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

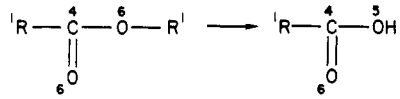
A recently proposed method for generating numeric identifiers for chemical reactions is discussed. It is shown that the method depends upon the exact form in which the reaction is described and also the method results in the same identifier being assigned to different reaction types.

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

the reaction is described, and secondly, the method results in the same identifier being assigned to different reaction types.

The numeric identifier for a reaction, $\Delta^1\chi^V$, is calculated by substracting the $^1\chi^V$ molecular connectivity index[3] for the reactant reaction site from that of the corresponding product site in a model of the generalized reaction. The procedure may be illustrated by the hydrolysis of an ester to an acid:



In the reaction equation R represents structural features which are not involved in the reaction, and the number appended to each nonhydrogen atom is its connectivity degree, δ^V.[3] The reactant $^1\chi^V$ is obtained from $1/(1.4)^{1/2} + 1/(4.6)^{1/2} + 1/(4.6)^{1/2} + 1/(1.6)^{1/2}$ and the product $^1\chi^V$ value from $1/(4.5)^{1/2} + 1/(4.6)^{1/2} + 1/(1.4)^{1/2}$ which results in a reaction connectivity index $\Delta^1\chi^V$ of −0.388.

This example is taken from the paper of Mosby and Kier[1] who state that "inclusion or omission of the carbonyl group in both structures of the model is immaterial, since the index value would be the same". If, however, the carbonyl group *is* omitted from the model, the reaction equation becomes

$$R\text{–}O\text{–}R \rightarrow R\text{–}OH$$

for which the $\Delta^1\chi^V$ is calculated to be −0.369.

This difference is caused by the arbitrary assignment of a δ^V value of 1 to all R groups, whereas the atom attached to the reaction site in a particular case may, in fact, have a different value. This leads to a different contribution to the overall $^1\chi^V$ values for reactant and product. It does not help to ignore the contributions of bonds to R groups. The value of the reaction connectivity index hence depends critically on the exact form of the general reaction model which is chosen for encoding.

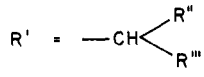
Another example is provided by the reaction model

$$R\text{—}CH\text{=}CH\text{—}CH_2\text{—}R' \rightarrow R\text{—}CH_2\text{—}CH\text{=}CH\text{—}R'$$

which gives a value of 0.0 for $\Delta^1\chi^V$. If, however, additional unchanged substructures are included in the model, a nonzero value may be obtained. Mosby and Kier give the model with R′ = —CH=CH—R, that is,

$$R\text{—}CH\text{=}CH\text{—}CH_2\text{—}CH\text{=}CH\text{—}R \rightarrow R\text{—}CH_2\text{—}CH\text{=}CH\text{—}CH\text{=}CH\text{—}R$$

which results in a value of 0.055, the same as that given by the rather different changes where



and

$$R' = -NH_2$$

Again, the four cases of

$$R' = C(=O)OH,\ —C\equiv R'',\ -NH\text{–}R'',\ \text{and}\ -CR''_3$$

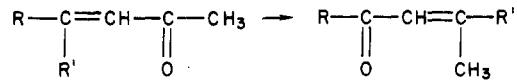
all result in a value for $\Delta^1\chi^V$ of 0.064, while R′ = $-NO_2$, –OH and $-NR''_2$ all give 0.072.

Mosby and Kier do refer obliquely to the different values produced by different models of the same reaction in the case of amine acylation. They suggest that by specifying (at least partially) the chemical nature of the R groups, more specific indexes can be calculated, which could be used to index separate files or subfiles under the general reaction model. This makes clear that the assignment of an index involves a subjective judgement by an encoder as to what should, or what should not, be included in the general model. Hence, as with manually assigned fragmentation codes, there may be a loss of recall if a searcher specifies a different model (or models) from that of the encoder. A comprehensive search for, e.g., all reactions involving the shift of a carbon–carbon double bond could only be carried out by specifying the $\Delta^1\chi^V$ values corresponding to all possible substructural environments for the basic reaction model. In their paper, Mosby and Kier list 12 different δ^V values for commonly occurring atoms in molecules, and hence even if attention is restricted only to these atom types and if variable substitution is allowed only at one end of the reaction model, 12 different $\Delta^1\chi^V$ values would need to be specified at search time (and proportionately more if additional substitution is permitted) to encompass the possible encodings of the general model.

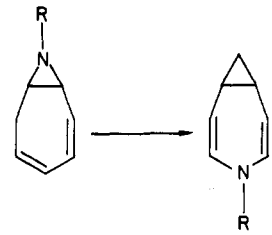
It is stated in the abstract and introduction to the paper by Mosby and Kier that their procedure results in a unique number, although it is later stated that this may not be so, and it would only be possible to find chemically feasible counterexamples by exhaustive testing. However, in one specific case, it can easily be shown that the same identifier is in fact assigned to quite different reaction types. This occurs with a variety of rearrangement and bond shift reactions, all of which have a value of 0.0 for $\Delta^1\chi^V$: examples include the double-bond shift described above, the comparable

$$R\text{—}N\text{=}CH\text{—}NH\text{—}R' \rightarrow R\text{—}NH\text{—}CH\text{=}N\text{—}R'$$

and the reactions



and



We also note that the substructures —C≡N and —NH—OH both contribute a value for $(\delta_i^V\delta_j^V)^{-1/2}$ of 0.2236, and this may lead to reaction models involving them being assigned equal $\Delta^1\chi^V$ values. Such equivalences are inevitable if substructures as small as the atom–bond–atom grouping are considered, though the problem was overcome in part by using higher order connectivities in studies of the use of molecular connectivity and other topological indexes in deriving numeric identifiers for complete structures.[4] However, even when these larger substructures were used with the molecular formula as a means of discriminating between molecules, it was not possible to obtain unique identifiers. Since it has not so far proved possible to derive a unique identifier for entire structures, it would seem most unlikely that a unique identifier could be obtained from consideration of pairs of substructures, that is, the generalized reaction model.

REFERENCES AND NOTES

(1) Mosby, M. A.; Kier, L. B. "Method for Generating a Chemical Reaction Index for Storage and Retrieval". *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 217–221.
(2) Willett, P. "Computer Techniques for the Indexing of Chemical Reactions". *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 156–158.
(3) Kier, L. B.; Hall, L. H. "Molecular Connectivity in Chemistry and Drug Research", New York: Academic Press, 1976.
(4) Evans, L. A.; Lynch, M. F.; Willett, P. "Structural Search Codes for On-line Compound Registration". *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 146–149.