books cited in references 31, 37, 38 and chapters 9 and 11 of Ref. 6 give a useful general discussion of pattern recognition methods. The important difference between unsupervised and supervised learning is that since the biological data does not enter the former, there are no restrictions on the ratio of parameters/compounds used in the analysis.

For supervised learning, while it is not possible to operate when degrees of freedom $\leqslant$ zero, there is an intermediate area in which, although there are adequate degrees of freedom, it is necessary to be aware of the risk of chance correlations. Jurs and co-workers investigated the problem for linear discriminant functions [39,40]. These studies suggested that the number of variables considered in an analysis using such discriminant functions should be kept to less than half, preferably less than one third, of the number of observations (compounds) in the data set. It would seem

TABLE 2
THE PRINCIPAL PATTERN RECOGNITION METHODS

Supervised learning	*Unsupervised learning*
Multiple regression	Non-linear mapping [35, 43]
Discriminant analysis [14]	Principal components [17, 43]
Linear learning machine [30]	Factor analysis [13]
PLS[a] [32]	Cluster analysis [12, 43]
SIMCA[a] [24]	κ-Nearest neighbour[b] [31]
Canonical correlation [33]	Correspondence analysis [36]
Adaptive least squares [34]	

[a] These methods are based on principal components analysis (*Unsupervised*) but use knowledge of class membership, hence *Supervised*.

[b] This method operates using knowledge of class membership (*Supervised*) but does not make use of this information to 'train' the analysis, hence the classification *Unsupervised*.

intuitively reasonable that, for other forms of supervised learning, similar caveats must be borne in mind. However, it is the demonstration of predictive ability which is the ultimate test of a QSAR analysis.

In unsupervised learning, the positions of the data points (compounds – whether existing or proposed) are calculated in a multidimensional space based on the chemical property matrix. This can be done even at the start of an exercise before there are any biological results. A convenient way of proceeding is through non-linear mapping [35]. A non-linear map is a 2-dimensional approximation to the true multidimensional interpoint distances; Fig. 2 shows a non-linear map of the data represented by Table 1. Two points which are close on the map should be more similar in terms of the input chemical property variables than two distant points. The map is then used in a purely passive way. As activity measurements are obtained, they are simply marked on the map and prediction of unknowns is achieved through spatial association on the map. In this and other unsupervised methods, the biological activity is not used to 'train' the chemical property matrix. The assumptions are (a) that the chemical properties originally considered are, on the basis of experience gained from other studies, likely to be relevant to biological interactions; (b) that compounds with similar chemical properties (as described) are more likely to have similar effects on a biological system.

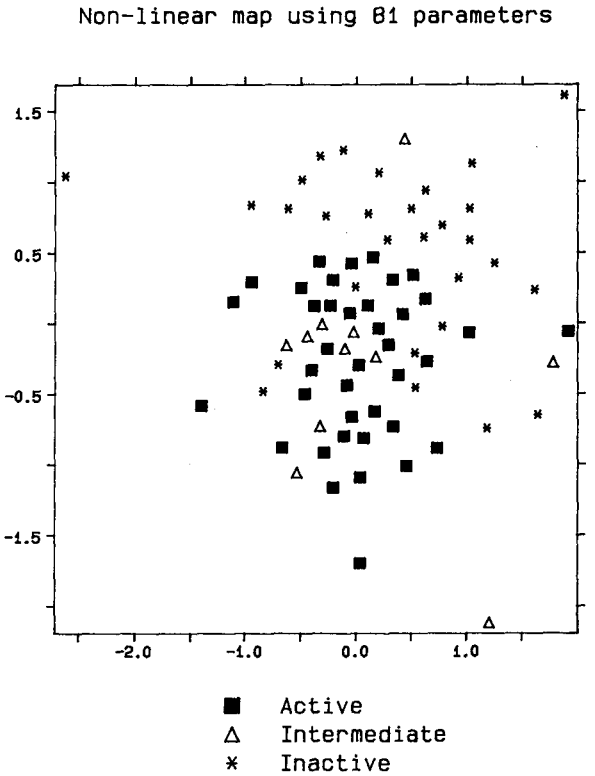

Fig. 2. Non-linear map of compounds from Table 1. (The authors thank Dr. S.G. Lister for providing this unpublished data as an example of the mapping technique.)

The principles behind this approach are an extension of those widely accepted for the planning of chemical series through substituent constant variation [41]. The objective is prediction through association rather than interpretation through correlation. At no stage is the biological activity used to select variables with which it is more highly correlated than others, and it is for this reason that unsupervised methods can operate outside the constraints which limit regression analysis. There can be no *false* correlations since there are no correlations *at all* to worry about. The main problem is one of signal–to–noise ratio, and it is for this reason that the data conditioning is even important for unsupervised learning. The method can be used visually and with great convenience, through the non-linear map. If convenience is sacrificed for rigour then it is possible to carry out the same predictive procedure in the full property space, by calculating euclidean distances to points of known biological activity [12, 30, 31].

Whilst unsupervised learning does not *appear* to suffer from the possibility of chance patterns emerging, there can be no certainty that an apparently 'useful' pattern has not occurred by chance. In this context a 'useful' pattern is one that can be used to correctly classify the majority of compounds in a training set. It is more likely, however, that the inclusion of parameters which contain no 'real' information will simply lead to 'noise' which will merely serve to obscure any useful underlying patterns.

DISCUSSION

For the QSAR analyst, computer chemistry is not only a valuable complementary technique, but also a prolific source of physicochemical parameters. In this second role, it has been exploited but little in the past. The potential benefits can be summarised as follows:

– Parameters may be readily calculated for almost any molecular environment.

– The variety of calculable parameters should give a better description of the molecules than the 'traditional' descriptors.

Some of the apparent problems associated with this approach are:

– The wide data matrices which result cannot be analysed using multiple regression analysis. This is *not* a problem since alternative methods exist.

– The determination of a common conformation may be difficult for flexible molecules. However, all that is required is a consistent and plausible basis for property calculation.

– Orientation of the molecules in a data set is critical since many parameters are orientation-specific. A number of orientation protocols have been described, Dean, for example, discusses the orientation of molecules in an electric field [42].

– Some of the computed parameters are not well understood. They may be readily calculated but it can be difficult to predict how their values change with changes in chemical structure. Nevertheless, the results of analysis can be used to evaluate the feasibility of proposed new compounds.

The multivariate or pattern recognition methods have been very sparingly used in QSAR compared with regression analysis. The value of unsupervised methods in handling over-square data matrices has been discussed earlier. However, even in cases where regression analysis would be 'allowed', the supervised learning methods would frequently offer a useful alternative. One advantage of pattern recognition is that the methods operate with either quantitative or qualitative biological data, although some of the methods do not allow quantitative predictions of biological

activity to be made even from quantitative biological data. Another potential advantage of some pattern recognition techniques is that they do not fit a pre-ordained mathematical function to the biological data, and could, in principle identify trends which would otherwise be lost.

In conclusion, the techniques of computer chemistry can extend the scope of QSAR both in the diversity of molecular environments for which properties can be assigned and in the detail with which the chemistry can be described. The ability to deal with diverse sets of flexible molecules is subject to the expected conformational and orientational constraints. The potential analytical problems associated with wide data matrices are soluble with the help of pattern recognition techniques. The enforced attention to these techniques in the context of wide data matrices serves as a valuable reminder of their worth even in situations which would otherwise be treated by regression analysis.

REFERENCES

1 Meyer, H., Arch.Exp.Pathol.Pharmakol., 42 (1899) 109.
2 Overton, E., Studien über die Narkose, Vol. 45, Fischer, Jena, 1901.
3 Hansch, C., Maloney, P.P., Fujita, T. and Muir, R.M., Nature, 194 (1962) 178–180.
4 Agin, D., Hersh, L. and Holtzmann, D., Proc.Natl.Acad.Sci. U.S.A., 53 (1965) 952–958.
5 Hansch, C., Muir, R.M., Fujita, T., Maloney, P.P., Geiger, F. and Streich, M., J.Am.Chem. Soc., 85 (1963) 2817–2824.
6 Franke, R., Theoretical Drug Design Methods (Pharmacochemistry Library, Vol. 7) Elsevier, Amsterdam, 1984.
7 Topliss, J.G. and Edwards, R.P., J.Med.Chem., 22 (1979) 1238–1244.
8 CLOGP and CMR, log P and Molar Refraction calculation routines, Part of the Pomona College, MEDCHEM software, Pomona College Medicinal Chemistry Project, Claremont, CA, U.S.A.
9 Hyde, R.M., J.Med.Chem., 18 (1975) 231–233.
10 Martin, Y.C. and Hackbarth, J.J., J.Med.Chem., 19 (1976) 1033–1039.
11 Kubinyi, H., Arzneim.-Forsch./Drug Res., 26 (1976) 1991–1997.
12 Hansch, C., Unger, S.H. and Forsythe, A.B., J.Med.Chem., 16 (1973) 1217–1222.
13 Weiner, M.L. and Weiner, P.H., J.Med.Chem., 16 (1973) 655–661.
14 Martin, Y.C., Holland, J.B., Jarboe, C.M. and Plotnikoff, N., J.Med.Chem., 17 (1974) 409–413.
15 Stuper, A.J. and Jurs, P.C., J.Am.Chem.Soc., 97 (1975) 182–187.
16 Hyde, R.M., Chem.Ind. (London), Nov. (1977) 859–862.
17 Streich, W.S., Dove, S. and Franke, R., J.Med.Chem., 23 (1980) 1452–1456.
18 Chen, B.K., Horvath, C. and Bertino, J.R., J.Med.Chem., 22 (1979) 483–491.
19 Yoneda, F. and Nitta, Y., Chem.Pharm.Bull., 12 (1964) 1264–1268.
20 Snyder, S.H. and Merril, C.R., Proc.Natl.Acad.Sci. U.S.A., 54 (1965) 258–266.
21 Neely, W.B., White, H.C. and Rudzik, A., J.Pharm.Sci., 57 (1968) 1176–1179.
22 Glen, R.C. and Rose, V.S., J.Mol.Graph., 5 (1987) 79–86.
23 Verloop, A., Hoogenstraaten, W. and Tipker, J., In Ariens, E.J. (Ed.), Drug Design, Vol. III, Academic Press, London, 1976, pp. 165–207.
24 Dunn, W.J., Wold, S. and Martin, Y.C., J.Med.Chem., 21 (1978) 922–930.
25 Dunn, W.J., Wold, S., Edlund, V., Hellberg, S. and Gasteiger, J., Quant.Struct.-Act.Relat., 3 (1984) 131–137.
26 Dunn, W.J. and Wold, S., J.Med.Chem., 21 (1978) 1001–1007.
27 Dunn, W.J. and Wold, S., Bio-org.Chem., 10 (1981) 29–45.
28 Hellberg, S., Wold, S., Dunn, W.J., Gasteiger, J. and Hutchings, M.G., Quant. Struct.-Act.Relat., 4 (1985) 1–11.
29 Hellberg, S., Sjöström, M. and Wold, S., Acta Chem.Scand., Ser.B, 40 (1986) 135.
30 Nilsson, N.J., Learning Machines, McGraw-Hill, New York, NY, 1975.
31 Varmuza, K., Pattern Recognition in Chemistry, Springer-Verlag, New York, NY, 1980.
32 Geladi, P. and Kowalski, B.R., Anal.Chim.Acta, 185 (1986) 1–17.

33 Szydlo, R.M., Ford, M.G., Greenwood, R. and Salt, D.W., In Dearden, J.C. (Ed.) Quantitative Approaches to Drug Design, Elsevier, Amsterdam, 1983, pp. 203–214.
34 Moriguchi, I., Komatsu, K. and Matsushita, Y., J.Med.Chem., 23 (1980) 20–26.
35 Kowalski, B.R. and Bender, C.F., J.Am.Chem.Soc., 95 (1973) 686–693.
36 Greenacre, M.J., Theory and Applications of Correspondence Analysis, Academic Press, London, 1984.
37 Stuper, A.J., Brügger, W.E. and Jurs, P.C., Computer Assisted Studies of Chemical Structure and Biological Function, Wiley, New York, NY, 1979.
38 Devijver, P.A. and Kittler, J., Pattern Recognition: A Statistical Approach, Prentice-Hall, London, 1982.
39 Stouch, T.R. and Jurs, P.C., Quant.Struct.-Act.Relat., 5 (1986) 57–61.
40 Whalen-Pederson, E.K. and Jurs, P.C., J.Chem.Inf.Comput.Sci., 19 (1979) 264–266.
41 Craig, P.N., J.Med.Chem., 14 (1971) 680–684.
42 Dean, P.M., Molecular Foundations of Drug–Receptor Interaction, Cambridge University Press, Cambridge, 1987, Ch. 7.
43 Bawden, D., Anal.Chim.Acta, 158 (1984) 363–368.