Production of Printed Indexes of Chemical Reactions Using Wiswesser Line Notations

MICHAEL F. LYNCH,* PETER R. NUNN, and JANET RADCLIFFE

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

Received July 21, 1977

A simple categorization of reaction types, based on comparisons of WLN symbol strings, is reported. This permits the development of simple analyses which provide WLN descriptions of 70% of the reactions. A prototype index is described.

The problem of classifying chemical reactions or producing indexes for the use of chemists is intrinsically a complex one.[1] None of the methods used to date has satisfied all the requirements of chemists. The reason for this is that there are so many attributes of a chemical reaction in which chemists may be interested, e.g., reactions of a specific compound, synthesis of a certain product, reactions of certain functional groups, and reactions of compounds containing specific rings or other substructures. Reactions may involve one or more reactants, one or more products, or may require the presence of catalysts as well as certain conditions of temperature and pressure.

The problem must be broken down into its constituent parts. Perhaps the most complex part of the problem is how to define the change from reactant to product—the reaction analysis. Once an analysis has been made, a suitable notation by means of which the analysis may be described must be devised. Finally, an index based on the notation scheme must be produced; this may either be in the form of a multiple key file or may be organized as a printed index.

Work has previously been carried out at Sheffield using structures encoded both as Wiswesser line notations (WLN)[2,3,5] and as connection tables.[4,5] Both methods have been tested on a small database of reactions taken from *Current Abstracts of Chemistry* and *Index Chemicus* (CAC&IC).[2,3] Results using WLN indicated that about 40% of the file could be analyzed successfully. However, this method gave descriptions which were unduly brief, particularly for reactions involving changes to ring systems; it was not possible to locate the position of the reaction center, only to state what change ostensibly occurred. The second approach using connection tables was designed to produce an accurate analysis of the changes involved in the reaction. Although this method produced analyses for a higher proportion of the reactions, the consistency of the analyses, and therefore the formulation of queries and searches, proved to be a considerable problem.

As a first stage in developing more adequate methods of analysis of line notation descriptions of reactions, a simple categorization was introduced. This was applied to a sample of reactions taken from 10 months' issues of CAC&IC, comprising 9197 one-reactant/one-product reactions from a total of 9882 reactions.

Three different circumstances are readily determined by a comparison of the ring size numerals for reactant and product notations, giving the following categories ("R", denoting the phenyl group, is obviously not considered as a ring numeral): (1) reactions with no apparent change in the number or sizes or rings, (2) reactions with a change in the number or sizes of rings, and (3) other reactions.

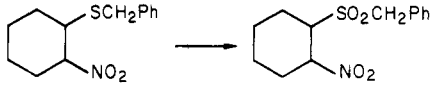
The proportion of reactions in each category was determined (Table I). This paper now reports the more detailed analysis of reactions in categories 1 and 2, and the preliminary design of an index for the former.

Table I. Categorization of 9197 One-Reactant/One-Product Reactions from 10 Months' Issues of CAC&IC

Reaction type	No. of reactions	% of database
1	4626	50.3
2	3063	33.3
3	1508	16.4
Total:	9197	100.0

TYPE 1 REACTIONS

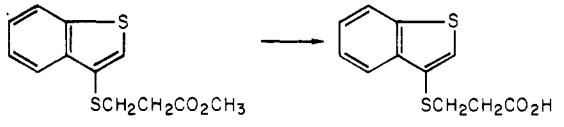
Reactions of this type involve changes to acyclic components of cyclic systems, as illustrated in eq 1. The character strings



(1)

L6TJ AS1R& BNW L6TJ ASW1R& BNW

within ring brackets (L or T, and J) were compared for both notations. If this comparison showed no differences, a table was established for each distinct ring system in each notation. For each locant position, the WLN string for substituents cited after the ring system was entered in the table. Comparison of the corresponding entries in the tables and extraction of the symbol strings denoting the changes provided a satisfactory characterization of the reactions in many instances, as illustrated in eq 2. While in the first case the characterization



(2)

T56 BSJ DS2VO1 T56 BSJ DS2VQ
Analysis S2VO1 S2VQ

might have been carried further to give O1→Q, loss of information about the immediate environment of the reaction site would result, and hence this further truncation was not undertaken.

This procedure failed to deal with a number of reactions in which the WLN symbols (and their locants) between the ring brackets showed only minor changes, particularly those involving:

(i) Interchange of the following pairs of symbols:

(a) M, N; H–N⟨ → –N⟨ Example 3
(b) U, Y; –C⟨ → =C⟨ Example 4
(c) V, Y; O=C⟨ → =C⟨ Example 5

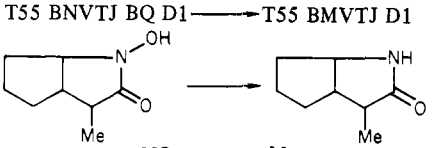
(ii) Insertion or deletion of WLN characters:

(a) V; O=C⟨ Example 6
(b) Y; =C⟨ Example 7

0095-2338/78/1618-0094$01.00/0

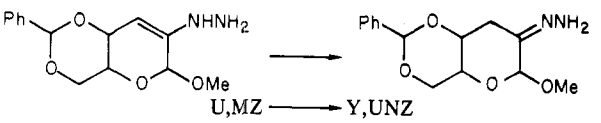
Tests for the conditions illustrated above were developed and included a comparison of locants cited both between and after ring brackets. If such a change was evidenced, the appropriate relationship was assumed to have been established. For instance, in reaction 3, the exchange of N and M is found at the B locant in the ring; this is further checked by identifying the loss of the symbol Q at the B locant.

T55 BNVTJ BQ D1 ⟶ T55 BMVTJ D1



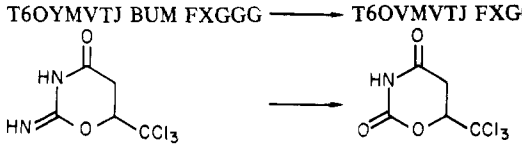
NQ ⟶ M (3)

T66 BO DO GO IUTJ CR& HO1 IMZ ⟶
T66 BO DO GO IYTJ CR& HO1 IUNZ



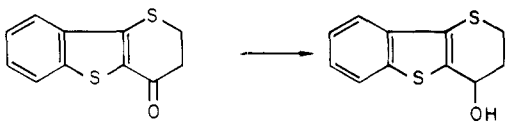
U,MZ ⟶ Y,UNZ (4)

T6OYMVTJ BUM FXGGG ⟶ T6OVMVTJ FXGGG



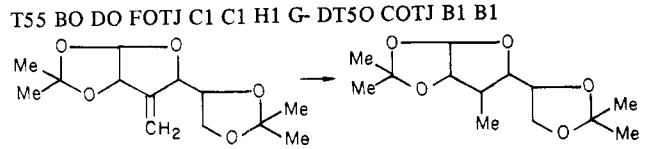
Y,UM ⟶ V (5)

T B656 CS FV HST&&J ⟶ T B656 CS HST&&J FQ



V ⟶ Q (6)

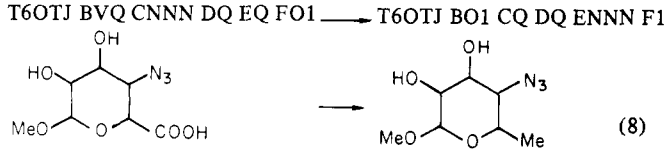
T55 BO DO FO HYTJ C1 C1 HU1 G- DT5O COTJ B1 B1 ⟶
T55 BO DO FOTJ C1 C1 H1 G- DT5O COTJ B1 B1



Y,U1 ⟶ 1 (7)

One problem with this type of analysis is caused by the ordering rules of WLN. For instance, in the simple reaction below (eq 8), the two notations are different at four of the locants on the ring. This occurs because the locants are lettered in different directions round the ring in the two notations.

T6OTJ BVQ CNNN DQ EQ FO1 ⟶ T6OTJ BO1 CQ DQ ENNN F1



(8)

The analysis for this reaction yields the following:

VQ → O1
Q → NNN
NNN → Q
O1 → 1

However, simply by eliminating identical pairs on opposite sides of the analysis, the overall reaction analysis VQ → 1 could be derived, and similar measures would be needed to deal with variations in ring paths.

Table II. Sample Reaction Index

Reactant	Product
	Q
1 E	1
OV1	Q
1 OV1	1 Q
OVR	Q
1 Q	1 OSW1
VQ	
VQ	1Q
S WG	S WMR D1
S WG	S ZW

No attempt has yet been made to analyze reactions involving changes in the saturation of the ring. Relatively simple algorithms could be devised to analyze reactions such as eq 9.

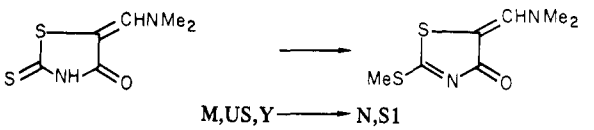
L5V BUTJ BX1&1&1 DE ⟶ L5VJ BX1&1&1

(9)

A deeper analysis of ring bonds would be required to deal with this and similar circumstances.

The simple algorithm developed to date gives the following results for the reaction 10. This analysis could be made more

T5MVYSYJ CU1N1&1 EUS ⟶ T5NVYSJ CU1N1&1 ES1



M,US,Y ⟶ N,S1 (10)

informative by linking the changes to show that there is only one reaction center.

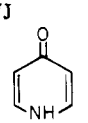
INDEX

A relatively simple index of type 1 reactions has been produced. The sort key for the index is the WLN symbols produced from the analysis. The WLN strings representing the reactant and product substituents are first compared character by character until a difference is found. The subsequent characters of the string provide the sort key, and the whole locant string is printed in the format illustrated in Table II, where the vertical line indicates the beginning of the sort key. This is for experimental purposes only; in an operational situation, one would also need to include the reaction itself in WLN form, with a reference to the source. Problems arise with this type of index when analyzing reactions involving two or more reaction centers. A more sophisticated solution is a computerized search system which can carry out logical procedures on more than one key simultaneously.

TYPE 2 REACTIONS

The analysis of type 2 reactions, i.e., those reactions showing a change in the number or size of rings, is carried out in two stages. The first of these is to identify the changes in the ring system itself, using the method described by Clinging.[3] This algorithm enables summaries of many simple ring changes to be produced and includes a canonical description of the atoms in the rings and their degree of saturation. The notation below gives the ring atom followed by a symbol for its state, e.g., H for a saturated carbon atom and U for an unsaturated carbon atom. The second stage of the analysis is to determine

T6M DVJ



(N,M)(C,U)(C,U)(C,V)(C,U)(C,U)

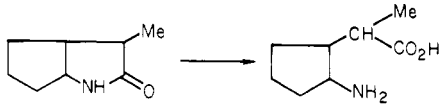
additional changes other than those occurring in the ring systems. This is carried out using the algorithm designed to analyze Type 1 reactions. Therefore, an analysis of the reactions including the acyclic parts of the molecule can be produced as illustrated in eq 11 and 12. These procedures

T6NN DNJ EVZ FMVH ⟶ T66 BNN EN GVM JNJ

Ring closure, forming: (N,N)(C,U)(N,M)(C,V)(C,U)(C,U) (11)

Substituents undergoing change: VZ, MVH

T55 BMVTJ D1 ⟶ L5TJ AZ BY1&VQ



(12)

Ring cleavage of: (N,M)(C,V)(C,H)(C,H)(C,H)

Substituents undergoing change: Z, Y1&VQ

provide analyses for 89% of type 2 reactions or 27% of the total file.

Although the successful percentage of analyses of type 2 reactions is high, the simpler ring changes predominate. Many of the ring system notations are too complex for analysis by software available to us, but for those not successfully analyzed, an index showing ring size changes alone may be produced.

CONCLUSIONS

It can be seen from this work that it is possible to analyze a substantial proportion of reactions using WLN. The algorithms, which entail manipulation of strings of WLN characters, are all relatively unsophisticated and have been developed in a short time. The programs could be developed to a considerably higher level to deal with a greater percentage of the database.

ACKNOWLEDGMENT

We thank British Library R&D Department for financial support, the Institute for Scientific Information for the data used, I.C.I. Pharmaceuticals Division for software, and G. W. Adamson, E. Hyde, and P. Willett for valuable discussions.

LITERATURE CITED

(1) J. Valls, and O. Schier, "Chemical Reaction Indexing in Chemical Information Systems", J. E. Ash and E. Hyde, Ed., Ellis Horwood, Chichester, 1975, pp 243–255.
(2) R. Clinging and M. F. Lynch, "Production of Printed Indexes of Chemical Reactions I. Analysis of Functional Group Interconversions", *J. Chem. Doc.*, **13**, 98–101 (1973).
(3) 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).
(4) Final Report to OSTI on the project "Computer Organisation and Retrieval of Chemical Structure Information", OSTI Report No. 5208, July 1974.
(5) Final Report to the British Library Research and Development Department on the project, "Development and Assessment of an Automatic System for Analysing Chemical Reactions", British Library R&D Dept. Report No. 5236, July 1975.

An Efficient Design for Chemical Structure Searching. II. The File Organization†

LOUIS HODES*
National Institutes of Health, Bethesda, Maryland 20014

ALFRED FELDMAN
Walter Reed Army Institute of Research, Washington, D.C. 20012

Received December 8, 1977

A novel file organization design for substructure search, based on hash coding is described. This organization permits search of only a portion of the file, the portion decreasing with increasing query specificity. The same organization is practical for full structure search, which is routinely used in searching for duplicates when updating the file. Statistics on a sample file of about 20 000 compounds are presented as well as its performance on typical queries.

INTRODUCTION

Part I[1] described a new method for generating molecular fragment screens based on heuristic and information theoretic principles.[2] It showed how the screens can be weighted and compacted through the use of superimposed coding. This paper continues the account by describing the file organization designed for this system. The delayed publication has awaited the generation of a sample file of over 20 000 compounds in order to present some of the preliminary results.[3]

Until now there have been two basic methods of designing files to facilitate substructure searching. The fragments or keys can be coded into a bit string so that the computer can rapidly reject compounds which do not match at least the pattern of the query structure. Alternatively, the fragments or keys can form an index and the corresponding inverted lists of structures can be intersected.

Although inverted file systems are generally considered superior for on-line applications, bit string systems have at least two advantages. The connection tables can be stored with the bit strings for atom-by-atom search of screen hits. Also, the bit strings can be used for full structure search as well as substructure search if one merely substitutes identity match for inclusive match of the query string. This kind of identity search can be very fast by binary search and even faster by hashing.

The advantages of inverted file systems lie in not searching the entire file in responding to a substructure query; only the

† Partially presented at the 168th National Meeting of the American Chemical Society, Atlantic City, N.J., Sept. 1974, and the Conference on Computer Graphics, Pattern Recognition, and Data Structure, Los Angeles, Calif., May 1975.