Computer Techniques for the Indexing of Chemical Reaction Information[†]

PETER WILLETT

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

Received October 30, 1978

Basic problems and possible approaches to the characterization of chemical reactions are considered and it is shown that structure-derived indexing is currently best suited to computer analysis. Two recently reported methods for automatic reaction indexing are compared, and it is concluded that both are necessary in a comprehensive reactions retrieval system.

SCOPE OF THE PROBLEM

The need for adequate means of retrieval for chemical reaction data has been apparent for many years; thus the preface of the first edition of Weyl's classic work on organic chemistry contained the statement that a scientist could hardly hope to be familiar with every one of the innumerable methods described therein.[1] More recently, both Meyer[2] and Valls[3] have called attention to the importance of providing access to reaction information, and this need is likely to become still greater with the introduction of computer-aided synthesis design systems.[4,5] As there are now over four million compounds known and as any one of them may be transformed into many others by suitable reactions, it is clear that the amount of potential data is quite enormous, and Hendrickson has pointed out that there are large classes of reaction for which there are as yet no known members;[6] at the same time, the number of reactions actually reported is steadily increasing.[7] There are often many ways in which a molecule or substructure may be synthesized and yet there are currently few aids to help the chemist in his search for a viable synthetic pathway. The difficulty of the problems involved may be evidenced by the widespread recognition of the achievements of synthetic chemists such as Corey and Woodward, and by the frequent use of terms such as "elegant" in reviews of syntheses: synthetic organic chemistry has been described as "an art in the midst of a science".[8]

It might have been expected that computers would provide a ready means for the control of reaction data, but this has not proved to be so. At least in part, this lack of success has been due to the limited amount of research carried out in the field—the documentation of a reaction presupposes quite sophisticated structure manipulation techniques, and these have only become generally available over the last 10 to 12 years—but the main problem is that, whereas a chemical molecule is a unique entity and thus susceptible to listing in a canonical form, such as via the CAS Registry System, a reaction has many characteristics, all of which may need to be stored for subsequent retrieval. Thus starting materials, products, reaction sites, catalysts, yields, mechanisms, experimental conditions, bond changes, and by-products may all need to be represented in the file. Many of these features may be indexed, via either free text or some form of coding, and then retrieved using standard bibliographical means while the reacting molecules can be handled by substructure search software; the crux of the problem is therefore the provision of a machine-readable representation of the changes engendered by the reaction. Apart from work carried out at Sheffield, almost all of the systems that have been examined or implemented to date, however effective in operation, have been very expensive to create since the identification and description of the changes, which we shall henceforth call the analysis, have been performed manually. Only when these intellectual tasks of content analysis and representation have been automated will the computer be useful in anything but a passive role as a repository and rapid searching tool.

POSSIBLE INDEXING APPROACHES

As with the characterization of compounds, the earliest forms of reaction indexing were based upon nomenclature, and to this day the most widely employed and most easily understood description is the use of a trivial name, usually that of the chemist(s) who originally discovered the reaction. Terms such as the Fischer indole synthesis, the Claisen rearrangement, and the Wolff–Kischner reduction are common in the literature, and several compendia of such names are now available. Nomenclature may occasionally prove very powerful in rapidly describing complex changes which may be quite difficult to characterize using more systematic methods, e.g. the Cope rearrangement, but in general the use of indexing terms which have no direct relationship with the reaction that they are supposed to describe may lead to severe problems in retrieval. Thus structurally similar transformations may be separated which might be considered more fruitfully in conjunction, and there may also be disagreement as to the exact extent of the reactions that should be considered under a single heading. However, the greatest deficiency is simply the lack of coverage offered by such a system since the great majority of reaction types has not been graced by a suitable appellation. Against this, we should note the large amount of work that has been carried out, primarily by CAS, in the use of systematic nomenclature for structure storage and search.[9,10] It may be that the availability of machine-readable files of compound names will lead to algorithmic means for reaction name generation using some form of word segmentation procedure.[11]

Nomenclature apart, Valls has identified two possible means of analyzing the reaction.[3] Either the reaction may be described in terms of the differences between the reactant and product structures or the transformation itself is described without regard to the actual (sub)structures involved. Structure based indexing methods are limited in that the initial and final states of the reacting molecules may not adequately describe the exact nature of the change that is taking place; thus an ester may be hydrolyzed to the corresponding acid by a variety of means, but the overall transformation is the same in all cases. Ideally a much deeper study of the exact nature of the change that has occurred is needed, this implying the application of some degree of mechanistic knowledge to the indexing. At least three objections may be brought against such an approach. Firstly it may be difficult to state the mechanism without a complex series of experiments. Next, if the mechanism can be identified, or is already known, some means must be found to represent the various charge-transfer complexes, electron shifts and bond migrations in some machine-readable form. Finally, at a purely pragmatic level, a synthetic chemist may be primarily interested in determining

[†] Presented in part in the symposium, "Retrieval, Analysis and Indexing of Chemical Reactions", 176th National Meeting of the American Chemical Society, Miami Beach, Fla., Sept 12, 1978.

0095-2338/79/1619-0156$01.00/0

whether a certain transformation has been reported in the literature, irrespective of the manner in which the reaction takes place.

For these reasons, and because structural data is readily available on a large scale in machine-readable form, we have chosen to investigate structure-based automatic indexing techniques. The aim of the work is to permit the use of reaction diagrams in a chemical information system in much the same way as we now manipulate structure diagrams; ideally, the methods of analysis should result in representations which can be manipulated using conventional structure-handling software with the minimum of alteration.

AUTOMATIC INDEXING METHODS

The basis for our work has been a paper by Vleduts[12] in which he pointed out that "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 molecules do not change their structures". This being so, a reaction may be characterized by a comparison of the structures of the reacting molecules which results in the discarding of those parts which have not been altered in the course of the reaction; we are thus led to the concept of the reaction site which is those substructures which have been involved in the change. Vleduts suggested that the sites alone would be sufficient to characterize a reaction, but this severely limits the utility of a reactions file since searches for, e.g., the reduction of a ring carbonyl while an acyclic β-keto ester is unchanged or the opening of the five-membered ring in a steroid, cannot be carried out owing to the absence of any information in the file as to the environment of the reaction sites.[13]

As well as the need to provide descriptions of the unchanged parts of the molecules, a major factor influencing the choice of indexing technique is the need for efficiency in operation if large files of reactions are to become computationally feasible. We have therefore tended to use simple procedures which deal rapidly and effectively with the overwhelming majority of reaction types rather than more sophisticated techniques which, although of wider potential applicability, make much greater demands on computer resources. In particular, rigorous graph theoretical approaches, such as maximal common subgraph detection alogrithms,[14] may not be usable on a large scale without drastic modification. Insofar as we are limited to approximate approaches, it is clear that they must contain detailed error-detection routines so as to identify both faulty analyses and structural occurrences which are beyond their scope: this latter point is of great importance in a notation-based algorithm where certain symbol juxtapositions may not prove amenable to simple processing. Finally, we should emphasize the differences between a general-purpose reaction indexing program and one designed for indexing the reactions used in a computer-aided synthesis system. The main purpose of the latter is to characterize some small number of reactions of proven synthetic value which are applicable to a wide range of structural environments; conversely a general indexing program must be able to process all reactions reported, irrespective of the extent of their synthetic utility. It is this difference that makes a large reactions file a natural complement to a synthesis design program since, having found that a goal molecule may readily be obtained by application of some general transform, search of a comprehensive file may well reveal a particular reaction variant directly applicable to the change of interest.

Work in Sheffield over several years[15-20] has recently led to the development of two methods for the automatic analysis of chemical reactions which seem to satisfy the constraints above.[13,21,22] Although designed for use with different types of structure representation, WLN and connection tables, and with different applications in mind, printed index production and machine search, they are based on a common principle. This is to identify in the reacting molecules substructures which are as large as possible, subject to the restriction that they correspond to features present upon both sides of the equation. Once these areas have been noted as common, the atoms or WLN symbols contained therein may be flagged in some way and the process repeated upon the unmarked parts of the molecules until no further common areas may be found; the remaining portions of the molecules will then correspond to the reaction sites. In the connection table approach[22] the common features are circular substructures which have been judged to be isomorphic using an approximate graph-matching procedure based upon an adaption of the Morgan algorithm; in the second method, the identification of identical WLN symbol strings is used to determine the common features after the application of a multi-level fragmentation procedure.[21]

The use of WLNs implies a categorization of reaction types upon the basis of symbol, rather than substructural, differences but in many cases there is found to be a close correspondence between the two. This is due to the especial prominence given by the notation to those features which are of prime importance in synthetic work, such as rings and many unsaturated linkages and functional groups; thus an analysis based on WLN may be expected to give a simple and precise result for many common reaction types. In other cases, however, there may be little similarity between the reactant and product notations even though large parts of the molecules are unchanged in the course of the reaction such as in the formation of a ring from purely acyclic precursors. Also, the fact that a few symbols may represent quite large numbers of atoms and bonds implies both that a change may be described in somewhat generalized terms and that quite complicated symbol manipulations may be required in the course of the processing. Despite these limitations, the ability to provide character descriptions of the substructures involved allows one to produce cheap printed indexes of reactions which could be used in a manner similar to permuted WLN compound lists; an evaluation of such an index showed it to possess a retrieval capability comparable to that obtainable from a commercially available reaction documentation service.[23]

The analyses resulting from the connection table approach, conversely, are only searchable in a wholly computerized system, access to the file being via a range of bit screens which allow the specification of molecular formula, atom, bond, and ring requirements for both the reaction sites and the parent molecules.[13] The need for mechanized search is compensated for by the ability to carry out simultaneous substructure searches for both reacting and nonreacting features, by the variety of retrieval access modes provided, and by the high degree of reaction site localization afforded by the analysis. The first of these facilities, dual access to both reacting and nonreacting features, is not available from the WLN analysis where the initial means of access is via the noncommon fragments; a subsequent inspection of the reaction site notations and the original WLNs is needed to identify the nonreacting features required by the query.

It is found that the two types of analysis are complementary in their coverage of the reactions in our file. Both methods deal satisfactorily with a wide range of functional group changes and acyclic addition and elimination reactions, the connection table approach in greater detail, but ring changes are processed quite differently. The WLN analysis has been designed to isolate complete monocycles involved in a reaction whereas the connection table analysis identifies first those individual ring atoms involved; the former approach is ideal for ring formation and cleavage reactions but insensitive to small changes within an individual ring, whereas the converse applies to the latter approach. As ring formation and cleavage reactions account for at least 20% of the file studied,[19] it can be seen that both analyses are needed if a comprehensive retrieval service is to be provided. The ring change information could be obtained using some form of ring perception algorithm but such techniques may prove quite expensive in terms of computer time, whereas the WLNs of the ring systems in the reacting molecules, if available, may be processed very rapidly because the smallest set of smallest rings has been previously isolated in the determination of the notation. The presence of the WLN symbol strings also provides a second level of search for those reactions which match the query bitstring; such multi-level searching is common in industrial substructure search systems.[24] Accordingly, the experimental reactions retrieval system now being developed in Sheffield will include the results of both types of analysis; this work will be reported shortly.

CONCLUSIONS

In this paper we have discussed some of the problems involved in the indexing of chemical reactions and basic considerations in the design of a computerized indexing system. Structure-derived techniques are, currently, most easily applied and a qualitative comparison is made of two automatic indexing methods which use the machine-readable structure representations of the reacting molecules as the basis for their characterizations. One method, based upon WLN, is particularly useful for the analysis of reactions involving ring changes and for the production of printed indexes; the connection table approach involves an analysis at the individual atom and bond level and is thus capable of a more precise localization of the reaction site given a suitable search system.

ACKNOWLEDGMENT

Thanks are due to the Institute for Scientific Information for assistance in the presentation of this paper, to Michael Lynch for helpful discussions, and to the Department of Education and Science for the award of a British Library Information Science Research Studentship.

REFERENCES AND NOTES

(1) T. H. Weyl, "Die Methoden der organischen Chemie", 3 vols., Georg Thieme, Leipzig, 1901-1911.

(2) E. Meyer, "Information Science in Relation to the Chemists' Needs" in "Chemical Information Systems", J. E. Ash and E. Hyde, Eds., Ellis Horwood, Chichester, 1975.

(3) J. Valls, "Reaction Documentation" in "Computer Representation and Manipulation of Chemical Information", W. T. Wipke, S. R. Heller, R. J. Feldman, and E. Hyde, Eds., Wiley, New York, 1973.

(4) M. Bersohn and A. Esack, "Computers and Organic Synthesis", *Chem. Rev.*, **76**, 269-282 (1976).

(5) W. T. Wipke and W. J. Howe, Eds., "Computer-Assisted Organic Synthesis", *Am. Chem. Soc., Symp. Ser.*, No. 61 (1977).

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

(7) E. Garfield, G. S. Revesz, and J. H. Batzig, "The Synthetic Chemical Literature from 1960 to 1969", *Nature (London)*, **262**, 307-309 (1973).

(8) J. B. Hendrickson, "Systematic Synthesis Design", *Top. Curr. Chem.*, **62**, 49-172 (1976).

(9) G. G. Van der Stouw, P. M. Elliott, and A. C. Isenberg, "Automated Conversion of Chemical Substence Names to Atom-Based Connection Tables", *J. Chem. Doc.*, **14**, 185-193 (1974).

(10) R. G. Dunn, W. Fisanick, and A. Zamora, "A Chemical Substructure Search System Based on *Chemical Abstracts* Index Nomenclature", *J. Chem. Inf. Comput. Sci.*, **17**, 212-219 (1976).

(11) F. H. Allen and W. G. Town, "The Automatic Generation of Keywords from Chemical Compound Names: Preparation of a Permuted Name Index with KWIC Layout", *J. Chem. Inf. Comput. Sci.*, **17**, 9-15 (1977).

(12) G. E. Vleduts, "Concerning One System of Classification and Codification of Organic Reactions", *Inf. Storage Retr.*, **1** (2/3), 117-146 (1963).

(13) P. Willett, "Computer Analysis of Chemical Reaction Information for Storage and Retrieval", unpublished Ph.D. thesis, University of Sheffield, 1978.

(14) J. J. McGregor and P. Willett, "Use of a Maximal Common Subgraph Algorithm in the Identification of the Ostensible Bond Changes Occurring in Chemical Reactions", in preparation.

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

(16) 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).

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

(18) 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).

(19) R. Clinging and M. F. Lynch, "Production of Printed Indexes of Chemical Reactions. II. Analysis of Reactions Involving Ring Formation, Cleavage, and Interconversion", *J. Chem. Doc.*, **14**, 69-71 (1974).

(20) M. F. Lynch, P. R. Nunn, and J. Radcliffe, "Production of Printed Indexes of Chemical Reactions Using Wiswesser Line Notations", *J. Chem. Inf. Comput. Sci.*, **18**, 94-96 (1978).

(21) M. F. Lynch and P. Willett, "The Production of Machine-Readable Descriptions of Chemical Reactions Using Wiswesser Line Notations", *J. Chem. Inf. Comput. Sci.*, **18**, 149-154 (1978).

(22) M. F. Lynch and P. Willett, "The Automatic Detection of Chemical Reaction Sites", *J. Chem. Inf. Comput. Sci.*, **18**, 154-159 (1978).

(23) D. Bawden, T. K. Devon, F. T. Jackson, S. I. Wood, M. F. Lynch, and P. Willett, "A Qualitative Comparison of Wiswesser Line Notation Descriptors of Reactions and the Derwent Chemical Reaction Documentation Service", *J. Chem. Inf. Comput. Sci.*, **19**, 90-93 (1979).

(24) J. E. Ash and E. Hyde, Eds., ref 2.