The Automatic Encoding of Chemical Structures*

By ALFRED FELDMAN, DONALD B. HOLLAND, and DAVID P. JACOBUS
Walter Reed Army Institute of Research, Washington, D. C.
Received June 7, 1963

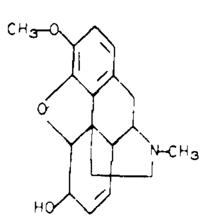
Many methods for the coding of chemical structures have been described in the literature. Some methods code a compound only partially; typical of these are the methods using descriptor codes. Other methods, among which the so-called ciphers are prominent, code compounds exhaustively. All these methods require cerebral effort; that is, a chemist is needed who must have learned the rules of the code, and who must known how to dismember correctly each structure to be coded. A disadvantage of code designations of structures is, furthermore, that they are not generally understood by chemists. No advantages accrue to the chemist from knowing how to generate and how to interpret a chemical code. Codes are needed only for the mechanical manipulation of chemical structures. Clearly then, if the coding of chemical compounds could be accomplished automatically, this automatic conversion would relieve the chemist of a considerable burden.

A conversion of this order has already been demonstrated by the Monsanto Chemical Company.[1,2] They have obtained a machine-intelligible code by doing nothing more than drawing structures of the shape illustrated in Fig. 1a, and punching these directly, with commercially available equipment, on standard Hollerith cards. The output from the computer, as obtained through a high-speed printer, has the same shape. Although it does not diminish the value of the Monsanto contribution, this distorted shape is unfortunate.

Chemical structures of attractive, undistorted shape (Fig. 1a) have been obtained since 1952, when Miller and Fletcher of American Cyanamid equipped a conventional typewriter with special keys[5] (Fig. 2). The machine is still in use, its sole purpose being to produce attractive structures for their card files. They claim that the time required for typing a structure averages to about two minutes.[6]

It now occurred to one of the authors (A. F.),[7] that the Miller-Fletcher typewriter might be modified to generate the code of the compound typed out, so that this code would be obtained indirectly, as a completely automatic by-product of the typing operation.

A typewriter, namely, might record sufficient information to satisfy all the requirements for chemical coding. These requirements were laid down by Calvin Mooers in 1950.[8] Drawing upon an earlier concept by Gordon, Kendall, and Davison,[9-11] Mooers stated that if a unique number were assigned to each atom in a given molecule, and if the connections between atoms were shown, then the molecule would be completely encoded, and the code suitable for computer manipulation. The code would be correct even if, at different times, different numbers were assigned to the same molecule. In other words, the sole limitation is that no number be used more than once.



(a)

```
C 1 1 1 1 1 C 1 C
1           2   2
C 1 N 1 C   2   C
1   1       2   1
1   C       2   1
1   1       2   1
1   C       2   1
1   1       2   1
C 1 C 1 1 1 C   1
1   1       1   1
C   C 1 O 1 C 2 C 1 O 1 C
2   1
C 1 C 1 O
```

(b)

Fig. 1.—Encoded structure for codeine, input and output forms as obtained by: (a) Walter Reed typewriter; (b) Monsanto Chemical Co.

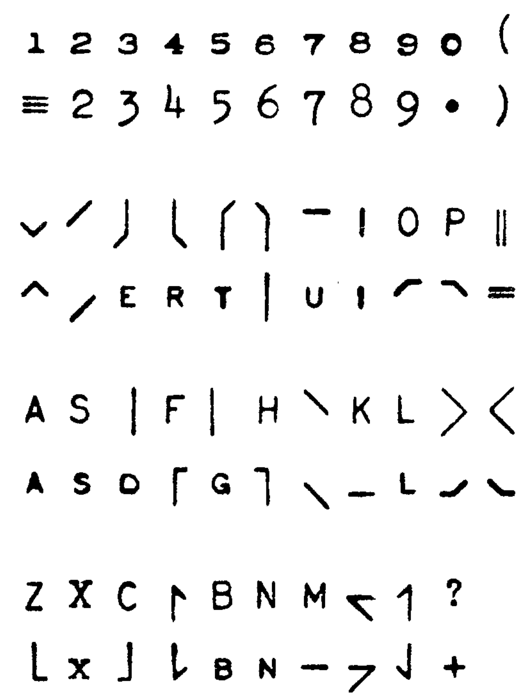


Fig. 2.—Strike-out of keyboard of the chemical typewriter.

* Presented before the Division of Chemical Literature, 141st ACS National Meeting, Washington, D. C., March 22, 1962.

These requirements are fulfilled on a typewriter if, in addition to the characters typed, it records also the coordinates of their locations on the typewritten paper. A typewriter is a digital device. The type-heads in a typewriter strike the paper only in discrete locations. Each of these locations can be uniquely designated by its Cartesian coordinates, *e.g.*, line 10, space 26. To those keys that represent atoms, the coordinates serve as unique numbers, in the sense specified by Mooers. For those keys that represent bonds, the coordinates will permit establishing which atoms they link.

Key codes and coordinates are recorded on paper tape. This constitutes the entire chemical code. If the typed structure is correct, then it is also encoded correctly. The typist may type as she pleases. There are no restrictions.

The code so obtained is exhaustive; that is, it contains the complete description of each chemical structure coded. The code is devoid of idiosyncrasies, as it does not introduce artificial groupings, in the fashion of descriptor-codes or ciphers. Since there are no rules according to which to number atoms within a structure, and since, accordingly, such numbering may be made even at random, it follows that the disposition of the structure on paper is irrelevant, too. Mooers' code can be encoded by hand and has been used in this fashion experimentally,[12] but, although it requires no acquaintance with rules, the sheer amount of labor necessary makes its manual encoding very time-consuming and subject to errors.

The unit built at the Walter Reed Army Institute of Research is a tape typewriter, modified by the addition of commutators with sliding brushes on both platen and carriage, by means of which the y- and x-coordinates are sensed each time a key is depressed (Fig. 3). By appropriate circuitry[13] all this information is punched onto paper tape. The punches for each character and its location appear in blocks (Fig. 4). The first row in a block represents the code of the typed character, the second row its y-coordinate, and the last two its x-coordinate. The typewriter can code either from manual input (striking the keys) or from tape input, and it is possible to use the tape-input facility to feed the typewriter

Fig. 3.—The chemical typewriter, built at the Walter Reed Army Institute of Research.

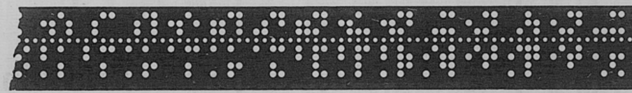


Fig. 4.—Typical tape produced with the chemical typewriter. The blocks, each containing the codes for the key struck, and its x- and y-coordinates, are clearly discernible.

fragments of structures which recur frequently, such as the benzene ring. These fragments may be positioned by using the tabulator stop. With the tape-feed input, it is possible to code the entire steroid nucleus in 35 seconds.

Techniques for searching chemical structures on an atom-by-atom basis have been reported by Ray and Kirsch,[12] Waldo and DeBacker,[1] Meyer and Wenke,[14] Opler,[15] and others.[16,17] All these techniques require the use of a substantial computer. We do not apologize for the need for a computer. We feel that the disadvantage of needing a computer is offset, first of all, by gaining facile access to an exhaustive code; secondly, by being able to conduct sophisticated searches, correlations, etc.; and lastly, by providing an automatic monitor for errors of implausibility.

If the routine use of a computer is inconvenient, a one-time operation might be resorted to, by means of which a deck of punched cards could be obtained, coded according to any specified descriptor code. These cards may be searched thereafter on inexpensive sorting equipment. If an out-of-the-ordinary search is desired, recourse to the computer by means of the original typewriter tapes is always possible. Also, there remains always the feasibility of obtaining new decks of cards from these tapes, punched according to new descriptor codes. As long as structures are used in chemistry, the original tape from the typed structures cannot become obsolete.

It should be mentioned, finally, that the typewriter may serve also to obtain an attractive pictorial printout from computer produced tapes.

ACKNOWLEDGMENT

To Mr. B. R. Moberly of Remington Rand, we wish to express our most appreciative thanks, for his invaluable and ready help in our struggle with the intricacies of the typewriter.

REFERENCES

(1) W. H. Waldo and M. DeBacker, "Printing Chemical Structures Electronically: Encoded Compounds Searched Generically with IBM-702," International Conference on Scientific Information, IV, NAS-NRC, Washington, D. C., 1958, p. 49.

(2) Other methods for the mechanical encoding of chemical structures have since been suggested. Meyer[3] is developing a machine that scans, by means of photo-cells, a structure hand-drawn on a grid; structures that can be coded in this way are limited, however, to carbon rings and chains, and other elements have to be introduced indirectly. Pantyukhina[4] describes an optical scanner which recognizes structures hand-drawn on a grid, and which may contain any elements.

(3) E. Meyer, statement provided for the National Science Foundation Report, "Current Research and Development in Scientific Documentation," No. 9, 1961.
(4) M. Ye. Pantyukhina, "Machine Input and Output of Information on the Structure of Chemical Compounds," in Foreign Developments in Machine Translation and Information Processing, Office of Technical Services, No. 31, U. S. Dept. of Commerce, Washington, D. C., June 22, 1961, pp. 77-91 (JPRS 8479).
(5) *Chem. Eng. News*, **30**, 2622 (1952). Another keyboard, described by F. Murphy in T. E. R. Singer's "Information and Communication in Industry," Reinhold Publishing Co., New York, N. Y., 1958, pp. 293-296, yields esthetically less pleasing chemical structures.
(6) Personal communication.
(7) The invention is therefore credited to A. F.
(8) C. N. Mooers, "Ciphering Structural Formulas—The Zatopleg System," Zator Technical Bulletin No. 59, The Zator Co., Boston, Mass., July, 1950.
(9) M. Gordon, C. E. Kendall, and W. H. T. Davison, "Chemical Ciphering: A Universal Code as an Aid to Chemical Systematics," Royal Inst. of Chemistry of Great Britain and Ireland, 1948.
(10) W. H. T. Davison and M. Gordon, *Am. Doc.*, 8, 202 (1957).
(11) Dr. Gordon, in a private communication, dated April 2, 1962, made the following comment: "I might briefly mention one point about the Gordon, Kendall, Davison System concerning its topological basis, especially as authors of other systems have made considerable play concerning the topology of their systems. Our method of tracing a path through a structure is essentially that given by Wiener, who first considered the old problem of labyrinths as a mathematical problem in Mathematische Annalen, **6**, 29-30 (1873), as quoted in the famous book on combinatorial topology by D. Köning, "Theorie der Endlichen und Unendlichen Graphen," Akademische Verlagsgesellschaft M. B. H., Leipzig, 1936, p. 36. The reason I did not quote this in the original papers is that I innocently re-discovered Wiener's solution. There are very minor logical differences between our treatment and Wiener's solution, but essentially they are identical. I doubt if any real improvement on this problem has been made since 1837."
(12) L. C. Ray and R. A. Kirsch, *Science*, **126**, 814 (1957).
(13) This circuitry will be described in another publication.
(14) E. Meyer and K. Wenke, *Nachr. Dokument.*, **13**, 13 (1962).
(15) A. Opler, "A Topological Application of Computing Machines," AIEE Special Publication T-85, Proceedings of the Western Joint Computer Conference, February, 1956; *cf.* also: A. Opler and T. R. Norton, "A Manual for Programming Computers for Use with a Mechanized System for Searching Organic Compounds," The Dow Chemical Co., Midland, Mich., April 25, 1956.
(16) G. E. Vleduts and E. D. Stotskiy, "On Certain Systems for Recording Structural Formulas in Organic Chemistry," ref. 4, No. 30, June 2, 1961, pp. 36-48 (JPRS 8372).
(17) V. B. Borshchev, G. E. Vleduts, and V. K. Finn, "An Algorithm for Translating Structural Formulas in Organic Chemistry into Canonical Notation," ref. 4, No. 58, December 12, 1961, pp. 63-125 (JPRS 11483).

A Linear Notation for Organic Compounds

By JOHN A. SILK
Imperial Chemical Industries Limited, Jealott's Hill Research Station,
Bracknell, Berkshire, England
Received May 9, 1963

While the notation outlined in this paper in some respects resembles the INPAC Notation,[1] its distinctive features justify its consideration as an independent and alternative system. Apart from semantic differences, it may be noted here that the notation uses only one alphabet and one set of numerals, so that the standard teleprinter keyboard and computer print-out system provide an adequate range of symbols. Since errors in ciphers are less obvious than those in names, the absence of subscript or superscript numerals or lower-case letters is also an important simplification in copying and checking ciphers. It will be appreciated, however, that additional symbols, such as subscript numbers for use as multipliers, could be introduced, if desired. It may also be pointed out that possibly the most attractive and legible ciphers are those with lower-case letters and small (but not subscript) numerals. Compare B4M(A)2.2F.AOQ—2A3N and b4m(a)2.2f.aoq—2a3n.

THE BASIC SYMBOLS

The first four letters of the alphabet are used to cipher the fundamental chains and ring systems of organic molecules, and for this reason they are referred to as *basic symbols*.

A denotes an Aliphatic (Alkane) carbon atom, together with the appropriate number of hydrogens.

B denotes the simple (unfused) Benzene ring. It is a specific symbol, and no other ring fused, hydrogenated, or heterocyclic, is ciphered with B.

C is the general symbol for a Cyclic structure. The number following it gives the number of rings in the structure. Ring systems are understood to contain the maximum number of noncumulative double bonds and rings to be six-membered, in the absence of other information.

D is used to cipher individual heterocyclic rings in a ring system. The number following it shows the size of ring, and it is followed by the symbols for each of the hetero atoms.

Thus, A*n* denotes a chain of *n* carbon atoms, C*n* a system of *n* fused rings, and D*n* an *n*-membered heterocyclic ring. A, C, D, and also H, are the main letter symbols for which numbers immediately following serve as multipliers. Multiplication of most other single letter symbols is shown by repeating the symbol.

The symbol A has been chosen for an aliphatic carbon atom because carbon is the prime element in organic compounds, and in alphabetical arrangements of ciphers it is