Evaluation of an Empirical Structure–Activity Relationship for Property Prediction in a Structurally Diverse Group of Local Anaesthetics

By **George W. Adamson** * and **Judith A. Bush**, Postgraduate School of Librarianship and Information Science, University of Sheffield, Western Bank, Sheffield S10 2TN

The local anaesthetic activities of 39 compounds belonging to different structural classes have been found to be highly correlated with a linear function of fragments derived from their structure diagrams in the form of connection tables. The relationship obtained by regression analysis has been used to simulate the prediction of the property of each structure, by using in each case the parameters obtained from a regression analysis on the remaining 38 structures. The strength of the relationship between structure and activity varies only slightly with the type of structural descriptor used. Quadratic functions did not give significantly better results than a linear function.

As the search for biologically active compounds by use of traditional screening methods becomes increasingly expensive,[1] it seems reasonable to try to develop approaches to the problem based on quantitative structure–activity relationships which might help reduce the amount of effort involved. A large proportion of the methods described in the literature use regression analysis to correlate biological activity either empirically with the structural features of the molecule [2] or semiempirically with known physicochemical parameters [3] which are, in turn, related to structure. The empirical method used in this investigation differs from most other published methods in its ability to handle a wide range of structural types in a single investigation. This is because the structural features of the whole molecule are used to give an explanation of activity. Most other methods involving the use of regression analysis consider the variation of activity with the side-chain structure only, and hence require the compounds studied in an investigation to be derivatives of the same parent compound. This allows only for property optimisation within a known lead series.[4,5] A similar method to that considered in the present investigation has been described by Nys and Rekker,[6] who define the partition coefficient data of a group of 87 compounds as a linear function of their structural features. However, in their case the fragments chosen to describe structures are not derived automatically but are based on a small number of ring

[1] J. E. Johnson and E. H. Blair, *Chem. Technol.*, 1972, Nov. 666.
[2] S. M. Free, jun., and J. W. Wilson, *J. Medicin. Chem.*, 1964, **7**, 395; A. Cammarata, *Ann. Reports Medicin. Chem.*, 1971, 245.
[3] C. Hansch in ' Drug Design,' ed. E. J. Ariëns, Academic Press, New York, 1971, vol. I, ch. 2, p. 271.
[4] G. Redl, R. D. Cramer, tert., and C. E. Berkoff, *Chem. Soc. Rev.*, 1974, **28**, 273.
[5] A. Verloop, in ' Drug Design,' ed. E. J. Ariens, Academic Press, New York, 1972, vol. I.
[6] G. G. Nys and R. F. Rekker, *Chim. ther.*, 1973, 521.

and chain features, together with some chemically important functional groups. Another difference is that they treat benzene rings as whole entities whereas saturated rings are broken down into fragments which are not differentiated from the corresponding chain fragments. In the work described below all saturated and unsaturated ring systems are fragmented, and all ring and chain fragments are differentiated.

The large amount of data generated in research directed towards the development of biologically active compounds is often stored as machine-readable files of properties and structure diagrams.[7] These files are a possible source of information for correlating chemical structure with biological activity. The method used in this investigation is one of a number which have been described for carrying out structure–activity investigations automatically by using this type of data,[8,9] although so far little has been published about their performance. The results of applying the method in the case of 39 structurally diverse local anaesthetics are reported below.

An attempt was also made to determine the power of the method to predict the log(MBC) values for the structures. The correlation obtained from a regression analysis carried out on all 39 structures would not necessarily be a good indication of the predictive power of the method. The reason for this is that the observed property value for a structure is included as a datum for the analysis and so the results of the analysis will depend upon this. The estimated property value for a structure obtained from the regression analysis will thus depend upon its own observed value. If a prediction were being carried out under real conditions then, of course, the observed property value for the structure would not be known.

A more reliable estimate of predictive power is obtained by the 'hold one out' technique. This method has been suggested as a means of determining the usefulness of regression analysis [4] or pattern recognition [10,11] methods in prediction. In applying this technique, the set of 39 structures was partitioned into two sets, the 'design set' consisting of 38 structures and the 'test set' consisting of one structure. The regression analysis was then carried out on the design set and the regression constant and coefficients obtained were used to 'predict' the property of the structure in the test set. Each of the structures in turn was selected to be the test set.

[7] V. B. Bond, C. M. Bowman, N. L. Lee, D. R. Peterson, and M. H. Reslock, *J. Chem. Documentation*, 1971, **11**, 168; E. Hyde, D. R. Lambourne, and L. A. McArdle, Abstracts of Papers, 163rd National Meeting of the A.C.S., Boston, April 1972; D. P. Jacobus, D. E. Davidson, A. P. Feldman, and J. A. Schafer, *J. Chem. Documentation*, 1970, **10**, 135.

[8] P. J. Harrison, *J. Appl. Statistics*, 1968, **17**, 226; G. W. Adamson and J. A. Bush, *Information Storage and Retrieval*, 1973, **9**, 561; R. D. Cramer, tert., G. Redl, and C. E. Berkoff, *J. Medicin. Chem.*, 1974, **17**, 533; D. T. Saggers, *Pesticide Sci.*, 1974, **5**, 341; K. Chu, *Analyt. Chem.*, 1974, **46**, 1181; B. R. Kowalski and C. F. Bender, *Chem. Eng. News*, 1974, **52**(7), 19; B. R. Kowalski and C. F. Bender, *J. Amer. Chem. Soc.*, 1974, **96**(3), 916; A. J. Stuper and P. C. Jurs, *ibid.*, 1975, **97**(1), 182.

[9] G. W. Adamson and J. A. Bush, *Nature*, 1974, **248**, 406.

METHOD AND RESULTS

It was first assumed that the log(MBC) (MBC = minimum blocking concentration) of a compound is a linear function (1) of the structural fragments contained in it,

$$\log(\mathrm{MBC})_i = \sum_{j=1}^{n} b_j x_{ij} + \mathrm{const.} \qquad (1)$$

where there are n types of structural fragment in the set of structures, b_j is a coefficient for the jth fragment type and indicates its contribution to activity, and x_{ij} is the number of times this fragment occurs in the ith structure. The values of b_j and the constant were determined by multiple regression analysis.[12] The structures and observed values were taken from the work of Agin *et al.*,[13] who derived an expression relating log(MBC) values to polarisability and ionisation potential. The quantum chemical approach led to a good correlation which could be of use in predicting local anaesthetic activity. The empirical method considered in the present investigation has the advantage of being more generally applicable, and of requiring fewer assumptions about the mode of action of the compounds.

The regression analysis was carried out by use of a variety of structure descriptors based on the 'simple pair' and 'augmented pair' fragment types.[14] A 'simple pair' is centred on each bond in each structure and consists of the bond type together with the atoms it links. An 'augmented pair' describes a larger region of the molecule by recording in addition the number of external connections at each end of the pair. Hydrogen atoms and bonds to hydrogen can almost always be inferred correctly and no explicit description of them is included. Attempts to use larger fragments failed, as in the cases studied the number of variables needed to represent the structures exceeded the number of structures.

The structures were fed into a computer as redundant connection tables and fragment generation, observation matrix generation, and regression analysis were carried out automatically.

An automatic analysis of the 39 local anaesthetics showed them to contain 16 different simple pair fragments and 43 different augmented pair fragments. In the case of the augmented pairs two groups of three and three groups of two fragments were found to have within group correlation coefficients exactly equal to one, and it was thus possible to reduce the number of variables to 36 by excluding all but one of the variables from each of these groups.

The regressions were carried out so that only fragments

[10] P. A. Lachenbruch, 'Estimation of Error Rates in Discriminant Analysis,' Ph.D. Dissertation, Univ. Southern California, Los Angeles, 1965.

[11] B. R. Kowalski and C. F. Bender, *J. Amer. Chem. Soc.*, 1972, **94**, 5632.

[12] Statistical Analysis Mark II Applications Package, International Computers Limited Technical Publication 4301, London, 1971.

[13] D. Agin, L. Hersch, and D. Holtzman, *Proc. Nat. Acad. Sci. U.S.A.*, 1965, **53**, 952.

[14] J. E. Crowe, M. F. Lynch, and W. G. Town, *J. Chem. Soc. (C)*, 1970, 990.

with coefficients significantly different from zero at the 10% level were considered. A summary of each analysis is shown in Table 1. In all cases a high correlation coefficient was obtained. Correlation by use of augmented pair descriptors gave the lower residual error and thus provided a slightly better explanation of the data ($R = 0.998$, $F = 277$, $\nu = 20$). The two correlations were shown by the F-test [15] to be significantly different at the 5% level. Following the investigation with these two fragment types, expressions including quadratic terms were considered to see if these would lead to any improvement. These higher order terms take account in a very approximate manner of the interaction between fragments. This type of investigation however was only possible when dealing with simple pair fragments, as augmented pair fragments gave rise to too many variables for regression.

First an approximate quadratic expression (2) based on simple pairs was considered which included squared terms in addition to the linear and constant terms,

$$\log(\mathrm{MBC})_i = \sum_{j=1}^{n} [b_j x_{ij} + c_j (x_{ij})^2] + \text{const.} \quad (2)$$

where b_j and x_{ij} are as defined for equation (1), and c_j is the regression coefficient of the squared term for the jth fragment.

As it was only necessary to introduce squared terms for the 10 variables for which more than two different values occurred, this gave a total of 26 variables. A summary of the analysis obtained in this case is given in Table 1, and shows that introduction of the appropriate squared terms leads to no significant improvement in the result at the 5% significance level.

TABLE 1

Summary of the results of the regression analyses by use of the different structural representations described in the text

Fragment type	Variables significant at the 10% level	Degrees of freedom	Multiple correlation coefficient	Residual error
Simple pairs	11 + const.	27	0.994	0.258
Augmented pairs	18 + const.	20	0.998	0.164
Simple pairs plus squared terms	9 + const.	29	0.993	0.280
Simple pairs (Quadratic)	13 + const.	25	0.996	0.240

The analysis based on simple pairs was finally repeated taking all quadratic terms into consideration, including cross terms, where x_{ij}, b_j, and c_j are as defined above,

$$\log(\mathrm{MBC})_i = \sum_{j=1}^{n} b_j x_{ij} + \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} d_{jk}(x_{ij}x_{ik}) + \sum_{j=1}^{n} c_j (x_{ij})^2 + \text{const.} \quad (3)$$

x_{ik} is the number of times the kth fragment occurs in the ith structure, where $j < k \leqslant n$, and d_{jk} is the coefficient for the cross product term relating to fragments j and k.

Of the 152 variables in expression (3), 69 occurred with the same frequency in every structure and could therefore be considered as constants, and another 39 were excluded as they belonged to groups of perfectly correlated variables, where it was only necessary to retain one variable from each group. This gave a total of 44 variables which was finally reduced to within the required limits by excluding an additional eight variables which belonged to groups of highly correlated variables, where the intra-group correlations exceeded 0.9. Here, as before, one variable from each group was retained. The exclusion of constants and perfectly correlated variables would have no effect on the result whilst the exclusion of highly correlated variables should affect the result only slightly. The results of the regression in this case are again shown in Table 1. The multiple correlation coefficient and residual error showed the agreement to be an improvement over that from expression (1), although no significant difference between these two correlations is indicated at the 5% level. The use of expressions (1)—(3) in the case of simple pairs, therefore, gives rise to only slight variations in the agreement between structure and property, and from a statistical point of view none of the correlations obtained differ significantly from each other at the 5% level. The correlations based on expressions (2) and (3) again differ significantly at the 5% level from the correlation obtained by using augmented pair fragments, and in the former case the difference is significant at the 1% level also. Table 1 also shows the augmented pair fragment to give the lowest residual error overall.

Thus, for the particular group of structures in question the larger, more specific fragment explains the variance in activity slightly better than a relationship based on the smaller fragment type which also considers quadratic terms. The larger fragment was subsequently used in the evaluation of the method for prediction.

More detailed results of the analysis obtained by using augmented pairs are shown in Table 2. All except two of the regression coefficients are significant at the 1% level. The coefficients show that the fragments containing carbon–carbon bonds tend to increase activity and carbon–oxygen containing fragments in general decrease activity. Fragments containing carbon and tertiary nitrogen also tend to increase activity; however, those containing carbon and primary or secondary nitrogen have coefficients which are not significantly different from zero at the 10% level and are not included in the regression. The chlorine-containing fragment gives a negative coefficient and thus increases activity although it only occurs in one structure. These results are consistent with those of other authors [13,16] who report that activity depends on the hydrophobic nature of the compound, and that aromatic groups increase activity and hydrogen-bonding groups decrease it. To test whether

[15] R. R. Sokal and F. J. Rohlf, ' Introduction to Biostatistics,' Freeman, San Francisco, 1969, p. 140.

[16] J. Zaagsma and W. Th. Nauta, *J. Medicin. Chem.*, 1974, **17**, 597.

the differences between regression coefficient values were significant expression (4) was used, where S_i and S_j are

$$S(b_i,b_j) = (S_i^2 + S_j^2 - 2s^2 C_{ij})^{\frac{1}{2}} \quad (4)$$

the standard errors of the regression coefficients b_i and b_j obtained for fragments i and j, respectively, s is the residual error of the regression and C_{ij} is the inverse

TABLE 2

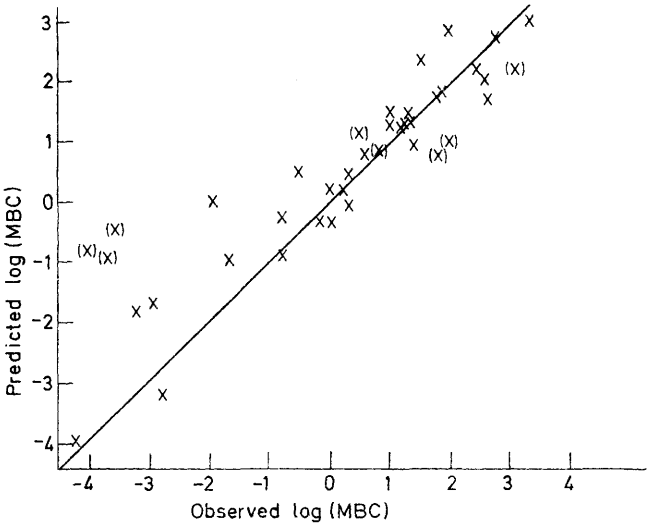
Results of the regression analysis by use of augmented pairs. The fragments are represented in the form n_aA–Bn_b where n_a and n_b are the numbers of non-hydrogen atoms bonded to atoms A and B, respectively. The symbol ⋯ is used for aromatic bonding and groups of fragments which are dependent in this sample of compounds are bracketed. The terms ring and chain given after the fragment indicate the positions of atoms A and B in the structure.

Fragment type	Regression coefficient	Student t (20 degrees of freedom)
0C–C1	−0.80	11.2
1C–C1 (chain)	−0.51	18.0
1C–C2 (chain)	−1.08	13.3
2C–C2 (chain)	−1.10	18.4
1C–C3 (ring)	−1.95	17.9
1C⋯C1 (ring)	−0.14	5.7
1C⋯C2 (ring)	−0.31	9.4
2C⋯C2 (ring)	−0.38	7.3
0C–N2	−0.53	6.4
{2C–N2 (chain) 2N=O0}	−0.68	3.4
2C–N2 (ring)	−0.58	5.6
0C–O0	0.74	4.0
1C–O0	1.24	10.1
1C–O1 (chain)	0.61	5.6
2C–O0	0.21	2.5
2C–O1 (chain)	0.25	2.0
{2C–O1 (ring) 2N–O1 (ring) 1N–N2 (ring)}	−0.57	3.0
2C–Cl0	−0.28	4.5
Regression constant	2.35	25.9

normalised cross-product matrix term relating to fragments i and j. The appropriate level of significance is then found by comparing the value $S(b_i,b_j)/s$ derived from this expression with values of Student's t distribution relevant at the number of degrees of freedom for the given regression. In this comparison of coefficient values particular attention was given to differences between like fragments in rings and chains, *e.g.* fragments 1C–C2 (chain) and 1C–C2 (ring), 2C–O1 (chain) and 2C–O1 (ring) *etc.*, and between carbon–carbon chain fragments with varying degrees of substitution. The comparison showed, however, that none of the fragment pairs differs significantly from another at the 5% significance level, or even at the 10% level. Thus, although most of the regression coefficients themselves are different from zero at the 1% significance level, the above comparisons show that it would be dangerous to make firm conclusions about the differences between pairs of fragments, even though these are mostly in agreement with expected trends.

[17] G. W. Adamson and J. A. Bush, *J. Chem. Inf. Computer Sci.*, 1975, **15**(1), 55.

By using augmented pair fragments, predictions were simulated by excluding each structure in turn from the analysis and estimating a log(MBC) value for it from the results of the regression analysis carried out on the remaining structures. Eight of the structures contained unique fragments, which meant that insufficient parameters for prediction were available from the analysis which excluded them. In this case, to obtain a prediction the missing information concerning the fragments in question may be either estimated or assumed to be zero. In the present investigation missing values were assumed to be zero, and as a result the predicted activities in these cases were not as good and led to a sum of squares ratio [where the sum of the squares of the difference between observed and predicted log(MBC) values is taken as a ratio of the sum of the squares of the deviations of the



Graph of observed *vs.* ' predicted ' log(MBC) values, with the predictions involving unique fragments in parentheses (see text for explanation)

observed values from their mean value [17]] of 0.27. The removal of these structures from the set led to a considerable improvement in the result and a sum of squares ratio of 0.13. The extent of the agreement between observed and predicted log(MBC) values is shown in the Figure, where the structures containing unique fragments are indicated in parentheses. The graph also gives the 45° line, upon which all points would lie if observed and predicted log(MBC) values were in complete agreement. The mean deviation between observed and predicted values for the group which excludes the structures containing unique fragments is 0.45 log(MBC) units, compared with a range of 6.95 and a mean deviation for observed values of 1.43.

DISCUSSION

In view of the large approximations involved and the wide range of structural types considered, the correlations obtained between structure and property are good, and estimated log(MBC) values are in reasonable agreement with observed values. The results indicate the

possible utility of the method for predicting biological activity, and when considered with the results obtained for a group of penicillins,[9] show the ability of the method to deal with both closely related and diverse structural types. The flexibility of the method in this respect may well lead to a better understanding of the contributions to activity of different structural features. The method could thus be used with suitable data to try to generate new leads.[4,16] The small size of the structural fragments used in this study leads to the loss of some information on the relative positions of groups. More information of this type could be included by use of larger structural fragments, although, because of the limitations on the use of regression analysis, the number of fragment types would have to be kept below the number of structures available. Larger fragments could be generated automatically from connection tables but the use of Wiswesser Line Notation may be easier because of the inclusion in the notation of explicit ring-substituent locant information.[18] In another study the same set of structures and properties was used to compare the performance of a number of different similarity and dissimilarity measures in the automatic classification of chemical structures.[17] The structures were classified and the performance of the similarity or dissimilarity coefficients was assessed by using them, or a classification derived from them, to predict the properties of the structures. The best set of predictions obtained gave a mean deviation of 0.79 and a sum of squares ratio of 0.34. The corresponding figures obtained from regression analysis are 0.45 and 0.27. The predictions obtained from the regression analysis are thus more accurate, and it must be expected that, where it can be applied,[19] regression analysis is capable of being a better method of prediction than classification methods. In the classification work augmented atom structural fragments were used, but this would not be expected to change the relative performance of regression and classification methods.

If the objective of an investigation is quantitative structure–property correlation or property prediction, regression analysis, where applicable, should give more accurate results than classification methods. The classification methods can present the information differently and this may be of value in itself. They also have potentially useful applications in the storage and retrieval of chemical structure information and property data.

EXPERIMENTAL

Computer programs were run on the Sheffield University ICL 1907 computer, which has a 24-bit word length and a cycle time of *ca.* 2 μs. The regression analysis was carried out by using the ICL statistical Analysis Package.[12] Supporting programs for obtaining the data in a form suitable for input to the ICL Package and for simulating property predictions by using the results of the regression analysis were written in PLAN (the ICL assembly language), FORTRAN, and ALGOL.

PLAN programs were developed to analyse connection tables and derive structural descriptors based on frequencies of occurrence of the appropriate substructural features. These descriptors were then used to set up a matrix of frequency values in a form suitable for the regression analysis package. Where necessary the PLAN programs incorporated FORTRAN subroutines for the derivation of the appropriate quadratic terms.

Predictions were batched, although it was not possible to interface the ICL statistical package with user programs and thus predictions were carried out in a stepwise manner. First, the appropriate frequency vectors for the structures to be predicted were excluded from the data matrix used as input to the regression program. The results of the regression analysis were then returned to the PLAN program used to generate the original data matrix, for calculation of the appropriate log(MBC) value. An ALGOL program was used to establish the extent of the agreement between observed and estimated log(MBC) values.

We thank Drs. M. F. Lynch and U. D. Naik for discussions and the British Library Research and Development Department, formerly the Office for Scientific and Technical Information (London), for a Postgraduate Research Studentship (to J. A. B.).

[5/1198 *Received*, 18*th June*, 1975]

[18] G. W. Adamson and D. Bawden, *J. Chem. Inf. Computer Sci.*, 1975, in the press.

[19] Ref. 15, p. 228.