(4) D. M. Murray "A Scatter Storage Scheme for Dictionary Lookup", *J. Libr. Autom.*, **3**, 173-201 (1970).

(5) A. Zamora and D. L. Dayton, "The Chemical Abstracts Service Chemical Registry System. V. Structure Input and Editing", *J. Chem. Inf. Comput. Sci.*, **16**, 219-222 (1976).

(6) L. C. Ray and R. A. Kirsch, "Finding Chemical Records by Digital Computers", *Science*, **126**, 814-819 (1957).

(7) J. H. R. Bragg, M. F. Lynch, and W. G. Town, "The Use of Molecular Formula Distribution Statistics in the Design of Chemical Structure Registry Systems", *J. Chem. Doc.*, **10**, 125-128 (1970).

(8) G. M. Dyson, "Studies in Chemical Documentation", *Chem. Ind. (London)*, 676-684 (1952).

(9) S. R. Shaw, "An Investigation of Some Methods of Improving the Performance of the Molecular Formula in Indexing", unpublished M.Sc. thesis, University of Sheffield, 1973.

(10) D. Lefkowitz, "A Chemical Notation and Code for Computer Manipulation", *J. Chem. Doc.*, **7**, 186-192 (1967).

(11) M. F. Lynch, J. Orton, and W. G. Town, "Organisation of Large Collections of Chemical Structures for Computer Searching", *J. Chem. Soc. C*, 1732-1736 (1969).

(12) G. L. Mishchenko, "Empirical Formulas of Bonds of Compounds and Their Possible Role in Retrieving Factographic Information in Chemistry", *Inf. Probl. Sovrem. Khim.*, 25-38 (1976); *Chem. Abstr.*, **86**, 170058 (1977).

(13) D. H. Rouvray, "The Search for Useful Topological Indices in Chemistry", *Am. Sci.*, **61**, 729-735 (1973).

(14) C. F. Wilcox, "A Topological Definition of Resonance Energy", *Croat. Chem. Acta*, **47**, 87-94 (1975).

(15) M. Randic, "On Characterization of Molecular Branching", *J. Am. Chem. Soc.*, **97**, 6609-6615 (1975).

(16) L. B. Kier, L. H. Hall, W. J. Murray, and M. Randić, "Molecular Connectivity. I. Relationship to Nonspecific Local Anesthesia", *J. Med. Chem.*, **64**, 1971-1974 (1975).

(17) L. B. Kier and L. H. Hall, "Molecular Connectivity in Chemistry and Drug Research", Academic Press, New York, N.Y., 1976.

(18) H. L. Morgan, "The Generation of a Unique Machine Description for Chemical Structures—a Technique Developed at Chemical Abstracts Service", *J. Chem. Doc.*, **5**, 107-113 (1965).

(19) W. T. Wipke and T. M. Dyott "Stereochemically Unique Naming Algorithm", *J. Am. Chem. Soc.*, **96**, 4834-4842 (1974).

(20) C. A. Shelley and M. E. Munk, "Computer Perception of Topological Symmetry", *J. Chem. Inf. Comput. Sci.*, **17**, 110-113 (1977).

(21) E. Hyde, F. W. Matthews, L. H. Thompson, and W. J. Wiswesser, "Conversion of Wiswesser Notation to a Connectivity Matrix for Organic Compounds", *J. Chem. Doc.*, **7**, 200-204 (1967).

(22) R. G. Freeland, S. J. Funk, L. J. O'Korn, and G. A. Wilson, "Augmented Connectivity Molform—a Technique for Recognition of Structure Topology Identity", 169th National Meeting of the American Chemical Society, Philadelphia, Pa., April 6-11, 1975, Abstract CHLT 29.

(23) L. J. O'Korn, personal communication, 1977.

(24) R. H. Penny, "A Connectivity Code for Use in Describing Chemical Structures", *J. Chem. Doc.*, **5**, 113-117 (1965).

The Production of Machine-Readable Descriptions of Chemical Reactions Using Wiswesser Line Notations

MICHAEL F. LYNCH* and PETER WILLETT

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

Received November 8, 1977

A method has been developed for the automatic analysis of chemical reactions by a consideration of the changes in the Wiswesser Line Notations of the reacting molecules. The notations are broken down by a multilevel fragmentation process which yields descriptions for all parts of the molecules. The two fragment lists are compared, duplicates eliminated, and the remaining fragments recombined to produce a reaction site. The output from the program consists of these reaction sites and a set of fragment descriptors derived from them. The method has been tested on a file of 9197 one-reactant, one-product reactions and analyses were produced for 7415 of them (80.6%); the success rate could be increased to ~90%.

INTRODUCTION

Techniques for the automatic retrieval of chemical structural information have now reached both a high level of sophistication and a wide range of applicability;[1] searches may be carried out both for individual molecules and for classes of compounds having certain substructural features in common. The development of comparable means of access to chemical reaction data has proved to be a continuing problem, although the provision of such information is of fundamental importance to the advancement of chemistry. At least part of the problem lies in the multifarious nature of the data, since reaction conditions, yields, mechanism, and the presence of substructural features not actively involved in the reaction may all be of interest. However, the main problem lies in the adequate representation of the reaction site, those parts of the reacting molecules involved in the change, in a machine-readable form.[2]

The most widely used device for representing chemical reactions is the reaction equation, a diagram in which the reactants are displayed upon one side of the equation and the products upon the other. However, the retrieval of reactions by the molecules involved is of limited utility since the main requirement is for substructural transformations such as the conversion of an α,β-unsaturated acid to the corresponding amide or the elimination reactions of dibromo compounds. Vleduts[3] has pointed out that chemical changes generally involve only a limited part of the participating molecules: "A distinctive feature of organic reactions, which involve complicated molecules containing almost exclusively covalent bonds, is the destruction and creation of a comparatively small number of bonds in such a way that, during the process, fairly extensive portions of the molecule do not change their structure." This being so, we should be able to eliminate those parts of the molecules that play no part in the course of the reaction, the remaining partial structures then being taken as describing the reaction sites.

Work in this department has led to two distinct approaches to the automatic identification of reaction sites. In the first[4,5] we sought to map the structures of the reactant and product molecules onto one another so as to identify the largest common fragments and thus, by subtraction, the differences. The work was abandoned owing to program complexity and the amount of processing time required; ten years later, the development of substantially faster computers and of new ways of identifying the common substructures has led us to a reexamination of the potential of such an approach.[6,7] Secondly, we have compared the reactant and product molecules to identify the differences directly; both connection tables[8] and Wiswesser Line Notations[9,10] have been used as the structure representation.

The earlier work[8] used the connection tables for the reactant and product molecules to generate two sets of small, bond-

0095-2338/78/1618-0149$01.00/0

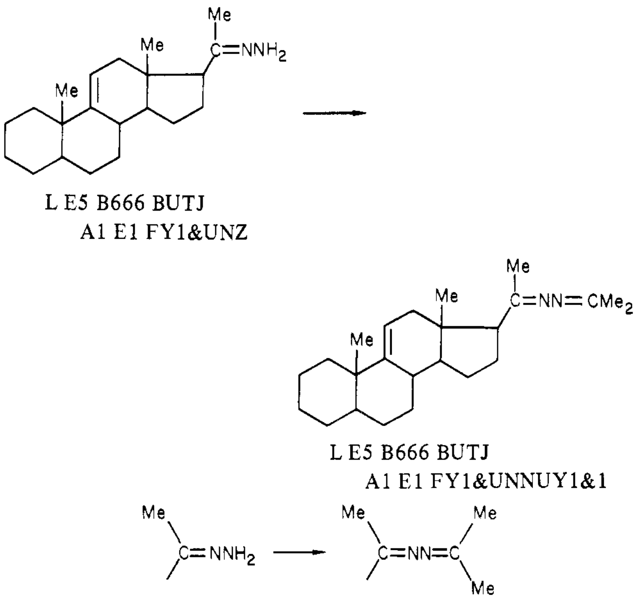
centered fragments. After elimination of the fragments common to both lists, the remainder were manipulated to synthesize a reaction site in much the same way as one might assemble a jigsaw puzzle. Although this proves to be a computationally efficient way of generating reaction sites, problems may arise at the search stage since the immediate environment of the reaction site may not be defined. In principle, this could be overcome by mapping the analysis fragments back onto the parent structures but, as a fragmentation process has been used, a degree of ambiguity is present and the very small size of the fragments (a bond, the atoms at each end, and their connectivity patterns) implies that there may be a large number of ways in which the mapping could take place. Such an approach would be more practicable using much larger fragments; each of these would only occur one or two times and thus the number of possible mappings would be much reduced. There is also the point that the analysis yields a fragment set at a single level of description and this level may not necessarily be the most useful for any given reaction; thus a ring formation would, perhaps, be better described by the entire ring formed rather than the atoms involved in the closure reaction.

Work was also carried out using Wiswesser Line Notation records (WLN). There are disadvantages here due to the lack of explicit connectivity information and to the fact that a few WLN symbols may represent quite large numbers of atoms and bonds, which implies that in some cases the changes may only be described in very general terms. A more serious objection is that the same substructural features may be described by different character strings depending upon the nature of the molecule containing the substructure, or upon other features present; thus in the cyclization of a long-carbon-chain dicarboxylic acid there will be almost no similarity between the reactant and product WLNs even though large parts of the reacting molecules are not involved in the change. In fact, the analysis describes the change in the WLN symbols, rather than in the molecules that they represent. Within these limits, it was found that relatively simple programs were sufficient to handle ring change and functional group interconversion reactions, and as the WLN symbols provide printable character representations of the structural features present, it would be reasonable to think in terms of printed indexes of reactions similar to KWIC and KWOC compound indexes.

2. WLN-DERIVED FRAGMENTATION CODE FOR REACTION INDEXING

Algorithmic fragment generation is routinely used as a means of obtaining possible screens for searching structure files; efficient screen sets are obtained by consideration of the distribution of fragment occurrence frequencies, the fragments being chosen so that they occur approximately equifrequently across the whole file.[11-13] The algorithmic fragmentation process developed in the present work is based upon a rather different criterion. We wish to produce fragments which are as large as possible, subject to the constraint that they represent features common to both sides of the reaction equation. Once these large common features have been identified, they may be discarded and a more specific fragmentation method adopted to remove further common features. The process continues until, hopefully, the remaining truncated structures represent the reaction sites. We have implemented four levels of fragmentation, their choice being governed both by a knowledge of the reaction types in our file and by the way in which WLN delineates substructural features.

Lynch et al.[10] found that in a large number of reactions the basic ring system remained unaltered, changes being restricted to the ring substituents. This being so, the features extracted in the first stage are ring systems *in toto*, benzene rings, and substituents. Thus for the reaction shown in eq 1 we obtain



L E5 B666 BUTJ
A1 E1 FY1&UNZ

L E5 B666 BUTJ
A1 E1 FY1&UNNUY1&1

*Y1&UNZ *Y1&UNNUY1&1 (1)

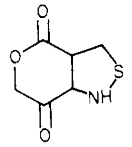
the fragment lists

(i) reactant: L E5 B666 BUTJ, *1, *1, *Y1&UNZ

(ii) product: L E5 B666 BUTJ, *1, *1, *Y1&UNNUY1&1

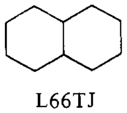
where * represents attachment to a ring of some kind. Elimination of the duplicate fragments yields the analysis shown in the lower part of eq 1.

More generally, there will also be changes in the ring system so the second-level analysis is a description of any ring systems remaining in terms of their component monocycles. WLN describes complete ring systems so that it is necessary to provide descriptions of the individual rings present. We have used the method of Granito et al.[14] to identify the atoms within each ring in the system. Once this has been done, a modified form of the WLN rules for encoding monocycles is applied to ensure that a particular monocycle will always lead to the same description, whatever its environment. The procedure is illustrated by the ring system shown below which gives rise to the monocycles #T5 AM BSTJ and #T6 AV BO DVTJ,

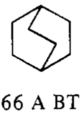


T56 BMS FVO IVTJ

where # represents a fused ring. The monocycles identified are only those expressly delineated by the WLN; thus both the systems shown below would be regarded as consisting of the two rings #L6TJ. Fusion patterns could be identified,



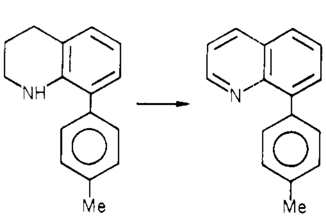
L66TJ



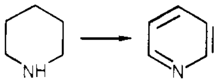
L66 A BTJ

e.g., by the method of Bedrosian and Milne,[15] but this was not considered necessary.

To illustrate the use of this level of fragment description, consider the reaction shown in eq 2. The fragment sets after



T66 BMT&J JR D1   T66 BNJ JR D1



T6 AMTJ   T6 ANJ

(2)

the first fragmentation and analysis are

(i) reactant: T66 BMT&J

(ii) product: T66 BNJ

(the benzene rings and their substituents having been eliminated). Analysis of the two ring systems yields #L6J, #T6 AMTJ and #L6J, #T6 ANJ where #L6J represents a fused, carbocyclic, unsaturated, six-membered ring. Elimination of the #L6J rings yields the analysis shown in the lower portion of eq 2.

The third fragmentation is an analysis of the acyclic portions of the reacting molecules and involves breaking the symbol string whenever a terminal atom or a branching symbol is detected in the WLN. Branching symbols for the present purpose are defined to be anything that disturbs the linearity of the WLN symbol string; thus "X", "Y", and "–SI–" are considered as branching symbols whereas "V" or "NU" are not. (A related method for obtaining acyclic fragments was developed by Lefkowitz[16] for screen generation from the Mechanical Chemical Code.) This mode of fragmentation was chosen for three reasons:

(i) As the resultant fragments are linear, it is easy to obtain canonical descriptions for them by a straight alphanumeric comparison of the fragment character string, as generated, and its reverse. The string sorting lower is arbitrarily chosen as the representative so that, e.g., the substructure

$$-CH_2O\overset{\overset{O}{\|}}{C}CH_2CH=CH_2-$$

will always be described by the fragment string /1OV2U1/, where "/" represents attachment to an acyclic branching symbol.

(ii) We need to be able to describe the whole structure so that we can regenerate parts of it if required (see section 3).

(iii) A high percentage of functional groups remains intact under the present fragmentation where we use functional group to describe any string of hetero- and/or unsaturated atoms; both Vleduts[3] and Hendrickson[18] have emphasized the importance of such features. Clinging and Lynch[9] showed that ~20% of the reactions in our file could be analyzed by a small dictionary of functional group interconversions; the more general definition used in the present work implies that a higher percentage of reactions should be susceptible to analysis.

The reaction in eq 3 illustrates the method; the noncanonical fragment lists given below yield the analysis shown in the lower part of eq 3.

$$Me_2CHCH=CH(CH_2)_4\overset{\overset{O}{\|}}{C}CMe_2CH=NOH$$

QNU1X1&1&V5U1Y1&1

$$\rightarrow Me_2CHCH=CH(CH_2)_4\overset{\overset{O}{\|}}{C}CMe_2CN$$

NC×1&1&V5U1Y1&1

$-CH=NOH \rightarrow -C\equiv N$

/1UNQ → /CN (3)

(i) reactant: QNU1/, X, /1, /1, /V5U1/, Y, /1, /1

(ii) product: NC/, X, /1, /1, /V5U1/, Y, /1, /1

This mode of fragmentation necessarily implies that long, unbranched carbon chains will remain intact so the fourth and final level of analysis involves the truncation of any such features to a fixed length of one methylene unit. Consider a reaction yielding the analysis

/5U1VO5 → /5U1VQ

i.e.

$$-(CH_2)_4CH=CH\overset{\overset{O}{\|}}{C}O(CH_2)_4CH_3 \rightarrow -(CH_2)_4CH=CH\overset{\overset{O}{\|}}{C}OH$$

Although the long carbon chains are important in describing the exact environment of the group that has changed, it does mean that in a printed index there will be an inevitable scattering of the entries describing the hydrolysis of unsaturated acid esters because of the variable lengths of the methylene chains. The final fragmentation truncates the symbol strings, in this case, to yield the much more general analysis

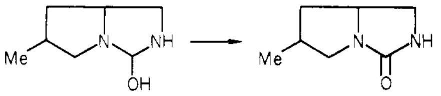
/1U1VO1/ → /1U1VQ

i.e.

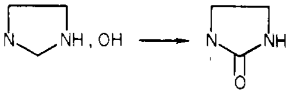
$$-CH=CH\overset{\overset{O}{\|}}{C}OCH_2- \rightarrow -CH=CH\overset{\overset{O}{\|}}{C}OH$$

3. PROGRAM FOR CHEMICAL REACTION ANALYSIS

The algorithmic fragmentation code described above has been used in a program to automatically index chemical reactions. It will be realized that the examples of reactions quoted so far have been carefully selected insofar as all of the analyses have consisted of only one fragment upon each side of the equation; in general this will not be so and the fragments that do remain are often linked together in the parent reacting molecule. Consider the reaction and analysis shown in eq 4;



T55 AN CMTJ BQ G1   T55 ANVMTJ G1



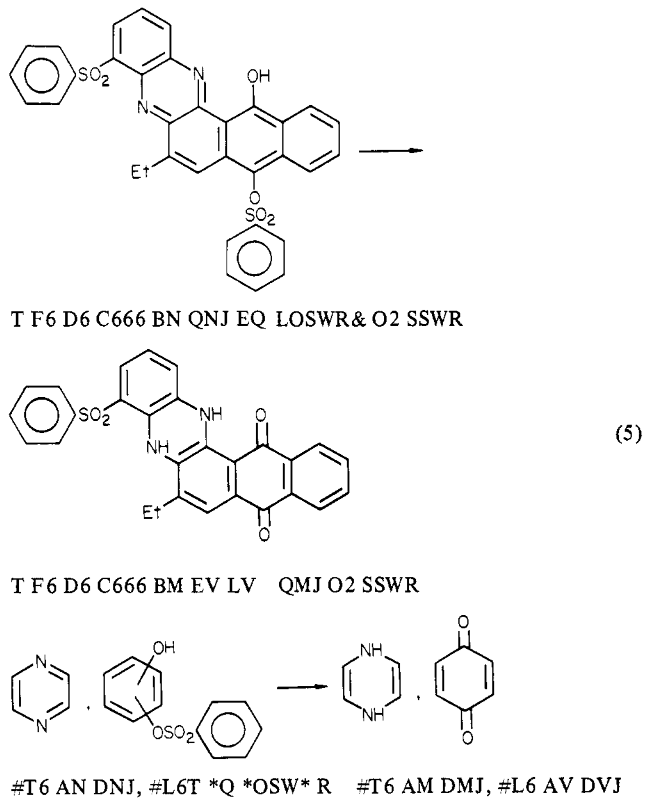
#T5 AM CNTJ, *Q   T5 AM BV CNTJ

(4)

it would obviously be useful if the two reactant analysis fragments could be recombined. Such an approach forms the basis of the work described in ref 8, and we have developed an analogous synthesis segment to reunite fragments left after the analysis. In principle, this could be done after each and every level of fragmentation, but in practice we have included synthesis routines only after the second and third levels. In the first case, the fragments to be considered are monocycles and ring subtituents, while in the second they are branching symbols and pendant linear chains; as reactions involving ring systems predominate in our file,[10] the first of these routines is much the more heavily used.

The choice of level at which synthesis takes place has been largely conditioned by the ease with which the requisite connectivity information can be obtained from the input notation. As WLN is a whole-structure representation, it is possible to generate a full adjacency matrix for a large percentage of structures,[17,19] but at the present stage of development it was decided that the incorporation of a full connection table generation segment would be uneconomic both in terms of programming effort and in the speed of operation of the rest of the program. A connection table would, however, possess many advantages both in the rebuilding segment and elsewhere; probably the best compromise would be a notation-derived record, in which it is possible to relate the connectivity data very easily to the input WLN symbols.[20] Instead of a full atom adjacency matrix we have used fragment connectivity lists which are built up during the running of the program. We shall consider first the ring-substituent list. During the first level fragmentation, while scanning the input WLN strings for ring brackets and benzene rings, a note is made of the locant position of all substituents; at the same time, a stack is operated to keep track of all benzene rings and later ring systems so that it is possible to match all substituents with their parent ring systems. If any such systems are left after the first analysis, the monocycle generation routines produce a list of all locant positions for each monocycle so that it is also possible to match analysis substituents with their parent monocycles. During rebuilding, the analysis fragments are joined together in a linear string using the information in the connectivity lists; where a choice is possible the program chooses the nonoverlapping site or sites which are most highly connected, i.e., those comprising the largest number of analysis fragments. The resultant sites are then compressed to a linear symbol string for output.

The rebuilding is done in two stages: firstly, where appropriate, substituents are joined to benzene rings and then these larger fragments are joined to their parent monocycles; the method is exemplified by the reaction in eq 5. No details are included as to the manner in which individual monocycles are joined together so that if two analysis monocycles were fused, the fact could only be noted by an inspection of the parent ring system WLN; the majority of reactions involving ring change in our file are found to be confined to a single monocycle so that we do not consider this to be a major problem. Potentially more worrisome is the fact that substituent positions upon rings are not specified; rather, we are dealing with a form of Markush structure, though whether this is a serious defect will only be found by actual usage.



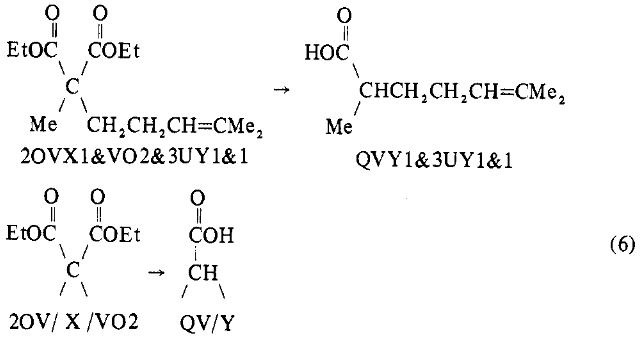
T F6 D6 C666 BN QNJ EQ LOSWR& O2 SSWR

T F6 D6 C666 BM EV LV QMJ O2 SSWR

#T6 AN DNJ, #L6T *Q *OSW* R #T6 AM DMJ, #L6 AV DVJ

(5)

The second set of synthesis routines is used for acyclic molecules after the third level fragmentation and analysis; during this fragmentation a record is generated, similar to the above, but rather than noting substituents attached to monocycles we list linear chains attached to branching symbols. As an acyclic molecule can be considered as a tree, it is relatively simple to reconnect all the fragments so that we have a true whole-structure representation. As an example consider the hydrolysis–decarboxylation reaction shown in eq 6.



2OVX1&VO2&3UY1&1 QVY1&3UY1&1

2OV/ X /VO2 QV/Y

(6)

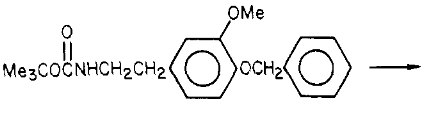
Application of the synthesis segment to the fragment lists resulting from the third level fragmentation produces the reaction site symbol strings in the lower half of eq 6.

The input to the program consists of the reactant and product WLNs; the output comprises the WLNs, the reaction sites, and the analysis fragments, these being the monocycles produced in the second fragmentation and/or the truncated linear chains generated in the fourth. The program contains about 3000 COBOL statements and has been run in 32K words on the University of Sheffield ICL 1906S computer.

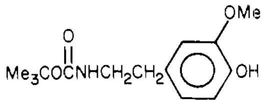
4. RESULTS AND DISCUSSION

The file of 9197 one-reactant, one-product reactions used in previous work[9] was processed in 587 cpu s, analyses being produced for 7415 reactions (80.6%). The remainder can be divided into two classes: those failures arising from limitations in the methodology and those from limitations in the program. A total of 1154 reactions (65%) of the cases are in the latter class, these arising from a variety of restrictions such as ring systems with nonconsecutive locant paths, too many analysis fragments for available storage, variable-valency heteroatoms, and the like. Such limitations could, of course, be overcome relatively easily. More important are the 628 failures due to limitations in the method. These can be subdivided into three classes: (i) 404 reactions for which a unique reaction site could not be generated using the criterion mentioned at 3 above, (ii) 119 reactions for which no common fragments could be detected between the reactant and product molecules, and (iii) 105 reactions for which all fragments on both sides of the equation were eliminated.

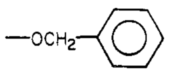
Examples of type (i) failures are shown in eq 7–9. In eq 7 there are two possible reaction sites in the reactant, these being represented by the character strings and substructures a and b. A frequent reason for failure is a substituent at a fusion point since if both rings change and if the substituent is involved in the change, the program cannot know to which monocycle the substituent should be allocated; this is exemplified by the methyl group in eq 8, which could be attached either to the #T3OTJ or to the #L9 AVTJ ring. Occasionally an ambiguity is noted where one does not actually



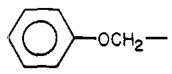
1X1&1&OVM2R CO1 DO1R



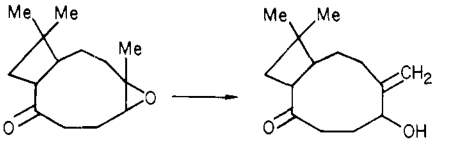
1X1&1&OVM2R DQ CO1



(a) *O1* R

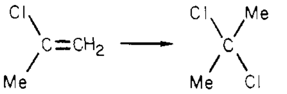


(b) R *O1* (7)



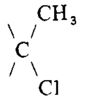
T D394 EO IVTJ - - D1 L1 L1

L49 EV IYTJ B1 B1- -HQ IU1 (8)



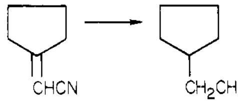
GY1&U1 GX1&1&9 (9)

exist; thus for the reaction shown in eq 9, the program generates the two product reaction site strings G/X/1 and X/1/G, both of which describe the substructure

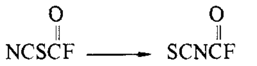


An ambiguity is therefore presumed to exist.

The second type of failure occurs mainly with small molecules producing only a limited number of fragments; two examples are shown in eq 10. Cases where all fragments are eliminated arise primarily from changes in ring saturation patterns since these are not explicitly defined in the monocycles produced in the second level fragmentation; the only information on monocycle-unsaturation is that obtained from the "T" or "&" symbols immediately prior to the "J" ring bracket in the parent system WLN.



L5YTJ AU1CN L5TJ A2Z

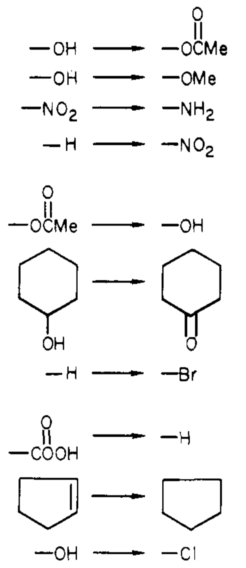


NCSVF SCNVF (10)

We have produced a printed index to the file of analyzed reactions. Each entry comprises three levels of description. The most general is the fragment which is the prime key in the index; more detailed information is provided by the reaction site while the search can be made still more specific by consulting the original WLNs. We have carried out trial searches upon this index with a fair degree of success in terms of relevant material retrieved; this may be taken to support the conjecture that the changes in the WLN symbol strings are a reasonable description of the actual structural changes involved. A more systematic evaluation will be reported shortly.

The 7415 reactions gave a total of 29 609 index entries, four per reaction. The frequency distributions, both of fragment and reaction site types, follow the general pattern noted by previous authors;[21,22] thus of the 1862 analysis fragment types identified, whether in reactant or product molecules, one, the substituent *OH, occurred 1534 times while 524 types occurred only once. The corresponding figures for the reaction sites, i.e., the symbol strings produced by the synthesis procedure, are one 54-fold and 4452 unique occurrences. The ten most frequent reaction sites are listed in Scheme I; all of these are very simple changes, in line with the findings of Garagnani and Bart.[22] The extremely specific descriptions provided by the reaction sites, 60% of them unique, suggest that the method of analysis described in this paper could be applied to significantly larger files; retrieval would be carried out using a WLN string search[23] upon the reaction sites, this being preceded by an automatically generated bit-screen.[24] As such facilities are widely used for searching WLN structure files, searches of reaction files generated as described in this paper could be implemented using currently available software with only minor modifications.

Scheme I[a]



[a] All substituents are upon rings and all monocycles are fused.

5. CONCLUSIONS

We have described an automatic method for detecting differences in the WLN symbol strings representing reactant and product molecules in a chemical reaction. A program has been used to process a file of 9197 one-reactant, one-product reactions for which analyses were obtained for 7415 (80.6%) though this success rate could be significantly increased. The descriptors derived from these analyses have been used to produce a printed index to this file and trial searches have been carried out upon it. The method is applicable to significantly larger files or, indeed, to much smaller ones using manual encoding.

ACKNOWLEDGMENT

We thank the Institute for Scientific Information for providing the WLN tapes from which the database was created and Dr. G. E. Vleduts (I.S.I.) and D. Bawden (Pfizer U.K.) for helpful discussions. P.W. thanks the Department of Education and Science for an Information Science Research Studentship.

REFERENCES AND NOTES

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

(2) J. Valls, "Reaction Documentation", in "Computer Representation and Manipulation of Chemical Information", W. T. Wipke, S. R. Heller, R. J. Feldmann, and E. Hyde, Ed., Wiley, New York, N.Y., 1974.

(3) G. E. Vleduts, "Concerning One System of Classification and Codification of Organic Reactions", *Inf. Storage Retr.*, **1**, 117-146 (1963).

(4) J. E. Armitage, and M. F. Lynch, "Automatic Detection of Structural Similarities among Chemical Compounds", *J. Chem. Soc. C*, 521-528 (1967).

(5) J. E. Armitage, J. E. Crowe, P. N. Evans, M. F. Lynch, and J. A. McGuirk, "Documentation of Chemical Reactions by Computer Analysis of Structural Changes", *J. Chem. Doc.*, **7**, 209-215 (1967).

(6) G. E. Vleduts, "Development of a Combined WLN/CTR Multilevel Approach to the Algorithmical Analysis of Chemical Reactions in View of Their Automatic Indexing", British Library, Research and Development Department, Report No. 5399, 1977.

(7) M. F. Lynch and P. Willett, "The Automatic Detection of Chemical Reaction Sites", following paper in this issue.

(8) J. M. Harrison and M. F. Lynch, "Computer Analysis of Chemical Reactions for Storage and Retrieval", *J. Chem. Soc. C*, 2082-2087 (1970).

(9) R. Clinging and M. F. Lynch, "Production of Printed Indexes of Chemical Reactions. I. Analysis of Functional Group Interconversions", *J. Chem. Doc.*, **13**, 98-102 (1973); "II. Analysis of Reactions Involving Ring Formation, Cleavage, and Interconversion", *J. Chem. Doc.*, **14**, 69-71 (1974).

(10) M. F. Lynch, P. R. Nunn, and J. Radcliffe, Final Report to the British Library, Research and Development Department on the Project "Development and Assessment of an Automatic System for Analysing Chemical Reactions", BLR&D Report 5236, 1975.

(11) M. F. Lynch, "Screening Large Chemical Files" in ref 1, pp 177-194.

(12) G. W. Adamson, J. Cowell, M. F. Lynch, A. H. W. McLure, W. G. Town, and A. M. Yapp, "Strategic Considerations in the Design of a Screen System for Substructure Searches of Chemical Structure Files", *J. Chem. Doc.*, **13**, 153-157 (1973).

(13) L. Hodes, "Selection of Descriptors According to Discrimination and Redundancy. Application to Chemical Structure Searching", *J. Chem. Inf. Comput. Sci.*, **16**, 88-93 (1976).

(14) C. E. Granito, S. Roberts, and G. W. Gibson, "The Conversion of Wiswesser Line Notations to Ring Codes. I. The Conversion of Ring Systems", *J. Chem. Doc.*, **12**, 190-196 (1972).

(15) S. D. Bedrosian, and M. B. Milne, "Graphical Representation for Automated Retrieval of a Class of Fused Six-Rings", *J. Chem. Inf. Comput. Sci.*, **17**, 47-49 (1977).

(16) D. Lefkowitz and A. R. Gennaro, "A Utility Analysis for the MCC Topological Screen System", *J. Chem. Doc.*, **10**, 86-94 (1970).

(17) E. Hyde, F. W. Matthews, L. H. Thomson, and W. J. Wiswesser, "Conversion of Wiswesser Notation to a Connectivity Matrix for Organic Compounds", *J. Chem. Doc.*, **7**, 200-204 (1967).

(18) J. B. Hendrickson, "Systematic Synthesis Design. IV. Numerical Codification of Construction Reactions", *J. Am. Chem. Soc.*, **97**, 5784-5800 (1975).

(19) A. Leo, D. Elkins, and C. Hansch, "Computerized Management of Structure-Activity Data. III. Computerized Decoding and Manipulation of Ring Structures Coded in WLN", *J. Chem. Doc.*, **14**, 65-69 (1974).

(20) J. E. Ash, "Connection Tables and Their Role in a System", in ref 1, pp 156-176.

(21) M. F. Lynch, "The Microstructure of Chemical Data-Bases and the Choice of Representation for Retrieval", in ref 2.

(22) E. Garagnani, and J. C. J. Bart, "Organic Reaction Schemes and General Reaction-Matrix Types. III. A Quantitative Analysis", *Z. Naturforsch., Teil B*, **32**, 465-468 (1977).

(23) J. E. Crowe, P. Leggate, B. N. Rossiter, and J. F. B. Rowland, "The Searching of Wiswesser Line Notation by Means of a Character-Matching Serial Search", *J. Chem. Doc.*, **13**, 85-92 (1973).

(24) C. E. Granito, G. T. Becker, S. Roberts, W. J. Wiswesser, and K. J. Windlix, "Computer-Generated Substructure Code (Bit Screens)", *J. Chem. Doc.*, **11**, 100-110 (1971).

The Automatic Detection of Chemical Reaction Sites

MICHAEL F. LYNCH* and PETER WILLETT

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

Received December 5, 1977

An approximate structure-matching algorithm is described which rapidly identifies substructures common to the reactants and products of a chemical reaction. The deletion of these features results in the identification of those parts of the reacting molecules that have been changed in the course of the reaction; at the same time it is possible to locate the reaction sites within their parent molecules so that substructural searches could be performed both for reacting and nonreacting features. The procedure has been tested upon a sample file of 340 reactions, and intuitively reasonable analyses were obtained for 315 of them (92.6%); a detailed failure analysis is given. Potential applications to computer-aided synthesis design and to the production of large files of chemical reactions are discussed.

1. INTRODUCTION

Work in this department[1,2] has led to the development of automatic methods for the detection of the overall structural changes occurring in organic reactions. This has been achieved by breaking the reacting species down into sets of fragments, eliminating the duplicate fragments, those parts of the molecules that ostensibly remain unchanged, and rebuilding the remaining features to synthesize a reaction site. As a fragmentation process is involved, a degree of ambiguity is present, and it has not proved generally possible to determine the exact location of the reaction sites within their parent molecules.

Vleduts has recently suggested a procedure whereby this might be achieved.[3] His approach consists of an atom-by-atom mapping of one reacting molecule onto another so as to identify the largest common substructures and, by subtraction, the differences engendered by the reaction. As no fragmentation is involved, a much more specific localization of the reaction site may be obtained, and it should be possible to identify the bonds which have been broken or formed in the course of the reaction. The algorithm involves the identification of maximal subgraphs common to the two sides of the equation; in contrast to the problem of graph isomorphism,[4-7] maximal subgraph isomorphism has been little studied due to the greater complexity of the problem.[8,30]

It is well known that isomorphism can be determined by a simple enumeration process;[7] in the present case a possible procedure would consist of generating all possible subgraphs (partial structures) from one graph (reacting molecule) and matching them against all possible subgraphs from the other.[9] The computation required may be substantially reduced if properties of such subgraphs which are invariant under isomorphism are taken into account; thus a reactant atom may not be mapped onto a product atom if the two atom types are different. Such "set reduction" techniques, initially described

0095-2338/78/1618-0154$01.00/0 © 1978 American Chemical Society