32. By retaining only 32 of the features it is possible to reduce storage requirements and computation times to the point where elaborate pattern recognition problems can be implemented on microcomputers. It should be noted that these results are dependent on the classification problem investigated and the type of discriminant function used.

In a second application of this feature selection technique, Woodruff[12] calculated the nearest neighbor to each of the compounds on the basis of their binary infrared spectra. He performed these calculations using both 139 features and the 32 best features as chosen by this technique. The results of this work demonstrate that the nearest-neighbor algorithm performs as well or better with the reduced set of features than with all 139. The simplicity of this feature selection technique and its success in these cases demonstrate that it should be considered in any situation where computations involve binary spectral data.

ACKNOWLEDGMENT

Support of this research by the National Science Foundation is gratefully acknowledged.

LITERATURE CITED

(1) Jurs, P. C., and Isenhour, T. L., "Chemical Applications of Pattern Recognition", Wiley-Interscience, New York, N.Y., 1975.

(2) Kowalski, B. R., Jurs, P. C., Isenhour, T. L., and Reilley, C. N., "Interpretation of Infrared Spectrometry Data", *Anal. Chem.*, **41**, 1945 (1969).

(3) Preuss, D. R., and Jurs, P. C., "Pattern Recognition Techniques Applied to the Interpretation of Infrared Spectra", *Anal. Chem.*, **46**, 520 (1974).

(4) Liddell, R. W., and Jurs, P. C., "Interpretation of Infrared Spectra Using Pattern Recognition Techniques", *Anal. Chem.*, **46**, 2126 (1974).

(5) Andrews, H. C., "Introduction to Mathematical Techniques in Pattern Recognition", Wiley-Interscience, New York, N.Y., 1972, p 15.

(6) Tou, J. T., "Feature Extraction in Pattern Recognition", *Pattern Recognition*, **1**, 3–11 (1968).

(7) Kowalski, B. R., and Bender, C. F., "Pattern Recognition II Linear and Nonlinear Methods for Displaying Chemical Data", *J. Am. Chem. Soc.*, **95**, 686 (1973).

(8) Jurs, P. C., "Mass Spectral Feature Selection and Structural Correlations Using Computerized Learning Machines", *Anal. Chem.*, **42**, 1633 (1970).

(9) Wilkins, C. L., Williams, R. C., Brunner, T. R., and McCombie, P. J., "Heuristic Pattern Recognition Analysis of Carbon-13 Nuclear Magnetic Resonance Spectra", *J. Am. Chem. Soc.*, **96**, 4182–5 (1974).

(10) Woodruff, H. B., Ritter, G. L., Lowry, S. R., and Isenhour, T. L., "Density Estimation and the Characterization of Binary Infrared Spectra", *Technometrics*, in press.

(11) Woodruff, H. B., Lowry, S. R., and Isenhour, T. L., "A Comparison of Two Discriminant Functions for Classifying Binary Infrared Data", *Appl. Spectrosc.*, **29**, 226 (1975).

(12) Woodruff, H. B., Lowry, S. R., Ritter, G. L., and Isenhour, T. L., "Similarity Measures for the Classification of Binary Infrared Data", *Anal. Chem.*, **47**, 2027 (1975).

A Method of Structure–Activity Correlation 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 June 19, 1975

Fragment sets generated manually from Wiswesser Line Notation have been used to correlate the chemical structures of a group of 79 penicillins with their serum binding activity, using multiple regression analysis. Statistically significant correlations were found, with results in accordance with the generally accepted nature of the binding. Algorithmic methods for the generation of such fragment sets are proposed and the use of various structural representations for structure–property correlation within chemical information systems is discussed.

INTRODUCTION

The investigation of the relationship between physical, chemical, or biological properties and chemical structures has recently been a field of considerable activity, particularly in the search for a methodology of rational drug design. Four major approaches to this problem may be distinguished:[1]

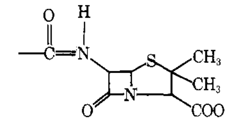
(i) semiempirical methods, correlating biological activity with physico-chemical properties[2]
(ii) additive mathematical modelling, for series of structurally related compounds[3]
(iii) correlations based on quantum-mechanical studies[4]
(iv) substructure analysis: in which the activity of a chemical species is correlated directly with structural features using methods related to substructure search procedures.

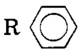
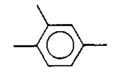
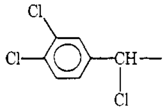
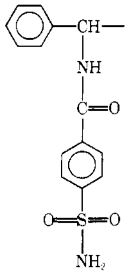
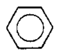
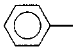
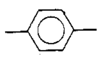
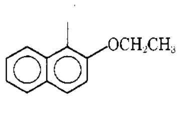
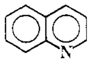
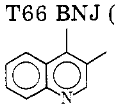
This last approach has the major advantage that, unlike the semiempirical and additive modelling methods, it can be applied to collections of structurally diverse compounds. Also its relative simplicity and compatability with substructure search techniques should enable its use as a routine, large-scale procedure in a manner not presently possible with the more sophisticated quantum-mechanical methods. The structural features used have included connection table fragments[5–7] and substructures from a fragmentation code.[8] Structural features such as standardized heteroatom counts have also been utilized.[9] Both regression analysis and pattern recognition techniques have been applied for structure–activity correlation and property prediction.

Wiswesser Line Notation (WLN)[10] is widely used for chemical structure representation in information storage and retrieval systems, and is also used in systems storing property data.[11] This notation thus has obvious potential for structure–property correlation, as has been noted by

Table I. Examples of WLN Fragmentation Method[a]

(a)



	Simple WLN set		Complex WLN set	
Side-chain structure	Fragment	No. present	Fragment	No. present
(b) $CH_3CH_2CH_2$ $CH_3CH_2CH_2C-$ $CH_3CH_2CH_2$	X -C-	1	As for simple set	
	(1c) $-CH_2-$ (1t) $-CH_3$	6 3		
	R 	1	R(A, C, D) 	1
(c) 	Y -CH-	1	Y -CH-	1
	G Cl	3	(loc) G Cl (ring)	2
			G Cl (alkyl)	1
(d) 	R 	2	R(A) 	1
	Y -CH-	1	R(a) 	1
	MV -NHCO-	1	Y -CH-	1
			MV -NHCO-	1
	SWZ SO_2NH_2	1	SWZ $-SO_2NH_2$	1
(e) 	T66 BNJ		T66 BNJ (D, E)	
		1		1
	O -O-	1	(loc) O -O-(ring)	1
	(1c) $-CH_2-$	1	(1c) $-CH_2-$	1
	(1t) $-CH_3$	1	(1t) $-CH_3$	1

[a] The WLN's are illustrative and do not indicate the possible permutations of the notation. t = terminal; c = connective; (loc) = ring locant in WLN.

several authors.[9,12,13] In the work reported here, fragments were generated manually from WLN. The manner of fragmentation was chosen in such a way as to allow for automatic generation of such fragment sets by computer program, if desired.

FRAGMENTATION AND CORRELATION PROCEDURE

An examination of the relationship between the serum binding properties of a group of 79 penicillins and their chemical structures[14] was carried out using these WLN fragment sets. A ready comparison was then possible between these results and those obtained from a similar analysis of these compounds which used fragments automatically derived from connection tables.[5]

Serum binding activity was assumed to be an additive function of the structural units present, so that its value, y, for the ith compound is given by:

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

where there are a total of n types of structural fragment in the set of structures, and x_{ij} is the number of times that the jth fragment occurs in the ith structure; b_j is the regression coefficient for the jth fragment and represents the effect of that fragment in increasing or decreasing the activity of the compounds in which it occurs. These coefficients were determined by multiple regression analysis, using a standard statistical package.[15]

The major features of the WLN fragmentation process are set out below:

(i) Ring systems (monocyclic and fused, including benzene rings). These are treated as whole entities, i.e., the string of symbols between T . . . J or L . . . J, or the R symbol was extracted as a single substructure. Substitution patterns of rings may be distinguished by considering ring locants.

(ii) Functional groups (including branching carbon atoms with symbols X, Y, and C, including halogens, and including unsaturated fragments). These may be single WLN symbols, e.g., Y for >CH–, G for Cl, or may be a combination of several WLN symbols. Thus, in the fragment sets below, adjacent Q and V symbols are treated together as QV, i.e., the –COOH carboxylic acid grouping. Certain groups whose properties may be affected by interaction with aromatic rings, e.g., –OH or $-NH_2$, may be distinguished according to whether they are attached to such rings or to saturated moieties.

(iii) Hydrocarbon fragments. Numeric symbols are identified as terminal or connective, and fragmented into either $-CH_3$ and $-CH_2-$ groups, or solely $-CH_2-$ groups, respectively. The difference in lipophilic properties between such groups has recently been discussed.[16]

(iv) Other symbols (for example elements) are treated as single fragments.

Examples of fragments derived according to these rules from WLN representations of some of the penicillins examined are shown in Table I.

Table I(a) shows the penicillin nucleus, common to all 79 compounds, and hence not included in the analysis except for the $-CH_3$ and –CONH– substituents.

Table I(b)–(e) show fragmentation sets produced from four side chains, with I(c), I(d), and I(e) demonstrating the use of alternative "simple" and "complex" fragment sets. These differ according to whether distinction is made between ring substitution patterns and between the ring or chain environment of heteroatomic functional groups. These examples also illustrate the combination of symbols to form "functional group fragments", e.g., ZWS ($-SO_2NH_2$) and MV (–NHCO–).

Multipliers and contractions of WLN were not taken into account in devising the fragmentation procedure, in view of the increasing tendency toward the use of the uncontracted notation within information systems, and the capability to convert contracted to uncontracted notation.[17,18]

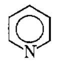
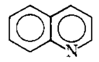
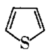
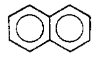
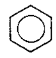
RESULTS

Two fragment sets, denoted "simple" and "complex", were produced from the penicillin structures, and together with their serum binding values[14] were analyzed by multiple regression. Log (b/f) was used as the dependent variable.[14] The full sets of structural fragments, with their WLN symbols, are set out in Tables II and III, together with the regression coefficients and Student t statistics from the analysis. Table IV shows for comparison the equivalent results from a connection table analysis of these compounds.[5,19]

The results of these three regression analyses are summarized in Table V. These analyses were carried out in such a way that as many as possible of the structural features used were included in the regression, subject only to the limitations of the accuracy of the calculation on the model of computer used.

From Table V it is apparent that essentially similar overall regression results, with high correlation coefficients, are obtained with fragment sets derived from both WLN and connection tables. Use of the F test showed the regression using the complex WLN set to be significantly better, from a statistical viewpoint, than that using the simple WLN set at the 5% significance level, but not at the 1% level. There was no significant difference between the regression using the complex WLN set and that using augmented atoms.

Table II. Simple WLN Fragment Set from Penicillin Structures with Regression Results[a]

Fragment	WLN	Regression coefficient	t statistic
	T6NJ	Not included by regression program	
	T66 BNJ	1.094	5.84
	T5SJ	0.843	4.69
	L66J	1.813	10.67
	R	1.067	8.28
$-CH_3$	(1t)	−0.043	0.44
$-CH_2$	(1c)	0.282	7.46
–CH	Y	0.339	3.37
–C–	X	0.679	2.70
–CHO	VH	0.266	0.83
–CONH	VM	−0.249	0.84
–OH	Q	−0.436	2.41
$-NH_2$	Z	−0.672	6.70
–O–	O	0.043	0.44
$-SO_2NH_2$	SWZ	−1.438	3.50
$-NO_2$	NW	−0.304	1.06
–Cl	G	0.362	7.06
–Br	E	0.420	2.35
–F	F	0.114	1.01

[a] The WLN's are illustrative and do not indicate the possible permutations of the notation. t = terminal; c = connective; (loc) = ring locant in WLN.

On consideration of the individual coefficients for each of the fragments in Tables II, III, and IV, it is apparent that essentially the same insight is gained from analyses using either WLN or connection table fragmentation. The fragments with positive regression coefficients (i.e., increasing serum binding activity) are hydrocarbon groups, aromatic rings, and halogens (i.e., hydrophobic fragments), while the hydrophilic, polar fragments (e.g., $-NH_2$ and –OH groups) have negative regression coefficients (i.e., reduce serum binding activity). This is in accordance with the well-known dependence of penicillin serum binding on the lipophilic nature of the side chain.[20]

The significance of the differences between regression coefficients for the fragments were tested for the complex WLN set using the formula

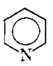
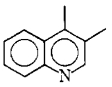
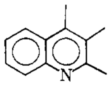
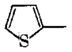
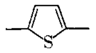
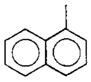
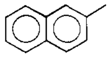
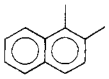
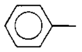
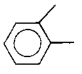
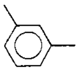
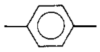
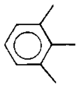
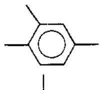
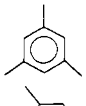
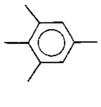
$$t(b_1,b_2) = \sqrt{S_1^2 + S_2^2 - 2r^2C_{12}}/r$$

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 then compared with values in tables of Student's t distribution.

The differences in regression coefficients for the majority of fragment pairs, for example, $-CH_2-$ (cf. >CH), $-NH_2$ (chain) (cf. $-NH_2$ (ring)), and –Br (ring) (cf. –OH (ring)), were found to be significant at the 25% level, but not at the 10% level. Some fragment pairs with widely differing coefficient values, for example $-SO_2NH_2$ (cf. 2-naphthyl), were significantly different at the 10% level. Other pairs of fragments, for example, Br (ring) (cf. F (ring)), did not differ significantly even at the 25% level.

These significance levels are higher than those usually accepted as indicating a significant difference, suggesting that it may be unwise to draw firm conclusions based solely on the differences between coefficient values for particular pairs of substructures.

Table III. Complex WLN Fragment Set from Penicillin Structures with Regression Results[a]

Fragment	WLN	Regression coefficient	*t* statistic
$-CH_3$	(1t)	−0.097	1.33
$-CH_2-$	(1c)	0.265	8.99
$-\underset{\vert}{C}H-$	Y	0.415	4.08
$-\overset{\vert}{\underset{\vert}{C}}-$	X	0.672	3.16
−CHO	VH	−0.102	0.43
$-SO_2NH_2$	SWZ	−1.325	3.96
$-NO_2$	NW	−0.134	0.51
−O− (chain)	O	−0.361	3.32
−O− (ring)	(loc) O	0.142	1.67
$-NH_2$ (chain)	Z	−0.927	7.81
$-NH_2$ (ring)	(loc) Z	−0.705	2.73
−OH (chain)	Q	−0.538	2.36
−OH (ring)	(loc) Q	−0.498	2.54
−Cl (chain)	G	0.002	0.01
−Cl (ring)	(loc) G	0.357	3.16
−Br (chain)	E	0.079	0.35
−Br (ring)	(loc) E	0.567	2.59
−F (ring)	(loc) F	−0.111	0.81
	T6NJ	Not included by regression program	
	T66 BNJ (D, E)	0.621	323
	T66 BNJ (C, D, E)	0.812	3.69
	T5SJ	0.55	3.82
	T5SJ (B, E)	0.960	4.49
	L66J (B)	1.507	8.40
	L66J (C)	1.621	9.03
	L66J (B, C)	1.198	7.08
	R (A)	0.735	7.16
	R(A, B)	1.058	6.71
	R (A, C)	0.909	5.78
	R (A, D)	0.954	6.32
	R (A, B, E)	0.591	2.71
	R (A, C, D)	0.979	4.07
	R (A, C, E)	0.891	2.76
	R (A, C, D, E)	0.967	2.50

[a] The WLN's are illustrative and do not indicate the possible permutations of the notation. t = terminal; c = connective; (loc) = ring locant in WLN.

In order to investigate the effect of relatively minor changes in the fragment sets used, additional regression analyses were performed with more detailed distinctions between the environments of halogen atoms and of naphthalene rings. In both cases there was negligible change in the overall regression results, indicating the insensitivity of the correlation procedure to such changes in the nature of the structural fragments. However, improved estimated values were obtained for a number of the compounds containing the modified substructures.

AUTOMATIC FRAGMENT GENERATION

The automatic generation of WLN fragment sets such as those above is necessary if the method is to be used routinely. Such an algorithm would take successive Wiswesser symbols and identify them as single-symbol fragments to be stored, or as part of a ring system or multi-symbol group to be combined to form a fragment.

Problems will obviously arise with the ordering of the notation in multi-symbol groups; thus the -COOH group, for example, may be extracted from the notation as QV or VQ. Some form of dictionary of such groups could be used to allow for such permutation. A similar problem exists with ring substitution patterns. Thus a (1,3,4)-substituted benzene ring as extracted from the notation could have locants (A,C,D), (A,D,C), (A,B,D), (A,D,B), (A,B,E), or (A,E,B) according to the nature of the substituents. Some form of canonicalization procedure would be necessary in such cases.

As an alternative to the algorithmic generation of a standard form of fragment set, particular substructures could be specified by a system's user and identified by procedures analogous to those used for WLN substructure searches. This approach would inevitably involve the problems inherent in WLN string searching, especially when parts of rings, single rings within fused systems, substructures occurring as either cyclic or acyclic fragments, or cyclic structures with several branching points are to be identified. Recent studies have suggested that the problems posed in searching for such substructures cannot be satisfactorily overcome without use of a connection table representation.[21]

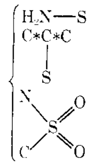
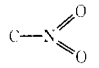
APPLICATIONS

The various types of structural representation available for structure–activity correlation would, in a practical situation, be complementary. WLN fragmentation is of most use in treating ring systems, functional groups, and possibly hydrocarbon chains as whole entities. Generation of WLN fragment sets appears to allow for a greater degree of user control, and hence of chemical knowledge and intuition, in the choice of fragment type, that can be readily achieved with connection tables. The primary advantage of the connection table is its explicit specification of each atom and bond present, allowing for the breakdown of ring systems, etc., into subunits if required. The use of connection tables also enables the automatic generation of the appropriate form of structural unit from a hierarchy of bond-cėntered and atom-centered fragment types.[22] The use of fragmentation codes for correlation purposes has also been demonstrated.[8]

An approach to structure–activity studies using more than one structural representation is thus likely to prove profitable. The existence of programs for the interconversion of notations[23,24] and the availability of standard statistical analysis packages make this a practical possibility for a comprehensive, integrated chemical information system.

The WLN correlation procedure described above is in principle applicable to any kind of additive property of chemical species, though its application is inevitably limited by the availability of adequate numerical data. In addition to the type of biological system described here, it could well be applied to other biological properties, to chemical reactivities, and to physico-chemical and thermodynamic properties. The inclusion of property data in addition to

Table IV. Augmented Atom Fragment Set[a] from Connection Tables with Regression Results

Fragment	Regression coefficient	t statistic	Perfectly correlated fragments
C*C*C	0.256	7.36	
C*C*C | C	−0.096	1.08	
C*C*C | O	0.125	0.75	
C*C*C | N	Not included by regression program		
C*C*C | Cl	−0.041	0.10	
C*C*C | Br	Not included by regression program		
C*C*C * C	0.049	0.84	
C*N*C	0.046	0.08	
C*C*N	Not included by regression program		
C*C*N | C	−0.186	0.77	
C*C*N * C	−0.192	0.36	
C*S*C	0.491	1.51	C*C*S | C
C*C*S	−0.098	0.28	
C*C*S | Cl	Not included by regression program		
C*C*S | Br	Not included by regression program		
C–C–C | C	−0.235	0.94	
C | C–C–C | C	−0.311	1.23	
C–CH$_2$–C	0.268	7.77	
H$_3$C–C	0.280	3.22	
C–O–C	0.027	0.14	
O | C–C–C | C	Not included by regression program		
C–C–O	0.042	0.41	
C–C–O | C	−0.103	0.61	
O=C	−0.329	2.31	C–N–C
H$_3$C–O	Not included by regression program		
HO–C	−0.233	0.99	
C–C–N | C	−0.107	0.38	
H$_2$N–C	−0.392	1.61	
N–C=O | C	−0.242	0.85	
O=S	−0.335	1.43	
O=N	0.113	0.93	
N–CH=O	Not included by regression program		
Cl–C	0.681	1.71	
C–C–Cl | C	−0.235	0.55	
Br–C	0.914	3.77	
Br | C–C–C	−0.523	1.56	
F–C	0.319	3.22	C*C*C | F

[a] Results from ref 18. * = aromatic ring bond.

Table V. Summary of Regression Analysis Results

Structure representation	No. of fragment types	No. of variables included in regression	Degrees of freedom	Multiple correlation coefficient	F value	Residual error
Simple WLN	19	18 + constant	60	0.931	21.68	0.276
Complex WLN	35	34	45	0.981	33.84	0.211
Augmented[a] atoms	44	29	50	0.976	35.87	0.225

[a] Results from ref 18.

structural features could improve performance; a recent example of the semiempirical method involved the use of structural parameters in addition to physico-chemical properties,[25] indicating a potential overlap of the techniques.

Pattern recognition techniques could also be applied to WLN fragment sets for correlation purposes. Such methods have advantages over parametric techniques in certain circumstances for dealing with chemical problems,[26,27] and could complement the more widely known regression analysis methods for structure–property investigations.

EXPERIMENTAL DETAILS

The regression analyses were performed using the ICL statistical analysis package and run on the Sheffield University ICL 1907 computer. Core storage required was 20K words, with CPU times ≤ 18 sec.

ACKNOWLEDGMENTS

We wish to thank Professor M. F. Lynch, Drs. U. D. Naik and M. J. Maher, and Miss J. Bush for helpful discussions, and the British Library Research and Development Department for the award of a Postgraduate Research Studentship to D. Bawden.

LITERATURE CITED

(1) Redl, G., Cramer R. D., and Berkoff, C. E., "Quantitative Drug Design", *Chem. Soc. Rev.*, **3**, 273-292 (1974).

(2) Hansch, C., "A Quantitative Approach to Biochemical Structure-Activity

Relationships", *Acc. Chem. Res.*, **2**, 232–239 (1969).
(3) Free, S. M. and Wilson, J. W., "A Mathematical Contribution to Structure-Activity Studies", *J. Med. Chem.*, **7**, 395–399 (1964).
(4) Kier, L. B., "Molecular Orbital Theory in Drug Research", Academic Press, New York, N.Y., 1971.
(5) Adamson, G. W., and Bush, J. A., "Method for Relating the Structure and Properties of Chemical Compounds", *Nature (London)*, **248**, 406–407 (1974).
(6) Adamson, G. W., and Bush, J. A., "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).
(7) Adamson, G. W., and Bush, J. A., "Evaluation of an Empirical Structure-Activity Relationship for Property Prediction in a Structurally Diverse Group of Local Anaesthetics", *J. Chem. Soc. Perkin Trans. 1*, in press.
(8) Cramer, R. D., Redl, G., and Berkoff, C. E., "Substructural Analysis. A Novel Approach to the Problem of Drug Design", *J. Med. Chem.*, **17**, 533–535 (1974).
(9) Kowalski, B. R., and Bender, C. F., "The Application of Pattern Recognition to Screening Prospective Anticancer Drugs. Adenocarcinoma 755 Biological Activity Test", *J. Am. Chem. Soc.*, **96**, 916–918 (1974).
(10) Smith, E. G., "The Wiswesser Line-Formula Chemical Notation", McGraw-Hill, New York, N.Y., 1968.
(11) Hansch, C., Leo, A., and Elkins, D., "Computerized Management of Structure-Activity Data. I. Multivariate Analysis of Biological Data", *J. Chem. Doc.*, **14**, 57–61 (1974).
(12) Brasie, W. C., and Liou, D. W., "Chemical Structure Coding", *Chem. Eng. Prog.*, **61**, 102–108 (1965).
(13) Osinga, M., and Verrijn Stuart, A. A., "Documentation of Chemical Reactions. II. Analysis of the Wiswesser Line Notation", *J. Chem. Doc.*, **14**, 196–198 (1974).
(14) Bird, A. E., and Marshall, A. C., "Correlation of Serum Binding of Penicillins with Partition Coefficients", *Biochem. Pharmacol.*, **16**, 2275–2290 (1967).
(15) "Statistical Analysis Mark II Applications Package", International Computers Limited Technical Publication 4301, London, ICL, 1971.
(16) Nys, G. G., and Rekker, R. F., "Statistical Analysis of a Series of Partition Coefficients with Special Reference to the Predictability of Folding of Drug Molecules. The Introduction of Hydrophobic Fragment Constants (f Values)", *Chim. Ther.*, **5**, 521–535 (1973).
(17) Saada, E., "Proposal for the Elimination of Contractions and Multipliers in the WLN Notation", *Bull. Chem. Notation Assoc.*, (1st ed), 39–41 (1974).
(18) Ash, J. E., and Hyde, E., "Chemical Information Systems", Ellis Horwood, Chichester, 1975, pp 106–108.
(19) Bush, J. A., Doctoral thesis, University of Sheffield, in preparation.
(20) Scholtan, W., "Die Bindung der Antibiotica an die Eiweisskorper des Serums", *Arzneim.-Forsch.*, **13**, 347–360 (1963).
(21) Crowe, J. E., Leggate, P., Rossiter, B. N., and Rowland, J. F. B., "The Searching of Wiswesser Line Notations by Means of a Character-Matching Serial Search", *J. Chem. Doc.*, **13**, 85–92 (1973).
(22) Adamson, G. W., Cowell, J., Lynch, M. F., McLure, A. H. W., Town, W. G., and Yapp, A. M., "Strategic Considerations in the Design of a Screening System for Substructure Searches of Chemical Structure Files", *J. Chem. Doc.*, **13**, 153–157 (1973).
(23) Heller, S. R., and Koniver, D. A., "Computer Generation of Wiswesser Line Notation. II. Polyfused, Perifused and Chained Ring Systems", *J. Chem. Doc.*, **12**, 55–59 (1972).
(24) Granito, C. E., Roberts, S., and Gibson, G. W., "The Conversion of Wiswesser Line Notations to Ring Codes. I. The Conversion to Ring Systems", *J. Chem. Doc.*, **12**, 190–196 (1972).
(25) Hansch, C., and Yoshimoto, M., "Structure-Activity Relationships in Immunochemistry. 2. Inhibition of Complement by Benzamidines", *J. Med. Chem.*, **17**, 1160–1167 (1974).
(26) Kowalski, B. R., and Bender, C. F., "Pattern Recognition. A Powerful Approach to Interpreting Chemical Data", *J. Am. Chem. Soc.*, **94**, 5632–5639 (1972).
(27) Kowalski, B. R., and Bender, C. F., "Pattern Recognition. II. Linear and Nonlinear Methods for Displaying Chemical Data", *J. Am. Chem. Soc.*, **95**, 686–693 (1973).

CRYSRC: A Generalized Chemical Information System Applied to a Structural Data File

JOSE VILLARREAL, JR.,[†] EDGAR F. MEYER, JR.,[*,†] ROGER W. ELLIOTT,[**] and CARL MORIMOTO[†]

Texas A&M University, College Station, Texas 77843

Received June 23, 1975

An interactive retrieval system, CRYSRC, based upon the modular design of input, query, and output routines has been implemented and tested on a well defined data base, the Cambridge Crystal Data Centre files. Use has been made of three-dimensional display facilities in this laboratory to create models of retrieved molecules. All routines have been implemented on a laboratory mini-computer, the PDP11/40, in Fortran IV. They are available for distribution and should be upwards expandable to a variety of computer systems.

INTRODUCTION

Chemistry, like other disciplines, has been caught up in the information explosion. Automated methods of data collection have contributed to the generation of numerous files of chemical information by both private and public sources that are becoming increasingly available in machine-readable form. Chemists have expressed interest in accessing information in these files to serve as an additional information source. Yet, existing files are rarely compatible in terms of their data structures.

The lack of compatibility between files is related to the mode of the representation of chemical structural data that is often included in available files. The most widely used representations of chemical structural data appear to be some type of linear notation or form of connectivity. The form of representation presents special problems to the design of chemical retrieval systems.

Although there is interest in accessing currently avail-

[†] Department of Biochemistry and Biophysics, Texas Agricultural Experment Station, Texas A&M University.

[*] To whom correspondence should be addressed.

[**] Computer Science Division, Department of Industrial Engineering, Texas A&M University.