(22) Select in M′ for a RN not alone in its group the other RN's to which it is directly bonded; select in S the group numbers for these RN's, arrange them in increasing order, and handle the result as Y. Repeat this for the other RN's of this type. If all Y's are equal go to (24); if not (23).

(23) Do (5), making a note if the highest group number given differs from the highest previous group number. If a note has been made, delete the note and repeat (22). If no note has been made go to (24).

(24) Count in the storage for the sections pertaining to RN's not alone in their group the number of atoms (including double or triple bonds) per term, arrange the numbers obtained per section in increasing order, and check whether the numbers obtained per section are equal. If equal, go to (25). If not, handle the result as Y and go to (23) reading for (24), (25).

(25) Arrange per section the terms without rear side RN in alphabetical order in front of the other terms. Arrange per section pertaining to those RN's which are not alone in their group the terms with rear RN and arrange these terms themselves in alphabetical order. Select from the sections pertaining to RN's which are not alone in their group the section which would come first alphabetically if the terms in that section were considered as one word. Allot to this section a preference number 1; repeat the selection for the remaining sections of this type, giving them also a preference number, which is equal to the previous one if the selected section is alphabetically identical with the previous one and one higher than the previous number if it is not identical. Test whether per group all preference numbers are equal. If equal, go to (26). If not equal, use them as Y and go to (23) reading for (24), (26).

(26) Test whether more than one group containing more than one RN is present. If not, go to (27). If present, allot to the RN's in one of these groups an ascending series of numbers starting with 1. Allot 1 to the other RN in these groups and handle the result as Y. Go to (23) reading for (24), "(26) but allotting the ascending number series to another group of that type."

(27) Allot to the RN's a number increasing by 1 and, starting with group 1, going to higher groups in S. Replace the RN's in the storage list by the numbers allotted to them. Arrange the sections of the storage list in increasing order of their front new RN. Arrange per section the terms in increasing order of their rear RN, taking no RN as zero, and, as far as two or more rear RN's in a section are equal, in alphabetical order of the terms in question.

REFERENCES

(1) H. Bouman, *J. Chem. Doc.*, 3, 92 (1963).

(2) G. M. Dyson, W. E. Cossum, M. F. Lynch, and H. L. Morgan, *Inform. Storage Retrieval*, 1, 69 (1963).

(3) G. Salton, and E. H. Sussenguth, private communication from The Computer Laboratory of Harvard University, Cambridge, Mass.

ChemSEARCh—An Operating Computer System for Retrieving Chemicals Selected for Equal, Analogous, or Related Character*

DAVID GOULD and EDWARD B. GASSER,
Colgate-Palmolive Research Center, New Brunswick, New Jersey

and

JOHN F. RIAN
Control Data Corporation, Rockville, Maryland

Received March 9, 1964

All chemists, and especially those whose responsibilities include synthesis of compounds having pharmacologic or medicinal effect, have felt the necessity for a reference file of structures and publications organized for their areas of interest. From this file, they wish to select: (1) specific compounds, (2) compounds related by containing specific subgroups of atoms, (3) compounds having a specified relation between groups, and (4) activities of compounds. From these components, they hope to derive structure-activity relationships to guide future research.

The organization of such a file has progressed through the alphabet, nomenclature as in *Chemical Abstracts*, and sorting by groups as in "Beilstein," The last still has considerable value to the chemist, particularly by use of edge-notched cards[1] with such systems as the first published by Frear.[2] When these approaches showed some difficulties owing mainly to ambiguity, Dyson[3] initiated complete, unique, and unambiguous codes. A serious problem with these is the complexity of the coding

* Presented at the Symposium on Advances in Chemical Structure Storage and Searching Systems. 147th National Meeting of the American Chemical Society, Philadelphia, Pa., April 6, 1964.

rules, generally requiring considerable study beyond the training of the average chemist. Wiswesser[4] perhaps best overcomes this problem by attempting to utilize the training of the chemist through extensive use of symbols familiar to chemists. Studies have indicated, however, that even experts cannot use such systems without some difficulty,[5] at least, in part owing to disagreement on interpretation of the coding rules for some of the more complex molecules. Thus, a compound coded "correctly" by one expert can be lost to another who thinks the correct code is different. Perhaps more serious, the chemist cannot easily use them, the codes are not readily recognized as a given structure for a visual error correction, and the use of clerks is unthinkable.

The most serious limitation is that a cipher is a condensed statement of the structure. The condensation required prejudgment as to what will be of interest in the way of chemical groupings to future generations of chemists, and even to those at present whose field is widely removed from that of the originator of the cipher. Of necessity, then, such a depiction is a compromise between these various interests, and while it may satisfy many, it cannot satisfy all.

The solution to this dilemma lies in the complete representation of the molecule, atom by atom, and a provision for the future searcher to define his own interests when he is making the search. This has its strength in the chemist's familiarity with structure. A chemical structure is, of course, a topological network, and in theory, the handling of such systems was solved in the nineteenth century.[6] In practice, the application to chemical structures was only recently initiated, depending on the ground work laid by Gordon, Kendall, and Davison[7] and the careful mathematical analysis of Mooers.[8] Since then, a number of systems have been proposed,[9] some of which apply elegant mathematics to the various problems, but very few of these systems have been organized into a practical operation. The obvious reason for this is that by its nature a complete topological representation of a structure involves a great mass of data, and a direct comparison of structures through network search requires a number of tedious operations in comparison to a condensed code. Only when an advanced computer is available for regular use can such a system show its full power by assigning to the rapid, errorless, and unbored machine the job of studying the mass of notational data and analyzing two statements for an exact correspondence of topological network.

When our group at Colgate started devising a mechanized system for handling chemical and biological activity in relation to chemical structure, we were influenced, aside from background in mathematics, programming, data handling, medicinal chemistry, and the history of chemical codes, by several major considerations. We lacked a reservoir of technical personnel available for input coding; we could use a clerical pool; and we had the use of a strong computer.[10] The background and the computer recommended a topological system, and for this we relied on two basic strengths: the training of the chemist in understanding, recognizing, and handling bonds and atoms in a structural formula; and the ability of a digital computer to collate data and to determine the equivalence (or difference) of two stated situations. Since construction of a file was *not* to utilize professional personnel, and also to diminish error, coding was made as simple and recognizable as possible, permitting clerical conversion of a structure drawn by a chemist to computer input. Once the structure was entered as a unique topological network, any user of the file could question this structure from the viewpoint required by his own area of investigation.

When a chemist writes a structural formula, he does not actually write the detailed topological structure of the molecule, but uses many abbreviations such as CH_3, as well as line depiction of rings and sometimes chains. Such a condensed structure, even though perfectly intelligible to the chemist, obviously could not serve for clerical coding, since a topological structure does not yet exist until special knowledge is applied to the depiction. The chemist inputting a structure to the file is therefore required to write all its atoms and all bonds between them. A compensating gain in the ChemSEARCh system is the omission of hydrogen bound to noncarbonyl carbon. The resulting structure is just as intelligible to the chemist as the more frequently used condensed forms. When one looks at this completely delineated structure, it is now apparent that a given point (node, group of letters, atom) is connected to another by certain symbols (bonds) whose codes and definitions are given in Table I. Acceptable atoms are listed in Table II. The input code for the structural formula is simply a statement of these observations.

Table I
Bond Types

Symbol[a]	Code[b]	Definition
—[c]	1	Single covalent bond; also used for ionic bonds in inorganic structures
=	2	Covalent or semipolar double bond
≡	3	Covalent triple bond
⧣	4	Aromatic bond; all aromatic bonds considered equal and all structures are written in the most highly aromatized form
	5	Salt, dot, or complexing bond; acid salts of organic bases, quaternary ammonium and other onium salts, betaines, and other coordinated complexes
∼	6	Neutral bond; used when the bond type is unknown or not specified
+	0	Plus bond; used to show salt formation of an organic acid with a basic cation

[a] Symbols used to write structures for computer coding. [b] Code numbers show the bond type in written codes. [c] The single bond is not always shown in standard structures. It *must* be shown in computer structures.

Table II
Element Order

Symbol	Name	Symbol	Name
X	Neuter	P	Phosphorus
Unlisted elements, alphabetically		S	Sulfur
Ad	Addend	N	Nitrogen
E	Empirical	O	Oxygen
F	Fluorine	C	Carbon
Br	Bromine	H	Hydrogen
Cl	Chlorine	Z	Zero

```
3999
Methyl isopropenyl Ether
EF C4 H8 Ø1

SF
C2  1Ø1-1C4-2C3
C3  2C2
C4  1C2
Ø1  1C2-1C1
C1  1Ø1
END
```

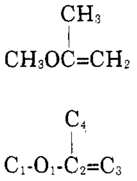


Figure 1.—Structure coding and format.

The illustration in Figure 1 for methyl isopropenyl ether shows: the usual type of depiction; the formula written as a topological network denuded of hydrogen; and the simple format of computer input created from the structure after the atoms are arbitrarily numbered to distinguish one from the other. The acquisition number signals the beginning of a group of compound codes and aids in identification on output. The chemical name (required from the chemist) also serves for identification on output. The empirical formula (EF, also requested from the chemist) is listed normally. SF signals that structure equations follow. Equations start with a subject atom for which all of the connected atoms are listed, immediately preceded by the bond code. END signals the completion of a set of codes. Other coding signals are given in Table III.

Since this system is aimed primarily for use by chemists interested in structure-activity relationships, certain modifications of salts were adopted to increase the ease of obtaining an entire related family. Of greatest importance, however, was elaboration of the concept of generalized atoms and bonds which permits generic questioning of the file with a high degree of sophistication. Illustration of the mechanics of various bonds and pseudoelements follows the list of rules which guide their use.

Table III
Coding Signals

Signal	Definition
ADD	Empirical formula of an addend follows
EF	Parent empirical formula follows
EMP	In structural codes, empirical formula of "E" or "Ad" follows
END	End of whole code record
RAD	Set of equations for repeated or repeating groups follows
SF	Set of equations for structure follows
SFEX	Equations for exclusive structure search
TAB	Tabular separation of equation codes on right from subject atom or signal on left
x	(a)—Under EF, x (tab) shows empirical formula of repeating unit follows (b)—Under SF, equated on right to RAD, unknown number of repeats (c)—Under SF, signal on left ends RAD equations
END WAIT	End of one batch of input; another to follow
END END	End of input; proceed to process

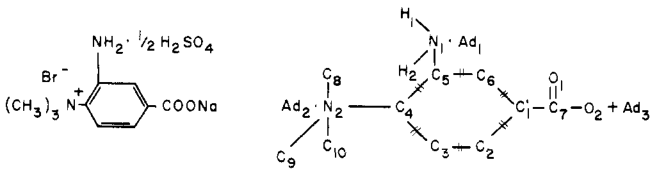
Conventions for Structure Writing

1. A conventional formula must be rewritten so that every atom and bond is shown. *No line depiction of rings is permitted. Exception*: Hydrogen atoms attached to noncarbonyl carbons are usually omitted.
2. Each atom is connected to its substituents by normal and special bonds as given in Table I (Bond Types).
3. The atoms of each different element, normal or artificial, are numbered from 1 to the highest number of such atoms, in any convenient order.
4. All structures are written in the most highly aromatized form. Bonds are placed in a ring when it is possible to make the ring aromatic.
5. The keto rather than the enol form and the imine rather than the enamine form are used except where rule 4 prevails.
6. When there is a choice of arrangement of bonding and bond type, the highest number of covalent rather than charged bonds is used.
7. Addends (Ad) used in place of salt-forming acids, cations, or anions are defined empirically alongside the structure.
8. Repeating units of one or more atoms are enclosed by parentheses or brackets followed by a subscript number of such units. If the number is unknown as in a polymer, the subscript is x. For the written code, the equations within the parentheses are signaled by RAD equated to x, or the exact number of repetitions. The repeating set of equations is ended by a subject x.
9. Repeated units of more than one atom, attached to one location, are enclosed in parentheses with the appropriate subscript. The equations for atoms within are signaled by RAD equaling the number of repetitions and ended by x.
10. If the structure is only partly known, the known portions are written out and attached to E. The E is defined empirically beside the structure.
11. When a metal is attached to C or H, or strongly bound to heteroatoms as in chelates, the bond is given as single (1). Additional coordinate bonds are dots (5). Addends (Ad) are not used.
12. If an unknown element or group is attached to a known structure, it is indicated by Z (the Zero element). Any definite but unspecified single element (in a question) may be indicated by X (the Neutral element).

The conventions have been established partly to require some thought on the part of the chemist inputting a structure, to emphasize the generic relationships of families (*e.g.*, salts and their parents), to establish some consistency of input format, and largely for simplicity and clarity of presentation, ease of coding, and avoidance of errors. It should be emphasized, however, that given any correct depiction of atoms and bonds in a topological network, all other conventions can be ignored and the structure is retrievable from the file. The flexibility characteristic of the ChemSEARCh system allows the professionally trained chemist, with sufficient thought and knowledge of possible alternate representations, to formulate a single question which will retrieve not only identical (perhaps mesomeric) structures entered with varying but accepted bonding, but even closely related structures such as tautomeric keto-enol pairs. Some of the conventions draw attention to the type of situation which may arise and thus serve as guides to the formulation of sophisticated questions.

Coding in ChemSEARCh mainly uses normal bond and element symbols, but Tables I and II also list several artificial constructs for handling special cases and for searching generically. Thus the 4-bond is used for all resonating or aromatic bonds which occur in cyclic mesomeric structures, instead of alternating single and double bonds. The dot (5) and plus (0) bonds are used with Ad in coding cations, anions, and acids in salts, to aid in finding all related salts. In contrast, the neutral (6) bond is used mainly in questions to achieve generality.

The pseudo-elements listed in Table II are Z, E, Ad, and X. These artificial elements are handled and numbered distinctively just like chemical elements. In practice, the letter symbols are transformed in the computer into code numbers, and the bonding equations for each subject element are sorted and searched in the order given. In general, therefore, the rarer elements are handled first and nonmatching structures are discovered and eliminated sooner.



EF	C10 H15 O2 N2
ADD	Na1
ADD	H2 S1 O4
ADD	Br1
SF	
C1	1C7-4C2-4C6
C2	4C1-4C3
C3	4C2-4C4
C4	4C3-4C5-1N2
C5	4C4-4C6-1N1
C6	4C5-4C1
C7	1C1-2O1-1O2
O1	2C7
O2	1C7-0Ad3
Ad3	0O2
EMP	Na1

SF	(cont'd)
N1	1C5-1H1-1H2-5Ad1
H1	1N1
H2	1N1
Ad1	5N1
EMP	H2 S1 O4
N2	1C4-1C8-1C9-1C10-5Ad2
C8	1N2
C9	1N2
C10	1N2
Ad2	5N2
EMP	Br1
END	

Figure 2.—Coding of addends.

Addends (Ad).—In order to draw forcible attention to closely related substances differing only by generally inconsequential parts, subsidiary portions of the molecule such as the acid of an amine salt, the anion of a quaternary, or the cation of an acid salt, are defined separately as addends. In the written formula, the addend is given as Ad and the empirical formula of each Ad is given alongside. The addends are not included in the parent empirical formula, but the empirical formula of each addend (without multipliers such as ½) is signaled by ADD immediately following the parent empirical formula (EF). In the special case of salts of acids, the parent EF given is that of the free acid, and this situation is signaled by connecting such cations (Ad) to their correct locations by plus (0) bonds. In the structure portion under SF, the subject equation for each Ad element is followed immediately by the subject signal EMP equated to the empirical formula of that Ad. Thus, in Figure 2, the rather unusual compound given contains all usages of addends (Ads). Ad_1, the bisulfate, is defined empirically (after EMP) as $H_2S_1O_4$. Ad_2, the anion of the quaternary, is defined empirically simply as Br_1. Ad_3, the salt of an acid, is defined empirically as Na_1, and at the same time one additional hydrogen is added to the parent empirical formula (EF). Also in the figure, the use of aromatic (4) bonds is illustrated.

Polymers and Repeated Groups.—In polymers with repeating units, the first atom preceding the parenthesis is coded normally to the first atom inside the parentheses, and the atom codes within are signaled by RAD. In Figure 3, compound 801 is a polyoxyethylene derivative of methanol with an unknown chain-ending substituent (Z, the zero element). The empirical formula (EF) is only that portion counted exactly, that is, only the atoms outside of the parentheses. Immediately following EF, x signals the empirical formula of the monomer within the parentheses, which is of course to be multiplied x times. Under the structure signal, SF, C_1 is bonded to O_1. The contents of the repeating units are signaled by RAD, followed by x since it is repeated an unknown number of times, and are defined normally by equations immediately thereafter. If this were a polymer of known structure, RAD would be equated to the exact subscript number, and the total empirical formula would be included in the parent empirical formula after EF. The end of the coding for the monomer unit is signaled by x, followed by any remaining codes outside the parentheses.

$C_1\text{-}(\text{-}O_1\text{-}C_2\text{-}C_3\text{-})_x\text{-}Z_1$

801

Methyl polyoxyethylene complex

EF	C1 H3 Z1
x	C2 H4 O1
SF	
C1	1O1
RAD	x
O1	1C1-1C2
C2	1O1-1C3
C3	1C2-1Z1
x	
Z1	1C3
END	

$C_1\text{-}(\text{-}O_1\text{-}C_2\text{-}C_3)_3$

802

Triethoxymethane

EF	C7 H16 O3
SF	
C1	1O1
RAD	3
O1	1C1-1C2
C2	1O1-1C3
C3	1C2
x	
END	

Figure 3.—Coding of polymers and repeated groups.

Also illustrated in Figure 3 is the coding of the repeated unit of triethoxymethane. For compound 802, EF includes the whole molecule since it is completely known. The methane carbon (C_1) outside the brackets is connected to O_1 normally. The content of the brackets, to be multiplied three times, is signaled by RAD = 3, and the end of the structure within the brackets is signaled by x.

The zero element (Z) represents: (1) an atom or group of unknown empirical formula; (2) an atom or group whose structure and composition does not have to be specified, as in questions; (3) the absence of an atom, when there may be a choice between an unknown substituent or none. There may be multiple bonds to Z, but if Z should be bound to more than one atom, the structure would be disconnected at Z.

The Empirical Element (E).—When the structure of a portion of the molecule is unknown but its empirical formula is known, that portion is indicated by the artificial element, E. Any number of single or multiple bonds may be connected to E. Each E is coded normally in equations but immediately following the subject equation for E, the signal EMP is given as the next subject and is equated to the empirical formula of E. EF, being known exactly, signals the empirical formula for the whole molecule.

The neutral element (X) is used only in questions to represent any definite element. It may be connected as desired with any bond to any element. Thus, specific structures or groups defined only by bonds may be constructed for certain generic searches.

Operation of ChemSEARCh.—After a compound is drawn by the chemist with all bonds explicitly given and appropriate hydrogens omitted, the structure with atoms numbered is coded in writing on a coding form (average time, 3 min.). The format, including number, name, empirical formula, and structural formula, is typed on a paper-tape-punching typewriter.[11] The Syntax Checker program for error determination (on paper tape) is inserted into the computer, followed by the punched paper tape corresponding to the type script, usually in combination with a number of other structures. Most corrections must be made before the computer will accept a code set for the main file.

When one or several sets of compounds are ready for insertion into the main file, the File Maintenance program is inserted into the computer. If the compounds on different tapes are all in increasing order of accession number, they may all be inserted in one operation. If they are not, a separate File Maintenance run is needed for each tape. As the insertion progresses, the master file on magnetic tape is scanned until the proper location for insertion of an accession number occurs, at which point the new record is inserted while all subsequent records are moved back one notch. The file is now ready for searching using the Question Reader program which compares structure according to the flow chart in Figure 4. Any reasonable number of questions may be asked simultaneously. If it cannot handle them at once, the computer stores some of them for a second run.

Topological Comparison.—The comparison of one topological structure with another requires that their networks match. For two identical structures, no matter how they were originally written, it must be exactly possible to trace a like path through each, with retracing where necessary to cover every node and linkage in both networks. In the computer, each piece of bonding information is stored at a different location in memory, and the comparison of a question structure with a file structure consists of matching moves to all locations in each network. There are five basic actions tracing such a path:

1. *Mark and Move.* The starting point is checked and both ends of a move to a new location (atom) are recorded.
2. *Mark and No Move.* When a move is attempted to a point previously checked (excluding return moves), as in completion of a ring, the bonding connection to such an atom is checked but no move is made to that atom.
3. *Back Code.* The existence of the back coding is recognized but no return move is made until all other possible operations have been exhausted.
4. *Return.* When all paths have been traced from one point in the network, a return move is made to the preceding point in the network, and the move marks are erased.
5. *Reverse Operation.* When a mismatch of atom, bond, or move occurs on one path, the paths in both structures are traced backward to the next earlier point where all paths have not been tried, and all marks must be deleted during this reverse operation.

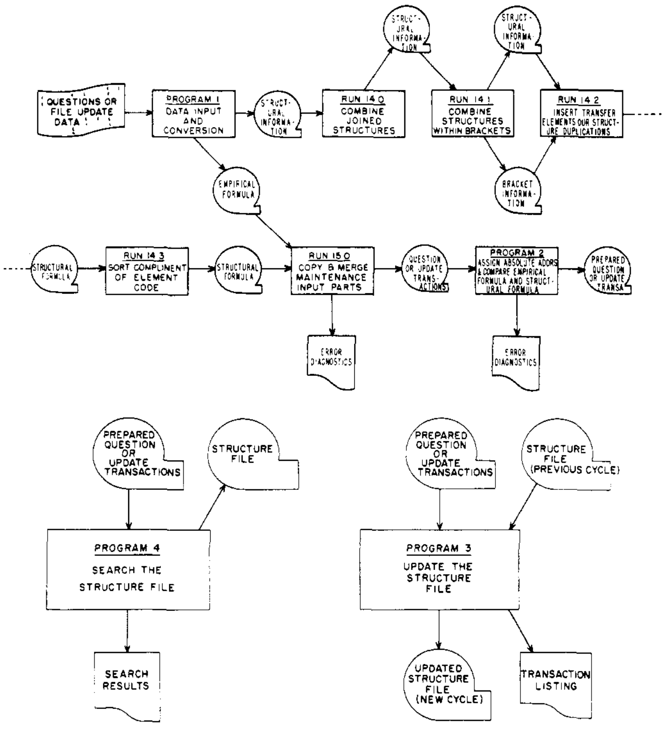
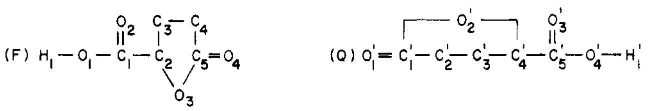

QUESTIONS OR FILE UPDATE DATA
PROGRAM 1
DATA INPUT AND CONVERSION
STRUCTURAL INFORMATION
RUN 14 0
COMBINE JOINED STRUCTURES
STRUCTURAL INFORMATION
RUN 14 1
COMBINE STRUCTURES WITHIN BRACKETS
STRUCTURAL INFORMATION
RUN 14 2
INSERT TRANSFER ELEMENTS OUR STRUCTURE DUPLICATIONS
BRACKET INFORMATION
EMPIRICAL FORMULA
STRUCTURAL FORMULA
RUN 14 3
SORT COMPLINENT OF ELEMENT CODE
STRUCTURAL FORMULA
RUN 15 0
COPY & MERGE MAINTENANCE INPUT PARTS
ERROR DIAGNOSTICS
QUESTION OR UPDATE TRANSACTIONS
PROGRAM 2
ASSIGN ABSOLUTE ADDRS & COMPARE EMPIRICAL FORMULA AND STRUCTURAL FORMULA
ERROR DIAGNOSTICS
PREPARED QUESTION OR UPDATE TRANSA
PREPARED QUESTION OR UPDATE TRANSACTIONS
STRUCTURE FILE
PROGRAM 4
SEARCH THE STRUCTURE FILE
SEARCH RESULTS
PREPARED QUESTION OR UPDATE TRANSACTIONS
STRUCTURE FILE (PREVIOUS CYCLE)
PROGRAM 3
UPDATE THE STRUCTURE FILE
UPDATED STRUCTURE FILE (NEW CYCLE)
TRANSACTION LISTING

Figure 4.—Flow of data—computer input and search routine.

An example given in Figure 5 shows network tracing to match a question structure (Q), α-hydroxyglutaric acid γ-lactone, against a structure on file (F), 5-keto-tetrahydrofuran-2-carboxylic acid. Ordinarily, the primary trace is that of the question (Q), and the file structure (F) must match any chosen path in Q. In this case, the trace starts with O_1' and proceeds through C_1', C_2', C_3', C_4', C_5', to O_3', where having examined all paths for O_3', the trace returns to C_5' and continues through O_4' to H_1'. Assuming that the F trace starts at O_2, exact match of element and bond is observed following the path C_1, C_2, C_3, C_4, C_5, O_4, O_3, until C_2 is reached. Here a mismatch between H_1' and C_2 occurs, and the Reverse Operation takes the primary trace (Q) back to C_2', the next prior junction point where all paths have not been followed in F. The only other path, however, leads to O_3 which is again a mismatch and the reverse operation goes back to C_1'.



Q[a]		O_1'	→	$2C_1'$	→	C_2'	→	C_3'	→	C_4'	→	C_5'	(↔ $2O_3'$)		→	O_4'	→	H_1'					
F[b]		O_2	→	$2C_1^*$	→	C_2^*	→	C_3	→	C_4	→	C_5	(↔ $2O_4$)		→	O_3	→	C_2					
Equality[c]		=	MM	=	MM	=	MM	=	MM	=	MM	=	MMR	=	MM	=		≠					
Q						C_2'	→	C_3'			←	(reverse to $C_2' + C_2^*$)											
F						C_2	→	O_3															
Eq.						=		↔															
Q				$2C_1'$	→	C_2'					←	(reverse to $C_1' + C_1^*$)											
F				$2C_1$	→	O_1																	
Eq.				=		≠																	
Q[d]		O_1'	→	$2C_1'$					AND			O_1'	→	$2C_1'$	(reverse to O_1' + NB)								
F	NB	O_1^*	→	C_1	or	H_1					NB	O_3^*	→	C_2	or	C_5							
Eq.		=		≠		≠						=		≠		≠							
Q		O_1'	→	$2C_1'$	→	C_2'	→	C_3'	→	C_4'	(	↔	O_2'	(↔ C_1'))		→	C_5'	(↔ $2O_3'$)		→	O_4'	→	H_1'
F	NB	O_4	→	$2C_5^*$	→	C_4	→	C_3	→	C_2	(	↔	O_3,	(↔ C_5))		→	C_1	(↔ $2O_2$)		→	O_1	→	H_1
Eq.		=	MM	=	MM	=	MM	=	MM	=		MM	=	NR	=	MM	=	MMR	=	MM	=	MMR	=

[a] Q (question) is master path—must be matched. [b] * unused branch in F (file). [c] =, match atom and bond; ≠, no match; MM, mark and move; R, return; N, mark and no move. [d] NB, try path with new starting point.

Figure 5.—ChemSEARCh network tracing and matching.

Here a mismatch again occurs between C_2' and O_1. All possible paths in F starting with O_2 therefore fail to match the chosen path in Q, and the F trace examines other possible starting points. Examination of O_1 and O_3 as starting points show a mismatch at the first step in each case. Finally, when the trace in the file structure (F) is started at O_4, an exact match is found at each step of the way for the complete trace through the question structure (Q).

Error Prevention.—A knowledge of the ease with which errors may occur in human transfer of information has been a fundamental guide to our thinking. Thus, for example, any departure from familiar easily recognized symbols has been avoided where possible and, for example, atoms are indicated by their letter symbols rather than the more easily confused atomic numbers. The computer can handle such transformations far more accurately than any human. So also, bond codes 1-3 correspond directly to their visual symbols. Coding and the format are simple and straightforward, and allow ready self-correction. The permitted input of parenthetical groups reduces counting errors and boredom simultaneously.

For searching, numbering need only be distinctive, not ordered. The convenient sequence in a structure is probably also the best for accurate coding. The rule for numbering of each element serially from 1 allows easy cross-check of empirical formula and structure, a process repeated by the computer which calculates H as necessary for each C. Again the order of coding is unimportant, but the convenient completion of all equations for one element numbered in sequence before beginning another is best for avoiding omissions. A check on complete coding is simply done by counting the set of statements for each element and all elements to match their count in the written structure and EF. In complicated structures, care is needed in following bonds and in completing operations at one atom before proceeding to the next. Back coding, however, introduces a useful degree of redundancy to permit a computer check of the validity of the code set.

Errors of a different order are avoided by invoking no seniority rules, preventing the necessity of making a sometimes ambiguous decision as to where to begin a structure or what is most important. The computer analysis of ChemSEARCh does not require any particular form of written structure, any particular numbering system (except that no two numbers for one element are the same) or that any point be chosen as an initial point of numeration. Also, there is no listing of groups such as the number of rings, chains, adjacent hydrogens, carboxyls, etc., even though they might be useful, if accurately entered, for screening in searches. Avoidance of greater requirements for judgement, accuracy in counting, and prerecognition of sometimes rather incognito groupings such as the ester of 2-ketotetrahydrofuran is a necessity for authentic input.

The Syntax Checker makes 47 different internal checks on the correct format and content of the coding, such as back coding, atoms bound to a subject atom which are not also subjects, correct empirical formula for a given structure, equality of bond between two atoms, and even whether the atomic codes used have been approved by the IUPAC. If this program does not discover errors, the computer makes no comment. If an error has occurred, the computer produces a paper tape for printing which states the number of the compound in error, the location of the error within the format, and the number (type) of the error. About 4% of structure codes are found incorrect, and another 4% of more easily correctable typographical, format, and counting errors are detected.

Assuming that such errors have been corrected, it is wise to consider that the paper tape is the last point for recovery from any human error, particularly in the drawn structure. It is highly advisable, therefore, to take the type-script set of codes and reconstruct the molecule from them (average time, 2 min.). The structure which has been regenerated is then checked visually, preferably by the chemist, for identity with the original structure submitted. In practice, about 0.1% of the structures have been found wrong at this point, and it has been our practice to insert structures into the main file immediately after syntax checking, deleting and replacing them later if necessary.

Formulation of Searches.—It is apparent that any completely invariant and unambiguous structure may easily be retrieved from the file simply by coding the exact structure as a question. It is only with ill-defined or incompletely specified structures or those with variable representation that any difficulty occurs. One cannot as yet expect a computer to define a structure which cannot be or has not been defined by a chemist. Such compounds, however, along with several other unusual situations may be considered in the ChemSEARCh system by a combination of the coding conventions, pseudo-elements and bonds, and ingenuity. Thus *o*-dichlorobenzene, commonly accepted as either of two different written structures, is converted to a single invariant substance, as it should be, by use of 4-bonds, and is therefore retrievable with a single exact question. Likewise, amine salts may not be given in ionic form, to avoid two file structures for the same compound. Other cases have also been considered and some are illustrated in Table IV. The philosophy of limiting acceptable representations appears desirable to us, but interchangeability with other files can always be achieved, if the ground rules are clearly delineated for the questioner.

There are, however, cases such as macromolecules, plastics, and mixtures, whose exact representation may never be possible. The input form determines what must be asked to retrieve the item. The relatively uncomplicated case of linear polymers is handled by convention, and the ChemSEARCh question for its recall is given in Table IV. To be able to store a complex condensation product of ill-defined structure such as Bakelite, a *Mixture* convention was developed. The product is then entered as a mixture of uncondensed monomers, and an example of such a retrieval is in Table IV. Also listed is a selection from an actual mixture (of fatty acids) in the file which is entered similarly. Notice that in all questions, each detached structure or group is signaled by SF separately and coded successively. The empirical formula (EF) is the sum of all the parts, and therefore a Z (Zero element) must be added to EF if only one part of a mixture is sought.

Format for most questions (Q) is identical with coding for file (F) input under SF (the structure codes), but EF (the empirical formula) must contain a Z to recall structures larger than the subfragment coded. Question numbers cannot be larger than four digits, and names are omitted. When questioning the file, any subsection of the total code except the name may be used for searching. Inclusion of a greater proportion of the total code record leads to greater and greater specificity in a question. Useful searches can be made on EF, particularly in the case of inorganic compounds which have generally, although not always, been entered without structure in F. Even without structure, since there are limited possible structural combinations of inorganics, a fairly specific answer can be obtained (See Table IV). Answers to a set of Q's are given as number and name alongside the number of each Q.

An exact EF as a question will call out all those compounds (with or without structure given) which have the exact EF, or have in addition addends (ADD's). If a question is asked with EF = Z_1, and an exact definition of ADD or a repeating unit, file compounds coded with such an ADD or x will answer (see Q for all hydrochlorides, Table IV). The entire file would be called forth by a Q formulated as EF equals Z_1.

Table IV

To seek (in F)	Code (Q)
Single organic compound	Exact EF and SF
Inorganic compound	Exact EF
Phenol–formaldehyde resin	EF C7 H8 O2 Z1 SF C_6H_5OH SF CH_2O
C_{12} and C_{14} fatty acids from a mixture having 12, 14, 16, and 18 C	EF C26 H52 O4 Z1 SF $CH_3(CH_2)_{12}COOH$ SFEX $CH_3(CH_2)_{10}COOH$
All hydrochlorides	EF Z1 ADD H1 Cl1
All nitrophenols	EF N1 O3 H1 C1 Z1 SF $Z-N(=O)_2$ SF $(Z4)_2C-O-H$
All salts of a specified quaternary	EF (exact) ADD Z1 SF $R_1R_2R_3R_4N \cdot Ad$ Ad = Z
Any amine and salts	N(6)Z
Any primary amine (no salts)	$Z-N(-H)_2$
A specified acid and any esters	$R_1-C(=O)-O(6)Z$
Any monovalent phenyl	C_6H_5-Z
All six-membered rings	$6X_6Z_5(6)Z$
Any with two phenyls separated by two atoms	$C_6H_5-X(6)X-C_6H_5$ (6) (6) Z Z
All with only one phenyl	0001N EF C12 H10 Z2 SF C_6H_5-Z SFEX C_6H_5-Z 0001 EF C6 H5 Z1 SF C_6H_5-Z
All phenylpolypropylenes $C_6H_5(CH(CH_3)CH_2-)_x-H$	EF C9 H12 Z1 SF $C_6H_5-CH(CH_3)CH_2Z$ SF $Z-CH(CH_3)_2$
All stannic triorganics	$(Z-)_3Sn(6)Z$
All sodio carbides	Na(6)C(6)Z

Structural questions may be *Independent*, that is, containing only one SF (an exact structure, or a fragment) with the corresponding EF. More general are *Dependent* Q's with two or more SF's, all of which must be included in the answers. The EF must be large enough to cover all answers, usually by including Z, but should not exceed quantitatively for a specific element the smallest answer possible, *e.g.*, the monomer of a polymer. These searches are *Inclusive*, in which case one or more of the specified atoms of each desired group can be included in another requested substructure.

Questions can also be *Exclusive* when none of the atoms of one requested group or structure can be used for a second requirement of structure. The first such group coded in a Q is signaled by SF and should be the largest. All subsequent groups start with SFEX and the EF is the quantitative sum of all specific elements except Z. One cannot, for example, make a generic request for basic amino amides Inclusive since the N of an amide would also satisfy the N of an amine.

Since question format is essentially the same as coding for file input, many can be designed from given input examples. Some modifications are illustrated in Table IV. For brevity, there are various condensations in Table IV not normally used in practice especially under SF, but the standard ChemSEARCh rules must be observed in actual operation. EF, even if omitted, is to be coded. SF, although given structurally, must be coded, and necessary signals are understood throughout the table. Under Q, R_n is a specifically coded radical, C_6H_5 is a phenyl with 4-bonds and no coded H, C_6Z_5 is a phenyl with Z's on each C, $6X_6Z_5$ is any 6-ring with Neutral bonds and Z's bonded with Neutral bonds, and (6) is a Neutral bond.

When Q is to be generic, the most important concepts to consider are Z, (6), and X. All substructures in Q are coded exactly as described but *must* be bonded to Z to indicate that the needed group can be attached to *anything*—single atom or large structure. Z's in EF are also usually necessary, and in all usage indicate that F may be greater than or equal to Q. Almost as frequently, (6) is needed to show that the bond may vary or sometimes (with Z) be nonexistent. X is used where an exact geometric spacing through any element is needed between specified parts of Q.

It is sometimes desirable to exclude from a possible answer an unwanted element, group, or even complete structure. This can be accomplished through the *NOT Question*. Any of the forms and variations of positive Q which have been described can be forbidden by the simple device of suffixing an N to a Q number coding the excluded item. Any F compounds answering the NOT Q will *not* be searched by the following Q (same number, no N), which is asking desired structure. The answers to the latter therefore can never contain excluded parts.

A final problem is shown in Figure 6. The simple structure, propionylacetone (1), is stored and retrieved easily by many systems. But which of the enol forms, (2), (4), or (5), being differently enciphered by most codes, should be stored in a file? And how should the closely related sodio derivatives, (3) and (6), be handled? Although (1) is a distinct compound, (2), (4), and (5) are actually identical, as are (3) and (6), and these three substances deserve three separate entries. Six or more acceptable codes, however, would cause retrieval failure on occasion. Examination of (5) or (6) shows the bonds for the cyclic C's and O's are resonating, so they are coded 4 in ChemSEARCh, while the other cyclic bonds are single. Plural codes for a singular substance are thus prevented, and these unique codes as Q will always bring out the right answer. If a chemist were studying any one of these, however, should he ignore the existence and properties of the others? If he wishes to be thorough, since certain components of all three structures are constant (7), he may code as Q the generic structure as given in (8) with 6-bonds and Z's, and he will withdraw from F not only the three correct structures of interest, but also all of the incorrect representations which may have been entered in the master file by chance or faulty logic.

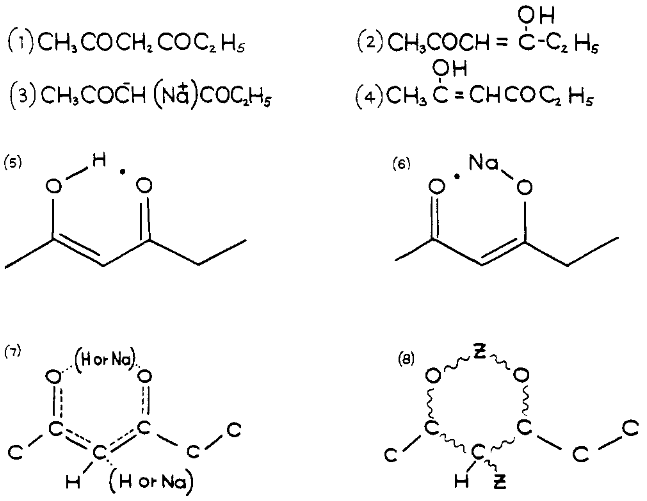


Figure 6.—Search for tautomers and relatives.

Operating Experience.—At present, the main file contains about 3000 structures of current company interest and is being expanded to cover all back files. Approximately 350 generic and specific searches have been carried out, in some cases, however, leading to excessive response when questions were formulated in overly general terms.

Manual coding and typing (for paper tape punching) is estimated to cost 20-30¢ per structure; and manual decoding for final checking is 20¢. Machine preparation of input and insertion on magnetic tape requires for 100 structures 15–22 min. if no expansion of repeated or repeating subunits is needed, and 24–32 min. if such expansions are present. The machine cost is therefore 14–30¢. Total input cost per verified structure amounts to 55–80¢. Expanded structure records average 1.3 in. on magnetic tape and a full reel thus holds 20,000 records.

Actual search time varies largely according to the number of answers, if punched paper tape output is used. Using a file of 2,500 compounds, a single question with no answers took 1.5 min., but a multiplexed search (four questions including the question above) giving 29 total answers took 2.2 min., or *ca.* 50¢/question. One multiplexed search (four questions: 7, 53, 4, and 3 answers, respectively) took 8.8 min. with paper tape output, but magnetic tape output for off-line printing, or direct on-line printing reduces this time toward 2 min. If a multiplexed search, for example, with 20 questions and 200 answers, is sorted by question number, about 1 min. more of machine time is needed.

Future Directions.—To achieve the structure–activity correlations for which the synthetic chemist aims, a computer file, keyed by number to the structure file, is being constructed to list alternative names, sources, trade names, literature references, and qualitative and quantitative properties. A search for a given value (or exceeding it) of a certain property would then elicit a set of accession numbers, whose structures are readily obtained from the structure file of the ChemSEARCh program. Conversely, a related set of structures obtained by ChemSEARCh could be used to obtain their properties for comparison. An as yet unsolved aspiration is the determination of a set of chemicals with a given property or properties, whose closest similarity in structure is then determined by a computer.

In the present program, the relatively cumbersome process of decoding is used for the necessary visual determination of accurately coded input. If an acceptable written structure could be converted directly to a mechanically correct set of topological codes and the structure regenerated from the data immediately before file input, an accurate check to the original could be made with relative facility. A solution can be achieved through the regeneration of conventional structures by the Walter Reed Chemical Typewriter of Feldman, *et al.*,[12] an adaptation of the chemical structure typewriter of the Cyanamid group [13] to permit recording the input for reproduction. To this end, a cooperative project has succeeded in transforming the codes of the Army Chemical typewriter into the complete topological description of a structure which is searchable by the ChemSEARCh system. This interface also allows commercially available coding typewriters[11] to be used for visual input through modification of the HECSAGON system of Horowitz and Crane.[14] The visual input–output can thus correspond closely to the written structures for coding in ChemSEARCh and can eliminate manual decoding for visual check with the candidate structure.

The use of an oriented structure as input, either from the Army typewriter or HECSAGON, permits the use of present written conventions for stereochemistry. It is apparent that three-dimensional, asymmetric structures can also be delineated in ChemSEARCh codes through the use of additional bond types (D and L, up and down, or right and left), when the need arises. The current operating system, however, is well adapted to most research needs, simple or complex. In addition to handling a number of borderline areas where more than one bonding configuration is acceptable or an explicit structure is unknown, ChemSEARCh provides the greatest challenge in formation of generic questions to obtain sets of compounds related in an interesting and hopefully revealing sense.

Acknowledgment.—We wish to express our appreciation to Drs. J. H. Brant and L. D. Apperson for their support and encouragement; and to E. Goldstein and C. W. Gregory for their contributions to the successful initiation of the program.

REFERENCES

(1) *Cf.* "Punched Cards," R. S. Casey, J. W. Perry, M. M. Berry, and A. Kent, Ed., Reinhold Publishing Corp., New York, N. Y., 1958.

(2) D. E. H. Frear, *Chem. Eng. News,* **23**, 2077 (1945).

(3) G. M. Dyson, "A New Notation and Enumeration System for Organic Compounds," Longmans, Green and Co., London, 1947.

(4) W. J. Wiswesser, "A Line-Formula Notation," Thomas Y. Crowell Co., New York, N.Y., 1955.

(5) *Cf.* A. D. Pratt and J. W. Perry, ASTIA Document No. AD245936, Western Reserve University, Cleveland, Ohio, Aug. 1, 1960; Staff Report, *Chem. Eng. News,* **33**, 2838 (1955).

(6) First treated by Wiener, *Mathemat. Ann.,* **6**, 29 (1873).

(7) M. Gordon, C. E. Kendall, and W. H. T. Davison, "Chemical Ciphering," The Royal Institute of Chemistry, London, 1948.

(8) C. N. Mooers, "Ciphering Structural Formulas—The Zatopleg System," Zator Technical Bulletin No. 59, The Zator Co., Boston, Mass., July 1950.

(9) See, for example, L. C. Ray and R. A. Kirsch, *Science,* **126**, 814 (1957); A. Opler, "Proceedings of the Western Joint Computer Conference," Feb. 1956, p. 86; W. H. Waldo and M. DeBacker, "Proceedings of the International Conference on Scientific Information," Vol. 1, Nov. 1958, p. 711; E. Meyer and K. Wenke, *Nachricht. Dokument,* **13**, 144 (1962).

(10) Control Data Corporation No. 160A, with basic memory of 8,192 words; No. 162-2 tape synchronizer; four No. 603 magnetic tape transports; No. 166-2 line printer.

(11) Standard Friden SPD Flexowriter.

(12) A. Feldman, D. B. Holland, and D. P. Jacobus, *J. Chem. Doc.,* **3**, 187 (1963).

(13) Miller and Fletcher, *Chem. Eng. News,* **30**, 2622 (1952).

(14) P. Horowitz and E. M. Crane, "HECSAGON," monograph, Eastman Kodak Co., Rochester, N. Y., 1961.