A Qualitative Comparison of Wiswesser Line Notation Descriptors of Reactions and the Derwent Chemical Reaction Documentation Service

DAVID BAWDEN,* TREVOR K. DEVON, FRANK T. JACKSON, and SANDRA I. WOOD
Pfizer Central Research, Sandwich, Kent, England

MICHAEL F. LYNCH and PETER WILLETT
Postgraduate School of Librarianship and Information Science, University of Sheffield, Western Bank, Sheffield, S10 2TN England

Received September 13, 1978

Two methods of retrieving chemical reaction information are compared. One involves the generation of reaction descriptors automatically by an analysis of the Wiswesser line notation of the reacting molecules. The other, Derwent's Chemical Reaction Documentation Service (CRDS), involves manual indexing and uses a bond-change code to describe the reaction, with Ringcode for structural description. A series of reaction queries was searched using both systems; the results were qualitative and indicative of the general nature of the descriptions provided. Both systems are found to perform effectively with queries involving a definite reaction site change. The WLN system gives greater precision in some cases, owing to the varying levels of structural representation provided. CRDS is valuable where particular bond changes are specified and could be valuable in synthetic planning. Neither system performs well with queries where no definite reaction site is specified, and both would require additional concept indexing for full effectiveness. The WLN system has a useful potential for producing printed indexes of reactions.

INTRODUCTION

The provision of access to chemical reaction information has been a continuing problem for chemical information workers, and a variety of approaches has been adopted.[1,2] One method involves the automatic generation of reaction descriptions from machine-readable representations of chemical structures. Such descriptions may then be searched by computer or used for the production of printed indexes. This approach is likely to be of particular value within computerized chemical information systems. Investigations along these lines have been carried out for some years at Sheffield, using both connection table and Wiswesser line notation (WLN) representations of structure.[3] This work has resulted in the development of a method of reaction analysis based on WLN.[4] The WLNs for the reactant and product molecules are fragmented algorithmically, the fragments compared and duplicates eliminated, and the remaining fragments then recombined to give a description of the reaction site. The fragments constituting the reaction site are the main entry points to the reaction file; further information may be obtained by considering the fuller reaction site notations, and then the original WLNs. In a printed index these latter stages are carried out by scanning the entries under the appropriate reaction site fragment(s). In a computerized form, a string search procedure would be used on the reaction site notation and/or full WLN. At present this approach to reaction indexing is at an experimental stage.

A reaction documentation system based on structural concepts is, however, commercially available at the present time. This is the Chemical Reactions Documentation Service (CRDS) based on a system originally devised by the Pharma Dokumentation Ring.[6] This system describes reactions according to a representation of bonds formed and broken, derived from the coding used in Theilheimer's "Synthetic Methods" series.[7] Structures of reactant and product molecules are represented by the fragmentation code developed by the Pharma Dokumentations Ring (Ringcode).[8] The service is amenable to computer searching in batch mode, the reactions being searched by the bond change codes and Ringcodes for the reactant and product structures. It has been proposed by Derwent that this system will be made available on-line, with added keyword indexing.

A comparison of these two systems appears to be worthwhile in order to determine whether one type of reaction description is markedly superior to the other.

METHODOLOGY

There are major differences in the current state of implementation of the two systems. CRDS is a fully operational computerized system with facilities for searching on reaction conditions, etc., and allowing searching of reactant and product structures using Ringcode; the Sheffield WLN system is still at an experimental stage and has a printed index output with manual scanning. Thus, the provision for whole structure searching in the two systems is so different that, for example, relative precision figures would be meaningless. For these reasons, and because the main objective of the study was a comparison of the basic reaction descriptions provided, rather than of overall system effectiveness, no formal quantitative evaluation was attempted. Rather, the aim was to produce a qualitative understanding of the strengths, weaknesses, and potentialities of each method; quantitative evaluation would only be appropriate in the context of two fully operational systems. Ease of use and other human factors were not specifically examined, but had to be taken into account to some extent. Using printed tools, especially with a relatively small database, it is easy to scan a large proportion of the possible results. Some subjective judgment as to what would be realistic in a practical application was therefore necessary.

The database used for the evaluation consisted of 273 abstracts, chosen randomly from each of Volumes 22, 24, and 30 of Theilheimer's "Synthetic Methods"; this series forms the bulk of the CRDS database. Each one-step reaction from these abstracts, including all possibilities in the case of multistep reactions, was selected, giving a total of 582 reactions. The reactant and product molecules were encoded in fully expanded WLN and a printed index was produced for the set of reactions using the Sheffield programs;[4] this index was searched manually. The CRDS file, which includes the set of reactions under consideration, was searched using programs written at Pfizer for that purpose. The appropriate volumes of Theilheimer were also searched, both by the manual

0095-2338/79/1619-0090$01.00/0

Scheme I[a]

(a) $R\text{-}NO_2 \rightarrow R\text{-}NH_2$

(b) $R\text{-}O\overset{\overset{O}{\|}}{C}CH_3 \rightarrow R\text{-}OH$

(c) $R\text{-}\overset{\overset{O}{\|}}{C}H \rightarrow R\text{-}CN$

(d) $R\text{-}H \rightarrow R\text{-}NO_2$

(e) $\begin{matrix} R_1 \\ \quad\backslash \\ \quad C{=}O \\ \quad / \\ R_2 \end{matrix} \rightarrow \begin{matrix} R_1 \\ \quad\backslash \\ \quad CH\text{-}OH \\ \quad / \\ R_2 \end{matrix}$

(f) $R\text{-}X \rightarrow R\text{-}H$ or $R\text{-}D$

(g) $R\text{-}Br \rightarrow R\text{-}D$

[a] In all of the schemes and equations, R, R_1, R_2, etc., represent any nonreacting moiety. In this case R_1 and R_2 may not be part of the same ring and X represents a halogen.

coding system and by the keyword index. The purpose of this was to ascertain whether the keywording or codes would be useful in a specific situation where the structural description did not perform well.

A set of 18 queries was then constructed which was intended to represent the variety of reaction searches which a general purpose system should deal with; both general and specific queries were included. Because of the small size of the database there were in general few examples of each reaction type; this is to be expected from the known distribution of reactions[9,10] and is not greatly deleterious to the qualitative evaluation attempted here. Each abstract in the database was examined to determine those reactions relevant to each of the queries. This provided the ideal response sets against which the performance of the systems could be measured.

One example of the searching procedures is given here by way of illustration. The example shown in Scheme 1c involves the replacement of an aldehyde by a cyano group.

Relevant reactions will be analyzed by the WLN algorithms[4] as

$$\text{-CHO} \rightarrow \text{-CN}$$

or

$$\text{-CH}_2\text{CHO} \rightarrow \text{-CH}_2\text{CN}$$

where the groups may be attached either to rings or to acyclic substructures. Therefore the printed index was scanned under the reactant site fragments *VH, /VH, *1VH, and /1VH,[4] and then the possible reactions were checked by consideration of the reaction site notations.

In CRDS the (formal) breaking of C—H and C═O bonds and formation of a C≡N bond were encoded. The codes for aldehyde in the reactant and cyano group in the product were also included. The search output was the Theilheimer abstract numbers.

RESULTS

Seven relatively simple functional group interchange reactions were first considered. These are shown in Scheme I. Both the WLN systems and CRDS worked well on these examples. The first five examples were searched straightforwardly and all relevant answers found by both systems. In examples f and g, one possibly relevant reaction was missed by the WLN; this reaction was of the form

$$\text{R-COCHBr} \rightarrow \text{R-COCHD}$$

Because the WLN analysis algorithms produce the most detailed description possible of the reaction site,[4] the analysis was

$$\text{-COCHBr} \rightarrow \text{-COCHD}$$

rather than the more general

$$\text{-CHBr} \rightarrow \text{-CHD}$$

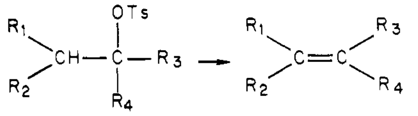
This is an example of the precision of the WLN approach. For a more general acyclic search, it is at present necessary to consider possible subsections of the reaction site character strings. Note that the reactions f and g require different coding in CRDS since the latter requires specification of both the bromine and the deuterium: the WLN searches are identical. The manual Theilheimer coding also proved reasonably efficient for these simple queries, although it involved a good deal of manual scanning.

A more specific query is shown in eq 1 (a ketonic group must

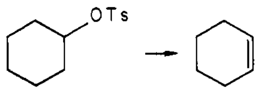
$$\text{R-NO}_2 \rightarrow \text{R-NH}_2 \qquad (1)$$

remain unreduced in R) which involves a consideration of the reactant/product structures. In the case of CRDS more specific Ringcoding than in the general case was needed; with WLN, scanning of the printed notations was sufficient (in a computerized system a stringsearch could be used). Both examples were straightforwardly searched in the two systems and the relevant reactions found.

Two elimination reactions, one with greater structural specificity, were tested (eq 2 and 3). Both systems found the relevant answers for the more general case, and eq 3 was found by product structure search.



(2)



(3)

Ts is *p*-toluenesulfonyl

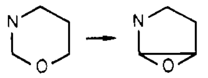
Two somewhat more complex reactions were then examined, as shown in eq 4 and 5. For eq 4, the addition of methyls to an unspecified substrate, the lack of information made it impossible to code any reactant or product structure for the CRDS. The large output resulting from use of the rather general reaction code included the relevant reactions. The relevant answers were found in WLN by scanning the full notations of those reactions involving the gain of two methyl groups. In equation 5 a WLN search was possible by looking through examples of formation of all appropriate heterocyclic rings, which retrieved the relevant examples. CRDS produced the relevant reactions, but with many spurious answers because of the ill-defined query.

$$\begin{matrix} \quad H \\ / \\ R \\ \backslash \\ \quad H \end{matrix} \rightarrow \begin{matrix} \quad CH_3 \\ / \\ R \\ \backslash \\ \quad CH_3 \end{matrix} \qquad (4)$$

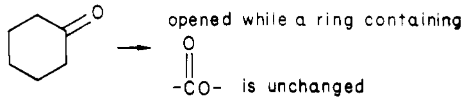
$R_1NH\text{-}X\text{-}NHR_2 \rightarrow$ a 5- or 6-membered ring with meta nitrogen atoms (5)

X is any carbon atom

Two ring reaction queries were considered (eq 6 and 7).



(6)



opened while a ring containing -CO- is unchanged (7)

The specific formation of a C–C bond within a defined heterocyclic ring in eq 6 presented no problems to either of

Scheme II

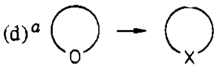
(a) $\text{-CH=CH-} \rightarrow \text{-CH}_2\text{-CH}_2\text{-}$

(b) $R_1 = \text{CH-CH}_2\text{-CH}=R_2 \xrightarrow{\qquad} R_1\text{-CH=CH-CH}=R_2$

(c) $\text{-CH=C=CH-} \rightarrow \text{-CH}_2\text{-}\overset{\overset{\text{O}}{\|}}{\text{C}}\text{-CH}_2\text{-}$

The three reactions above may occur in any environment

(d)[a]



[a] The circle represents any sort of ring and X represents nitrogen or sulfur.

the systems, both of which produced the relevant reactions from a straightforward search. The same was found with eq 7, where the presence of a carbonyl linkage in a ring in both reactant and product gives structural specificity. It is worth noting that in both these cases a search in the Theilheimer volumes via the reaction coding would be highly inefficient, since all the sections corresponding to formation or breaking of C–C bonds would have to be scanned. The CRDS system allows specification of reactant and product structure, while in the WLN system the reaction site fragments include the whole ring formed or broken.

Finally, four more general queries were selected, as shown in Scheme II. These in general caused the greatest problems to the systems. In the first three examples the structural environment is ill-defined. For all of these queries only acyclic WLN searches were made since in cyclic structures the reaction site fragments would comprise the whole monocycles involved in the change and each of these would have to be separately searched. No coding at all can be produced for a CRDS search for (b), while (a) can only be coded for "formation of a C–H bond", giving rise to many errors.

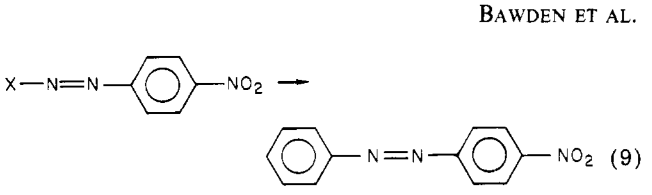
Keywording for general concepts, e.g., "hydrogenation" or "double bond migration", seems a more feasible way of dealing with general concepts of this sort. Thus in example c, the relevant reactions may be readily found from the index to the Theilheimer volumes, under the heading "ketones from allenes", emphasizing the value of keywording for concepts of this sort.

For query d in Scheme II, the relevant answers can be found from the WLN only by scanning all appropriate heterocyclic rings formed, an impractical procedure for a large database; similarly the CRDS search produces a very large output, because of the generality of the structure change. The relevant reactions are readily found from the Theilheimer volumes by the keyword phrase "replacement of oxygen, cyclic, by nitrogen/sulphur, cyclic".

Several of the queries were retested, using only the reaction (bond change) coding of CRDS, without reactant or product structures. In all cases a very large number of answers, many erroneous, resulted from the general coding. The specification of reactant and/or product structures by Ringcode is obviously an essential component of this system in a practical situation. The erroneous output from CRDS used in this way represents, as might be expected, a wide variety of reactions involving the same type of bonds broken or formed. When CRDS is used with structures specified, relatively few erroneous results appear. These are usually due to a bond change in a different part of the structure; thus the reaction shown in eq 8 was

$$R_1\text{-C=C=C-}R_2 \rightarrow R_1\text{-}\overset{\overset{\text{O}}{\|}}{\text{C}}\text{-CH}_2\text{-CH}_2\text{-}R_2 \qquad (8)$$

retrieved as an answer to (c) in Scheme II and that in eq 9 as an answer to (d) in Scheme I.



(9)

X is any halogen

It is difficult to make a direct comparison with likely errors in the WLN system, where the search was carried out by scanning a printed index. It is evident, from the number of occurrences of the various fragment keys from the WLN analysis, that some form of structure search may be necessary to limit the output. However, increased precision in searches may be obtained from the fact that a structural feature may be specified as being actually involved in the reaction, rather than merely being present in one of the reacting molecules, by its presence in the reaction site notation.

DISCUSSION

The most immediate impression gained from the results of this comparison is the great similarity between the performance of the two systems. In general, reactions occurring in well-defined structural environments may be searched efficiently by either system, whereas more generally stated queries are poorly dealt with. There are, nonetheless, distinct differences, as will be noted below.

An evaluation of this sort makes clear the large extent to which a reaction information system requires a structure search capability. The CRDS system requires the specific coding of reactants and products to reduce output to a manageable level. The WLN system to some extent incorporates structural information by including larger fragments in its reaction site analysis, but may still require some examination of reactant and product structures for maximum effectiveness. In many cases, however, the reaction site notations are sufficient to characterize the change. In a computerized system based on WLN a substructure search procedure would be required, operating on the reaction site and/or full reactant and product notations. The relative merits of Ringcode and WLN for substructure search would then have to be considered in a comparison of these reaction systems.[11]

The inclusion of considerable structural information in the reaction site notation often enables the WLN system to give a more precise analysis than CRDS. This is exemplified by the search for reaction f (Scheme I), where the presence of a ketone adjacent to the reaction site gave a different analysis, and in the ring formation and closure reactions where the monocycles involved were delineated both by the fragments and the reaction site notation. This is a very powerful feature of this type of analysis. Frequently reaction queries are specified in just this way, i.e., in terms of precise groups and ring systems, and an analysis based on WLN gives a rapid and reliable result. This is due to the extent to which such analyses retain the ability of the notation to describe structures in accordance with chemical intuition. In other cases, however, the two types of analysis are comparable.

Both systems are currently poorly equipped for handling the more general queries, i.e., those involving particular structural modifications in a variety of environments. For a manually indexed file, these problems could be alleviated by the use of intellectually assigned keywords similar to those used in the indexes of Theilheimer; this has been proposed by Derwent for the on-line version of CRDS. In the case of the WLN system, generic search capabilities could be obtained algorithmically both from the reaction site and parent compound notations.[12]

In summarizing the performance of the two systems, it is useful to consider the various access points to reaction information provided by the systems, and their appropriateness to the several types of reaction query likely to be encountered.

The two systems, as noted above, allow for approaches to reactions at different levels of structural specificity. CRDS allows searching directly for bonds formed and broken: in the WLN system this can only be achieved indirectly, by considering the possible reaction site changes brought about by a given bond change. The WLN system allows for a direct search on reaction sites; in CRDS an indirect search, combining bonds changed with structural features present or not present in reactant and product, is required.

The WLN approach gives three levels of structural description:[4] reaction fragments, reaction site notations, and parent structure WLNs. CRDS allows two levels: bond changes and bond changes plus structural features of reactant and product (which may or may not form part of the reaction site).

It is helpful, accepting some degree of over-simplication, to consider possible reaction enquiries as falling into three classes: structural concept related, reaction site related, and bond related.

Concept related questions are typified by the more general test queries above. They are expressed as structural concepts, such as those in eq 6 and 7, but are not restricted to anything other than a very general structural environment. Such questions are poorly dealt with by the structural reaction descriptions of both systems, and some form of concept indexing is desirable.

Reaction site related queries involve specification of the bond changes, with sufficient information on surrounding atoms to give a description of the reaction in chemically significant units: functional groups, ring systems, etc. It may well be that this type of query will predominate for a general organic reaction information service. These queries are dealt with by using the reaction site information from WLN, or the bond change with reactant and product structures in CRDS. As noted above, both systems dealt effectively with test questions of this sort, with the WLN system having some advantages.

Bond related queries involve specification of a bond or bonds broken or formed, without full specification of the reaction site change; such information could be particularly useful for synthetic planning. Direct access at the bond level would be a valuable component of a comprehensive reaction information system; this could be provided by manual indexing, as in CRDS, or by algorithmic means.[13,14]

Any of these three types of query may involve specification of structural features of the reacting molecules, not involved in the reaction site. This may be achieved in both systems, using the substructure searching capabilities of WLN and Ringcode, respectively.

A comparison of ease of use of the two systems would not be entirely meaningful, since much of the complexity of use of CRDS is due to structure searching via Ringcode. The corresponding computerized WLN searching was not undertaken; consideration of this factor again brings up the comparison of WLN and Ringcode as structure representation.

A WLN system is inherently more flexible than a fragmentation based system, since it allows production of printed indexes with whole structure representations provided as well as computer searches. This means that a WLN reaction system can provide both hard-copy output and printed desk-top tools which, given a working knowledge of WLN, could be used by the bench chemist. A fragmentation code system can only be efficiently used via computer, unless it relies upon a restricted coding such as the Theilheimer code. It may be that a printed index WLN reaction analysis system would be best used as an aid to immediate synthetic problems, perhaps with relatively small files. In this way full advantage could be taken of its ability to rapidly answer precise questions of the kind often encountered in day-to-day synthetic work. A useful application would be reaction indexing of internal data banks, where structures are already coded in WLN and where existing WLN handling programs could be utilized. This would give a reaction searching capability entirely compatible with in-house structure searching. For larger files, a computerized search system probably would be required.

CONCLUSIONS

The two systems, based on WLN analysis and on bond change descriptors with Ringcode, were both found to deal effectively with queries defined in terms of reaction site change, involving functional groups, ring systems, etc. Such queries may well predominate in general purpose reaction information systems. The WLN system provided greater precision in some cases, owing to the varying levels of structural representation provided. For some questions the bond change information in CRDS is valuable; this may be particularly useful for synthetic planning. Both systems perform poorly with concept related queries, where there is no specific reaction site indication. They both require some form of concept indexing for full overall effectiveness. A WLN-based system may be valuable in providing printed indexes of reactions.[15]

ACKNOWLEDGMENT

We thank Mr. D. A. Faulkner for valuable discussion and assistance with WLN encoding, and Dr. A. Finch of Derwent Publications for advice on CRDS usage. Peter Willett thanks the Department of Education and Science for an Information Science Research Studentship.

REFERENCES AND NOTES

(1) J. Valls, "Reaction Documentation". In "Computer Representation and Manipulation of Chemical Information", Wipke, W. T.; Heller, S. R.; Feldman, R. J.; Hyde, E., Eds., Wiley: New York, 1974.

(2) Valls, J.; Schier, O., "Chemical Reaction Indexing". In "Chemical Information Systems", Ash, J. E.; Hyde, E., Eds.; Ellis Horwood: Chichester, 1975.

(3) Willett, P., "The Automatic Analysis of Chemical Reaction Data". *Inf. Sci.* **1974**, *11*(4), 125–135.

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

(5) Derwent Publications Ltd., Rochdale House, 128 Theobalds Rd., London, WC1 England.

(6) Schier, O.; Nübling, W.; Steidle, W.; Valls, J. "A System for the Documentation of Chemical Reactions". *Angew. Chem., Int. Ed. Engl.* **1970**, *9*, 599–604.

(7) Theilheimer, W., "Synthetic Methods of Organic Synthesis", Basel: New York, 1946; Vol. 1.

(8) Nübling, W.; Steidle, W., "Dokumentationsring der chemisch-pharmazeutischen Industrie: Aims and Methods", *Angew. Chem., Int. Ed. Engl.*, **1970**, *9*, 596–598.

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

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

(11) Sasamoto, M.; Kubota, T.; Hamano, T.; Shinba, T.; Nakai, M. "A Qualitative Comparison of Wiswesser Line Notation with Ringcode", *J. Chem. Doc.* **1973**, *13*, 206–211.

(12) Granito, C. E.; Becker, G. T.; Roberts, S.; Wiswesser, W. J.; Windlinx, K. J. "Computer-Generated Substructure Codes (Bit Screens)", *J. Chem. Doc.* **1971**, *11*, 106–110.

(13) Vleduts, G. E., "Development of a Combined WLN/CTR Multilevel Approach to the Algorithmic Analysis of Chemical Reactions in View of Their Indexing", BLR&D Report 5399, London, 1977.

(14) McGregor, J. J.; Willett, P. Unpublished results.

(15) Since submission of this paper, CRDS has become available on-line, with added keywording.