Searching Two-Dimensional Structures by Computer[1]

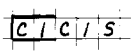
By W. H. WALDO
Monsanto Chemical Company, St. Louis 66, Missouri
Received August 24, 1961

By far the most challenging problem solved by the IBM 704 computer system[1,2] was the storage of two-dimensional structures of organic compounds in such a way that the file could be searched for any structural fragment or moiety that could be drawn and have the computer print the structure is such a way that the chemist could recognize it without translating or decoding.

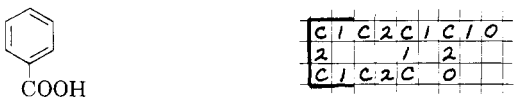
A chemical compound is stored by first writing the structure clearly and accurately in the conventional form. Then the structure is rewritten on common cross-hatched paper with numerals written for chemical bonds and single character letters for the elements where

C = carbon
N = nitrogen
O = oxygen
S = sulfur
P = phosphorus
X = chlorine
Y = bromine
I = iodine
B = boron
L = silicon
F = fluorine
R = any wholly covalent moiety
M = all other elements
Hydrogen is not in the structure but only "understood"

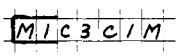
1 = single bond
CH_3CH_2SH



2 = double bond
(also coordinate covalent bond)

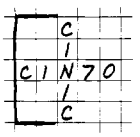


3 = triple bond
CuC ≡ CCu



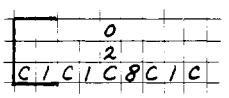
7 = ionic bond, designating salts, complexes, and similar addition compounds

$[(CH_3)_2NH]^+ OH^-$



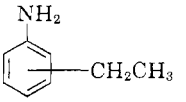
8 = a single bond, where there are an indeterminate number of carbon atoms in a chain, and the lowest number of carbon atoms possible is written

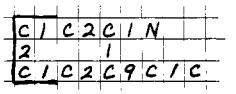
$CH_3CH_2C(=O)-(CH_2)_nCH_3$



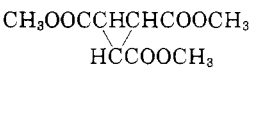
[1] Summary of presentation given before the Division of Chemical Literature, ACS St. Louis Meeting, March 1961. The invited paper and this brief note were prepared to make the information given in References 1 and 2 more available and up to date. Following the presentation of this paper the author has become aware of two similar systems, one developed by Pepinsky at the Groth Institute and the other by Crane, *et al.*, at the Kodak Research Laboratories.

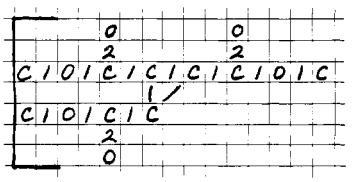
9 = only a single bond where the point of attachment is in doubt; where there exists indeterminate geometrical isomerism

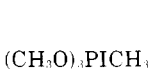


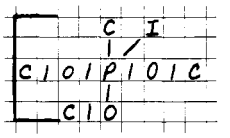


/ = a special bond symbol used in special cases to indicate single bonds only; such as three-membered rings and penta-substituted elements

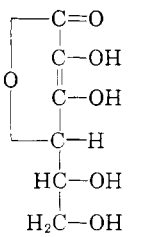


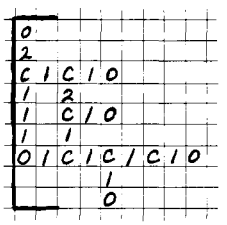






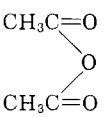
Compounds with an odd number of elements in the ring (5, 7, 9, etc.) are handled by repeating the appearance of one of the bonds three times so that the rectangular configuration is maintained.

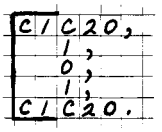




In this structure writing, (1) the aromatic resonance is frozen into the alternating sequence of double and single bonds; (2) the benzene hexagon becomes a rectangle, and (3) there are no written hydrogen atoms in the system; thus a single bonded oxygen represents a hydroxyl group—single bonded sulfur, a sulfhydryl—a single bonded nitrogen, a primary amine, etc.

This rewritten form of the organic structure is now ready for a keypunch operator. To make it convenient for the keypunch operator I suggest boxing and punctuating each structure thus





The "box" is simply limits, left-hand, top, and bottom, of the structure. The punctuation is the right-hand limit. This makes it easier for the keypunch operator to enter the structure onto a punch card.

Arbitrarily I have limited the number of characters in the horizontal line on cross-hatched paper to 98 since the computer is limited in the length of line it can print. Furthermore, I have limited the number of rows arbitrarily to 13 to accommodate our particular report-writing program.

As the record is being transformed from the punched card format to the output format, the 704 calculates certain control information it needs in order to remember how to reproduce the structure when called upon. It counts the number of horizontal rows in the structure (three or five for the benzene ring); it also counts the number of columns in each row of the structure beginning at the left edge as indicated by the box line and terminating with the punctuation. Editing is then performed by the 704 to ensure that the structure does not contain an incorrect number of rows and that each row does not contain an improper number of characters.

Incorrectness occurs, of course, as a result of human errors. I have noted two empirical coincidences; both of which are used in this editing process: (1) the number of rows in the rewritten structures must of necessity always be odd, 1, 3, . . . , 13; and (2) the number of squares from the box in any row must be odd, since all elements are single-character symbols as are the bonds. However, the last position in every row is signaled with a punctuation mark. Thus the number of characters, including blanks, in a row is always even, two through 98. When errors are detected, the structure is rejected from entrance to the permanent file. If correct, the control numbers are stored as a part of the structure record. After the structure is stored the computer proceeds to calculate the molecular formula and insert it in its proper place in the record.

As the structure is being analyzed, further checking by the machine is made to determine the accuracy of the input data. The rules for rewriting structures are integrated in the program so that the computer is able to take a sophisticated *look* at the chemist's rewritten structure and the keypunch operator's work. It will not allow any atom to have too many or too few bonds, nor is a 7-bond permissible with atoms for which ionic bonds are not "legal." Improper atom and bond symbols and misplaced characters are recognized by the computer.

A chemist proposing a machine search of the chemical structure files must state precisely the elements and bonds he wants, how they should be connected, and what he does *not* want. These search specifications are transformed to IBM cards, and become the set of rules by which the computer will perform the search.

Search questions such as finding all the derivatives of resorcinol are meaningless to the computer. The chemist posing the question must determine for the computer what is meant by a derivative. It can pull out all phenols, all compounds having a benzene ring structure, or the computer can pull out all compounds containing two hydroxyl groups. Control data include the molecular formula requirements and the substructure, rewritten in the same manner as the structure records. Further control information is utilized to control the switching network during the search.

REFERENCES

(1) W.H. Waldo, R.S. Gordon, and J.D. Porter, *Am. Document.*, **9** (1), 28 (1958).

(2) W.H. Waldo and M. DeBacker, Preprints of the International Conference on Scientific Information, Washington, D. C., Nov. 1958, Area 4, pp. 49–68.

Application of a Line Formula Notation in an Index of Chemical Structures*

By. H. T. BONNETT and D. W. CALHOUN
G. D. Searle & Co., Chicago 80, Illinois
Received August 24, 1961

Any laboratory engaged in synthetic organic chemistry finds an index to past efforts essential. Over a period of years the total number of compounds studied often becomes large enough to make inviting the use of machinery in the creation and use of indexes to such files of compounds.

Such was the situation in the Searle Laboratories. Over a period of about twenty-five years the laboratories had made and screened thousands of compounds for biological activity. New compounds were being entered into the file at an increasing pace. This experience, of course, is not unique.

Our initial effort was directed to a chemical structure index. While many types of information could be put in such an index, the information we chose included the following: (a) the structure of the compound; (b) the name of the chemist who submitted the compound; (c) the identifying number of the compound; (d) a code for rough classification; (e) a functional group index. To this list should be added the desire that the format chosen be capable of extension to compounds taken from published literature.

The design of any index must be tailored to the facilities available. Until accounting type equipment for research purposes could be justified by use, it was determined to use the facilities of the accounting department tabulating section. The equipment available included the usual IBM key punch, verifier, sorter, collator, duplicating punch, and tabulator machines widely used in accounting operations. The basic indexing principles adopted, using this equipment, was to prepare one punched card per compound, and to translate questions asked of the index into the corresponding manipulations of the punched cards. But the machinery is located in another building and is not available at all times. These factors obviously invited development of a physical form of index which would not require manipulation of cards for all searches. To meet this need, it was desirable to sort the file into