as the random function in distributing structures evenly in the file. This shows that there are many possible ways of hashing the key which give equally good results, as long as all bits in the key are involved. We described the theoretically perfect hash function as a standard against which other hash functions could be compared. We also defined a practical definition of work required for retrieval which facilitates comparison of results on the basis of expected performance. These techniques have been implemented in the SECS program for computer-assisted design of organic synthesis.

ACKNOWLEDGMENT

This work was supported by a National Cancer Institute Contract N01-CP-75816 and by an allocation of the SUMEX-AIM resource at Stanford (RR00785, J. Lederberg, principal investigator).

LITERATURE CITED

(1) For a review of chemical notation systems see M. F. Lynch, J. M. Harrison, W. G. Town, and J. E. Ash, "Computer Handling of Chemical Structure Information", American Elsevier, New York, N.Y., 1971.
(2) W. T. Wipke, "Computer-Assisted Three-Dimensional Synthetic Analysis", in "Computer Representation and Manipulation of Chemical Information", W. T. Wipke, S. R. Heller, R. J. Feldmann, and E. Hyde Ed., Wiley, New York, N.Y., 1974, pp 147–174.
(3) W. T. Wipke, H. Braun, G. Smith, F. Choplin, and W. Sieber, "SECS—Simulation and Evaluation of Chemical Synthesis: Strategy and Planning", in "Computer-Assisted Organic Synthesis", W. T. Wipke and W. J. Howe, Ed., *ACS Symposium Series*, 1977, pp 98–129; W. T. Wipke, G. I. Ouchi, and S. Krishnan, "Simulation and Evaluation of Organic Synthesis—SECS. An Application of Artificial Intelligence Techniques", *Artificial Intelligence*, in press.
(4) W. T. Wipke, G. I. Ouchi, and S. Krishnan, in preparation.
(5) (a) W. D. Maurer and T. G. Lewis, "Hash Table Methods", *Comput. Surveys*, **7**, 5–19 (1975); (b) D. E. Knuth, "Sorting and Searching", *Art Comput. Programming*, **3**, 506–549 (1973); (c) D. E. Knuth, "Algorithms", *Sci. Am.*, **236**, 63–80 (April 1977).
(6) R. J. Feldmann, "Interactive Graphic Chemical Structure Searching", in "Computer Representation and Manipulation of Chemical Information", W. T. Wipke, S. R. Heller, R. J. Feldmann, and E. Hyde, Ed., Wiley, New York, N.Y., 1974, pp 55–81.
(7) W. T. Wipke and T. M. Dyott, *J. Am. Chem. Soc.*, **96**, 4825, 4834 (1974).
(8) J. A. Katzenellenbogen, "Insect Pheromone Syntheses: New Methodology", *Science*, **194**, 139–148 (1976).
(9) M. Suzuki, E. Kurosawa, and T. Irie, "Three New Sesquiterpenoids Containing Bromine, Minor Constituents of *Laurencia glandulifera* Kützing, *Tetrahedron Lett.*, 821–24 (1974).
(10) Note that the theoretical efficiency of the random function varies as a function of L, the table loading factor, and is complicated because of the discrete nature of the perfect function (for example, the first derivative of $W_{perfect}$ as a function of L is discontinuous). Therefore, the $\bar{W}$ is better for comparing real hash functions at different load factors.

A New Linear Representation of Chemical Structures

LUCIANA QUADRELLI*
Montedison—Istituto Ricerche "G. Donegani", Novara, Italy

VITTORIO BAREGGI
Montedison, Milano, Italy

SERGIO SPIGA
Montedison—Centro Ricerche Antiparassitari, Milano, Italy

Received November 8, 1976

A system of recording, analyzing, and displaying chemical structures by a new linear notation system is described.

I. INTRODUCTION

The problem of filing, reading, and analyzing structural formulas by using a computer in a simple and fast way is considered important in any fine chemical research center where relationships between structures of many new chemical substances and their physical-chemical or/and biological properties are studied. This is considered one of the most important problems in the Pesticide Research Center of Montedison's Agriculture Division. This is due both to the need to control the heavy quantity of data that are being collected and to correlate statistically the properties of a substance and its chemical structure. Therefore we designed a new method of chemical structure representation which we call the "linearization method" because the structural formula (two dimensional) takes the form of a linear sequence of characters, as in other methods (Dyson, Wiswesser). This linearization method, is more simple and economic than the older ones.

In the following we shall present the constraints and the criteria used in the linearization method (section II), the technique to represent a chemical structure on a special format (section III), the rules selected in order to write the special format (section IV), a special set of rules to rapidly characterize ring structures (section V), and finally the techniques adopted to determine substructure in a set of structures.

II. LINEARIZATION METHOD

A. Our constraints for the method were: the linearized formula has to maintain all the information shown by the structural formula; the linearized formula has to hold letters normally used in the normal input/output units of computers.

B. The following principles were established. Every formula may be written on a grid with small and insignificant deformations as compared to the classical representation. Each atom or atomic group of the structural formula must be on a junction. Bonds must be disposed on the sides; the formula so arranged can be written line after line in a linear succession of character, with respect, for every line, to the following conventions: (a) all the elements (atom or atomic group, empty junctions, and horizontal bonds) must be listed sequentially, from left to right, using the appropriate coding, row after row; (b) the possible vertical bonds starting from any atom must be expressed in a way that permits one to single out the above starting points; (c) a special character permits one to express in the linearization the end of a row condition of the grid without the necessity to start a new coding line.

0095-2338/78/1618-0037$01.00/0

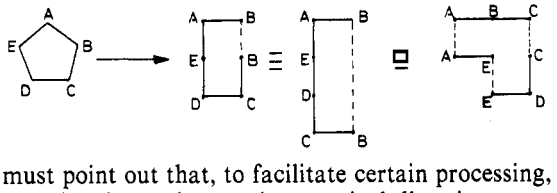
Table I. Decode Table

	Character	Decode		
	*			
	A	C	H	3
(+)	B	C	H	2
(+)	C	–	C	H
(+)	D	–	C	–
	E	–	C	L
	F	–	F	–
	G	–	B	R
	H	–	J	–
	K	–	C	O
	J	N	O	2
	L	–	C	S
	M	–	N	H
	N	–	N	–
	O	–	S	H
	P	–	P	–
	Q	–	O	–
	R	–	O	H
	S	–	S	–
(‡)	T	C	H	2
	U	–	S	O
(‡)	V	–	C	H
	W	N	H	2
	X	C	H	O
	Y	–	C	N
(+)	Z	–	C	–
	I Ø	–	B	A
	I 1	–	C	A
	I 2	–	F	E
	I 3	–	M	N
	I 4	–	N	1
	I 5	–	K	
	I 6	–	C	U
	I 7	–	S	N
	I 8	–	N	A
	I 9	–	Z	N

(+) Open chain
(+ +) Ciclic order

III. GRID REPRESENTATION

The grid is dimensioned on 12 lines and 30 columns. The bonds have only horizontal and vertical directions. This causes "small and insignificant deformations" because the ring structures are represented by geometrical figures with horizontal and vertical sides. When the ring has an odd number of sides it is necessary to add fictitious links to enclose the figure. This figure can afterwards change its appearance and adapt to the more or less complex structure; for instance, a ring of five atoms can be represented in different ways:



We must point out that, to facilitate certain processing, the fictitious bonds can have only a vertical direction.

IV. LINEARIZATION STANDARDS

After placing the structural formula on the grid, in conformity with the above criteria, linearization is carried out according to the following standards:

Each atom or atomic group is represented by one or two alphanumeric letters (see Table I).

Each empty junction is represented by a "*" (it is advisable to neglect all empty knots which will be found after the last atom (or atomic group).

Each horizontal bond is represented by a special character placed between the two concerned atomic groups. The special characters are: a hyphen "-" for a single bond; a full stop "." for a double bond; a colon ":" for a triple bond.

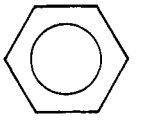
Each vertical bond is represented by a special character which precedes the atom or atomic group, from which the bond originates downwards; the special characters are: a symbol "¢" for a single true bond; a question mark "?" for a double true bond; a symbol "%" for a triple true bond; a symbol "#" for a fictitious bond. Note that the last symbol can precede the symbol "*" representing an empty bond (see example below).

Each end of a line is represented by the symbol "&" following the last character in the linearized description of the same.

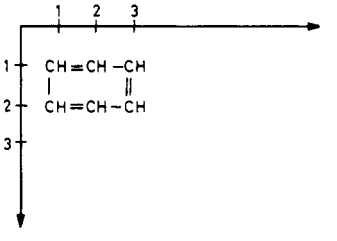
Each end of a formula is represented by the symbol "@" which replaces the "&" of the last line.

Every lack of horizontal bond between any two elements, including the empty ones, which occupy consecutive junctions of a line, is represented by an exclamation mark "!" put between them.

Example: Benzene



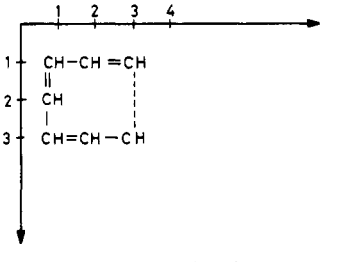
The usual description on the grid is:



and the corresponding linearization is:

¢ V.V-?V&V.V-V@

a different description on the grid, using the fictitious bond, can be:



and the corresponding linearization is:

?V-V.#V&¢ V!*!#*&V.V-V@

If an atom or atomic group is repeated consecutively *n* times on the same line, and if along this chain will be found only simple horizontal bonds (also lacking vertical bonds of any kind), it can be represented only once if the symbol of the atom or atomic group is immediately preceded by the number *n*. If the mentioned conditions are not respected, the symbol representing this atom or atomic group has to be repeated as many times as necessary. The number *n* should be made up by only one figure. For instance, a chain of 12 atoms or atomic groups, can be resolved in two chains (e.g., 6 + 6 or 9 + 3, etc.). The same thing can be said if on the same line an empty junction is consecutively repeated *n* times.

An example referring to the linearization of the structure formula of auroxanthin (Table II) is appended in Table III where all the explained standards find an application. The special and alphanumeric characters which are used represent a possible choice among the various symbols given by the IBM

Table II. Auroxanthin: Empirical Formula and Structural Formula $C_{40}H_{56}O_4$

$C_{40}H_{56}O_4$

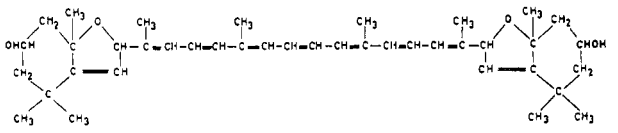
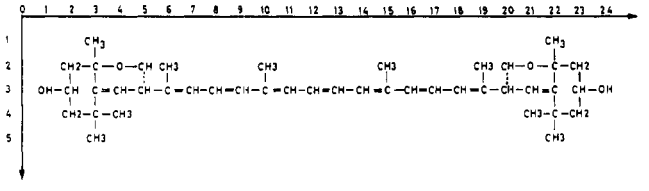


Table III. Auroxanthin: Lattice Formula and Linear Formula



2*!¢A!9*!9*!¢A&*!¢T-¢Z-Q-#V!¢A!3*!¢A!4*!¢A!3*!¢A!
#V-Q-¢Z-¢T&R-¢V!¢Z.V-V-D.C-C.C-D.C-C.C-C.D-C.C-C.
D-V-V.¢Z!¢V-R&*!T-¢Z-A!9*!7*!A-¢Z-T&2*!A!9*!9*!
A@

input/output units, with special regard to the already stated principles of economy and simplicity of application. It is also possible to use a completely different codification, with respect to the above-described rules. The linearization of a structural formula can be carried out manually by arranging the formula on the grid and applying all the described standards.

The linearization can also be carried out automatically by arranging the formula on the grid and supplying it on a card to the linearization program. The linearization program applies to the actual linearization as well as to the decodification.

There is also a program "Print Formula" which is able, by reading a linearized formula, to print the corresponding structural formula. The print retains the distortion the formula has undergone when it was disposed on the grid. The structural formula can also be written in its classical format with greater but also more expensive sophistication.

V. RAPID SEARCH OF RING SUBSTRUCTURES

Ring structures are often meaningful for correlations between properties and structures. It was decided to ease their search by registering on the side of each linearized formula an appropriate code, showing the ring structure and its main characteristics.

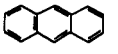
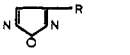
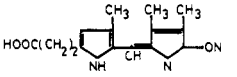
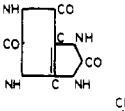
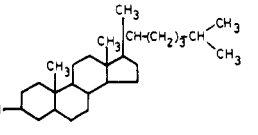
The alphanumeric code is conceived (beginning from the left) as follows: the characters 1° and 2° (alphanumeric and compulsory) indicate the number of elementary cycles which compose a structure (the range of the number is between 01 and 99); the characters 3° and 4° (numerical and compulsory) indicate the number of peripheral elements of the structure (the range can be between 01 and 99); the subsequent characters are alphanumeric and optional.

If at least one part of the structure is aromatic, an "X" must be written in 5th position. If the structure contains one or more atoms different from C, two alphanumeric characters must be written for each of them. The first represents the type of the atom and the second represents the quantity. For each cyclic structure one code, as described above, is written within parentheses (Table IV).

VI. DEFINITION, IMPORTANCE, AND MODALITY FOR SEARCH OF A SUBSTRUCTURE

A substructure is defined as one or more atomic groups at a certain distance from one another. This distance is measured

Table IV. Ring Substructures Coding

Substructure	Notation
	(Ø1Ø6)
	(Ø1Ø6X)
	(Ø21ØX)
	(Ø2Ø9XO1)
	(Ø314X)
	(Ø1Ø5N2Ø1)
	(Ø1Ø5XN1)(Ø1Ø5XN1)
	(Ø2Ø9N4)
	(Ø417)

by the number of bonds along the shortest way which unites them.

If we consider the previously described problem of the correlations between biological, chemical, and physical properties of a substance and its chemical structure, it is essential that the computer be able, by enquiring the database, both to select the substances having the same functional groups and to measure the distance between them. This is complex problem of "information retrieval" that can be logically divided into: (a) presence control of the searched for atomic groups; (b) measure of their distance.

The first phase is carried out through the usual search by keyword and uses a language which is already in the trade and which is also particularly suited for these kinds of searches, the RAMIS (Rapid Access Management Information Systems); a linearized formula can, in fact, be considered as a normal phrase where this or that word can be searched. It is possible to choose, among all the recorded formulas, those which contain whatever combinations of atomic groups by appropriate questions and taking advantage of the classic logic operators AND and OR.

The second phase is worked out by using an originally developed system which is based on the observations that: in a formula, generally, there are several paths which connect one atomic group to another; it is impossible to establish, a priori, which one of these paths is the shortest and along which path the distance has to be calculated.

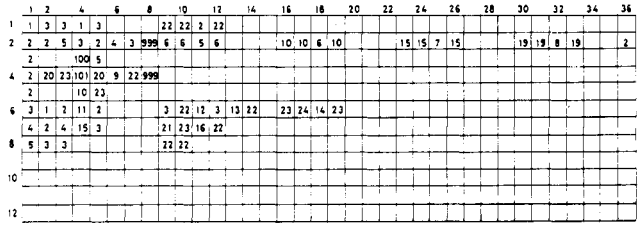
If we refer to the formula's representation in the lattice, we can define "piece" as the ordered succession of atomic groups written on the same row and bound to one another by horizontal links (simple or multiple). One example in the second row of Table III is

$$CH_2-\overset{|}{\underset{|}{C}}-O-CH$$

We define "coordinates of a piece" as the coordinates of its extreme atoms.

The coordinates of the above first example are (2,2), (5,2). We define coordinates of a vertical bond as the coordinates of the atom from which the bond starts toward the bottom of the sheet.

Table V. Auroxanthin: Distance Calculation Matrix



To help the calculation of distance between two atomic groups a table was created that contains for every row the coordinates of its piece and, for every one of them, the coordinates of its vertical bonds. The logic used to compute the distance between two atomic groups is as follows: (i) one looks for their coordinates on the lattice of the formula's representation; (ii) one find in the table the piece that contains one of them; (iii) if also the other atomic group appears on the same piece, the distance is simply computed as a difference between their abscissas; (iv) if they appear on different pieces, one starts from one of them and simulates all the possible paths moving along the vertical and horizontal bonds; (v) one eliminates through appropriate checks both the closed paths and the longer ones and, finally, finds out the shortest possible way.

The really innovative part of this system is formed by the table, built up and memorized definitively when the formula is recorded; this table represents a second and more simplified linearization (see Table V). The table looks like a matrix, but during this phase of automatic generation this is memorized in a compacted way as a vector with the aim to avoid any waste of space. This second representation does not contain the chemical definition of the various groups but describes only the whole structure of the formula. The small quantity of processing necessary for the setting up this schedule permits one to eliminate the work necessary to carry out the distance calculation directly on the linearized formula.

Comparison with Other Methods. From a comparison with other linearization methods, we notice that this method of chemical structure presents the greatest number of affinities with Horowitz and Crone's Hecsagon system.[12] The fundamental value of this method consists of its extreme simplicity and in the limited number of syntactic rules on which it is based. Therefore, when one executes "manually" the linearization, one need not be trained to carry it out, unlike other systems such as WLN[6] and IUPAC.[16]

The main difference with the above-mentioned methods lies in the fact that our method reflects the structure as it is sketched without hierarchical characteristics. Moreover, our notation appears more flexible than others. As to the length of the notation, this system requires almost the same as the linearization systems of Hiz[13] and Eisman,[14] which also describe a molecule through an absolutely linear representation, atom by atom, and a length which is less than that of the topologic representations having the form of a matrix.

An important advantage is represented by the fact that our linearization permits the search of substructures on a large file of compounds. A disadvantage of the same is represented by the fact that the notation is not univocal and therefore, when using this linearization, it is not possible to match directly a complete compound in a file with a mask. Anyway it is possible to identify a compound by using a set of structural screens. On the other hand, univocal notation was not felt as fundamental when what is asked for is the description of the connection between the atoms of a molecule in a complete, brief, and nonobscure way.

REFERENCES AND NOTES

(1) D. I. Gluck, "A Chemical Structure, Storage and Search System Designed at Du Pont", *J. Chem. Doc.* **5**, 43–51 (1965).
(2) J. F. Feeman, "A Novel Organizational Code for Organic Structures Based on Functional Groups", *J. Chem. Doc.*, **6**, 184 (1966).
(3) E. Meyer, "The IDC System for Chemical Documentation", *J. Chem. Doc.* **9**, 109 (1969).
(4) G. Palmer, "Wiswesser Line-Formula Notation", *Chem. Brit.* **6**, No. 10 (Oct 1970).
(5) D. P. Jacobus et al., "Experience with the Mechanized Chemical and Biological Information Retrieval System", *J. Chem. Doc.*, **10**, 135–140 (1970).
(6) R. J. Feldmann and D. A. Koniver, "Interactive Searching of Chemical Files and Structural Diagram Generation from Wiswesser Line Notation", *J. Chem. Doc.* **11**, 154 (1971).
(7) C. E. Granito, G. T. Becker, S. Roberts, W. J. Wiswesser, and K. J. Windlinx, "Computer-Generated Substructure Codes (Bit Screens)", *J. Chem. Doc.* **11**, 106 (1971).
(8) S. J. Tauber and K. Rankin, "Valid Structure Diagrams and Chemical Gibberish", *J. Chem. Doc.* **12**, No. 1 (1972).
(9) J. Figueras, "Substructures Search by Set Reduction", *J. Chem. Doc.*, **12**, No. 4 (1972).
(10) B. E. Holm, M. G. Howell, H. E. Kennedy, J. H. Kuney, and J. E. Rush, "The Status of Chemical Information", *J. Chem. Doc.*, **13**, No. 4 (1973).
(11) D. R. Eakin, E. Hyde, and G. Palmer. "The Use of Computers with Chemical Structural Information: ICI CROSSBOW System", *Pestic. Sci.*, **5**, 319–326 (1974).
(12) P. Horowitz and E. M. Crone, "Hecsagon: A System for Computer Storage and Retrieval Chemical Structure", Eastman-Kodak Co., Rochester, N.Y., 1961, 33 pp.
(13) H. Hiz, *J. Chem. Doc.*, **4**, 135 (1964).
(14) J. E. Dubois and H. Veillard, Laboratoire de Chimie Organique de la Faculté des Sciences de Pari, "Theorie de generation – description – Systhême DARC", 1967.
(15) Mathematica Inc., Princeton Station Office Park, P.O. Box 2392, Princeton, N.J.
(16) IUPAC Nomenclature of Organic Chemistry, Butterworths, London, 1969.