Use of a Maximal Common Subgraph Algorithm in the Automatic Identification of the Ostensible Bond Changes Occurring in Chemical Reactions

JAMES J. MCGREGOR† and PETER WILLETT*

University of Sheffield, Western Bank, Sheffield, S10 2TN, United Kingdom

Received December 2, 1980

A fast procedure is described for the discovery of the largest substructure common to the reactant and product molecules in a chemical reaction. Once this substructure has been found, it is possible to identify automatically those bonds in the reacting molecules which have apparently been broken or formed in the course of the reaction; these bond changes could be used as indexing terms for the retrieval of chemical reaction information.

INTRODUCTION

A possible approach to the classification of chemical reactions involves the identification of the bonds broken or formed in the course of a reaction. This technique, first proposed by Weygand,[1] was developed by Theilheimer and used as the basis for the indexing of the series *Synthetic Methods of Organic Chemisty*,[2] which has now been incorporated in the Derwent Chemical Reactions Documentation Service (CRDS)[3] by using methods originally developed by the Pharma Documentation Ring.[4] A recent comparison of two methods for indexing reactions emphasized the importance of bond-change information for synthetic planning,[5] and bond-change codes were included in the experimental reaction retrieval system described by Eakin and Hyde.[6]

Systems to date have generally required the manual identification of the bond changes that have occurred, and the question arises as to whether such identifications could be performed automatically. Early work in this area was carried out by Mishchenko et al.[7] and Harrison and Lynch[8] following suggestions made by Vleduts;[9] more recently, Vleduts has described a more sophisticated procedure by which the required information could be obtained.[10] His algorithm involves the atom-by-atom matching of the set of reactant molecules involved in a reaction with the set of product molecules so as to identify the largest substructure common to the two sides of the reaction equation. The identification of bonds not included in this maximal common substructure (MCS) but having one or both of their atoms included in it then enables the detection of those bonds which have been broken or formed in the course of the reaction. Cone et al. point out[11] that the detection of the MCS is equivalent to the derivation of the maximal common subgraph of two undirected labeled graphs, and these workers described a method for the identification of such subgraphs by using a procedure developed by Levi;[12] similar or related algorithms have been described by Bersohn and Mackay[13] and Varkony et al.[14] As in all isomorphism problems, computational requirements rise rapidly with an increase in structural complexity: thus Cone et al. found that the comparison of pairs of steroids required an average of 100-s CPU time on a PDP-11/45 for the identification of all substructures common to the two molecules,[11] and Varkony et al. required 45 s on a PDP-10 to identify the largest substructure common to three marine sterols.[14] Such timings are not acceptable if many sets of structures are to be processed, as would be required for the inclusion of bond-change information in a large chemical reaction data base. This paper describes a fast procedure for the provision of such information.

† Department of Computer Science.

* To whom correspondence should be addressed at the Department of Information Studies.

AN APPROXIMATE PROCEDURE FOR THE IDENTIFICATION OF MAXIMAL COMMON SUBSTRUCTURES

The identification of all substructures containing k atoms which are common to two structures containing m and n atoms, respectively, requires, at most,

$$\frac{m!n!}{k!(m-k)!(n-k)!} \tag{1}$$

atom-by-atom comparisons.[12] If the common substructures are to be identified in a reasonable amount of time, means must be found of reducing this very large number of comparisons. The normal means of improving the efficiency of graph isomorphism procedures is the inclusion of heuristic procedures which limit the number of comparisons required by consideration of various atom or bond properties which are invariant under isomorphism.[14-16] Examples of such techniques for the maximal common subgraph problem have been described by Cone et al. and Varkony et al., but a more drastic reduction in computation is possible if some of the atoms and bonds in the structures could be entirely eliminated from consideration prior to the iterative search, i.e., if m and n could be reduced.

Lynch and Willett have described a rapid method for the detection of certain of the subgraph isomorphisms present in the pair of graphs representing the sets of reactant and product molecules involved in a chemical reaction.[17] The procedure, which is derived from the set reduction technique introduced by Sussenguth,[15,16] identifies equivalent circular substructures on the two sides of a reaction equation without a detailed examination of the constituent atoms and words. While efficient in execution, the procedure is approximate in that nonisomorphic substructures may be identified as equivalent in a small number of cases: however, since the establishment of graph isomorphism is generally a fast procedure (at least in comparison with the detection of subgraph and maximal common subgraph isomorphisms), it would be possible to eliminate such mismatches at a relatively small computational cost by using an isomorphism detection procedure. The identification and subsequent deletion of the common substructures permits a rapid localization of the reaction sites, that is, the parts of the reacting molecules which have been involved in the course of the reaction and which contain the bonds which have been broken or formed, the atoms added or eliminated, and certain of the unchanged atoms and bonds.[17] The localization of the reaction sites in the reactant and product molecules has been shown to be sufficiently precise to enable searches to be made for both reacting and nonreacting substructures in a file of chemical reactions.[18] The number of atoms in the reaction sites will generally be significantly less than the number in the original molecules, and

0095-2338/81/1621-0137$01.25/0

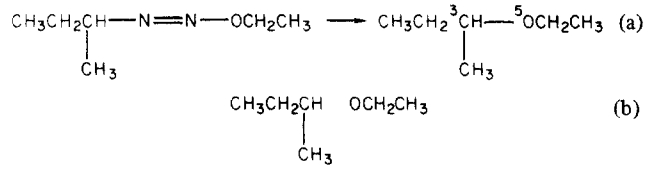


Figure 1. (a) Reaction. (b) Maximal common subgraph.

the reduction in number is such that it is possible to identify the maximal common substructures (MCS) for the two reactions sites relatively quickly, given an appropriate algorithm.

McGregor[19] has described a backtrack search procedure for extracting the maximal common subgraph of two graphs. In that paper, Levi's definition of the term "subgraph" and hence of the term "maximal common subgraph" was discussed and shown both to be a poor measure of similarity when used to compare the structure of two graphs and to be inappropriate for use in enumerating bond changes in a reaction. Levi's definition of a subgraph (substructure) requires that if two nodes (atoms) are included in the subgraph, then so must any arc (bond) connecting these nodes (atoms). In order for a maximal common subgraph algorithm to be used in detecting bond changes in a reaction, a weaker definition of a substructure is required, since it must be possible for two atoms of a molecule to be included in a substructure even if the bond connecting them is not included. For example, the MCS of the reacting molecules in Figure 1 is shown in the lower part of the Figure, where it will be seen that the bond connecting atoms 3 and 5 in the product molecule has been omitted in the MCS, although these two atoms have been included. Once such a maximal common subgraph has been found, a simple comparison of it with the reactant molecule will indicate which bonds have been broken in the reaction, and a comparison with the product molecule will indicate which new bonds have been formed. The definition we have used for an MCS is essentially the same as that used by Vleduts[10] for what he called a "maximal overlapping substructure".

The algorithm[19] performs a backtrack search for the set of correspondences between the atoms of one molecule and the atoms of the other molecule which result in the correspondence of the maximal number of bonds. Such a systematic search would be grotesquely inefficient if all possible mappings of the atoms of one molecule onto the atoms of a second molecule were considered. However, the search is organized in such a way as to construct possible mappings one pair of atoms at a time, and whenever a new atom pairing is tentatively added to the current partial mapping, an upper bound can be calculated for the number of bond correspondences which could appear in any complete mapping based on the current partial atom mapping. If this upper bound does not exceed the number of bond correspondences permitted by the best complete atom mapping found so far, then no atom mappings which include the current partial mapping need be considered. This can eliminate the need to examine large numbers of possible combinations of atom pairs.

The effectiveness of such a search pruning technique depends heavily on the order in which pairs of atoms are selected for inclusion in the mapping. If good solutions (i.e., mappings involving as many bond correspondences as possible) can be found early on in the search, pruning will subsequently occur more frequently. Methods of ordering the search with a view to doing this were extensively discussed in ref 19. At each step, first consideration should be given to the atom pairing which adds the maximum number of potential bond correspondences to the current partial mapping (such an atom pairing will not necessarily be included in the best complete solution but is likely to be included in a good solution). The molecular structure representations used in this work for the reacting molecules were Crossbow connection tables in which each atom is represented by a symbol which describes both the atomic type and the pattern of the bonds surrounding it; these symbols were used as a further guide for the selection of possible atom pairings since preference was given to pairs of atoms which had the same symbol in addition to being of the same atomic type.

We can summarize the criteria used for selecting the next tentative atom pairing (i,j), where i is an atom from the reactant and j is an atom from the product, as follows:

(1) Atom i is chosen to maximize the number of bonds connecting atom i to other atoms in the reactant which have already been tentatively paired.

(2) Alternatives for atom j are then considered in an order determined according to the following priorities: (a) Atoms i and j have the same Crossbow labeling, and the structure of bonds connecting atom i to other tentatively paired reactant atoms matches the structure of bonds connecting atom j to the corresponding product atoms. (b) Atoms i and j have the same atom type, and the structure around i matches the structure around j as in (a). (c) Atoms i and j have the same Crossbow labeling. (d) Atoms i and j have the same atom type.

We achieve (1) by ordering the atoms in the reactant prior to conducting the search. The set of atoms identified by the Lynch and Willett algorithm is selected as a starting point, and the remaining atoms are ordered by selecting at each step the atom that is connected to the maximum number of previously selected atoms. This ordering is then used to determine the choice of atom i at each stage in the search. We achieve (2) by associating, with each atom i in the reactant, a "priority subset" of the atoms in the product to which atom i could correspond. Atom j is a member of this subset only if the atoms in the current partial correspondence that are connected to atom i in the reactant have been mapped to atoms that are connected to atom j in the product. These subsets are updated each time a new tentative atom pairing is made. The set of atoms in the product with the same Crossbow symbol as a given atom from the reactant is easily constructed prior to backtracking and the representation of such sets as bit patterns makes for their efficient storage and manipulation during the search process. A detailed discussion of the use of priority subsets is given by McGregor.[19]

Care must be taken in handling reactions which have involved bond-order changes. A mapping involving a correspondence between two double bonds is better than a mapping with the same number of bond correspondences but in which a double bond corresponds to a single bond in the other molecule. In cases where a bond order has changed as a result of the reaction, it will be found that the correspondence of two bonds, one double and one single say, permits a mapping with more bond correspondences than would be the case if the double bond corresponded to any available double bond in the other graph. In counting the number of bond correspondences in a given solution, correspondences between the second and subsequent bonds of multiple bonds are counted separately but are given a weight which is a small fraction of the weight given to other bond correspondences. This ensures that a mapping involving a correspondence between two multiple bonds does not take priority over some other mapping which preserves more of the structure.

The reaction sites found by the algorithm of Lynch and Willett are used as the starting point for the backtrack search algorithm which thus needs to consider only pairings from the atoms in the reaction sites to identify the MCS. The efficiency of the search is hence considerably increased, since both m and n in eq 1 are much smaller than when the entire molecules are used for the detection of the MCS; as will be shown below,

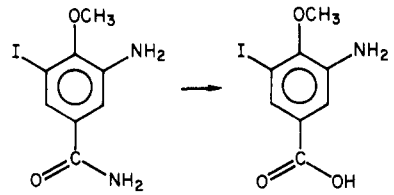


Figure 2.

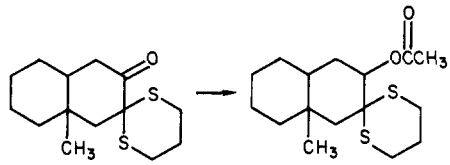


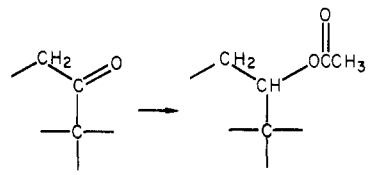
Figure 3.

Table I

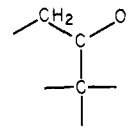
reactions processed	292
ring bonds	32
successful analyses	237
unsuccessful analyses	23

the increase in efficiency is sufficient to allow the search algorithm to process large files of reactions in an acceptable amount of processor time.

For an illustration of the use of the reaction sites, consider the reaction shown in Figure 2. The reaction site change for this transformation is —C(=O)—NH_2 → —C(=O)—OH, and application of the search algorithm yields the MCS — C=O. Comparison of the MCS with the reaction sites shows that a carbon–nitrogen bond has been broken (C—N) and a carbon–oxygen bond formed (C↔O). Similarly, if we consider the reaction shown in Figure 3, the reaction site change is



and the MCS is

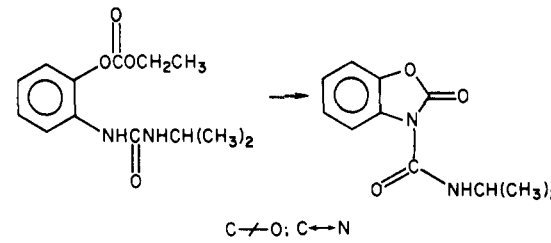


and the bond changes may be denoted by C≠O.

RESULTS AND DISCUSSION

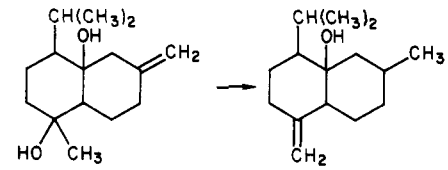
The backtrack search algorithm was implemented in ALGOL68 on the University of Sheffield ICL 1906S computer and was used to process a sample file of 292 pairs of reaction sites produced by the algorithm of Lynch and Willett. The results obtained are detailed in Table I.

The 237 successful analyses are those which have been judged as giving an accurate notification of the ostensible bond changes which have occurred in the course of the reaction. The term "ostensible" is used since it must be emphasized that our procedure will identify only those bonds which appear to have been altered and not necessarily those which were in fact broken or formed. For an exemplication of this, consider the following reaction which involves the hydrolysis of an ester: —C(=O)OCH_2CH_3 → —C(=O)OH. In such a case, the MCS algorithm will decide that the alkyl–oxygen bond has



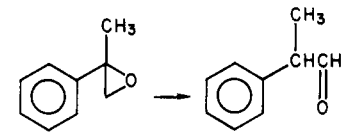
C—O; C←N

Figure 4.



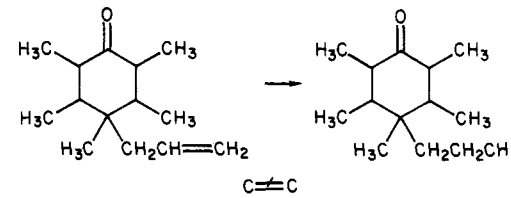
C—O; C≠C; C≡C

Figure 5.



C—O; C=O

Figure 6.



C≠C

Figure 7.

been broken irrespective of the actual mechanism of the reaction. This may be felt to be a limitation of the procedure, but in many cases, the mechanism may be in doubt or may be strongly dependent upon reaction parameters such as pH or temperature. Moreover, apart from the problems involved in building mechanistic knowledge into an information retrieval system, synthetic searches may not require mechanistic considerations, and, finally, neglect of the mechanism ensures that all such hydrolysis reactions may be retrieved without the need to search for several different bond changes. Analogously, all reactions of the form $-NO_2$ → $-NH_2$ will be considered to involve the rupture of two nitrogen-to-oxygen double bonds even if the reaction is known to have involved a nucleophillic substitution. Use of the ostensible bond changes does mean that searches where the mechanism is of importance may yield many false drops. With this proviso, some of the successful analyses are shown in Figures 4–7: in each case we give the full reaction equation and the derived bond changes using the notation given earlier.

There are occasions where the procedure does result in less successful analyses. Errors in the identification of the reaction site have been discussed briefly in the previous section and in more detail by Lynch and Willett.[17] We consider here the errors arising from the second part of our procedure, i.e., during the identification of the MCS. Such errors appear to be due in large part to the fact that the basic data available for these experiments consist of reactions involving one reactant and one product molecule only without any reagent being specified. Consider the simple reaction —CH=O → —CH=N—OH involving the action of hydroxylamine on the carbonyl group. However, since the MCS algorithm is una-

ware of the reagent, the product oxygen atom is presumed to be the same as that in the reactant so that the bond changes in the reaction are analyzed as C ≠ O; C ⇄ N; N ↔ O rather than C ≠ O; C ⇄ N. Given a full record of the reactant and product molecules, these erroneous analyses should not arise.

The connection table software used in this work cannot discriminate between single and multiple bonds within ring systems, and thus any reaction involving changes in ring bond multiplicity cannot be analyzed: there were 32 such reactions in the sample file studied. This limitation is a limitation of the structure representation used, not of the technique itself.

For a demonstration of the efficiency of the proposed two-step procedure, a sample file of 140 reactions was taken in which both the reactant and the product molecules in each case contained at most 24 atoms or bonds. The median time for the identification of the bond changes using the reaction sites was 0.04 s of central processor time; conversely, the median time for the identification of the bond changes using the complete molecules was 9.04 s, a drastic increase in computational requirements. The restriction in molecular size was imposed since the implementation of the backtrack search algorithm makes extensive use of bit handling procedures, and the computer used for this work is based on a word length of 24 bits. If reacting molecules of any size had been allowed in this comparison, the increase in computation when reaction site information is not available would have been still greater: the reaction site analysis is not generally affected by this limitation since there are few reactions in which the reaction sites contain more than 24 atoms or bonds.

CONCLUSIONS

We have described an efficient, but approximate, procedure for the automatic identification of the bonds which have been ostensibly broken or formed in the course of a chemical reaction. Apparently successful analyses were obtained for 237 of the 292 reactions in a sample file, but the success rate could be increased considerably if reagents were included in the reacting molecules and if a more detailed structure representation was to be used.

ACKNOWLEDGMENT

Our thanks are due to the Institute for Scientific Information for data, ICI Ltd. (Pharmaceutical Division) for software, and George Vleduts for helpful discussions in the early stages of this work.

REFERENCES AND NOTES

(1) C. Weygand, "Organisch-Chemische Experimentierkunst", Vol. 1-3, Barth, Leipzig, 1938.
(2) W. Theilheimer, "Synthetic Methods of Organic Chemistry", Vol. 1, Karger, Basel, 1946.
(3) Derwent Publications Ltd., Rochdale House, 128 Theobalds Road, London, W.C. 1, England.
(4) O. Schier, W. Nübling, W. Steidle, and J. Valls, "A System for the Documentation of Chemical Reactions", *Angew. Chem., Int. Ed. Engl.*, **9**, 599-604 (1970).
(5) 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).
(6) D. R. Eakin and E. Hyde, "Evaluation of On-Line Techniques in a Sub-Structure Search System", in "Computer Representation and Manipulation of Chemical Information", W. T. Wipke, S. R. Heller, R. J. Feldmann, and E. Hyde, Eds., Wiley, New York, 1974.
(7) G. P. Mishchenko, G. E. Vleduts, and A. M. Shefter, "Automatic Indexing of Reactions in an Information Retrieval System for Organic Chemistry", *Nauk. Tekh. Inform.*, **10**, 13-17 (1965).
(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) G. E. Vleduts, "Concerning One System of Classification and Codification of Organic Reactions", *Inf. Storage Retr.*, **1**, 117-146 (1963).
(10) 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, London, 1977.
(11) M. M. Cone, R. Venkataraghavan, and F. W. McLafferty, "Molecular Structure Comparison Program for the Identification of Maximal Common Substructures", *J. Am. Chem. Soc.*, **99**, 7668-7671 (1977).
(12) G. Levi, "A Note on the Derivation of Maximal Common Subgraphs of Two Directed or Undirected Graphs", *Calcolo*, **9**, 1-12 (1972).
(13) M. Bersohn and K. Mackay, "Steps toward the Automatic Compilation of Synthetic Organic Reactions", *J. Chem. Inf. Comput. Sci.*, **19**, 137-141 (1979).
(14) T. H. Varkony, Y. Shiloach, and D. H. Smith, "Computer-Assisted Examination of Chemical Compounds for Structural Similarities", *J. Chem. Inf. Compt. Sci.*, **19**, 104-111 (1979).
(15) E. H. Sussenguth, "A Graph-Theoretic Algorithm for Matching Chemical Structures", *J. Chem. Doc.*, **5**, 36-43 (1965).
(16) J. Figueras, "Substructure Search by Set Reduction", *J. Chem. Doc.*, **12**, 237-244 (1972).
(17) M. F. Lynch and P. Willett, "The Automatic Detection of Chemical Reaction Sites", *J. Chem. Inf. Comput. Sci.*, **18**, 154-159 (1978).
(18) P. Willett, "The Evaluation of an Automatically Indexed, Machine-Readable File of Chemical Reactions", *J. Chem. Inf. Comput. Sci.*, **20**, 93-96 (1980).
(19) J. J. McGregor, "Backtrack Search Algorithms and the Maximal Common Subgraph Problem", *Software Practice and Experience*, in press.

Comments on a "Method for Generating a Chemical Reaction Index for Storage and Retrieval of Information"

JOHN M. BARNARD and PETER WILLETT*

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

Received February 13, 1981

A recently proposed method for generating numeric identifiers for chemical reactions is discussed. It is shown that the method depends upon the exact form in which the reaction is described and also the method results in the same identifier being assigned to different reaction types.

Mosby and Kier[1] have recently described an indexing system for chemical reaction information. Unlike previous work, which has concentrated on the identification and description of the substructural changes occurring in a reaction,[2] their method produces a single number which is claimed to provide an unambiguous and unique characterization of the change. However, the system as described appears to have two limitations which might restrict its usefulness for the retrieval of chemical reaction information. First, the value of the numeric identifier is strongly dependent upon the exact form in which

0095-2338/81/1621-0140$01.25/0