A Computer Technique for the Retrieval of Related Chemical Structures Utilizing a Special Topological Cipher*

By DELBERT L. BALLARD**
National Cash Register Co., Electronic Division, Hawthorne, California

and

FRANCES NEELAND
System Development Corp., Santa Monica, California

Received April 9, 1963

This paper represents the results of a study of the problems which face the chemist when he attempts to locate all known compounds bearing some desired structural relationship. Several of the procedures presented herein need further refinement in the mill of operating experience, although the major concepts can be shown to be feasible and completely within the state-of-the-art of computer applications.

It is indeed a trite statement to claim that currently published chemical indexes and compendia are inadequate to meet the majority of chemical structure searching problems. The numerous papers describing various chemical codes and their applications, ranging from the printing of special indexes through utilizing simple cards, mechanized punched card machinery, and electronic computers, attest to the fact that the scientist is not only aware of his problems but is making a very serious effort to solve them.[1,2,9]

Chemical structure ciphering schemes have fallen into two basic classes: (1) the fragmentation codes, and (2) some form of topological description of the two-dimensional representation of the molecule. The fragmentation codes such as those devised by Frear, Seiferle, and King[6] and by The National Research Council, Chemical-Biological Coordination Center,[4] to mention but two, make little or no attempt to define the precise *geometrical relationships of the several fragments* which may make up the molecule. Such codes seldom yield a cipher unique to that compound alone, and often are not reversible in nature to reconstitute from the cipher a representation of the chemical structure. They are, however, relatively easy for a chemist to encipher, and are generally amenable to simple manipulation for the location of related chemical structures. Such devices as edge-notched cards, simple card-sorting machines, the IBM 101 Electronic Statistical Machine, the IBM Collator, or some combination of these or other punched card equipment have sufficed for practical applications of these and other such codes.[1,2]

The existence in these fragmentation codes of ambiguity as to the precise geometry of the molecule results in selecting *all* molecules containing the desired fragments with little or no regard for their spatial relationships. This requires post editing of the selections by a person familiar with the actual requirements of the interrogation.

On the other hand, topological ciphers, or the hybrid codes which combine some of the shorthand of the fragmentation codes with the positional relationships of topological description, often yield a unique cipher and may also be reversible in nature. The codes of Waldo and De Backer[14] or of Ray and Kirsch[13] are examples of topological codes, while those of Dyson,[5] Wiswesser,[15] and Norton and Opler[10] may be considered as hybrids between fragmentation and topological codes.

However, the rules for topological encipherment are generally more complex than those for the fragmentation codes. Encoding, in general, is not without its problems, as indicated by the study by Pratt and Perry[12] on the Dyson and Wiswesser notation systems. Retrieval of related chemical structures by topological cipher generally requires greater effort and/or the use of more complicated equipment.

It is obvious that any course which we take will be the resultant of numerous compromises among conflicting requirements, such as input effort per entry, output effort per answer, input and output rate requirements, the nature and quantity of the data to be recorded, and, of course, the type of equipment, the caliber and number of people, and amount of money we may have at our disposal. The probable ratio of items entered into the system to the number of items likely to be retrieved may also affect our decisions as to the effort we will expend on both input and output.

Without discussing all the above factors, which of course vary somewhat from one application to another, we shall proceed on the assumption that it is desirable to minimize all human effort on both the input and the output of this system in order to eliminate, insofar as possible, human error. The technique has been slanted toward separating the intellectual tasks from the manipulative tasks by placing the latter burden upon a computer. It should be noted that the speed and data storage capacity of the

[*] Presented before the Division of Chemical Literature, 144th ACS National Meeting, Los Angeles, Calif., April 3, 1963.

[**] Space and Information Systems Division, North American Aviation, Inc., Downey, Calif.

computer system is determined primarily by a combination of the number and length of ciphers stored in the system and the anticipated inquiry rate and permissible time to produce the answers, and secondarily by the processing program requirements. The cipher input processing will be discussed separately from the interrogation techniques.

To summarize, the dual objectives of the described technique are the provision of an automated means for retrieving unambiguously from among a large collection of organic and inorganic chemical ciphers those which possess in common one or more desired structural similarities, while at the same time preserving the characteristic of being parsimonious of human effort on both the data input and the preparation of search specifications.

Approach.—We have chosen a topological description for enciphering the compounds rather than a fragmentation coding scheme to obviate the necessity for a detailed knowledge of chemistry or nomenclature by the encoders. The rules which govern the human coder are limited to a few simple instructions requiring a minimum of technical decisions on the part of the encoder. For these reasons encoding may be performed by careful nontechnical personnel with a minimum of professional assistance. The major burden of the input effort is left to the computer.

This approach unquestionably results in a much longer initial cipher than any other coding scheme known to the authors, but furnishes the computer with complete operating data and enough redundant statements to permit automatic checking upon the self-consistency of each cipher.

We cannot, of course, assume that the computer can be indiscriminately burdened during the period when we are seeking to obtain answers to our questions, so a second important part of the input processing is the preparation of various summary type data for coarse screening. These include the empirical formula, the number and types of ring structures, if any, number and types of acyclic groups, and the number of each type of bond present. Failure of a chemical cipher to meet any of the above coarse screens for a given question will result in bypassing the detailed search of that particular cipher for that inquiry. However fast a computer may be, such short-cuts can be expected to effect significant savings in search time.

A third part of the input processing is concerned with a reorganization of the input cipher to decrease the number of computer program steps necessary to keep track of the atom-by-atom search which would follow successful passing of the coarse screens.

It is assumed that we can assign a definite structure to each compound which we desire to encode. For example, resonant bond structures will be "frozen," and each possible configuration will be separately encoded but given a common identification number, thereby permitting location of the desired compound regardless of the inquirer's viewpoint.

Provision has not been made for the encoding of "Markush" structures, although searches may be made upon the file of ciphers on a Markush basis.

We shall now discuss the cipher as it applies to the human encoder. Each molecule is treated as a network of points (the elements). Each point possesses an identity and is joined to one or more neighboring points by some sort of bond. This much information alone could be used to produce the end result (*i.e.*, the storable cipher); however, it leaves much to be desired from a practical operating point of view. Redundant information which could lead to checking of cipher consistency is meager. By requesting the encoding personnel to record certain self-evident interrelationships of the constituent elements in each point coding statement, we increase the redundancy of information available to the computer.

Taking advantage of the fact that hydrogen is certainly one of the most common elements in chemical compounds[15] and that it is, except in a hydrogen bond situation, always a terminal element when viewed as part of a chain, there is included as a portion of each point description a count of the adjacent hydrogen atoms. Also included is a count of the total number of adjacent atoms other than hydrogen. It is unnecessary to name all the adjacent elements because the related statements contain that information.

The encoder is also asked to state whether the element at a given point is within one or more ring structures or part of one or more chains, or both. The method by which these determinations are made leads to a somewhat unconventional view of a branch point on a ring, but permits the computer to compare its own conclusions directly against the human observations, and ultimately to distinguish between simple side chains, chains joining rings, ring fusions, bridges and spiro atoms.

The one portion of the encoding process which will require some knowledge of chemistry is the determination of what kind of bond exists between two elements. This is usually self-evident in properly drawn structural formulas, but it may prove necessary for a trained chemist to annotate those structures which contain the more unusual bonds before passing the job along to the encoder.

When a multiple bond occurs between two branch points, the chemist should indicate any *cis* or *trans* relationships of the branches. Also, any tracer element should be assigned its appropriate atomic weight for encoding.

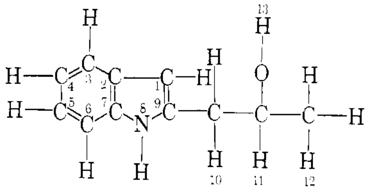
We thus find that each point-encoding statement may be entered on a topological cipher worksheet as shown in Fig. 1.

									Part of		
Statement No.	Point No.	Atomic No.	Isotope Weight	Bond Type	*cis*	*trans*	Adjacent Non-H's	Adjacent H's	Chain(s)	Ring(s)	Ring Closure to Point No.
1											
2											
3											
4											
5											
6											

Fig. 1.—Topological cipher worksheet.

THE ENCODING RULES

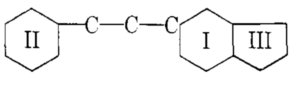
1. *Point No.* enumeration. Uniquely number each atom in the molecule except the H atoms in any convenient manner.
2. *Atomic No.* This number may be obtained from a conversion chart.
3. *Isotope Weight.* Leave blank unless the isotope weight for a tracer element has been shown.
4. *Bond Type.* Obtain code from table.
5. *cis-trans.* If both branch-points adjacent to a multiple bond are asymmetric, determine the *cis* or *trans* relationship of the shorter branch (total non-H atoms) on each side; or if they are of equal length, then determine it for the less branched, or the lower branch weight. (This particular rule should be thoroughly studied before application. We are not presenting it as a tested rule.)
6. *Adjacent Atoms other than H.* Count and record the total number of non-H atoms connected by any type bond to the atom at this point. Deuterium and tritium, while isotopes of hydrogen, are here considered to be non-H atoms to permit redundant coding and subsequent retrieval from either point of view.
7. *Adjacent H's.* Count the total number of H's attached to the atom at this point. Include the H in hydrogen bonds, and count as H any deuterium or tritium atoms attached to this atom.
8. *Part of Ring(s) and/or Chain(s).* Indicate under the appropriate column(s) the number of rings and/or chains in which the atom at this point appears.



In the example above, the C's at positions 1, 3, 4, 5, and 6 are each part of one ring and no chains, as is also the N at position 8. The C's at positions 2 and 7 are each part of two rings and no chains. The C at position 9 is part of one ring and one chain and those C's at positions 10 and 12 are each part of no rings and one chain. The C at position 11 is part of no rings and two chains, and the O at position 13 is part of no rings and one chain.

9. *Ring Closure to Point No.* In coding the closure of ring-containing structures, do not repeat the coding for a point which has already been coded for the ring being closed, but do include in the code for the adjacent element which completes the ring, the point number which has been assigned to the previously coded point.
10. *Coding Sequence.* In coding nonring-containing structures, start at either end of the longest chain of points (or heaviest, if length is equal) and encode through the first branch point. Code all points for all branches at that point, then return to coding the longest chain at the next point following the branch point, and so on until all points are coded.

In coding ring-containing structures, consider the rings as major (*i.e.*, to be coded first) to any chains and major to each other by the number of points within the ring. Note that in the case of two or more rings joined by a chain as



the chain must be coded in between the codes for the rings. The coding is to start from the largest group of ring points and proceed through the chain to the other ring(s). Note that six-membered ring I is in the larger group of ring points so that coding starts away from one of its junction points with five-membered ring III and proceeds around to the branch point where the chain starts, thence out through the chain to ring II, around ring II to its closure and returns to complete ring I. Redundant coding of the two common points of rings I and III is followed by coding ring III to completion.

When encoding a group of fused and/or spiro connected rings, start with a peripheral ring (*i.e.*, a ring having at least one side in common with the outside boundary of the molecule) which if possible is an end ring in the longest straight chain of rings. Code each successive ring spirally from the periphery toward the inside, exhausting first all peripheral rings. Follow a path through fused ring interfaces where possible, or as a secondary choice, through a "spiro" point of attachment. Do not code any complete ring twice, but do code each ring in its entirety even though several atoms may be coded more than once as parts of other rings. Code all rings in the same rotational sense as the gross spiral which has been applied to the molecule.

If all rings are peripheral and no linear trace is practical, the major ring is that ring which has the largest number of ring attachments, and all others are coded as substituents, considering ring fusions as major to spiro attachments.

Among other coarse screens, the computer might be programmed to develop a description in terms of number and kinds of rings comprising the border of a large block of fused rings in addition to the screening inventory of rings which were previously mentioned.

Input Processing.—The data encoded on the Topological Cipher Worksheet are punched into cards and verified, or recorded on punched paper tape, depending upon the equipment available at a particular installation. As an example of input code checking, the element at the first point is checked to determine whether it is part of a ring or part of a chain. It may prove to be both, as would certainly be the case in a fully substituted ring structure. If it is only part of a chain, the compound by definition of the coding sequence has no rings. Later appearance of a point which is part of a ring in this compound would cause print-out of the compound serial number and rejection of the data as input.

Further checks may be performed as the encoded points are read into the computer to verify that the accepted valences or isotopes of each element are not being violated by the encoding process. The fixed data for each element such as valences and permissible isotopes, in that case, would be stored as tables within the computer.

While the data are being read into the computer, the program will direct the preparation of the empirical formula and count the numbers and types of rings, chains, and bonds for the coarse screening of interrogations.

The computer should be programmed to perform, after checking for acceptability of input data, a conversion from the elementary human-oriented view of the structure to a computer, search-oriented organization of the data. It appears profitable to store the encoded cipher as a series of triplets consisting of two points, with their associated data, and of the code for the type of bond uniting these two atoms. This will require more storage space for a complete cipher, but has the advantage that the interrogation can make a direct comparison of each triplet without performing the needed association each time a question is posed. The associations will require considerable maneuvers in the program and hence will consume appreciable time. It is far better to go through this association operation completely during input processing rather than waste precious time during each interrogation search. A summary of the number of each kind of triplet would, of course, be added to the coarse screen.

After code conversion, the cipher should be checked for front-to-back symmetry so that the computer will not waste time performing a backward search on a symmetrically ciphered molecule if it fails to answer an interrogation on the forward pass. In no event will the cipher actually be stored in both the forward and backward sense. Such storage would be highly inefficient because all that actually needs to be reversed automatically is the sequence of interrogation specifications.

It should now be obvious that this ciphering method will yield the same answer regardless of the original orientation of the two-dimensional structural formula.

Up to this point, we have been discussing various aspects of the input to this system. Although some of the operating techniques may have appeared to be rather arbitrary, recognition of the inseparable interaction of the input parameters with the output capabilities has guided each choice as a part of the total system.

The hoped for gain in input accuracy through reduction of the level of intellectual effort is paid for in terms of a rather complex input and error detection program. There are, however, compensating benefits to be derived from this program. The summary data for the coarse screens are simply tallied as they are recognized by the input routine with very few additional program requirements.

Interrogations.—With a relatively small further sophistication of the input program, the input data can be converted from the restricted viewpoint of the input cipher, which treats each atom as an encoding point, to the next higher level of organization of the facts. At this higher level, two adjacent nonhydrogen atoms and their joining bond, are treated as a building block.

The triplet (A bond B) is the basis of all detailed interrogation searches. Most of the data that were read into the computer in the original point cipher accompany the atoms as they are grouped into the triplets. The triplets have the advantage that the associations of the various atom pairs are intrinsic to the triplet, whereas in the original encoding statements, the associations were implicit between statements. The encoding of branches upon a particular structure would separate the statements pertaining to that structure by some distance in the list, depending upon the nature of the branches.

Each interrogation will be prefaced with the minimum acceptable atom count and the minimum acceptable number and types of rings, chains, and bonds. These will, of course, be checked against the empirical formula and the summary data for screening prior to attempting a detailed search.

The atom count will be listed as a group of statements consisting of the atomic number, isotope weight (if necessary), and the minimum acceptable count for that particular atom. All statements will be listed in ascending atomic number order, subdivided by isotope weight if required. The coarse screen will reject any cipher not meeting these minimum requirements. In order to reject ciphers containing some unwanted element, the atom count for that element will be made zero rather than being left blank. The same approach will apply to the listings of the minimum acceptable count of rings, chains, bonds, and triplets.

If any of the desired counts are recorded as negative numbers, the computer will interpret this to mean the maximum acceptable count. In this way we may establish upper and lower limits by including, respectively, negative and positive counts.

The interrogation triplets are entered into the computer as a string of logical operations. Each statement, or group of statements, may be treated as logical AND, OR, or NOT operations. The AND and OR statements may take either of two forms. The *assertive* form of the AND statements requires that the next-to-be-considered pair of cipher statements must match the AND interrogation statements, without intervention of any other triplets. The *permissive* form of AND between two interrogation statements means that the next cipher statement to be processed must match the first of the interrogation statements, but that any number of intervening cipher triplets not matching the second interrogation statement will not terminate the search. The search will try cipher triplets one after another until the second interrogation triplet is matched, or until the end of the cipher is reached.

The *assertive* OR statements require a match of one of the OR interrogation statements with the next cipher triplet to be searched. The *permissive* OR will, like the permissive AND, continue its search until one of the OR interrogation statements is found or until the end of the cipher is reached.

A NOT statement can only be *assertive* and applies only to the next in line cipher triplet. If a match is found, the search conditions will return to those of the last branch point and another branch will be searched. If a match is not found for the NOT statement, the search will continue with the next interrogation statement.

So far it may appear that we have been concerned only with the basic (A bond B) triplet. It must be remembered that the input statements also carried over to the triplet a count of the adjacent H atoms and adjacent non-H atoms for *each* atom named in the triplet. The same convention is used for specifying the quantities of these adjacent atoms as was applied to the coarse screen data. A positive number represents the acceptable minimum, a negative number specifies an acceptable maximum, a zero specifies that there must be none of that field, and a blank denotes a "don't care" condition.

The content of the stored cipher triplet, which is shown below, indicates the number of digits required for each field therein.

	A						B					
Field	Atom No.	Iso-tope	R	C	H	Non H	Bond	Atom No.	Iso-tope	R	C	H	Non H
Dig. Req.	2	3	1	1	1	1	1	2	3	1	1	1	1

Although the encoding rules limit quite severely the number of ways in which any compound may be encoded, there still exists the possibility of similar structures within different compounds being coded in a different directional sense. For that reason we must be prepared to engage in a reverse search if the normal search has failed to detect the appropriate sequence of characteristics forming the interrogation. The computer will, in effect, perform both a forward and a backward search of all except symmetrical molecules, which will have been tagged by the computer during input processing to inhibit the reverse search.

For those interested in the logic of the detailed search process, we are presenting the broad flow chart of Fig. 2 which illustrates the application of each of the five logical operators used in composing the interrogations.

Among the unusual features which we have presented are the following:

1. Encoding reduction to clerical level.
2. Extensive redundancy included in input code.
3. Extensive error checking routines.
4. Automatic translation from human-oriented cipher to machine-oriented cipher.
5. Unified program for logical operations of interrogations.
6. Extensive coarse screens leading to a high degree of rejection of nonproductive detail searches.

As a final comment, we should like to pose the following problem regarding the use of computers. The atom-by-atom encoding process may be used as the basis for recording from tables stored in the computer the interatomic distances and bond strengths for each triplet.

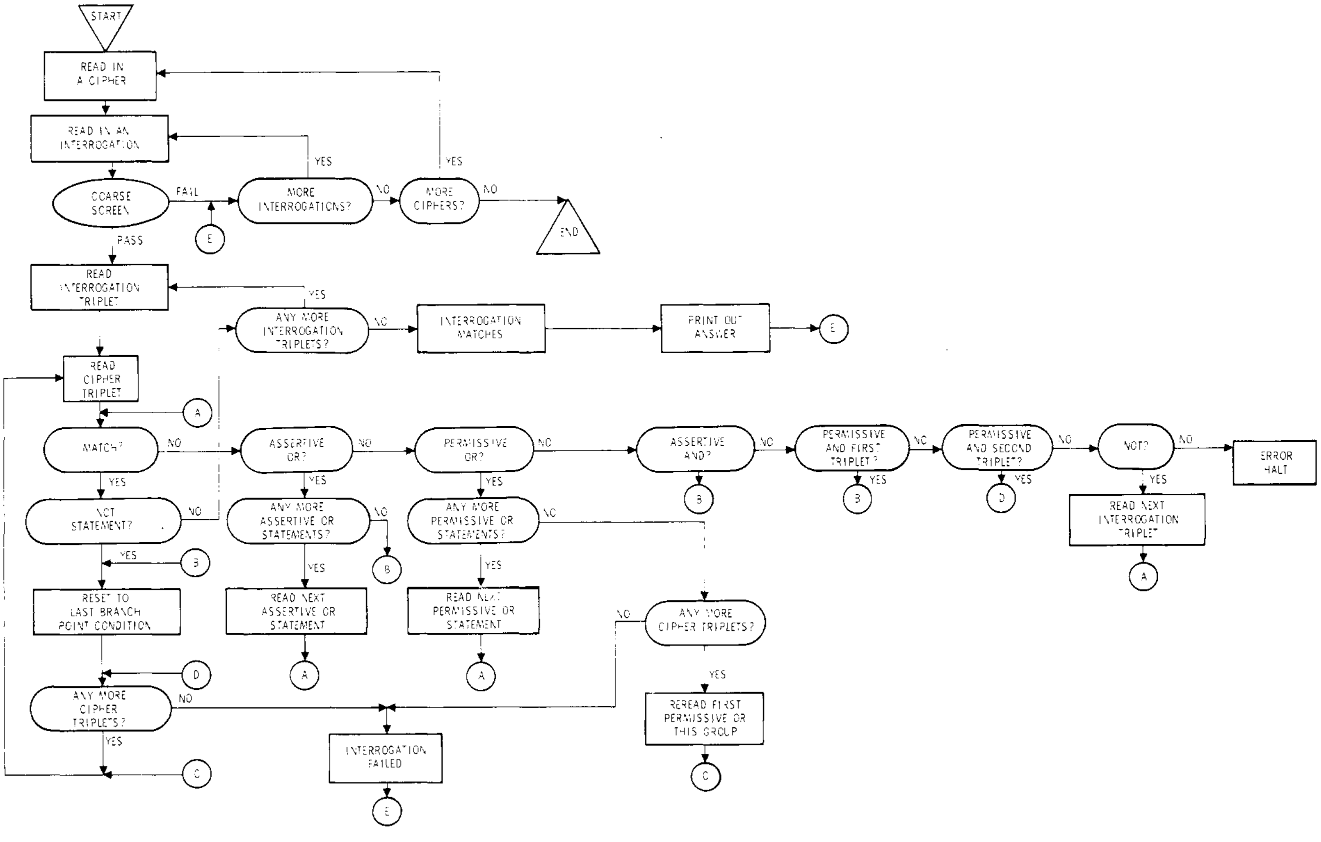

START
READ IN A CIPHER
READ IN AN INTERROGATION
COARSE SCREEN
FAIL
PASS
E
YES
MORE INTERROGATIONS?
NO
YES
MORE CIPHERS?
NO
END
READ INTERROGATION TRIPLET
YES
ANY MORE INTERROGATION TRIPLETS?
NO
INTERROGATION MATCHES
PRINT OUT ANSWER
E
READ CIPHER TRIPLET
A
MATCH?
NO
YES
ASSERTIVE OR?
NO
YES
PERMISSIVE OR?
NO
YES
ASSERTIVE AND?
NO
B
PERMISSIVE AND FIRST TRIPLET?
NO
YES
B
PERMISSIVE AND SECOND TRIPLET?
NO
YES
D
NOT?
NO
YES
ERROR HALT
NOT STATEMENT?
NO
YES
B
ANY MORE ASSERTIVE OR STATEMENTS?
NO
B
YES
ANY MORE PERMISSIVE OR STATEMENTS?
NO
YES
READ NEXT INTERROGATION TRIPLET
A
RESET TO LAST BRANCH POINT CONDITION
READ NEXT ASSERTIVE OR STATEMENT
A
READ NEXT PERMISSIVE OR STATEMENT
A
NO
ANY MORE CIPHER TRIPLETS?
YES
D
ANY MORE CIPHER TRIPLETS?
NO
YES
C
INTERROGATION FAILED
E
REREAD FIRST PERMISSIVE OR THIS GROUP
C

Fig. 2.—Search flow chart.

From these characteristics it is hoped that someone better qualified than the present authors will be motivated to devise an algorithm leading to a complete spatial-kinetic description of the molecule. This could well be a first step toward computer simulation of chemical reactions with all its attendant implications.

BIBLIOGRAPHY

(1) R. S. Casey and J. W. Perry, Ed., "Punched Cards—Their Applications to Science and Industry," Reinhold Publishing Corp., New York, N. Y., 1951.

(2) R. S. Casey, J. W. Perry, M. M. Berry, and A. Kent, Ed., "Punched Cards—Their Applications to Science and Industry," 2nd Ed., Reinhold Publishing Corp., New York, N. Y., 1958.

(3) R. S. Casey, "Annotated Bibliography on Uses of Punched Cards," ref. 2, pp. 637-672.

(4) Chemical-Biological Coordination Center, "A Method of Coding Chemicals for Correlation and Classification," National Research Council, Washington, D. C., 1950.

(5) G. M. Dyson, "A New Notation and Enumeration System for Organic Compounds," Longmans, Green and Co., London and New York, 1947; 2nd Ed., 1949.

(6) D. E. H. Frear, E. J. Seiferle, and H. L. King, *Science*, **104**, 177 (1946).

(7) H. W. Hayward, "A New Sequential Enumeration and Line Formula Notation System for Organic Compounds," Office of Research and Development, Patent Office, U. S. Department of Commerce, Washington, D. C., November, 1961 (Patent Office Research and Development Reports, No. 21).

(8) L. A. Lederman, K. Taylor, J. W. Perry, and M. E. W. Torok, "Bibliography on Uses of Punched Cards," ref. 1, pp. 457-488.

(9) E. Marden, and H. R. Koller, "A Survey of Computer Programs for Chemical Literature Searching," U. S. Department of Commerce, National Bureau of Standards, Washington, D. C., May 16, 1960 (National Bureau of Standards Report 6865, available from Office of Technical Services, PB Report 161586).

(10) T. R. Norton and A. Opler, "A Manual for Coding Organic Compounds for Use with a Mechanized Searching System," revised, Research Dept., Western Division, Dow Chemical Co., Pittsburg, Calif., March 15, 1956.

(11) A. Opler, and T. R. Norton, "A Manual for Programming Computers for Use with a Mechanized System for Searching Organic Compounds," Research Dept., Western Division, Dow Chemical Co., Pittsburg, Calif., April 25, 1956.

(12) A. D. Pratt, and J. W. Perry, "Chemical Notation Study: Dyson-Wiswesser Notation Systems Encoding Operations; Phase Report," revised, Center for Documentation and Communication Research, Western Reserve University, Cleveland, Ohio, August 1, 1960, ASTIA Document No. AD245936.

(13) L. C. Ray, and R. A. Kirsch, *Science*, **126**, 3278 (1957).

(14) W. H. Waldo, and M. DeBacker, "Printing Chemical Structures Electronically: Encoded Compounds Searched Generically with IBM-702," preprints of papers for the International Conference on Scientific Information, Washington, D. C., Nov. 16-21, 1958; National Academy of Sciences, National Research Council, Washington, D. C., 1958, Area 4, pp. 49-68; also published in "Proceedings" of the Conference, Vol. 1, pp. 711-730.

(15) W. J. Wiswesser, "A Line-Formula Chemical Notation," Thomas Y. Crowell Co., New York, N. Y., 1954, pp. 125-126.

A Cooperative Project in New Drug Reporting*

By PATRICIA GRAHAM BOHR and KATHERINE CRAWFORD OWEN
Warner-Lambert Research Institute, Morris Plains, New Jersey

Received February 25, 1963

The New Drug Information project (NDI) is an experiment among pharmaceutical companies in the exchange of information on new chemical compounds reported in the current literature to have biological properties. This project was designed to provide an alerting service for the scientists of each firm that would be more comprehensive than any one company could provide without greatly increased costs. Promptness of reporting and inclusion of the chemical structure of the compound were unique features of the plan.

The idea was discussed informally during the Gordon Conferences of 1961; its champions were Mr. Walter Southern of Abbott and Dr. Joe Clark of Lederle. It was originally planned to have a single format to which all would conform, but it might have taken months of conferences to hammer out such a format, and stringent entry requirements might have prevented the participation of some companies.

In the fall of 1961, Abbott and Lederle, joined by Schering, Mead Johnson, and Squibb, began to exchange information on new drug structures in whatever forms they provided this information internally. Warner-Lambert became a member of what has come to be called the "alerting ring" in December, 1961. Reflecting changes in personnel, Schering dropped out of the ring, and Wyeth joined. Late in 1962, Squibb which had been inactive for a time, resumed participation, and Searle became the seventh member.

* Presented before the Division of Chemical Literature, 143rd ACS National Meeting, Cincinnati, Ohio, January 14, 1963.