Substructural Analysis Techniques for Empirical Structure-Property Correlation. Application to Stereochemically Related Molecular Properties

GEORGE W. ADAMSON* and DAVID BAWDEN

Postgraduate School of Librarianship and Information Science, University of Sheffield, Western Bank,
Sheffield, S1U 2NT, England

Received April 4, 1979

The application of a method of substructural analysis for structure-property correlation to a set of data dependent upon three-dimensional structure is described. Structural descriptors were derived automatically from Wiswesser Line Notation and correlated with molecular properties by multiple regression analysis. The technique was applied to the boiling points of derivatives of cyclohexane and dioxane giving good correlations and results interpretable in physicochemical terms. The method could be useful for an analysis of a complex data set or as a method of determining unknown property values by extrapolation.

The importance of dealing adequately with the three-dimensional structure of chemical substances in attempting structure-property correlation has long been recognized. This applies to analyses of both physicochemical properties and biological activities: in either case inter- or intramolecular interactions may play an important part. Molecular properties may be affected by simple steric bulk, or by complex conformational factors.

A number of techniques have been devised or adapted to deal with this particular problem, particularly with a view to the design of biologically active compounds. Molecular orbital[1,2] and molecular mechanics[3,4] techniques have been widely used for studies involving relatively few compounds. These techniques allow investigation of conformational factors and indication of "most favored" conformations. Computer modeling of structure to identify three-dimensional "pharmacophoric patterns" in substances with similar activity has been described.[5]

Linear free-energy studies in physical organic chemistry commonly involve the use of parameters representing steric effects.[6] In the study of biological quantitative structure-activity relationships, where the biological property of interest is correlated with physicochemical properties, a number of steric parameters have been used.[7-9] Molar refractivity in particular has been widely applied as a measure of steric bulk, and has been used to include stereoisomers within a single correlation equation.[10] These have largely used variables derived from calculated or experimental molecular dimensions. However, the use of stereochemical descriptors in the additive modeling (Free-Wilson) methodologies[13] has not been reported.

A substructural analysis methodology involving automatic derivation of structural feature descriptors from computer-readable structure representations, followed by a multiple regression or cluster analysis procedure, has been described.[14-19] The structural representations used were connection tables or Wiswesser Line Notation (WLN), both being computer-readable equivalents of a structure diagram. In neither case did the description *explicitly* represent three-dimensional structure. This type of substructural analysis is very similar to the Free-Wilson analysis but is designed to be integrated with computerized structure-property or structure activity information system.

It is evident that there must be an implicit allowance for some stereochemical factors in the type of structural features used as variables in this kind of analysis, and in similar thermochemical additivity schemes.[20] At the simplest level, atom and bond counts will give a rough indication of molecular size, and the occurrence of variables corresponding to branching points, or alternatively to chains, in a structure gives information as to molecular shape. Nevertheless, there could be value in an ability to derive structural descriptors completely denoting three-dimensional factors.

The work described below was aimed at investigating the usefulness of substructural analysis techniques for sets of structures including stereoisomers, and for molecular properties likely to depend directly on three-dimensional structure. WLN was used as the computer-readable representation of structure. Provision, albeit "tentative", is made in the WLN rules for specification of stereochemistry.[21-22]

The objective of this work was twofold: firstly, to assess to what extent explicit stereochemical descriptors are needed (i.e., how adequately the method works without them); secondly, to determine how readily stereochemical descriptors, useful for correlation, can be derived directly from WLN in its present state, or whether more elaborate stereochemical coding rules would be valuable in this respect. This work also extended the application of this type of substructural analysis to a series of saturated cyclic compounds.

METHOD

Multiple regression analysis was used, as in earlier work[14,16-19] for correlating the occurrence of structural features with property, in order to assess quantitatively any improvement brought about by use of stereochemical descriptors. This limited the type of data which could be used; much biological data, of particular interest because of the great importance of conformation in biology systems, could not be used, because relatively large numbers of compounds with accurate and consistent property values could not be obtained from published sources.

The property chosen for analysis was boiling point. Values were available for a set of 60 cyclohexane and dioxane derivatives.[23] This property depends upon intermolecular effects and upon stereochemistry. In this case, equatorial and axial substitution affects conformation and hence boiling point. This data set should provide a good test for the need and feasibility of generating explicitly three-dimensional variables from WLN. The set of compounds used, with their property values, is given in ref 23.

ANALYSIS PROCEDURE

The structures were encoded in WLN, and structural features derived by computer program. The occurrence of structural features was correlated with property by multiple regression analysis, assuming an additive model of the form:

* Author to whom correspondence should be addressed at ICI Pharmaceuticals Division, Alderley Park, Macclesfield, Cheshire, England.

0095-2338/80/1620-0097$01.00/0 © 1980 American Chemical Society

$$Y_i = \sum_{j=1}^{j=n} b_j x_{ij} + c$$

where Y_i is the property value for the ith structure, b_j is the regression coefficient for the jth feature (where there are a total of n features in the set structure), x_{ij} is the number of times the jth feature occurs in the ith structure, and c is the regression constant. This mathematical formulation corresponds to a simple assumption of additivity of contributions to property from constituent structural features. A special variable is used to represent the effect of geminal substitution, which necessarily involves one equatorial and one axial substituent.

ANALYSES OF BOILING POINTS OF ALICYCLIC STRUCTURES

The data set used comprised 29 methyl cyclohexanes and 31 methyl derivatives of 1,3-dioxane, for which correlations of structural features with boiling point had been described.[23] These structures were encoded in WLN, according to the tentative rules for describing cis–trans isomerism in such systems. Since these compounds are known to exist exclusively in the chair form,[24] and since only one type of substituent is present, it is straightforward to determine algorithmically the most stable conformation from the principles of conformational analysis[24] using the stereochemical information from WLN. For more complex cases, e.g., several types of substituent or varying conformational preferences, the stereochemical information in the present WLN rules would still be adequate for analysis.

Structural features were generated by computer program to represent equatorial and axial methyl groups, and their relative positions: geminal (gem), ortho, meta, and para. The WLN gives explicit information on the relative position of the substituents above and below the ring. A substituent was arbitrarily set to be equatorial; then from the WLN information, and the conformational analysis principle of alternation of equatorial and axial substitution above or below the ring, the orientations of the other substituents were calculated. The procedure was repeated, setting the first substituent to be axial. Since the isomer with most equatorial substituents is most stable, and methylcyclohexanes will exist almost entirely in this form,[23] this isomer was used in the derivation of structural features. For the dioxane derivatives, substituent positions relative to the ring oxygens were also represented. These structural descriptors were correlated with boiling point by the usual multiple regression procedure.

(a) Cyclohexane Derivatives. A total of eleven sets of structural features were used. These were:

Not including the equatorial/axial distinction

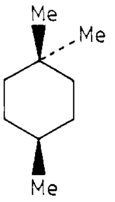
Set A	number of methyls only
Set B	including gem substitution
Set C	including gem and ortho substitution
Set D	including gem, ortho, and meta substitution
Set E	including gem, ortho, meta, and para substitution

Including the equatorial/axial distinction

Set F	number of equatorial/axial methyls
Set G	including gem substitution
Set F	including gem and ortho substitution
Set I	including gem, ortho, and meta substitution
Set J	including gem and para substitution
Set K	including gem, ortho, meta, and para substitution

Examples of derivation of structural descriptors for sets A, C, H, and K are shown in Tables I and II.

Table I. Structural Feature Derivation for Cyclohexanes[a]

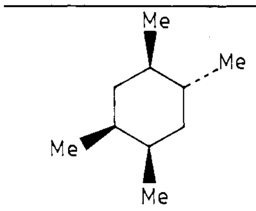


eq indicates equatorial
ax indicates axial
gem indicates geminal

	Structural Features
set A	three Me
set C	three Me one Me-gem-Me
set H	two eq Me one ax Me one eq Me-gem-ax Me
set K	two eq Me one ax Me one eq Me-gem-ax Me one eq Me-para-eq Me one ax Me-para-eq Me

[a] The ring system is not included, as it is common to all structures.

Table II. Structural Feature Derivation for Methylcyclohexanes[a]



eq indicates equatorial
ax indicates axial

	Structural Features
set A	four Me
set C	four Me two Me-ortho-Me
set H	three eq Me one ax Me one eq Me-ortho-ax Me one eq Me-ortho-eq Me
set K	three eq Me one ax Me one eq Me-ortho-ax Me one eq Me-ortho-eq Me one eq Me-ortho-ax Me one eq Me-meta-eq Me one eq Me-para-ax Me one eq Me-para-eq Me

[a] The ring system is not included as it is common to all structures.

The overall results of the regression analysis using these sets of structural descriptors are shown in Table III. Good correlations, significant at the 1% level, were obtained with all the sets. Improvements in correlation were assessed by the F test. For the sets which do not include terms representing relative positions of substituents, set A and set F, no improvement in correlation results from distinguishing equatorial and axial methyl groups. For all other pairs of sets including the same types of relative position term, e.g., set B and set G, set E and set K, the equatorial/axial distinction gives an improvement significant at the 5% level.

Inclusion of gem and ortho terms gives significant improvements in correlation, whether or not the equatorial/axial distinction is made. For the most part, inclusion of meta and para interaction gives no significant improvement in correlation, except that the most complex set, set K, shows an improvement at the 5% level over any other.

These results are generally in accord with the findings of previous workers.[23] The contributions for equatorial and axial methyl groups, and for the gem and ortho methyl groups, are similar to those of Kellie and Riddell. Significant contributions were also found for some meta and para related groups. This indicates that the major structural influences on boiling point are due to equatorial/axial effects and geminal and ortho

Table III. Regression Results for Cyclohexane Boiling Points

structural feature set	no. of structural features	no. included in analysis	degrees of freedom	multiple correlation coeff	residual error	*F* value
A	2	1 + constant	27	0.964	5.61	354.87
B	3	2 + constant	26	0.980	4.31	315.28
C	4	3 + constant	25	0.991	2.90	456.72
D	5	4 + constant	24	0.991	2.96	328.84
E	6	5 + constant	23	0.993	2.72	325.13
F	3	2 + constant	26	0.964	5.71	170.87
G	4	3 + constant	25	0.991	2.92	456.72
H	6	5 + constant	23	0.998	1.54	1146.55
I	9	8 + constant	20	0.998	1.61	623.13
J	9	8 + constant	20	0.998	1.50	623.13
K	12	11 + constant	17	0.999	1.03	771.57

number of structures = 29; range of boiling point values = 89 units

Table IV. Boiling Points of Cyclohexanes. Regression Results for Structural Feature Set K[a]

structural feature	regression coefficient	*t* statistic (17 deg of freedom)
eq Me	19.99	28.77
ax Me	25.53	22.26
eq Me-gem-ax Me	−5.97	8.30
ax Me-ortho-eq Me	3.05	4.99
eq Me-ortho-eq Me	3.04	5.34
eq Me-meta-eq Me	−0.79	1.69
ax Me-meta-eq Me	−2.35	4.25
ax Me-meta-ax Me	−1.32	1.46
ax Me-para-eq Me	−2.48	4.06
eq Me-para-eq Me	−1.97	3.52
ax Me-para-ax Me	−0.73	0.67
regression constant	80.91	87.31

[a] The ring system is common to all structures. *t* (17df) = 2.9 at the 1% level.

substitution. It is, however, worth noting that some significant improvement in correlation is brought about by the use of more complex descriptor sets, indicating the potential value of this sort of automated procedure in handling complex data sets.

The structural features of set K are shown in Table IV, together with their regression coefficient and *t* statistics. The contributions to boiling point of equatorial and axial methyls, and geminal and ortho effects are, as noted, similar to those of previous workers.[23]

(b) Dioxane Derivatives. The analysis of the set of dioxane derivatives was carried out using similar sets of structural features to those used for cyclohexane derivatives. These structural features differ from those of Kellie and Riddell, in that the relative position of pairs of methyl groups are used without taking account of their relationship to the oxygen atoms. All the multiple regression analyses gave highly significant correlations. The specification of position of methyl substituents relative to heteroatoms (as shown below) gave improvements significant at the 1% level. Equatorial–axial effects, and gem and ortho interactions were, as with the cyclohexanes, predominant structure influences on property. An analysis was carried out distinguishing between 2–4 and 4–6 meta substitution, i.e., with the methyls separated by C–O–C and C–C–C, respectively. A worse correlation resulted, indicating the inappropriateness of this distinction for this data set.

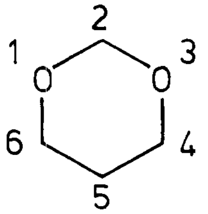


Table V. Regression Coefficients for Dioxane Boiling Points[a]

structural feature	regression coefficient	*t* statistic (15 deg of freedom)
2-eq Me	5.57	4.65
2-ax Me	21.92	13.49
4-eq Me	12.28	12.63
4-ax Me	24.39	16.74
5-eq Me	14.86	12.00
5-ax Me	15.15	9.68
eq Me-gem-ax Me	−6.92	7.30
eq Me-ortho-eq Me	2.13	2.92
eq Me-ortho-ax Me	2.91	4.14
ax Me-ortho-ax Me	5.42	4.03
eq Me-meta-eq Me	−1.54	2.27
eq Me-meta-ax Me	−3.27	4.32
eq Me-para-eq Me	1.32	1.07
eq Me-para-ax Me	−2.37	2.06
ax Me-para-ax Me	1.00	0.49
regression constant	103.77	90.52

[a] The ring system and O-meta-O feature is common to all structures. *t* (15df) = 2.95 at the 1% level.

The analysis using the most complex structure feature set, including the equatorial and axial distinction and gem, ortho, meta, and para interactions, with a total of 17 variables, gave a good correlation: $R = 0.998$, $r = 1.36$, df = 15, $F = 249.3$. The structural features of this set with their regression coefficients and *t* statistics are shown in Table V. The values are broadly in agreement with those of the original investigators.[23] The difference between contribution to boiling point of equatorial and axial groups for 2- and 4-substituents is noteworthy. Not surprisingly, therefore, the simple correlation using only the number of methyls, and not indicating position, is significantly worse than the best regression: $R = 0.859$, $r = 8.28$, df = 29, $F = 81.64$. Even so, this simple model does give a significant correlation.

(c) Cyclohexane and Dioxane Derivatives. The sets of cyclohexanes and dioxanes were combined to give a set of 60 compounds with a range of boiling points of 89°. An analysis was carried out with a set of 19 structural descriptors including equatorial–axial distinction and all relative position terms. The ring oxygens of the dioxanes were treated as substituents. A significant correlation was obtained: $R = 0.990$, $r = 3.14$, df = 40, $F = 103.69$, with coefficient generally in accord with the factors discussed above. This indicates the feasibility of studying sets of more diverse structures using variables of the sort. This method of accounting for the substitution of heteroatoms in rings was also used in a study of the p*K* values of some heterocyclic compounds.[19]

DISCUSSION

Significant correlations were obtained using structural features which did not take explicit account of three-dimensional structural information. However, significant improvements were usually obtained when three-dimensional generated features were used as variables. The results using automatically notation-based fragments were similar to those obtained by previous workers using manually generated features. This shows that the stereochemical descriptors derived from WLN are adequate for this purpose and that correlation studies or substructural analysis including explicit three-dimensional information is feasible. Although this study used WLN, the same information could be included in connection tables. In this type of analysis the ratio of structural features to structures, when using more detailed structural features, is sometimes higher than desirable on statistical grounds. In these cases the results should be treated with more caution.

The regression coefficients and constants may be used for extrapolation. A balance must then be struck between using detailed structural descriptors, which produce more accurate prediction, but require a large data base, and using simpler descriptors, which produce less accurate prediction, but require a smaller data base.

EXPERIMENTAL

The programs for fragmentation of WLN were written in ICL COBOL, and run on the University of Sheffield ICL 1907E computer. The multiple regression analyses were carried out using the ICL XDS3 statistical package.

ACKNOWLEDGMENT

We thank Professor M. F. Lynch, Dr. G. E. Vleduts, Dr. J. A. Bush, and Dr. P. Willett for helpful discussions and the Department of Education and Science (London) for the award of an Information Science studentship to D. B.

REFERENCES AND NOTES

(1) B. Pullman, Ed., "Quantum Mechanics of Molecular Conformation", Wiley, New York, 1976.

(2) L. B. Kier, "Molecular Orbital Theory in Drug Research", Academic Press, New York, 1971.

(3) E. M. Engler, J. D. Anduse, and P. von R. Schleyer, "Critical Evaluation of Molecular Mechanics", *J. Am. Chem. Soc.*, **95**, 8005–8025 (1973).

(4) N. L. Allinger, "Calculation of Molecular Structure and Energy by Force Field Methods", *Adv. Phys. Org. Chem.*, **13**, 1–82 (1976).

(5) P. Gund, W. T. Wipke, and R. Langridge, "Computer Searching of a Molecular Structure File for Pharmacophoric Patterns", Proceedings of an International Conference on Computers in Chemical Research and Education, Zagreb, 1973.

(6) J. Shorter, "The Separation of Polar, Steric, and Resonance Effects by the Use of Linear Free Energy Relationships", in "Advances in Linear Free Energy Relationships", N. B. Chapman and J. Shorter, Eds., Academic Press, New York, 1972, Chapter 2.

(7) S. H. Unger and C. Hansch, "Quantitative Models of Steric Effects", *Prog. Phys. Org. Chem.*, **12**, 91–118 (1976).

(8) Z. Simon and Z. Szabadai, "Minimal Steric Differences Parameter and the Importance of Steric Fit for Structure–Biological Activity Correlations", *Stud. Biophys.*, **39**, 123–132 (1973).

(9) A. Verloop, W. Hoogenstraaten, and J. Tipher, "Development and Application of New Steric Substituent Parameters in Drug Design", in "Drug Design", Vol. 7, E. J. Ariens, Ed., Academic Press, New York, 1976.

(10) M. Yoshimoto and C. Hansch, "Quantitative Structure–Activity Relationship of D- and L-*N*-Acyl-α-aminoamide Ligands Binding to Chymotrypsin. On the Problem of Combined Treatment of Stereoisomers", *J. Org. Chem.*, **41**, 2269–2273 (1976).

(11) W. E. Brugger, A. J. Stuper, and P. C. Jurs, "Generation of Descriptors from Molecular Structures", *J. Chem. Inf. Comput. Sci.*, **16**, 105–110 (1976).

(12) D. S. Dierdorf and B. R. Kowalski, "Three-Dimensional Molecular Structure–Biological Activity Correlation by Pattern Recognition", Report AD-785 863, 1974.

(13) S. M. Free and J. W. Wilson, "A Mathematical Contribution to Structure–Activity Studies", *J. Med. Chem.*, **7**, 395–399 (1964).

(14) G. W. Adamson and J. A. Bush, "Method for Relating the Structure and Properties of Chemical Compounds", *Nature (London)*, **248**, 406–408 (1974).

(15) G. W. Adamson and J. A. Bush, "A Comparison of the Performance of Some Similarly and Dissimilarly Measures in the Automatic Classification of Chemical Structures", *J. Chem. Inf. Comput. Sci.*, **15**, 55–58 (1975).

(16) G. W. Adamson and J. A. Bush, "The Evaluation of an Empirical Structure–Activity Relationship for Property Prediction in a Structurally Diverse Group of Local Anaesthetics", *J. Chem. Soc. A*, 168–172 (1976).

(17) G. W. Adamson and D. Bawden, "A Method of Structure–Activity Correlation Using Wiswesser Line Notation", *J. Chem. Inf. Comput. Sci.*, **15**, 215–220 (1975).

(18) G. W. Adamson and D. Bawden, "An Empirical Method of Structure–Activity Correlation for Polysubstituted Cycle Compounds Using Wiswesser Line Notation", *J. Chem. Inf. Comput. Sci.*, **16**, 161–165 (1976).

(19) G. W. Adamson and D. Bawden, "A Substructural Analysis Method for Structure–Activity Correlation of Heterocycle Compounds Using Wiswesser Line Notation", *J. Chem. Inf. Comput. Sci.*, **17**, 164–171 (1977).

(20) G. Janz, "Thermodynamic Properties of Organic Compounds", Academic Press, New York, 1967.

(21) P. A. Baker, G. Palmer, and P. W. L. Nichols, "The Wiswesser Line Formula Notation", in "Chemical Information Systems", J. E. Ash and E. Hyde, Eds., Ellis Horwood, Chichester, 1975, Chapter 9.

(22) E. G. Smith and P. A. Baker, "The Wiswesser Line Formula Chemical Notation", 3rd ed, Chemical Information Management Inc., Cherry Hill, N.J., 1976.

(23) G. M. Kellie and F. G. Riddell, "The von Auwers Boiling Point Rule. A New Approach", *J. Chem. Soc. A*, 740–744 (1975).

(24) E. L. Eliel, "Stereochemistry of Carbon Compounds", McGraw-Hill, New York, 1962.