(7) G. W. Adamson et al., "Relationship between Query and Data-Base Microstructure in General Substructure Search Systems", *J. Chem. Doc.*, **13**, 133 (1973).

(8) G. W. Adamson et al., "Distributions of Fragment Representations in a Chemical Substructure Search Screening System", *J. Chem. Doc.*, **14**, 72–74 (1974).

(9) V. H. R. Bragg et al., "The Use of Molecular Formula Distribution Statistics in the Design of Chemical Structure Registry Systems", *J. Chem. Doc.*, **10**, 125 (1970).

(10) A. J. Feldman and L. Hodes, "An Efficient Design for Chemical Structure Searching. I. The Screens", *J. Chem. Inf. Comput. Sci.*, **15**, 147–152 (1975).

(11) M. F. Lynch in "Computer Representation and Manipulation of Chemical Information", W. T. Wipke et al., Ed., Wiley, New York, N.Y., 1974, pp 31–53.

(12) E. Meyer, "Topological Search for Classes of Compounds in Large Files—even of Markush Formulas—at Reasonable Machine Cost", in ref 11, pp 105–122.

(13) AA = Augmented Atoms (see caption to Figure 10).

(14) Every 4th of the 40 000 CBAC structures and every 36th connection table contained in the 1972 Registry file (1.9 million structures at that time).

(15) Candidates = structures resulting from the mask search. Hits = structures resulting from the iterative search.

(16) "System Documentation for the Chemical Abstracts Service Registry System", Chemical Abstracts Service, Columbus, Ohio, 1968.

(17) Obtained from the REG/CAN file by Computer Output on Microfilm (COM).

(18) Polymer structures recorded in Registry II cannot be handled by our iterative search.

# An Empirical Method of Structure–Activity Correlation for Polysubstituted Cyclic Compounds Using Wiswesser Line Notation

GEORGE W. ADAMSON* and DAVID BAWDEN

Postgraduate School of Librarianship and Information Science, University of Sheffield, Western Bank,
Sheffield, S10 2TN, England

Received February 17, 1976

A method of substructural analysis for structure–property correlation and property prediction allowing representation of the effects of positional isomerism and substituent interaction is described. Rate constants for the bromination of 44 substituted benzenes are correlated by means of multiple regression analysis using sets of structural features derived automatically from Wiswesser Line Notation. The best set of structural features gives a multiple correlation coefficient >0.999. Property predictions are simulated for 24 compounds, with up to 5 substituents. The technique could be carried out automatically with large machine-readable structure–property files, and may be generally applicable to the properties of substituted cyclic compounds.

The correlation of properties with chemical structure and the consequent prediction of unknown values has long been a major goal of physical organic chemistry and is currently of considerable importance in such fields as drug design.[1] Several approaches to this problem have been employed, varying from purely empirical to highly theoretical.

Quantum mechanical methods have been applied to practical problems[2] but have not yet been applied very widely to the correlation of structure and biological activity. The recently developed MINDO/3 methodology[3] has been suggested to be of wider applicability, while calculations based on molecular mechanics[4] have been used successfully in some cases.

Semiempirical correlation methods, generally known as linear free energy relationships, have been very widely employed. The well-known Hammett equation and its derivatives[5,6] have been principally applied to structure–reactivity correlation for organic compounds. Its main successes have been in summarizing and clarifying experimental data and in aiding the elucidation of reaction mechanisms, though some property predictions have been made using such methods.[7] The Hansch methodology,[8] which aims to correlate biological activities with physicochemical molecular properties, has also been widely used.

Empirical relationships have long been used for the estimation of unknown property values, particularly for thermodynamic quantities.[9] More recently statistical modelling has been employed for the prediction of biological properties,[10] while pattern recognition techniques have qualitatively predicted Hammett values.[11]

A range of methods generally described as "substructural analysis", involving the correlation by computerized statistical analysis of structural features with property values, have enabled both qualitative and quantitative predictions of various biological properties.[12,13] Such methods have two major advantages: they are applicable to compounds of diverse structural type as well as to series of structurally similar compounds, and they can be used with large computer-based files containing chemical structures and property data,[14] with automatic derivation of appropriate structural features from computer-readable structural representations.[15] Thus, a recent example of substructural analysis[13c] gave simulated property predictions for a group of structurally diverse local anaesthetics by regression analysis, using structural features automatically derived from connection tables. As yet, however, no systematic method of general applicability has been developed for representing positional isomerism adequately for such analyses. Structural features derived from Wiswesser Line Notation (WLN) and used for structure–property correlation have included substituent patterns as part of the features representing ring systems,[13d] while structural features automatically derived from connection tables for application in information retrieval have included substituents as part of ring system features.[16] A recent study of boiling point variation within homologous series[17] used structural features which allowed for steric and dipolar interactions between identical substituents on a ring system, and obtained very high correlations which enabled simulated assignment of unknown stereochemistries, showing that chemical insight may be gained from such empirical methods.

It is evident that the ability to derive structural features so as to take account both of the position of a substituent relative

* To whom correspondence should be addressed.

Table I. Observed, Estimated, and Predicted Rate Constants for Substituted Benzenes

| Substituents | Reaction site position | Log $k^a$ obsd | Log $k^b$ estd | Log $k^c$ predicted |
|---|---|---|---|---|
| H | 1 | −5.569 | −5.730 | −6.03 |
| 1-Me | 2 | −2.745 | −2.550 | −2.40 |
| 1-Me | 3 | −4.328 | −4.413 | −4.53 |
| 1-Me | 4 | −1.959 | −1.844 | −1.75 |
| 1,3-$Me_2$ | 4 | 0.903 | 1.070 | 1.14 |
| 1,4-$Me_2$ | 2 | −1.469 | −1.454 | −1.43 |
| 1,2,3-$Me_3$ | 4 | 2.182 | 2.053 | 1.93 |
| 1,3,5-$Me_3$ | 2 | 3.954 | 3.719 | 3.34 |
| 1,2,3,4-$Me_4$ | 5 | 2.663 | 2.717 | 2.76 |
| 1,2,3,5-$Me_4$ | 4 | 4.439 | 4.481 | 4.49 |
| 1,3,5-$Me_3$-2-Et | 4 | 4.288 | 4.288 | |
| 1,3,5-$Me_3$-2-Cl | 4 | 0.447 | 0.447 | |
| 1,3,5-$Me_3$-2-Br | 4 | 0.531 | 0.531 | |
| 1,2,3,4,5-$Me_5$ | 6 | 4.954 | 4.978 | 5.04 |
| 1-MeO | 4 | 3.980 | 4.131 | 4.31 |
| 1-MeO | 2 | 2.300 | 2.265 | 2.22 |
| 1-MeO-2-Me | 4 | 4.685 | 4.616 | 4.57 |
| 1-MeO-2-Me | 6 | 2.837 | 2.750 | 2.65 |
| 1-MeO-3-Me | 4 | 5.757 | 5.766 | |
| 1-MeO-4-Me | 2 | 2.922 | 2.839 | |
| 1-MeO-2-F | 4 | 1.586 | 1.586 | |
| 1-MeO-3-F | 4 | 3.157 | 3.157 | |
| 1-MeO-3-F | 6 | 2.799 | 2.799 | |
| 1-MeO-4-F | 2 | −0.173 | −0.173 | |
| 1-MeO-2-Cl | 4 | 1.399 | 1.399 | |
| 1-MeO-3-Cl | 4 | 2.766 | 2.766 | |
| 1-MeO-3-Cl | 6 | 1.570 | 1.570 | |
| 1-MeO-4-Cl | 2 | −0.447 | −0.447 | |
| 1-MeO-2-Br | 4 | 1.635 | 1.635 | |
| 1-MeO-3-Br | 6 | 1.162 | 1.162 | |
| 1-MeO-4-Br | 2 | −0.404 | −0.404 | |
| 1-MeO-3-I | 6 | 1.129 | 1.129 | |
| 1,4-$(MeO)_2$ | 2 | 2.315 | 2.315 | |
| 1-MeO-2,3-$Me_2$ | 4 | 6.010 | 6.084 | 6.16 |
| 1-MeO-2,4-$Me_2$ | 6 | 2.816 | 3.059 | 3.49 |
| 1-MeO-2,5-$Me_2$ | 4 | 6.190 | 6.030 | 5.79 |
| 1-MeO-3,4-$Me_2$ | 6 | 5.174 | 5.014 | 4.76 |
| 1-MeO-3,5-$Me_2$ | 4 | 7.140 | 7.136 | 7.13 |
| 1-MeO-3,5-$Me_2$ | 6 | 5.854 | 5.977 | 5.85 |
| 1-OH | 4 | 4.608 | 4.608 | - |
| 1-$NMe_2$ | 4 | 8.336 | 8.311 | 8.18 |
| 1-$NMe_2$-3-Me | 4 | 8.992 | 9.041 | 9.07 |
| 1-$NMe_2$-3-Br | 4 | 7.772 | 7.772 | - |
| 1-$NMe_2$-3,5-$Me_2$ | 4 | 9.531 | 9.506 | 9.75 |

[a] Data from ref 19. [b] By the method described here, including all compounds in the regression. [c] By the hold-one-out method described here.

to a reaction site and of interaction between substituents is of considerable importance, if substructural analysis methods are to be of practical use. This is most easily achieved using WLN as the structural representation, since the ring locants permit a very convenient assignment of relative positions. The work described below involves the application of such a direct method for structure–property correlation and property prediction to a set of data involving position-dependent substituent effects and interactions between several substituents of different types, a situation which has caused problems for other correlation procedures.[18]

## METHOD

The data used were taken from a study[19] of substituent effects on the rates of electrophilic bromination of 44 substituted benzenes, with up to five substituents in each compound, listed in Table I. The structures were coded manually in WLN, and sets of structural features of varying complexity were derived by computer program. These were correlated with log $k$ by multiple regression analysis, using a computer manufacturer's statistical analysis package.[20] Log $k$ was assumed to be an additive function of the structural features present, so that its value for the $i$th compound is given by

$$\log k_i = \sum_{j=1}^{n} b_j x_{ij} + \text{constant}$$

where there are a total of $n$ types of structural feature in the set of compounds, and $x_{ij}$ is the number of times that the $j$th feature occurs in the $i$th structure. The regression coefficient for the $j$th feature, $b_j$, represents the effect of that structural feature in increasing (positive coefficient) or decreasing (negative coefficient) the reactivity of those compounds in which it occurs.

Three sets of structural features were tested in order of increasing complexity:

Set A log $k$ was assumed to be affected only by the type and number of substituents.

Set B log $k$ was assumed to be affected by the type and number of substituents, and by the positions of the substituents relative to the reaction site.

Set C log $k$ was assumed to be affected by type, number, and position relative to reaction site of all substituents, and also by the interaction between each pair of substituents.

Several examples of these feature derivations are shown in Table II. The benzene ring was not included as a structural feature, since it is common to all structures.

Each interaction between the substituents was accounted for by a single term, e.g., a compound with Br meta to Me would have the term Br-*m*-Me assigned, rather than two terms Br-*m*-Me and also Me-*m*-Br. Use of multiple terms would unnecessarily increase the number of structural features, without affecting the correlation or predictions achieved.

The regression analyses were carried out in a stepwise fashion, with structural fragments being included in decreasing order of their pivot elements,[20] i.e., approximately in decreasing order of the magnitude of their effect on the variation in the observed log $k$ values. Some structural features from set C were not included in the regression, as they had no effect, within the accuracy of the calculation on log $k$. These included "perfectly correlated" structural features, i.e., features occurring only together in a fixed ratio in the same structures. Only one feature from each perfectly correlated group was included in the calculation.

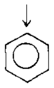
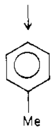
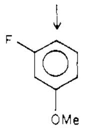
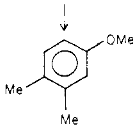
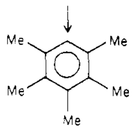
The results of the three regressions are summarized in Table III.

Set B gave a regression with a lower residual error than set A and was shown to be significantly different at the 1% level by the F-test, and hence can be said to be statistically superior. Set C was found in the same way to give a statistically superior regression to that with set B. For this reason set C was used for the subsequent simulation of property prediction.

The structural features incorporated in set C are shown in Table IV, together with the values of the regression coefficients and the $t$ statistics, which give a measure of the significance of the coefficient values. Overall the results are in accordance with the generally accepted reaction mechanism.[21] The positional effects on substituent activation and deactivation are clearly demonstrated, as are the effects of substituent interactions.

It is evident from the effects of the individual substituents, as measured by the regression coefficients, that various substituent interaction mechanisms are important. Inductive effects cause activation by alkyl groups and deactivation by halogens, while mesomeric effects are shown in the strong activation by methoxy and dimethylamino groups, and in the

**Table II.** Examples of Structural Feature Derivation

| Structure[a] | WLN | Reaction site locant | Structural features Set A | Set B | Set C |
|---|---|---|---|---|---|
|  | R | A | None | None | None |
|  | 1R | D | One Me | One *p*-Me | One *p*-Me |
|  | FR CO1 | F | One F<br>One OMe | One *o*-F<br>One *p*OMe | One *o*-F<br>One *p*-OMe<br>One F-*m*-OMe interaction |
|  | 1OR C1 D1 | F | One OMe<br>Two Me | One *o*-OMe<br>One *m*-Me<br>One *p*-Me | One *o*-OMe<br>One *m*-Me<br>One *p*-Me<br>One Me-*o*-Me<br>One Me-*m*-OMe interactions<br>One Me-*p*-OMe |
|  | 1R B1 C1 D1 E1 | F | Five Me | Two *o*-Me<br>Two *m*-Me<br>One *p*-Me | Two *o*-Me<br>Two *m*-Me<br>One *p*-Me<br>Four Me-*o*-Me<br>Four Me-*m*-Me interactions<br>Two Me-*p*-Me |

[a] ↓ indicates the reaction site.

**Table III.** Summary of Regression Results[a]

| Structural features | No. of structural features | No. of variables included in regression | Degrees of freedom | Multiple correlation coefficient | *F* values | Residual error |
|---|---|---|---|---|---|---|
| Set A | 9 | 9 + const | 34 | 0.905 | 17.09 | 1.587 |
| Set B | 19 | 19 + const | 24 | 0.987 | 47.70 | 0.702 |
| Set C | 44 | 31 + const | 12 | >0.999 | 196.16 | 0.175 |

[a] 44 structures were included in each regression.

effects of a substituent in the ortho and para compared with the meta positions. The reduced activation for substituents with positive coefficients in the ortho position, compared with the same substituents in the para position, is likely to be at least partly due to steric factors. The negative coefficients for structural features involving the simultaneous presence of two activating substituents show the effect of interactions between electron-releasing groups which are strongest for the dimethylamino and methoxy groups.[19]

The significance of the difference between coefficients was tested using the formula

$$t(b_1,b_2)=\sqrt{S_1^2+S_2^2-2r^2C_{12}}\,/r$$

where $S_1$ and $S_2$ are the standard errors of the coefficients, $r$ is the residual error of the regression, and $C_{12}$ is the corresponding term from the inverse cross-product matrix. The value of $t(b_1,b_2)$ is compared with values in tables of Student's $t$ distribution with the same number of degrees of freedom as the regression. None of the pairs of coefficients differed significantly at the 5% level, which is often taken as the limit of significance for statistical purposes. Thus, it would be unwise to attempt to draw very firm conclusions from a comparison of any single pair of coefficient values, though trends may be treated with more confidence. This ability to test statistically the difference in coefficient values, i.e., the difference in substituent effects, is a particularly useful feature of this correlation method.

The estimated values from the regression are listed in Table I. It is evident that the correlation for polysubstituted compounds is not inferior to that for mono- and disubstituted.

In order to simulate the prediction of an unknown property value, the "hold-one-out" method[1,22] was used. This involved omitting one compound at a time from the regressions and carrying out an analysis on the remaining compounds in the set, then deriving its log $k$ value by summing the regression coefficients for the structural features present in that compound, as in the example in Table V, i.e., a predicted value for log $k$ of 5.04, of observed value 4.95.

The regression coefficients differ from those in Table IV owing to the effect of omitting this compound from the analysis.

This procedure was possible for only 24 of the 44 compounds, since the remaining compounds contained unique structural features, for which no regression coefficient was obtainable when that compound was omitted from the analysis. In a practical situation, this problem could be solved approximately by estimating a value for such coefficients.

The predicted values of log $k$ are listed in Table I. Polysubstituted compounds are dealt with as adequately as simpler structures. Measures of discrepancies between predicted and observed values were calculated as below:

Mean discrepancy

$$\frac{\sum_{i=1}^{n}|x_i-\hat{x}_i|}{N}=0.24$$

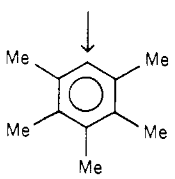
i.e., mean discrepancy between observed and predicted values.

**Table IV.** Regression Coefficients for Structural Features in Set C

| Structural feature | Regression coefficient | $t$ statistic[a] | Perfectly correlated structural features |
|---|---|---|---|
| *o*-Me | 3.179 | 23.81 | |
| *m*-Me | 1.317 | 8.08 | |
| *p*-Me | 3.886 | 26.96 | |
| *m*-Et | 0.569 | 2.56 | |
| *o*-OMe | 7.995 | 43.55 | Me-*p*-Et, Me-*o*-Et |
| *m*-OMe | 0.050 | 0.23 | OMe-*p*-OMe |
| *p*-OMe | 9.861 | 53.47 | |
| *p*-$NMe_2$ | 14.042 | 63.03 | |
| *o*-Cl | −0.670 | 2.34 | |
| *m*-Cl | −2.712 | 12.48 | |
| *p*-Cl | Excluded by regression program | | |
| *o*-Br | −0.539 | 2.27 | Br-*m*-$NMe_2$ |
| *m*-Br | −2.669 | 12.29 | |
| *p*-Br | −1.103 | 5.08 | Br-*m*-OMe |
| *o*-F | −1.508 | 5.26 | |
| *m*-F | −2.438 | 11.22 | |
| *p*-F | Excluded by regression program | | |
| *p*-I | −1.136 | 5.23 | I-*m*-OMe |
| *p*-OH | 10.338 | 45.97 | |
| Me-*o*-Me | −0.167 | 1.90 | |
| Me-*m*-Me | −0.265 | 3.13 | |
| Me-*p*-Me | −0.221 | 1.75 | |
| Me-*o*-OMe | −0.832 | 4.66 | |
| Me-*m*-OMe | −1.544 | 13.60 | |
| Me-*p*-OMe | −0.743 | 3.69 | |
| Cl-*o*-Me | −0.280 | 1.78 | Cl-*p*-Me |
| Br-*o*-Me | −0.259 | 1.65 | Br-*p*-Me |
| Cl-*o*-OMe | −0.020 | 0.07 | |
| Cl-*m*-OMe | −0.695 | 3.20 | |
| Cl-*p*-OMe | Excluded by regression program | | |
| Br-*o*-OMe | 0.173 | 0.60 | |
| Br-*p*-OMe | Excluded by regression program | | |
| F-*o*-OMe | −0.107 | 0.37 | |
| F-*m*-OMe | 0.534 | 2.46 | |
| F-*p*-OMe | Excluded by regression program | | |
| Me-*m*-$NMe_2$ | −2.449 | 15.17 | |
| Regression constant | −5.730 | 40.54 | |

[a] Twelve degrees of freedom.

**Table V**

| | Structural feature | Multi-plicity | Regression coeff × multiplicity |
|---|---|---|---|
|  | *o*-Me | 2 | 3.165 × 2 |
| | *m*-Me | 2 | 1.298 × 2 |
| | *p*-Me | 1 | 3.869 × 2 |
| | Me-*o*-Me | 4 | −0.155 × 4 |
| | Me-*m*-Me | 4 | −0.254 × 4 |
| | Me-*p*-Me | 2 | −0.201 × 2 |
| | Regression constant | | −5.178 × 1 |
| | Summation | | 5.039 |

Sum of squares ratio

$$\frac{\sum_{i=1}^{n}(x_i-\hat{x}_i)^2}{\sum_{i=1}^{N}(x_i-\bar{x}_i)^2}=0.0057$$

i.e., ratio of sum of squares of discrepancies between observed and predicted values and sum of squares of discrepancies between observed and mean observed values. The range of observed values is −5.57 to 9.53. $N$ is the number of structures, i.e., 24, $x_i$ is the observed property value, $\bar{x}_i$ is the mean observed property value, and $x_i$ is the predicted property value.

## DISCUSSION

The procedure described here has several characteristics which make it suitable as a widely applicable method of structure–property correlation and property prediction. The use of multiple regression, a well-established method, gives the analysis a good statistical basis and allows for significance tests on the overall correlation, on individual coefficients, and on the differences between coefficients. The results of the analysis are in a form which enables ready interpretation in chemical terms. That good correlations and a satisfactory explanation of the results should be possible in this case is not perhaps surprising, in view of the well-known importance of substituent effects on this reaction. However, it well illustrates the usefulness of this method of correlation and prediction in dealing with the interactions of different types of substituent.

It is similar to the Hammett approach in that it produces coefficients which are in principle transferrable to other reaction series, and may be compared with other examples of the statistical analysis of reactivity data.[23] It differs in that this method is designed primarily for the automatic processing of structure–property data. In addition, this type of analysis considers overall group contributions, rather than partitioning into, e.g., steric and electronic factors, as is common practice with the Hammett equation.[24] However, it appears to be at least as successful as LFER treatments in accounting for the overall effects of multisubstitution.

It is evisaged that this kind of substructural analysis will be used mainly for empirical structure–activity correlation and property prediction, particularly for biological properties, rather than for the examination of reaction mechanisms. Since this method involves WLN which is widely used in industrial information systems,[25] it is applicable to any cyclic structure, to a wide variety of properties, and is computationally economical. It would be well suited for application to large computer-based structure–property files[14] and could have value in lead generation or optimization in drug design.

## EXPERIMENTAL SECTION

The programs were run on the University of Sheffield ICL 1907 computer. The WLN fragmentation program was written in ICL COBOL and required 10K words of core storage, with CPU times ≤ 45 sec. The WLN strings representing substituents were stored with their locants, and relative positions were derived by considering the pairs of locants for each pair of substituents. Thus pairs of substituents with locants "A" and "D", "B" and "E", or "C" and "F" are evidently "para". The reaction site was assigned a WLN locant, in order to derive its position relative to substituents. The substituent groups were not broken down into smaller units. This procedure is applicable to all six-membered ring structures. For other ring systems an appropriate dictionary of locant pairs could be used, or a more general algorithm based on the spanning-tree could be utilized.

The multiple regression analyses were performed using the ICL statistical analysis package. Core storage required was 20K words, with CPU times ≤ 19 sec.

## ACKNOWLEDGMENTS

We wish to thank Professor M. F. Lynch, Drs. U. D. Naik and G. E. Vleduts, and Miss J. A. Bush for helpful discussions, and the Department of Education and Science (London) for the award of a Postgraduate Research Studentship to D. Bawden.

## REFERENCES AND NOTES

(1) (a) G. Redl, R. D. Cramer, and C. E. Berkoff, "Quantitative Drug Design", *Chem. Soc. Rev.*, 3, 273–292 (1974); (b) F. D. Kover, "Structure-Activity Correlation Bibliography", NTIS Report, PB-240 658, 1975.

(2) (a) L. B. Kier, "Molecular Orbital Theory in Drug Research", Academic Press, New York, N.Y., 1971; (b) L. Farnell, W. G. Richards, and C. R. Ganellin, "Conformation of Histamine Derivatives. 5. Molecular Orbital Calculation of the $H_1$-Receptor "Essential" Conformation of Histamine", *J. Med. Chem.*, **18**, 662–666 (1975).

(3) (a) R. C. Bingham, M. J. S. Dewar, and D. H. Lo, "Ground States of Molecules. XXV. MINDO/3. An Improved Version of the MINDO Semiempirical SCF–MO Method", *J. Am. Chem. Soc.*, **97**, 1285–1293 (1975); (b) M. J. Dewar, "Prediction of Properties and Behaviour of Materials", NTIS Report, AD-A003 698, 1974.

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

(5) O. Exner, in "Advances in Linear Free Energy Relationships", N. B. Chapman and J. Shorter, Ed., Plenum Press, London, 1972, Chapter 1, p 1.

(6) J. Hine, "Structural Effects on Equilibria in Organic Chemistry", Wiley, New York, N.Y., 1975, Chapter 3, p 55.

(7) O. Exner, ref 5, p 46.

(8) C. Hansch, in "Drug Design", Vol. 1, E. J. Ariens, Ed., Academic Press, New York, N.Y., 1971, Chapter 2, p 271.

(9) G. J. Janz, "Thermodynamic Properties of Organic Compounds", Academic Press, London, 1967; (b) J. Hine, ref 6, Chapter 1, p 1.

(10) (a) S. M. Free and J. W. Wilson, "A Mathematical Contribution to Structure-Activity Studies", *J. Med. Chem.*, **7**, 395–399 (1964); (b) P. J. Harrison, "A Method of Cluster Analysis and Some Applications", *J. Appl. Statistics*, **17**, 226–236 (1968).

(11) J. R. Koskinen and B. R. Kowalski, "Structure-Reactivity Correlations for Organic Molecules by Pattern Recognition", NTIS Report AD-785 913, 1974.

(12) (a) S. A. Hiller, V. E. Golender, A. B. Rosenblit, L. A. Rastrigin, and A. B. Glaz, "Cybernetic Methods of Drug Design. 1. Statement of the Problem – The Perceptron Approach", *Comp. Biomed. Res.*, **6**, 411–421 (1973); (b) B. R. Kowalski and C. F. Bender, "The Application of Pattern Recognition to Screening Prospective Anticancer Drugs. Adenocarcinoma 755 Biological Activity Test", *J. Am. Chem. Soc.*, **96**, 916–918 (1974); (c) R. D. Cramer, G. Redl, and C. E. Berkoff, "Substructural Analysis. A Novel Approach to the Problem of Drug Design", *J. Med. Chem.*, **17**, 533–535 (1974); (d) A. J. Stuper and P. C. Jurs, "Classification of Psychotropic Drugs as Sedatives or Tranquilizers Using Pattern Recognition Techniques", *J. Am. Chem. Soc.*, **97**, 182–187 (1975); (e) K. C. Chu, R. J. Feldman, M. B. Shapiro, G. F. Hazard, and R. I. Greran, "Pattern Recognition and Structure-Activity Relationship Studies", *J. Med. Chem.*, **18**, 539–545 (1975).

(13) (a) G. W. Adamson and J. A. Bush, "Method for Relating the Structure and Properties of Chemical Compounds", *Nature*, (*London*), **248**, 406–408 (1974); (b) "A Comparison of the Performance of Some Similarity and Dissimilarity Measures in the Automatic Classification of Chemical Structures", *J. Chem. Inf. Comput. Sci.*, **15**, 55–58 (1975); (c) "The Evaluation of an Empirical Structure–Activity Relationship for Property Prediction in a Structurally Diverse Group of Local Anaesthetics", *J. Chem. Soc. Perkin Trans. 1*, 168–172 (1976); (d) 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).

(14) (a) V. B. Bond, C. M. Bowman, N. L. Lee, D. R. Peterson, and M. H. Reslock, "Interactive Searching of a Structure and Biological Activity File", *J. Chem. Doc.*, **11**, 168–170 (1971); (b) E. Hyde, D. R. Lambourne, and L. A. McArdle, Abstracts of Papers, 163rd National Meeting of the American Chemical Society, Boston, Mass., April 1972; (c) C. Hansch, A. Leo, and D. Elkins, "Computerized Management of Structure-Activity Data. 1. Multivariate Analysis of Biological Data", *J. Chem. Doc.*, **14**, 57–61 (1974).

(15) M. F. Lynch, J. M. Harrison, W. G. Town, and J. E. Ash, "Computer Handling of Chemical Structure Information", Macdonald-Elsevier, London-New York, 1971.

(16) G. W. Adamson, S. E. Creasey, J. P. Eakins, and M. F. Lynch, "Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File. Part V. More Detailed Cyclic Fragments", *J. Chem. Soc. Perkin Trans. 1*, 2071–2076 (1973).

(17) G. M. Kellie and F. G. Riddell, "The von Auwers Boiling Point Rule. A New Approach", *J. Chem. Soc. Perkin Trans. 1*, 740–744 (1975).

(18) (a) O. Exner, ref 5, p 41; (b) J. Hine, ref 6, p 16.

(19) J. E. Dubois, J. J. Aaron, O. Alcais, J. P. Doucet, F. Rothenberg, and R. Ucan, "A Quantitative Study of Substituent Interactions in Aromatic Electrophilic Substitution. I. Bromination of Polysubstituted Benzenes", *J. Am. Chem. Soc.*, **94**, 6823–6828 (1972).

(20) Statistical Analysis Mark II Applications Package, International Computers Limited Technical Publication 4301, London, 1971.

(21) P. B. D. de la Mare and J. H. Ridd, "Aromatic Substitution", Butterworths, London, 1959.

(22) B. R. Kowalski and C. F. Bender, "Pattern Recognition. A Powerful Approach to Interpreting Chemical Data", *J. Am. Chem. Soc.*, **94**, 5633–5639 (1972).

(23) M. Sjostrom and S. Wold, "Statistical Analysis of the Hammett Equation. II. A Unified Inductive Sigma Scale", *Chem. Scripta*, **6**, 114–121 (1974).

(24) J. Shorter in "Advances in Linear Free Energy Relationships", N. B. Chapman and J. Shorter, Ed., Plenum Press, London, 1972, Chapter 2, p 71.

(25) J. E. Ash and E. Hyde, "Chemical Information Systems", Ellis Horwood, Chichester, 1975.

# Documentation of Chemical Reactions. III. Encoding of the Facets

M. OSINGA* and A. A. VERRIJN STUART**

Gist-Brocades N. V., Research & Development, Haarlem, Holland
and Centraal Rekeninstituut der Rijksuniversiteit Leiden, Leiden, Holland

Received June 27, 1975

A computer program is described for the automatic encoding of chemical reactions into the following faceted classification: (1) the chain facet, (2) the ring facet, (3) the rearrangement facet, and (4) the unusual element facet. The main problem for the chain facet is to determine the actual change in starting material and end product. The first step in the solution is to compare the pairs of DEAN's of two bonded atoms. Sometimes this leads to inconsistent equivalences. It was necessary to develop tests to find them and guidelines to find the correct equivalences. When encoding the ring facet, the main problem encountered was the delocalization of double bonds. When the correct equivalences are established, the rearrangement facet and the unusual element facet do not present serious problems.

## INTRODUCTION

The automatic encoding of chemical reactions is the ultimate purpose of our research. The reactions to be encoded make use of the faceted classification described before.[1] In this classification five facets are distinguished: (1) the chain facet (present in all reactions); (2) the ring facet; (3) the rearrangement facet; (4) unusual elements (elements other than C, H, O, N, S, P, F, Cl, Br, I); and (5) other facets (these are not dealt with in the automatic analysis).

The Wiswesser Line Notation (WLN) was selected as the chemical coding system for the starting material and end-product. A computer program for analyzing a WLN has been described.[2] This analysis leads to a bond table, in which a bond and two atoms on each end of the bond is described, as illustrated in Figure 1. In the bond table the numbers on a horizontal line represent a "bond pair".

The type of atom, after a first level analysis, is expressed as an auxiliary number (see left part of table of Figure 1). From this auxiliary number the definitive DEAN or Direct

* Author to whom correspondence should be addressed at Gist-Brocades N.V.
** Centraal Rekeninstituut der Rijksuniversiteit Leiden.