

## LITERATURE CITED

- (1) Granito, C. E., Roberts, S., and Gibson, G. W., "The Conversion of Wiswesser Line Notations to Ring Codes. I. The Conversion of Ring Systems," *J. Chem. Doc.*, **12**, 190-196 (1972).
- (2) Smith, E. G., "The Wiswesser Line-Formula Chemical Notation," McGraw-Hill, New York, N. Y., 1968, p 187.
- (3) Reference 2, p 118; note that the WLN for No. 6 in this reference is incorrect.
- (4) Reference 2, p 122, No. 15.
- (5) Reference 2, p 211, No. 26, with double bonds added.
- (6) Bowman, C. M., Landee, F. A., Lee, N. W., and Reslock, M. H., "A Chemically Oriented Information Storage and Retrieval System. II. Computer Generation of the Wiswesser Notations of Complex Polycyclic Structures," *J. Chem. Doc.*, **8**, 133-138 (1968).
- (7) Reference 2, p 105, No. 9.

## Production of Printed Indexes of Chemical Reactions. II. Analysis of Reactions Involving Ring Formation, Cleavage and Interconversion

ROBERT CLINGING and MICHAEL F. LYNCH\*

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

Received February 25, 1974

Simple algorithms designed to identify ring changes in records of chemical reactions are described. They operate on the WLN notations of reactant and product molecules sampled from *Current Abstracts of Chemistry and Index Chemicus*. They enable summaries of ring changes, including formation, cleavage, and interconversion, to be produced, and account for approximately 22% of reactions in the sample.

In a previous paper of this series<sup>1</sup> a method was described by which functional group interconversion reactions could be identified by analysis of the Wiswesser Line Notations (WLN's) of the reactant and product molecules of organic reactions. Using a specified list of reactions it was shown that about 20% of a file based on *Current Abstracts of Chemistry and Index Chemicus* could be successfully analyzed. Further work has since shown that this can be considerably enhanced if a more comprehensive list of reactions is used. The proportion of reactions analyzed is highly dependent on the source of the data base, although it is largely consistent over three monthly files from *Current Abstracts of Chemistry and Index Chemicus*.

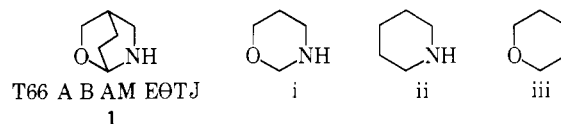
In order to extend the usefulness of this system, an approach has now been made to the analysis of reactions involving changes in rings (formation, cleavage, and interconversion).

At this stage, the description of the rings has been limited to the information explicit in the parts of the notations describing the ring systems.

### IDENTIFICATION OF INDIVIDUAL RINGS

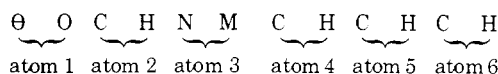
WLN's describe complete ring systems, so that it was necessary to develop a routine to analyze these to provide descriptions of the individual rings present. The routine developed here uses the method of Granito, Roberts, and Gibson<sup>2</sup> to identify the atoms within each ring; each is described by its atomic symbol and its degree of saturation or substitution, if this information is available within the ring notation. More specific information identifying fusion points, bridge atoms, etc., is also determined but is not used in the work described here. Only those rings described explicitly by the notation are dealt with; for instance, in compound 1, only rings i and ii are identified; ring iii is also present but is only described implicitly.

\* Author to whom correspondence should be addressed.



The routine detects certain illegal character sequences, and these cause it to set error codes and exit. In addition, there are certain restrictions on the size and complexity of the rings and ring systems involved (e.g., systems containing nonconsecutive locant paths or rings of greater than 20 atoms are excluded), and these are also indicated by exception codes. Each atom in a ring system is described by three characters. The first two characters contain the atomic symbol (for two character symbols the Wiswesser symbol is used if it differs from the standard atomic symbol), and the third defines its degree of saturation. For carbon atoms, this is H for a saturated atom and U if the atom is unsaturated regardless of whether the double bond is endo- or exocyclic. For nitrogen atoms it is not always possible to define the degree of unsaturation unambiguously from the cyclic part of the notation, and, therefore, the Wiswesser symbols M, N, and K are used unaltered. For other atoms the third character contains a zero.

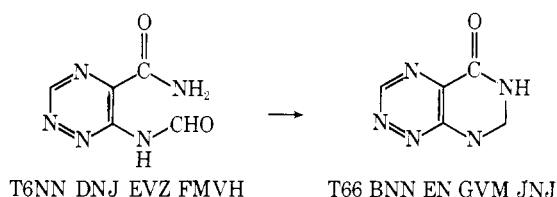
The atoms within the ring are listed starting with the atom whose atomic symbol has the latest alphabetic position and proceeding around the ring so as to give the latest possible position for the whole ring description. If this fails to give an unambiguous result, the characters describing saturation are also taken into account. The result of this is that a particular ring will always lead to the same description, whatever its environment. For example, the description of the ring i is



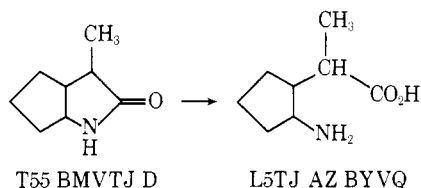
## RING FORMATION AND RING CLEAVAGE REACTIONS

Ring formation and ring cleavage reactions are identical as far as manipulation is concerned, except that the products and reactants are reversed, so that only the former need be described here.

The program for identifying ring formations requires firstly that only one ring may be gained. If this is so, the analyses of the rings in the product and reactant molecules (including unfused benzene rings) are compared, and those rings which occur on both sides of the equation are eliminated. If this fails to eliminate all reactant rings, a further comparison is made, this time ignoring information regarding ring saturation or substitution. If this still fails to eliminate all reactant rings; the reaction is not dealt with; otherwise the remaining ring on the product side of the equation must have been formed during the reaction. Examples of this type of analysis are shown in Figure 1.



Ring closure forming NNCUNMCVCUCU



Ring cleavage of NMCVCHCHCH

Figure 1.

One problem which can lead to unsatisfactory results is the addition of intact rings to the molecule, as, for instance, in condensation reactions such as that shown in Figure 2. This would be analyzed as a ring closure forming an N-substituted piperidine. This problem can be eliminated largely by checking the change in molecular formula which takes place during the reaction and rejecting any example where the number of carbon atoms increases by more than two. This leads to an inevitable loss in recall, but tests have shown it to be necessary to eliminate large numbers of unsatisfactory analyses.

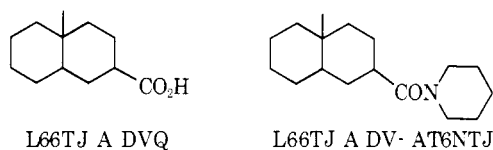


Figure 2.

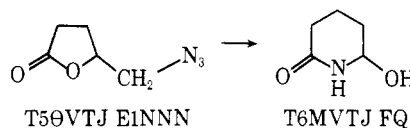
As stated earlier, ring cleavage reactions are dealt with by using the same procedure except that the roles of the reactants and products are reversed.

## RING INTERCONVERSION REACTIONS

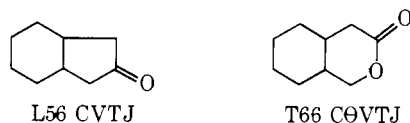
Before proceeding further, it is necessary to define a ring interconversion reaction in this context. Strictly the term should apply only to those reactions where one ring is converted into another while retaining most of the ring atoms.

For the purpose of this study, this definition is relaxed to include any reaction in which one ring is destroyed and another formed, irrespective of their relationship.

The first requirement for a reaction to fall into this category is that there must be no change in the number of rings; if this situation occurs, a further check is made to find whether the change in the number of carbon atoms is greater than two: if so, the reaction is not dealt with. This test is less vital when dealing with ring interconversions than with ring closures and cleavages, but attempts to obtain satisfactory results without it prove that it is still useful. The remaining reactions are analyzed using a similar procedure to that described above for ring closures, those acceptable to the algorithm being classified by means of the ring descriptions of the old and new rings, as shown in Figure 3.



Ring interconversion of  $\Theta O C V C H C H C H$  to  $N M C V C H C H C H C H$



Ring interconversion of  $C V C H C H C H C H$  to  $\Theta O C V C H C H C H C H$

Figure 3.

## RESULTS

A sample of 5444 reactions<sup>1</sup> was analyzed as described above, and the results were assessed manually to decide whether they were chemically satisfactory. These results are summarized in Table I. The term "miscoded" implies

Table I. Summary of Results

	Reactions extracted	Reactions miscoded	Satisfactory	Unsatisfactory
Closure	660	11	635	14
Cleavage	366	2	346	18
Interconversion	172	15	149	8
Total	1198	28	1130	40

that there is a serious error in the input data (e.g., an illegal notation); minor errors which did not affect the validity of the results (e.g., misplaced substituents) were ignored.

The proportion of the file extracted was 22.0% with an overall accuracy of 96.6% after allowing for miscoded reactions.

The unsatisfactory analyses fall into two categories: those which occur due to the use of multipliers in the notations, and those due to inadequacies in the algorithms. The former are of minor importance as an existing procedure could be used to remove all contractions before analysis.<sup>3</sup> They account for 15 of the 40 unsatisfactory answers recorded. It is worth noting that multipliers cause errors only when they are used differently in the reactant and product notations.

The more serious errors are those caused by inadequacies in the algorithms. These are invariably caused by the limitations of the molecular formula checks, for instance, in the example shown in Figure 4. The addition of the allyl group to ring D adds three carbon atoms to the molecule, while the loss of the pyrrolidine removes four, resulting in a net loss of only one carbon atom so that the incorrect anal-

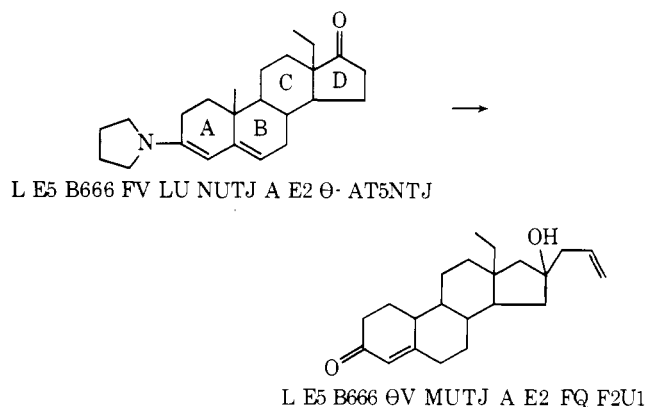


Figure 4.

ysis as the cleavage of a pyrrolidine ring is produced. A similar effect is sometimes observed when only one reaction center is involved (*e.g.*, the reaction shown in Figure 5 which is described as the conversion of a morpholine ring to a benzene ring).

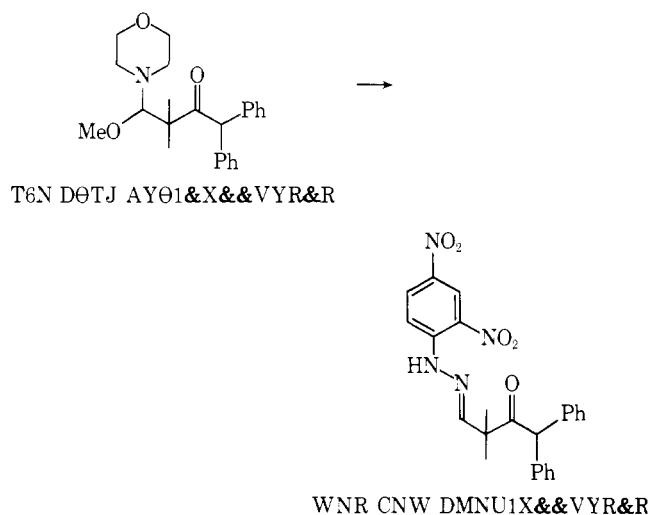


Figure 5.

Sometimes the ring formed during the reaction and a ring which changes its saturation pattern have the same arrangement of atoms, and in such circumstances the program is unable to identify which ring is the product ring. It gives a result in which details of the saturation, etc., of the ring in the product are not specified, *e.g.*, formation of a C<sub>6</sub> ring in Figure 6.

One notable feature of the unsatisfactory results is the frequency with which unfused benzene rings occur in the

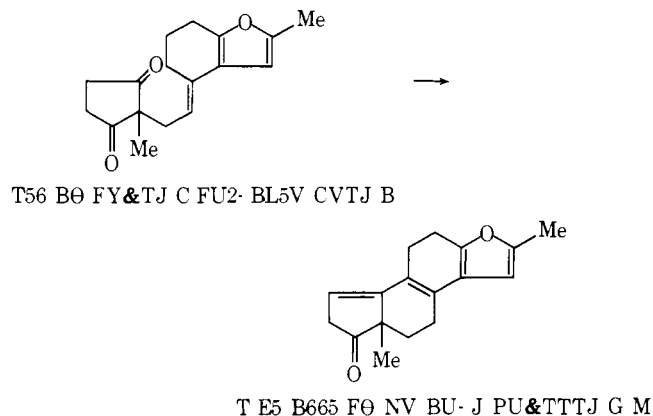


Figure 6.

final analysis. This fact could be used to increase precision further, although, with precision already very high, it is doubtful whether it could be justified.

## EXPERIMENTAL METHOD

Programs were written in the ICL 1900 series assembler language PLAN and run on the University of Sheffield ICL 1907 computer.

## CONCLUSIONS

Programs have been written which classify, at a relatively unsophisticated level, chemical reactions which involve the destruction and/or formation of rings. These reactions comprise 22% of a data base derived from *Current Abstracts of Chemistry* and *Index Chemicus*. The problems of devising representations of the reactions for inclusion in indexes are at present under consideration.

## ACKNOWLEDGMENTS

One of us (R. C.) wishes to thank the Office for Scientific and Technical Information, London, for the award of a Postdoctoral Fellowship in Information Science. We also thank E. Hyde, D. R. Eakin, and J. Radcliffe for valuable discussions.

## LITERATURE CITED

- (1) Clinging, R., and Lynch, M. F., "Production of Printed Indexes of Chemical Reactions. I. Analysis of Functional Group Interconversions," *J. Chem. Doc.*, **13**, 98-102 (1973).
- (2) Granito, C. E., Roberts, S., and Gibson, G. W., "The Conversion of Wiswesser Line Notations to Ring Codes. I. The Conversion of Ring Systems," *J. Chem. Doc.*, **12**, 190-196 (1972).
- (3) Hyde, E., and Eakin, D. R., personal communication.