transition, and inner-transition elements, respectively, and the NGRP value 00 for any element group. Thus classes of elements can readily be given numerical identifiers, *e.g.*

NATM	
1100	representative element
0100	transition element
−0900	inner-transition element
1700	second row representative element (Li-Ne)
0500	first row transition element (Sc-Zn)
−0700	lanthanide (Ce-Lu)
−0800	actinide (Th-Lw)
1108	halogen (F-At)
0102	group Ib element (Cu-Au)

and so on for other similar classes of elements.

EXPERIMENTAL

Computer programs were run in Fortran IV on a Burroughs 3500 computer (XFORTS complier). A subroutine LEMENT[1] has been written that will supply the NROW and NGRP values for any one- or two-letter element symbol argument (H to Lw). The subroutine matches letters in descending order of probability of occurrence so as to match the argument in the minimum time.

A frequency of occurrence table for chemical element symbols was generated by randomly sampling an index of molecular formulas[2] (index page numbers were generated with a random number generator GRBG). When LEMENT was applied to the sample element list (53 elements from 617 compounds), a frequency-weighted average of only 3.73 character-tests was required for identification of any element symbol.

ACKNOWLEDGMENTS

I wish to thank Lawrence Scacciaferro for practical assistance in this work. The work was supported by a Faculty Research Fellowship and Grant-in-Aid from the Research Foundation of the State University of New York.

LITERATURE CITED

(1) A listing of LEMENT subroutine is available from the author.
(2) Chemical Abstracts 8th Collective Formula Index, 1967–1971.

The Connectivity Stack, a New Format for Representation of Organic Chemical Structures

YOSHIHIRO KUDO

JEOL, Toyko, Japan

SHIN-ICHI SASAKI*

Miyagi University of Education, Sendai, Japan

Received September 16, 1974

A new type of format to represent the enumerated structural formula consistent with given information is discussed.

A system for the automated structure elucidation of organic compounds,[1] CHEMICS,[2] has been developed by the authors. One of the functions involved in the system is the enumeration[3] of all possible chemical structures (the informational homolog) which are consistent with given structural information. Although several studies[4,5] concerning the enumerations of the structures have already been published, the method developed by the authors is more widely applicable and more completely guaranteed compared to others. In this paper a new type of format to represent the enumerated structural formula is discussed.

From a practical point of view the format to be adopted in CHEMICS has to contain "dynamic" as well as "static" features. That is, it has to make execution time as short as possible and at the same time ensure complete and unique enumeration. Enumeration methods, such as in the heuristic DENDRAL,[4] and the one by Balaban,[5] are elegant but fix the limits of their scopes for application in terms of types of structures. On the other hand, description methods, such as IUPAC/Dyson notation,[6] CA/Morgan notation,[7] and Wiswesser line notation,[8] treat all types of compounds, but have great regard for unambiguous and unique representation of individual structures.

The connectivity stack** is a sequence of elements of the connectivity matrix. When the elements of the matrix are a_{ij}'s, the stack $a_1, a_2, a_3, \ldots, a_k, \ldots$, uniquely corresponds to one of the matrices after establishment of a correspondence rule. Here, the function of correspondence rules will be explained in, for practical reasons, two cases:

Case 1

$$k = N(i - 1) - i(i + 1)/2 + j \qquad (i < j)$$

(N stands for the order of the matrix)

Case 2

$$k = i + (j - 2)(j - 1)/2 \qquad (i < j)$$

Usually sets of equivalent stacks are possible to represent one structure. For example, in the case of propane, three, 101, 011, and 110, corresponding to three numberings, 1–2–3, 1–3–2, and 2–1–3, respectively, which are numbered to the carbon atoms, exist. In the case of propane, the same

* Author to whom correspondence should be addressed.

** In general, the word "stack" means a set of data of which the elements are sequentially arranged (in contrast with, *e.g.*, a tree arrangement), and where conversion always takes place at the end of the set (this in contrast with, *e.g.*, queue).

Table I. All Stacks of Isomeric Butanes[a]

Stack	Content	Case 1	Case 2	Stack	Content	Case 1	Case 2
S 1	111000	1	2	S11	011100	3	3
S 2	110100	2	1	S12	011010	3	3
S 3	110010	3	3	S13	011001	2	1
S 4	110001	3	3	S14	010110	3	3
S 5	101100	3	3	S15	010101	1	2
S 6	101010	2	1	S16	010011	3	3
S 7	101001	3	3	S17	001110	3	3
S 8	100110	1	2	S18	001101	3	3
S 9	100101	3	3	S19	001011	1	2
S10	100011	3	3	S20	000111	2	1

[a] Numerals 1, 2, and 3 in the columns of cases 1 and 2 express isobutane, cyclopropane + methane, and *n*-butane, respectively. The connectivity between atoms *i* and *j* is represented by the *k*th numeral of the stack decimal, where *k* is obtained *via* the correspondence rule (which see in text). If the numeral is 1, there is a connection between *i* and *j*, and if 0, there is no connection.

Table II. Number of the Structural Isomers of C_9H_{10}

	C	CH	CH_2	CH_3	Number
1	5	1	0	3	437
2	5	0	2	2	812
3	4	2	1	2	4,628
4	4	1	3	1	4,429
5	4	0	5	0	429
6	3	4	0	2	2,165
7	3	3	2	1	10,828
8	3	2	4	0	4,416
9	2	5	1	1	4,418
10	2	4	3	0	5,575
11	1	7	0	1	343
12	1	6	2	0	1,554
13	0	8	1	0	106
				Total	40,140

stacks are obtained in cases 1 and 2 by their respective correspondence rules. On the contrary, different stacks are derived by the rules for isomeric butanes. That is, one stack has a different meaning depending upon the correspondence rule as shown in Table I. Twenty stacks are divided into three groups, the first corresponding to isobutane, the second to cyclopropane + methane, and the third to *n*-butane. For isobutane, case 1 affords S1, S8, S15, and S19, and case 2 S2, S6, S13, and S20. Similarly, case 1 gives S2, S6, S13, and S20, and case 2 gives S1, S8, S15, and S19 for cyclopropane + methane. The rest, which are the same in both cases, correspond with *n*-butane.

In order to enumerate the informational homolog completely and uniquely, only one stack should be selected from equivalent stacks *via* a canonicalization rule. One rule selects the canonical stack as follows: a stack is compared with a decimal number and a_k of the stack is a number at the *k* place of the decimal. Then the stack with the greatest decimal value is adopted as the canonical one among equivalent stacks. The canonicalization rule leaves S1, S2, and S3 as the canonical stacks for butanes in both cases 1 and 2. Furthermore, another rule is required to eliminate nonconnective structures. A cluster of components which consists of atomic groups necessary to construct any organic structure is fixed in the system CHEMICS. The simplest component set consists of the elements themselves, and a higher hierarchical set comprises CH_3, CH_2, CH, C, OH, and O for the compounds containing C, H, and O. The enumeration carried out by the connectivity stack is adapted to compounds with heteroatoms or compounds described in terms of components. Table II gives the numbers of members of the informational homolog of C_9H_{10}. The numbers are not calculated directly from the molecular formula but by using the component set (CH_3, CH_2, CH, and C). The

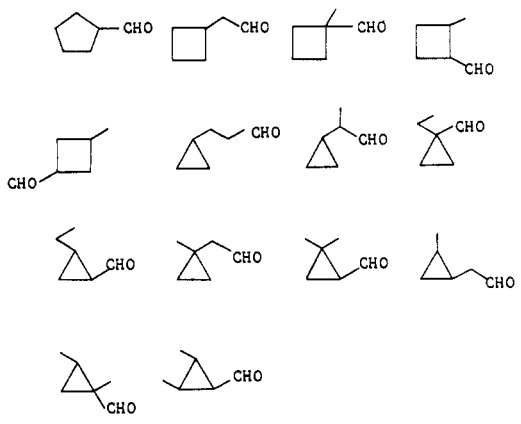

Figure 1. Fourteen cyclic structures of $C_6H_{10}O$ with a formyl group.

Table III. Acyclic and Cyclic Structures of $C_6H_{10}O$ and $C_7H_{12}O$ with a Carbonyl Group[a]

Molecular formula		$C_6H_{10}O$ Cyclic	$C_6H_{10}O$ Acyclic	$C_7H_{12}O$ Cyclic	$C_7H_{12}O$ Acyclic
Ketones	C	19	13	58	39
	D	19	13	57	40
Aldehydes	C	14	21	44	56
	D	13	21	47	56

[a] C and D express the results afforded by the CHEMICS and the DENDRAL, respectively.

Table IV. Influence of the Hierarchical Orders between Atoms (or Components) on the Execution Time

No.	Hierarchical orders				Time, sec[a] (±2 sec) (a) $C_2H_2N_2O_2$	(b) $C_2H_4N_2O_2$
1	C	N	O	H	20	39
2	C	N	H	O	22	59
3	C	O	N	H	25	45
4	C	O	H	N	31	77
5	C	H	N	O	26	83
6	C	H	O	N	35	115
7	N	C	O	H	19	38
8	N	C	H	O	21	56
9	N	O	C	H	29	48
10	N	O	H	C	43	102
11	N	H	C	O	26	118
12	N	H	O	C	50	182
13	O	C	N	H	24	42
14	O	C	H	N	29	70
15	O	N	C	H	29	46
16	O	N	H	C	41	97
17	O	H	C	N	35	242
18	O	H	N	C	50	350
19	H	C	N	O	29	325
20	H	C	O	N	37	410
21	H	N	C	O	29	365
22	H	N	O	C	53	690
23	H	O	C	N	38	550
24	H	O	N	C	54	829

[a] (a) The enumeration of 506 structures of $C_2H_2N_2O_2$ and (b) the enumeration of 807 structures of $C_2H_4N_2O_2$.

combination of five components C, one component CH and three components CH_3 gives 437 structural isomers, and so on.

Examples of enumeration computed by heuristic DENDRAL[9] and CHEMICS[3] are shown in Table III. It is a little curious that the results given by the DENDRAL are not always equal to the authors' results. For example, the DENDRAL enumerates thirteen structures for the cyclic structure of $C_6H_{10}O$ with a formyl group, while the CHEMICS

gives fourteen as shown in Figure 1. The number of isomers calculated by Balaban for C_6H_6 amounts to 217, which is identical with ours.[3] Table IV shows the influence of the hierarchical order of atoms or components on the execution time. The hierarchical order between atoms (or components) is one of the factors determining execution time, because the different orders result in different arrangements of the stacks (Table I) belonging to a certain chemical structure. Proper combination of the connectivity stack and rules for canonicalization, elimination of nonconnective structures, as well as the hierarchy of components, remarkably contributes to the rapid enumeration.

In IUPAC notation, CA notation, and WLN, various types of determinations are performed "globally," whereas in CHEMICS they are done as "locally" as possible. From this point of view, a relation between WLN and CHEMICS is similar to that between Fischer's conventional rule and the Ingold–Cahn–Prelog rule on the absolute configuration in the stereochemistry. CHEMICS' effort for local determination forms its dynamic features.

LITERATURE CITED

(1) Sasaki, S., Kudo, Y., Ochiai, S., and Abe, H., *Mikrochim. Acta*, 726 (1971).

(2) Sasaki, S., Abe, H., Kudo, Y., Ochiai, S., and Ishida, Y., *Kagaku No Ryoiki*, **26**, 981 (1972).

(3) Kudo, Y., and Ochiai, S., *Bunseki-kiki* (Analytical Instruments), **11**, 654 (1973).

(4) Lederberg, J., Sutherland, G. L., Buchanan, B. G., Feigenbaum, E. A., Robertson, A. V., Duffield; A. M., and Djerassi, C., *J. Amer. Chem. Soc.*, **91**, 2973 (1969).

(5) Balaban, A. T., *Rev. Roum. Chim.*, **18**, 635 (1973).

(6) "Rules for IUPAC Notation for Organic Compounds," Longmans, Green and Co., Ltd., London, 1961.

(7) Morgan, H. L., *J. Chem. Doc.*, **5**, 107 (1965).

(8) Smith, E. G., "The Wiswesser Line Formula Chemical Notation," McGraw-Hill, New York, N. Y., 1968.

(9) Sheikh, Y. M., Buchs, A., Delfino, A. B., Schroll, G., Duffield, A. M., Djerassi, C., Buchanan, B. G., Sutherland, G. L., Feigenbaum, E. A., and Lederberg, J., *Org. Mass Spectrom.*, **4**, 493 (1970).