Linearly Organized Chemical Code for Use in Computer Systems (Locus)

By H. BOUMAN
Shell Internationale Research Maatschappij N. V.,[1] 30, The Hague, Holland
Received July 3, 1962

This paper describes a system for the computer-based documentation of chemical structures. Other data may also be included in the system. Clerical staff can write the code number for the structures concerned provided structural formulas are given. The code number obtained generally contains as much information as the structural formula itself. The appendix describes how to use the computer for the retrieval of substructures. Another article is being prepared in which it will be shown that by processing the data fed into the machine beforehand the retrieval can be reduced to direct comparisons for most of the questions.

INTRODUCTION

The shortcomings of manual documentation methods in chemistry are becoming more and more evident. The availability of computers suggested the idea of seeing whether they could help to reduce the difficulties encountered.

Several ingenious systems of linear notations for chemical structural formulas suitable for handling by computers have been published (see bibliography). Generally a considerable chemical knowledge is required to use these systems, and this makes them costly. An exception is the system devised by W. H. Waldo, under which a clerical staff can prepare the data for a computer provided the structural formula is given. This system, however, is not based on linearly written notations. Further, the existing systems are intended for use both by those who want to retrieve information by machines, and also by those who want to make linearly written recognizable index entries for chemicals. The author believes that the latter aim unnecessarily complicates the code, thus making coding more cumbersome. There is therefore still a need for a simpler system which can be used by clerical staff provided the structural formulas are given.

A simpler system can indeed be devised if indexing is left out of account, which means that one and only one code notation for each compound is no longer necessary, although the reverse requirement, that one compound only must be covered by each code notation, is retained.

The system suggested here thus encompasses the possibility of "synonymous" notations for one and the same compound. This unavoidably increases the task of the computing machinery, but the human part of documentation work always takes such a large share of the total cost that the reduction in the cost of the human effort obtained under this system should outweigh the increase in the cost of the machine work. In addition, the computer input can be transformed into more readily retrievable data by using spare machine capacity, as will be shown in another article, and, as a result, the cost of a search can be reduced considerably. In fact, if several computer centers cooperate, only one need do the transformation work, the others simply feeding the data obtained directly into their machines. Further study will be required, however, to assess the machine time and cost involved, as a detailed machine program has not yet been developed. Questions involving whole compounds, and also most questions involving substructures, *i.e.*, compounds having certain structural characteristics, can be dealt with by means of the transformed data.

The system also provides for the straightforward coding of so-called general formulas (as are often encountered in patent specification) which cover a number of alternatives and it is not necessary to mention the same alternative group more than once during coding.

The system has not yet been developed sufficiently to cover the coding of compounds which are only partly defined (*e.g.*, alkylhalide, chlorinated styrene, etc.), but this is thought to be possible.

The code notations of chemical structures can be combined in one system with completely different data, *e.g.*, even if such data are not coded, but are indicated by index words only. These index words can be either free words or selected standard words or a combination of free and standard words. The difficulties which arise in index word systems from the huge number of chemical compounds to be covered can be eased considerably by such combinations, since by using them much of the data is covered systematically instead of by the index words. Some suggested combinations are given below.

The principles used for coding chemical structures can also be profitably applied in other fields of documentation involving structures, *e.g.*, electric wiring diagrams, distillation plant schemes, etc.

THE SYMBOLS USED

Altogether there are 44 symbols, so as to make it possible for ordinary computers to use the system (most computers can process at least 44 symbols).

Atoms are indicated by their usual chemical symbols, except that capital letters only are used. Exceptions are

[1] For a more detailed description please apply to: Shell Internationale Research Maatschappij N. V. RSP (Patent Division), Carel van Bylandtlaan 30, The Hague, Holland.

carbon atoms in aromatic rings, denoted by M, and divalent carbon atoms (carbene compounds), which are indicated by X. Hydrogen atoms are not mentioned if they are bonded to carbon and if they do not form one of the alternatives in general forumulas. A special symbol ˆ is used to indicate that the next two symbols belong together (*e.g.*, ˆCO cobalt while CO is carbon bonded to oxygen; 14 is one and four, but ˆ14 is fourteen).

Double bonds (not in aromatic rings) are indicated by the symbol =, triple bonds by the symbol : .[2]

Electric charges and electrovalent bonds are indicated by the symbol E, and free radical electrons by the symbol R, both being treated as if they were atoms.

Identical groups of more than one atom in a linear chain are written in parentheses followed by the number of times the group occurs (*e.g.*, -C-O-C-C-O-C-C-O-C- in a compound is given as (COC)3, etc.).

The letters Q and Z followed by figures or letters are used to indicate substances involved in a reaction (*e.g.*, starting material, catalyst, end-product, etc., reaction conditions, etc.

Two letters, G and L, are reserved for indicating "repetitions," for example, if a number of steroids have to be coded the identical part common to all steroids is placed between G and L the first time and the following times is replaced by GL, the machine being left to write out the coding in full. The symbols indicating the whole series of steroids are enclosed in parentheses.

The symbol, (comma) is used to indicate that the next two symbols are to be read as: from... to ... inclusive. Thus 14 means one and four but , 14 means: one, two, three and four.

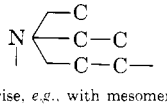
The use of other symbols such as ! and / is discussed below.

STRUCTURAL INFORMATION

The atoms in a linear chain are given in the sequence in which they occur preceded by the symbol /; thus normal hexane is /CCCCCC, which is shortened to /C6, and diethyl ether is /C20C2 (ignoring hydrogen, as already said).

Non-linear structures contain so-called branching point atoms, *i.e.*, atoms bonded to more than two other atoms, disregarding hydrogen. Each branching point atom is given a reference number, in an arbitrary sequence, but preferably with atoms of the same kind having successive numbers.

The code number is constructed by giving the reference number(s) followed by the atom type(s) pertaining to those reference number(s); then the linear chains of atoms bonded to the branching point atoms are given, each chain preceded by the symbol / and by the reference number of the branching point to which it is bonded and starting with the atom bonded to the branching point atom. Thus methylethyl *n*-propyl tertiary amine



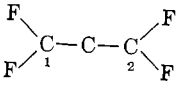
is coded as:

1N/1C/1C2/1C3.

The sequence of the parts coming after / is irrelevant, although coding is easier with an ascending series of reference numbers.

If, when following an atomic series, another branching point atom is encountered, it is indicated by the symbol ! followed by the reference number of the atom (the atomic symbol itself is not give, as it has already been indicated in front of the first /).

Thus, 1,1,3,3,-tetrafluoropropane

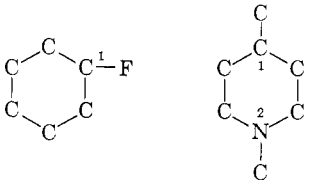


is coded as:

12C/1F/1F/1C!2/2F/2F.

When two branching point atoms are directly connected the same notation is used (thus if the central C-atom were absent the symbols /1!2 would be used instead of /1C!2).

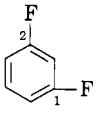
The same notation is used if, when a certain atomic series is followed, the branching point atom with which the series started is met again, as is the case with certain ring structures. Thus monofluorocyclohexane and *p*-methyl *N*-methylpiperidine



are coded as:

1C/1F/1C5!1 and 1C2N/1C/1C2!2/1C2!2/2C

while *m*-difluorobenzene

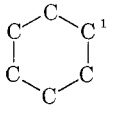


is coded as:

12M/1F/1M!2/1M3!2/2F

(since M stands for aromatic carbon atoms).

In single ring compounds without substituents, an arbitrarily chosen atom is taken as the branching point atom. Thus cyclohexane



is coded as:

1C/1C5!1.

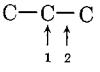
ALTERNATIVE GROUPS

If several alternative atoms or groups may be bonded to a certain part of a structure (*i.e.*, to the so-called

[2] Although some difficulties may arise, *e.g.*, with mesomers, these notations are short and do not cause much "noise" in a search. The difficulties can be avoided by mentioning the hydrogen atoms, or by following Gordon, *et. al.*, using a letter for CH_3, CH_2 and CH groups, but for the time being this is not considered necessary.

constant part) at least two branching points are considered to be present, *viz.*: (1) one coinciding with the atom of the constant part to which the alternatives may be bonded (irrespective of whether or not this atom is a branching point atom as defined above); (2) one between that atom and the alternative groups.

Thus in the general formula C—C—R, where R may be Cl or Br or F the (fictitious) branching points are considered to be in the positions indicated by 1 and 2 below



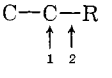
As "atom type" indication (given in the first part of the code number before the symbol /) the symbol : is used for "branching points" not corresponding to an atom (point 2 above).[3]

The code number for the above general formula will be

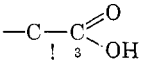
1C2:/1C/1!2/2^CL/2^BR/2F

If there are also branching point atoms *in* the alternative groups, they are indicated in the usual way in the first part of the code number (before the first symbol /), each of them being preceded by the symbol ! to indicate their position.[4]

Thus if in the above example a carboxy methyl group ($-CH_2-COOH$) could also be bonded to the ethyl radical, *i.e., in the formula*



where R may be Cl, Br, F or



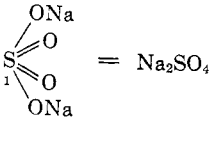
the code number would be:

1!3C2:/1C/1!2/2C!3/2^CL/2^BR/2F/3=0/30H.

SYNONYMS AND NON-STRUCTURAL DATA

As it would be cumbersome to code frequently occurring compounds in accordance with their structural formula each time they occur, a simpler notation is used. If this simpler notation is correct as to the number of atoms present (*e.g.*, in Na_2SO_4) the symbol ZA is added, which can be interpreted as: "if more than the number of atoms is required for a test, look up the synonymous structural notation in the memory."

The structural notation is fed into the computer together with the ZA notation. Thus in the example



yielding the code number

1S(/10^NA)2(/1=0)2ZA^NA2SO4

structural notation | synonymous notation

Only the last part preceded by ZA is used subsequently.

If the synonym is not in the usual atom notation (*e.g.*, if the compound is indicated by its name, which is often better), ZZ is used instead. This means: "Use the memorized structure even if a test for the atoms only is required."

Other non-structural data (*e.g.*, a function of a compound such as "insecticide") are added to the symbols for one or more compounds by giving to the word(s) (*e.g.*, insecticide) containing that data and to the code number for the compound an identical indication, *e.g.*, Z followed by a number. To indicate that the meaning of Z followed by a number does not apply to later information, the last compound for which a certain series of Z-number indications is relevant is given a Z-zero indication in addition to the other Z-number indication.

Thus if a patent specification claims a copolymer of vinyl chloride and acrylonitrile used in glass fiber laminates and mentions in the description a copolymer of butadiene and acrylonitrile used as a leather oil, this can be indicated as follows

Z1 claimed	Z6 leather oil
Z2 copolymer	Z1Z2Z4Z5/C=C^CL
Z3 copolymer	Z1Z2Z4Z5/C=C2:N
Z4 glass fiber	Z3Z6/(C=C)2
Z5 laminate	Z3Z6Z0/C=C2:N

The index words used will have to be included in a dictionary, so that questions put to the computer can be properly formulated.

If one does not want to write down the code number for acrylonitrile twice (as would be the case if it were a complicated compound with a long code number) the "repetition" symbols (G and L) can be used. Thus the last three indications will be

(Z1Z2Z4Z5G/C=C2:NL Z3Z6/(C=C)2 Z3Z6Z0GL)

These Z indications have a function which is more or less comparable to the conventional links which are often included in mechanical documentation systems.

Other indications similar to the conventional rôle indicators, can also be used; compounds participating in a reaction can, for example, be specified by numbered Q indications

Q 1 : starting product
Q 7 : end product
Q 9 : catalyst
etc.

These indications can precede the normal code number of the compound in question.

If desired, certain Z or Q indications can be given fixed meanings (like "standard words" in index systems), for which purpose combinations of letters are used. They

[3] This symbol : means a triple bond in the second part of the code number (*i.e.*, after the symbol /), but confusion cannot arise.

[4] The symbol ! has another meaning in the second part of the notation (after the symbol /), *viz.*: "this linear chain of atoms ends in a branching point, having the following reference number," but confusion cannot arise.

can be stored in the memory and should not be deleted therefrom by the Z-zero indication. Certain types of isomerism (*cis*, *trans*, *etc.*) can, for example, be indicated by ZBA, ZBB, *etc.*

RETRIEVAL

The LOCUS system makes it possible to code detailed information. This means that a complicated machine program will be required for the retrieval. As each symbol, however, is used according to stringent rules and has a clear-cut and simple meaning, it is not impossible to make such a program. One of the characteristics of the notation described is that the numbering sequence of the branching points is arbitrary. This is because there is no advantage in laying down rules for this sequence, since they would complicate the process of coding a compound. The I.U.P.A.C. system, for instance, contains a number of rules which are difficult to memorize and to handle and can be applied only by people with a thorough knowledge of chemistry. The disadvantage, however, of the notation described, resulting from the arbitrary numbering of the branching points, is that many different code numbers can be formed for the same compound, and it is not easy to recognize that they indicate the same substance. A computer which has one code number in storage and into which another code number for the same compound is fed as a question must be able to relate the two code numbers to the compound. Therefore, it would be advantageous if the code numbers of all compounds fed into the storage and of all compounds fed in during retrieval could be transformed into standard notations for each compound (another article is being prepared in which the formation of such a standard notation is discussed). The process of recognizing the identity between a stored compound and a compound asked for would in that case be reduced to mere comparison.

Some of the more simple transformations that could be made by a computer to the structural code indications fed into it so as to simplify later retrieval have been investigated in a preliminary and empirical way. Only a few seconds per compound probably will be required on an average for these transformations with an adequately programmed computer (computer spare capacity could be used). Retrieval itself is simplified considerably by these transformations and the total machine time is reduced, since each compound is stored only once, although it may be screened over and over again.

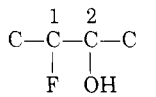
The author hopes that further development will make it possible to satisfy both of the nearly contradictory wishes of the information worker: simple coding and rapid retrieval.

Acknowledgment.—The author wishes to express his gratitude to Mr. B. de Graaf for frequent fruitful discussions and suggestions, and to Shell Internationale Research Maatschappij N. V. for permission to publish this article.

APPENDIX

Retrieval of Substructures.—The computer can transform a code number fed into it for storage into a list containing all the atomic series either from branching point to branching point inclusive, or from a branching point to the end of a chain.

The atomic indications for the branching points themselves can be taken from the first part of the code number. The compound



can thus give the symbols

12C/1C/1F/1!2/20H/2C

which are transformed into

1CC 1CF 1CC2 2COH 2CC

For those terms in the list which also have a branching point indication at the end (the term 1CC2), an additional term is inserted, which is obtained by reading the series backwards. Thus the list will be arranged according to front reference number:

1CC	2CC1
1CF	2C0H
1CC2	2CC

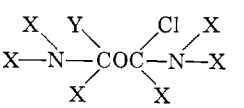
in which there are two sections, one for each front reference number.

Although the sequence of the terms and the numbers given to the branching points may vary, the terms themselves and their combinations cannot vary.

Searches involving certain types of questions can easily be answered from the storage list.

The presence of any univalent substituent(s) (*e.g.*,—OH group) can be found by simple subtraction of the atomic series involved in the storage list from the question list: when there is no positive remainder a compound containing that substituent has been found. Only terms with no reference number at the end need be checked if no branching point occurs in the substituent. If one branching point only occurs in the substituents only sections with one branching point at the end need be checked; in addition, the atomic series occuring in the question must be present in one section of the storage list. If two or more branching points occur in the structure sought, the univalent groups given in the question are sought, as mentioned above, and a check is made to see whether the branching points found in front of these series are bonded to each other as required.

If, for example, the structure required is:



irrespective of whether it is present wholly or partly as a ring structure, in which X and Y denote any group, but Y is not hydrogen, those compounds are selected which in the storage list are represented by one section containing

I CN. .
I COC II

in which roman ciphers indicate any reference number, and by another section containing

III CN. .
III C CL
III COC IV.

A check is then made to see whether I and II are identical with IV and III.

Questions in which it is not clear whether a certain atom is a branching point or not are more complicated. If in the above example Y can also be hydrogen, compounds in which the first carbon atom is not a branching point are also required. A question of this type may be split into two different questions, the first as given above, the second assuming that Y is hydrogen, and asking both on an "either-or" basis.

If the number of atoms which may or may not be branching points is more than one, several permutations are possible and the number of questions to be put increases considerably. Usually a maximum of about four dubious branching points can be catered for by "either-or" type questions. In the majority of such cases the amount of information already known about the compound will be so large that there will be very little "noise," if one neglects the dubious branching point indications altogether and searches only for atomic series which are known to be present.

If certain details can be stored beforehand together with the lists, the possibilities of the system are increased even more.

It is, for example, simple to assess the number of rings in the compound: those terms which have a reference number at the end and which are not identical with an earlier term when read backwards are arranged in such a way that (apart from the first front reference number) all reference numbers are mentioned earlier in the list as an end reference number than as a front reference number; in the list obtained the number of times that an end reference number is repeated is counted and this number is added to the number of terms in which front and end reference number are identical. If this number is calculated beforehand and used as a skipping indication, questions directed to compounds with for instance five rings will require a considerable reduced searching time.

It is also possible (as will be shown in another article) to store together with the storage list all the separate atomic series in each of the rings present, so that the rings can be checked separately. Chains of atoms outside rings and larger than those running from branching point to branching point can also be stored beforehand if desired. Such ring and chain notations then can be checked together with the other known requirements. The check consists of matching the required ring with those stored, each time shifting an atom from the back to the front until either a match is obtained or all possibilities have been tested. For example, the tetrahydrofuran ring can be written as:

CCCCO	or
OCCCC	or
COCCC	or
CCOCC	or
CCCOC	

and can be found in 5 shifts and subtractions at the most. If desired, each possibility can be stored beforehand, so that searching consists simply of subtraction. Also only the notation which has alphabetical priority could be stored and used in the questions.

BIBLIOGRAPHY

(1) M. Gordon, C. E. Kendall and W. H. T. Davison, "Chemical ciphering; a universal code as an aid to chemical systematics," Roy. Inst. Chem. Gt. Brit. Ireland, 30 Russel Square, London W. C. 1, 1948.

(2) G. M. Dyson, "A new notation and enumeration system for organic compounds," 2nd edition, Longmans, Green & Co., New York, N. Y., 1949.

(3) W. Gruber, "Die Genfer Nomenklature in Ziffern und ihre Erweiterung auf Ringverbindugen," Auszug: *Angew. Chem.*, 429, 61p. November, 1949; Ausfuhrlich: Beiheft 58 *Angew. Chem.*, (DM. 4.50) 1950 Verlag Chem., 127 Hauptstrasse Weinheim Bergstrasse.

(4) "A method of coding chemicals for correlation and classification," Chem.-Biological Coordination Center, National Research Council, Washington, D. C., 1950.

(5) M. M. Berry and J. W. Perry, Notational System for structural formulas *Chem. Eng. News*, p. 407, Feb., 1952.

(6) W. J. Wiswesser "The Wiswesser line formula notation," *Chem. Eng. News*, p. 3523, Aug., 1952.

(7) W. J. Wiswesser, "A line-formula chemical notation," W. Y. Crowell Co., New York, N. Y., 1954.

(8) E. T. Crane and M. M. Berry, "Which notation? (The composite notation system for molecular structural formulas)," *Chem. Eng. News*, **33**, 2842 (1955).

(9) Ascher Opler and Ted R. Norton, "New speed to structural searches," *Chem. Eng. News*, p. 2812, June 4, 1956.

(10) Ascher Opler "Dow Refines Structural Searching" *Chem. Eng. News*, p. 92, Aug. 19, 1957.

(11) W. H. T. Davison and M. Gordon, Sorting for chemical groups using Gordon-Kendall-Davison ciphers *Am. Doc.*, **VIII**, 202 (1957).

(12) W. H. Waldo, R. S. Gordon and J. D. Porter, "Routine report writing by computer," *Am. Doc.*, **9**, (1), 28 (1958).

(13) W. H. Waldo and M. DeBacker, "Preprints of the International Conference on Scientific Information," Washington, D. C., Nov. 1958, Area 4, p. 49-68.

(14) Ascher Opler and Norma Baird "Display of chemical structural formulas as digital computer output," *Am. Doc.*, p. 59, Jan., 1959.

(15) G. M. Dyson and E. F. Riley, Mechanical Storage and Retrieval of Organic Chemical Data," *Chem. Eng. News*, p. 72, April 17, 1961.

(16) G. M. Dyson and E. F. Riley, "Mechanical Storage and Retrieval of Organic Chemical Data," *Chem. Eng. News*, pp. 74-80, Nov. 20, 1961.

(17) Int. Union of Pure and Appl. Chemistry, "Rules for I.U.P.A.C. notation for organic compounds," Longmans, Green and Co. Ltd., London, pp. 1-107.

(18) W. H. Waldo "Searching two dimensional structures by computer," *J. Chem. Doc.*, **2**, 1 (1962).

(19) H. T. Bonnett and D. W. Calhour, Application of a line formula notation in an index of chemical structures," *J. Chem. Doc.*, **2**, 2 (1962).

(20) A. Gelberg, W. Nelson, G. S. Yee and E. A. Metcalf, "A program retrieval of organic structure information via punched cards," *J. Chem. Doc.*, **2**, p. 7 (1962).

(21) E. Meyer and K. Wenke, "Ein System zur topologischen Verschlüsselung organischchemischer Strukturformeln für die mechanisierte Dokumentation," *Nachr. Dok.*, p. 13, März, 1962.