Table VI. Computer Processing Times for Principal Runs

| | CPU (Min:sec)[a] | Elapsed (mins.)[a] |
|---|---|---|
| ISF-Chemical Substance Conversion | 342:20 | 397.29 |
| ISF-General Subject Conversion | 217:40 | 246.69 |
| *CA-Condensates* Merge | 7:45 | 42.09 |
| CA-ISF Merge | 177:44 | 199.21 |

[a] IBM 360/65 MVT with HASP.

## COMPUTER RUNS

All of the computer runs made as part of this study were done on the University's IBM 360/65, operating under OS MVT with HASP, in high speed core. All programs were written in PL/1 Level F, using numerous macro and subroutine facilities which are equally suitable for either SDF or SFF. Run times in terms of both CPU and elapsed time for the major computer jobs are given in Table VI.

## ACKNOWLEDGMENT

The authors acknowledge the cooperation of Chemical Abstracts Service in making available some of the supplemental resources used in this portion of the ISF study. Sincere thanks also go to Howard Petrie, of Sheffield, England, who undertook much of the preliminary analysis of the General Subject segment of the ISF while a Visiting Foreign Scientist at the University of Georgia.

## LITERATURE CITED

(1) "Evaluation of Alternative Retrieval Techniques for the Integrated Subject File," Computer Center, University of Georgia, Athens, Ga., 1971.
(2) "Standard Distribution Format Technical Specifications Revised," Chemical Abstracts Service, Columbus, Ohio, 1971.
(3) "Data Content Specifications for the CA Integrated Subject File in Standard Distribution Format," Chemical Abstracts Service, Columbus, Ohio, 1971.
(4) "UGA Text Search System," 4 Volumes, Computer Center, University of Georgia, Athens, Georgia, January 1971.
(5) Park, M. K., Carmon, J. L., and Stearns, R. E., Jr., "Chemical Compound Retrieval Based on CA Formula Index Nomenclature," Computer Center, University of Georgia, Athens, Georgia, July 1971.

# Production of Printed Indexes of Chemical Reactions. I. Analysis of Functional Group Interconversions

R. CLINGING and M. F. LYNCH*
Postgraduate School of Librarianship and Information Science, University of Sheffield, Western Bank, Sheffield, S10 2TN, England

Received November 27, 1972

**A set of programs is being developed for the purpose of producing printed indexes of chemical reactions from a simple reactant/product data base. A program is described which identifies functional group interconversion reactions, hydrogenations, and dehydrogenations in a data base containing structures encoded as Wiswesser Line Notations. These reactions account for about 20% of a sample of 5104 reactions. Production of the data base is briefly described.**

Because of the great variety of organic reactions, indexing them is a complex process and although several methods have been evolved, none fully solves the problems. The least logical approach is to name the type of reaction after the author who first reported it—e.g., Claisen Condensation and Diels-Alder Reaction—a method which has been much used in the literature.[1-3] As this method presupposes a knowledge of the name of the reactions, its usefulness, especially for general searching, is severely limited, and so it was not considered suitable for computer indexing. A second, more systematic, approach is to base the name of the reaction on the functional groups which are different in the reactant and product molecules. This approach comes close to that used by most chemists when searching outside their own specialized fields and so deserves special consideration. Patterson and Bunnett[4] have taken this a stage further by combining the names of the functional groups involved to form reaction names, but this becomes unwieldy in all except the simplest cases. A third approach[5,6] looks at the bonds which change during the reaction; although this is amenable to computerization,[6,7] it produces indexes which are still difficult to search.

A number of systems[7-10] have been, or are being developed, for the automatic or semiautomatic searching of various types of chemical reaction data bases. In addition, Corey *et al.*[11-15] are developing systems to predict the best ways of tackling complex syntheses. Still lacking is an automatic method for indexing reactions, especially with a view to producing easily used printed indexes. The first stage of an attempt to solve this deficiency follows.

Chemical reactions are analyzed by comparing the records of the reactant and the product molecules, and seeing which entities are changed. A more thorough approach would need to take into account neighboring groups which may affect the course of the reaction, but this is outside the scope of the present work. Initially it was intended to approach the problem using structures encoded as connec-

* To whom correspondence should be addressed.

tion tables, largely because considerable experience[7] in their manipulation was already available. On the other hand, the only large file which can be easily used in the construction of a reaction data base is *Current Abstracts of Chemistry* and *Index Chemicus* (CAC & IC), which has the structures encoded as Wiswesser Line Notations (WLN's) and, although these can be converted to connection tables, there will inevitably be some loss. Connection tables allow a more thorough approach to the problem but their manipulation is more complex, and hence more time-consuming. In addition, line notations contain character strings which are equivalent to certain functional characteristics of the molecule and so can be used as nomenclatural aids in organizing indexes. Before finally deciding which was the best approach, a manual assessment of part of the data base was carried out; this indicated that a considerable portion of the file consisted of either ring formation/cleavage reactions or acyclic functional group interconversions. These types of reaction can be dealt with reasonably efficiently using WLN string searches, and so it was decided that the best approach would be to deal with the simpler reactions using the WLN's and only resort to connection tables in the most complex cases. As acyclic functional group interconversions represent the simplest cases, they have been dealt with first, and constitute the work reported in this paper.

## THE DATA BASE

The first requirement was to obtain a substantial data base with a minimum of effort, the most convenient source being CAC & IC. All new compounds recorded in CAC & IC are numbered and their structures encoded as WLN's, which are available on magnetic tape. Ten months' issues[16] of the hard copy version were scanned manually, selecting the reactants and products associated with reactions. For instance, in the example shown in Figure 1, three reactions (1 → 2, 2 → 3 and 3 → 4) were selected. The WLN's and molecular formulas of the compounds involved were extracted from the magnetic tape version, creating the required data base.

## CHARACTERIZATION OF THE DATA BASE

Before tackling the problem directly, it was necessary to know something of the characteristics of the data base, and for this a subroutine was developed to count the symbols which indicate functional features in the notations. This subroutine does not count numeric characters; it sums together those elements which are represented by pairs of characters. It also takes into account most forms of contraction, including multipliers, but ring of ring contractions are not dealt with.

Three months' issues of the data base, containing 10,275 compounds, were scanned to give an indication of the occurrence of the WLN symbols; these are listed in Table I.

Table I. Occurrence of WLN Symbols in 10,275 Compounds. Total Number of Symbols 75,041

| Symbol | No. of Compds. Containing Symbol | Total No. of Occurrences of Symbol |
|---|---|---|
| O | 5222 | 10130 |
| V | 5954 | 9830 |
| N | 4414 | 8308 |
| R | 3910 | 6379 |
| T | 5296 | 6061 |
| U | 3349 | 4892 |
| Q | 3082 | 4576 |
| Y | 3016 | 4523 |
| L | 2753 | 2912 |
| M | 2163 | 2801 |
| G | 1317 | 2583 |
| X | 1378 | 2193 |
| S | 1542 | 2054 |
| H | 1337 | 1475 |
| F | 372 | 1300 |
| W | 881 | 1145 |
| Z | 885 | 1041 |
| C | 543 | 872 |
| E | 468 | 629 |
| Rare element | 431 | 500 |
| P | 334 | 374 |
| K | 266 | 284 |
| I | 79 | 88 |
| B | 69 | 86 |
| D | 4 | 4 |
| A | 3 | 3 |
| J | 0 | 0 |

The results are more or less what was expected and differ little from those published by Granito *et al.*,[17] although direct comparison is not possible because of slightly different rules for symbol selection.

The same subroutine was used to determine the changes in the WLN symbol count which occur during reactions. This gave a very wide spread of results, even when the diagnostically unimportant symbols, R, X, and Y, were ignored. The section of the data base used consisted of 1707 reactions and gave, ignoring R, X, and Y, 785 different analyses of which 482 (61.5%) were unique and only 27 (3.4%) represented more than six reactions. None of the more common analyses (Table II) represent any changes in atoms other than carbon, hydrogen, and oxygen.

## THE ANALYSIS

For the actual analysis of the data base into reaction types, a new subroutine was developed to locate the character strings associated with functional groups. Only certain changes in the cyclic parts of the notation were considered at this stage; for instance, a ring count, excluding

147877

A NEW SYNTHESIS OF ALPHA-METHYLENE-GAMMA-BUTYROLACTONES.
MILLER R B. BEHARE E S.
UNIV CALIFORNIA, DEPT CHEM, DAVIS, CA 95616, USA.
J CHEM SOC D CHEM COMMUN 1970(7),402-3. RECD FEB 11, 1970.

A high yield of α-methylene-γ-butyrolactones can be obtained by treatment of a 3-methoxycarbonyl-lactone successively with formaldehyde–dimethylamine hydrochloride, methyl iodide, and heating in dimethylformamide.

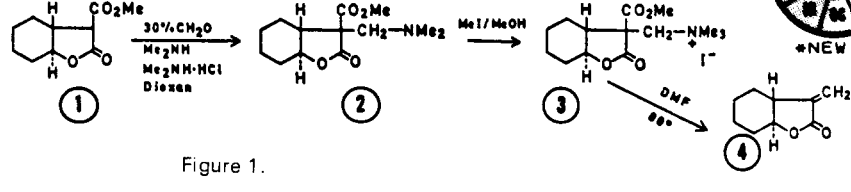



Figure 1.

Table II. Changes in WLN Symbol Counts during Reactions—Analyses Occurring in More Than 1% of the File

| Analysis | | No. of | |
|---|---|---|---|
| Group lost | Group gained | Occurrences | Percent of File |
| No change | | 61 | 3.58 |
| O | Q | 41 | 2.40 |
| U | — | 37 | 2.17 |
| V | — | 23 | 1.35 |
| V | Q | 23 | 1.35 |
| Q | O | 22 | 1.29 |
| Q | U | 22 | 1.29 |
| Q | V | 20 | 1.17 |
| Q | — | 19 | 1.11 |

unfused benzene rings, was incorporated. In addition, the carbonyl symbol (V) was counted wherever it occurred, whereas any other symbols were ignored if they occurred in the cyclic part of the notation. The character strings to be searched for where held in a dictionary, in store, as three-word records. If the group was terminal, the first word contained the string as it occurs within a notation, the second its equivalent if it starts the notation, and the third was kept for a count. The single character string V must always be the first one in the dictionary, but otherwise there were no restrictions. Non-terminal groups which have more than one representation—e.g., ester: V$\theta$ or $\theta$V—were located in consecutive three-word records. The strings were located in the first word of each record while the second word contained a positive integer which was the same for each representation of the same group. These later allowed the subroutine to identify pairs of strings which represented the same functional group so that they could be combined. Strings from the notation were compared with the dictionary in such a way that the same character would not be recorded in more than one string, and the longest string containing it would always take precedence; for instance, acetic acid (QV1) would be recorded as a carboxylic acid (QV) but not as an alcohol (Q) or a ketone (V). The exception is V within a ring notation which would always be recorded as such whatever its neighbors; for instance, butyrolactone (T5$\theta$VTJ) would be recorded as a ketone but not as an ester ($\theta$V) or an ether ($\theta$). This was done only because many cyclic carbonyl compounds behave atypically, owing to quinone-type character, and it was thought better to deal with them all together rather than as a series of separate but similar reactions. Also simple lactones, lactams, etc. tend to be opened during reactions and hence will not be dealt with by this program. The decision has been justified by the results obtained so far. Special allowance was made for locants, metallic elements, etc., but not for multipliers as the additional programming was not considered worthwhile.

Analysis of a reaction involves first an analysis of the reactant and product notations to determine the functional groups present. If at this stage there has been a change in the number of rings (excluding unfused benzene rings) during the reaction, it is passed on for further analysis, otherwise the reactant analysis is subtracted from the product analysis to give a reaction analysis. This technique is aimed only at simple reactions, and so if more than one different group is gained, or more than one different group is lost, the reaction is considered at a later stage. If this still leaves one type of group on each side of the reaction, a check is made to see whether the numbers of groups of these types are equal. If not, the reaction is considered later. Because this is a very simple technique, the resulting analyses, although they have high recall, have low precision, so further checks are required to obtain satisfactory results.

Table III. Reactions Used in Assessing the Program

| Reactant | Product | Reactant | Product |
|---|---|---|---|
| $RNO_2$ | $RNH_2$ | ROH | RCl |
| RCN | $RCO_2H$ | >CH—CCl< | >C=C< |
| >CHOH | >C=O | RH | RBr |
| >C=O | >CHOH | RBr | $RNH_2$ |
| >C=NOH | >C=O | RCHO | RCH=NOH |
| >C=O | >C=NOH | ROH | RI |
| >C=O | $>CH_2$ | RCN | $RCONH_2$ |
| RCN | $RCH_2NH_2$ | $RCONH_2$ | $RCO_2H$ |
| RH | $RNO_2$ | $RCO_2H$ | $RCO_2R^1$ |
| $RCO_2H$ | RH | ROH | $RO_2CR^1$ |
| RCl | ROH | $RO_2CR^1$ | ROH |
| $RCO_2H$ | $RCH_2OH$ | $RCO_2R^1$ | $RCO_2H$ |
| RH | RCl | ROH | $ROR^1$ |
| $RCH_2OH$ | $RCO_2H$ | $RNH_2$ | $RNHR^1$ |
| RCHO | $RCO_2H$ | $RCO_2H$ | $RCONHR^1$ |
| RCHO | $RCH_2OH$ | $RCONHR^1$ | $RCO_2H$ |
| >CH—C(OH)< | >C=C< | $RNH_2$ | $RNH.COR^1$ |
| $RCO_2H$ | RCHO | $RNHCOR^1$ | $RNH_2$ |
| ROR | RBr | $ROR^1$ | ROH |
| $RN_3$ | $RNH_2$ | $RCO_2R^1$ | RH |
| >C=NOH | $>CHNH_2$ | | |

If no change in the functional groups is detected, the molecular formulas are checked to see if any element other than hydrogen has been gained or lost; is so, it is passed on for further analysis. If not, and if hydrogen is gained or lost, the reaction is classed as a hydrogenation or dehydrogenation.

To test the remaining analyses, two dictionaries of reactions were set up, one to deal with reactions involving non-terminal groups and another to deal with reactions involving only terminal groups. In the latter case, the change in molecular formula during the reaction offers an efficient check as to whether the analysis is correct. Otherwise it is necessary to resort to a slightly less efficient method based on the change in the total WLN symbol count. Any analysis which is not found in either dictionary or fails to agree with the checks laid down in that dictionary is passed on for further analysis.

## EVALUATION OF THE RESULTS

The program was tested against a section of the data base which consisted of 5104 reactions, searches being carried out for the types of reaction shown in Table III together with hydrogenations and dehydrogenations.

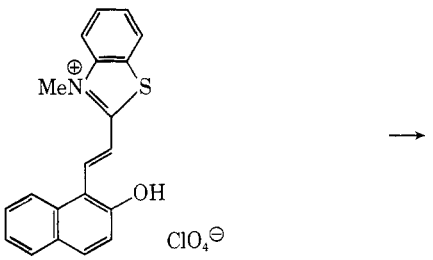
The program identified 996 (19.5%) reactions as being successfully analyzed. Manual checking showed 21 (0.41%) of these to be miscoded in the data base and only 15 (0.29%) to be incorrectly analyzed. In addition, a number of reactions, especially hydrogenations and dehydrogenations, involved double bond migration and in some cases, where non-terminal functional groups were involved, it was not possible to discriminate between two rather different reactions. For instance, the hydrolysis of an ester to an alcohol (Reaction 1) and the mixed anhydride reduction of an ester to an alcohol (Reaction 2) were not differentiated.

$$R.O.CO.R^1 \rightarrow ROH \quad (1)$$

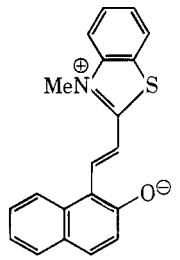
$$R.CO.O.R^1 \rightarrow R.CH_2OH \quad (2)$$

The analyses which manual checking showed to be incorrect are rather varied, and although some could be eliminated very easily, it is doubtful whether this would

prove economical in terms of the additional processing time involved. Several of the reactions involve terminal oxygen functions (Reactions 3 and 4) which are wrongly identified as ethers.

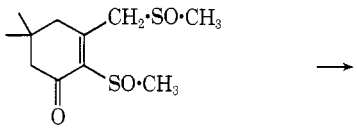


T56 BK DSJ B C1U1- BL66J CQ & G-θ4,

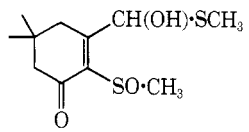


(3)

T56 BK DSJ B C1U1- BL66J Cθ &6/27



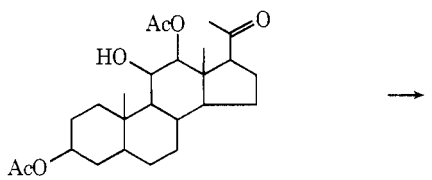
L6V BUTJ BSθ&1 C1Sθ&1 E E



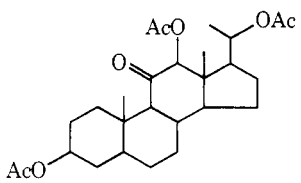
(4)

L6V BUTJ BSθ&1 CYQS1 E E

This could be eliminated if the strings θ∇ and θ& were included in the dictionary of functional groups, but this scarcely seems justified bearing in mind the rarity of such errors. In some cases, where the change taking place involves two reactions, the product of one and the reactant of the other mask one another giving erroneous results. For instance, in Reaction 5 the two carbonyl groups mask one another and the program decides that the change involved is the conversion of an alcohol to an ester. Only two such cases were found.



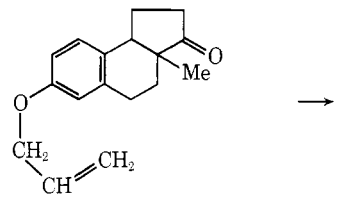
L E5 B666TJ A CQ Dθ V1 E FV1 θθV1



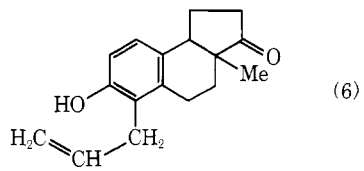
(5)

L E5 B666 CVTJ A Dθ V1 E FYθV1 θθV1

The most serious problem occurs when rearrangements which involve a change in the functional groups present are analyzed, for instance, the Claisen allyl ether rearrangement (Reaction 6) is classified as an ether-to-alcohol conversion. There seems no clear way at present of eliminating this type of discrepancy without also excluding some correct analyses.



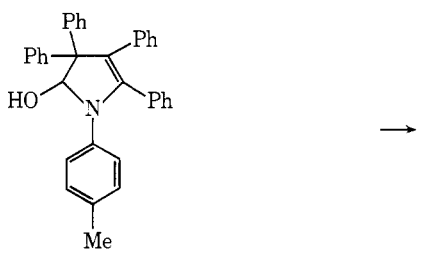
L B566 EVTT&J F Kθ2U1



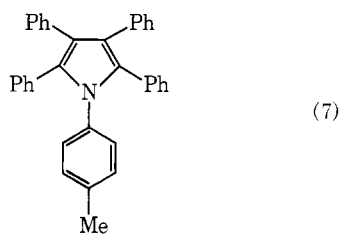
(6)

L B566 EVTT&J F J2U1 KQ

Nearly all errors occurred when the final check used the change in the WLN symbol count, but a few errors have occurred after molecular formula checks, for instance, dehydration with phenyl group migration in Reaction 7, and



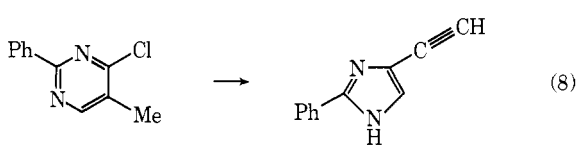
T5N BUTJ AR D& BR& CR& DR& DR& EQ



(7)

T5NJ AR D& BR& CR& DR& ER

ring contraction with HCl elimination in Reaction 8. These faults are not thought to be serious at this stage in the work, though it may be necessary to reassess them at a later date.



(8)

T6N CNJ BR& DG E T5M CNJ BR& D1UU1

## CONCLUSION

Strings of characters occurring within WLN's often correlate with functional groups present within the molecule. By comparing the strings present within reactant and product notations some of the simpler reactions, namely functional group interconversion, hydrogenations, and dehydrogenations, have been identified and analyzed into their reaction types. Such reactions constitute about 20% of those recorded in the data base, and the analysis is carried out with good recall and precision. This program will

be incorporated in a set of programs which is being developed to produce printed indexes of chemical reactions.

It is hoped that this will take the form of an alphabetical list of reactions listed according to the groups which change, but the final form will depend upon how efficiently cyclic reactions, rearrangements, etc. can be handled. For the more common transformations it is possible that details of reaction conditions will need to be considered in order to break up the listing for more convenient searching.

## EXPERIMENTAL

Programs were written in the ICL 1900 Series assembler language, PLAN, and run on the University of Sheffield's ICL 1907 computer. No accurate assessments of the running times of the programs have been made but, as an indication, a file of 1747 reactions was processed in 41 cpu secs., including program compilation and print out of results.

## ACKNOWLEDGMENT

One of us (R. C.) thanks the Office of Scientific and Technical Information, London, for the award of a Postdoctoral Fellowship in Information Science. We also thank the Institute of Scientific Information, Philadelphia, for giving permission to use files prepared by them, and the Experimental Information Unit, University of Oxford, who supplied these files. Thanks are also due to J. A. Bush, J. M. Harrison, A. H. W. McLure, and J. Radcliffe for preparing the reaction data base.

## LITERATURE CITED

(1) Krauch, H., and Kunz, W., "Organic Name Reactions," (translated by J. M. Harkin), Wiley, New York, London, and Sydney, 1964.

(2) Gowan, J. E., and Wheeler, T. S., "Name Index of Organic Reactions," Longmans, London, 1960.

(3) "The Merck Index of Chemicals and Drugs," 7th ed., pp. 1399–481, Merck and Co., Rahway, N. J., 1960.

(4) Patterson, A. M., and Bunnett, J. F., "Systematic Names for Substition Reactions," *Chem. Eng. News* **32,** 4019 (1954).

(5) Theilheimer, W., "Synthetic Methods of Organic Chemistry," S. Karger A. G., Basle, Vol. 1, 1947 to Vol. 26, 1972.

(6) Vleduts, G. E., "Concerning One System of Classification and Codification of Organic Reactions," *Inform. Stor. Retr.* **1,** 117–46 (1963).

(7) Armitage, J. E., Crowe, J. E., Evans, P. N., Lynch, M. F., and McGuirk, J. A., "Documentation of Chemical Reactions by Computer Analysis of Structural Changes," *J. Chem. Doc.* **7,** 209–15 (1967).

(8) Harrison, J. M., and Lynch, M. F., "Computer Analysis of Chemical Reactions for Storage and Retrieval," *J. Chem. Soc.* (C), 2082–7 (1970).

(9) Meyer, E., "The IDC System for Chemical Documentation," *J. Chem. Doc.* **9,** 109–13 (1969).

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

(11) Corey, E. J., "Computer-Assisted Analysis of Complex Synthetic Problems," *Quart. Rev.* **25,** 455–82 (1971).

(12) Corey, E. J., Wipke, W. T., Cramer, R. D., and Howe, W. J. "Computer-Assisted Synthetic Analysis. Facile Man-Machine Communication of Chemical Structure by Interactive Computer Graphics," *J. Amer. Chem. Soc.* **94,** 420–30 (1972).

(13) Corey, E. J., Wipke, W. T., Cramer, R. D., and Howe, W. J., "Techniques for Perception by Computer of Synthetically Significant Structural Features in Complex Molecules," *ibid.*, **94,** 431–9 (1972).

(14) Corey, E. J., Wipke, W. T., and Howe, W. J., "Computer-Assisted Synthetic Analysis for Complex Molecules. Methods and Procedures for Machine Generation of Synthetic Intermediates," *ibid.*, **94,** 440–59 (1972).

(15) Corey, E. J., and Petersson, G. A., "An Algorithm for Machine Perception of Synthetically Significant Rings in Complex Cyclic Structures," *ibid.*, **94,** 460–5 (1972).

(16) *Index Chemicus,* Vol. 35 (1969); *Current Abstracts of Chemistry and Index Chemicus,* Vols. 36, 37 and 38 (Nos. 1–5) (1970).

(17) Granito, C. E., Becker, G. T., Roberts, S., Wiswesser, W. J., and Windlinx, K. J., "Computer-Generated Substructure Codes (Bit Screens)," *J. Chem. Doc.* **11,** 106–10 (1971).