(9) Sharaf, M. H.; Kowalski, B. R. "Quantitative Resolution of Fused Chromatographic Peaks in Gas Chromatography/Mass Spectrometry". *Anal. Chem.* **1982**, *54*, 1291-1296.
(10) Malinowski, E. R.; McCue, M. "Qualitative and Quantitative Determination of Suspected Components in Mixtures by Target Transformation Factor Analysis of Their Mass Spectra". *Anal. Chem.* **1977**, *49*, 284-287.
(11) McCue, M.; Malinowski, E. R. "Target Factor Analysis of the Ultraviolet Spectra of Unresolved Liquid Chromatographic Fractions". *Anal. Chem.* **1983**, *37*, 463-469.
(12) Kalivas, J. H. "Precision and Stability for the Generalized Standard Addition Method". *Anal. Chem.* **1983**, *55*, 565-567.
(13) Malinowski, E. R. "Determination of the Number of Factors and the Experimental Error in a Data Matrix". *Anal. Chem.* **1977**, *49*, 612-617.
(14) Malinowski, E. R.; Howery, D. G. In "Factor Analysis in Chemistry"; Wiley: New York, 1980; pp 50-52.
(15) Malinowski, E. R. "Theory of Error for Target Factor Analysis with Applications to Mass Spectrometry and Nuclear Magnetic Resonance Spectrometry". *Anal. Chim. Acta* **1978**, *103*, 339-354.

A Convenient Notation System for Organic Structure on the Basis of Connectivity Stack

HIDETSUGU ABE, YOSHIHIRO KUDO,† TOHRU YAMASAKI,‡ KAZUO TANAKA,§ MASAHIRO SASAKI, and SHIN-ICHI SASAKI*

Laboratory for Chemical Information Science, Toyohashi University of Technology, Toyohashi, Aichi, Japan 440

Received August 8, 1983

A convenient notation system for organic structures has been developed for the application of the connectivity stack. A notation arbitrarily encoded for a structure by a user through a rather simple procedure using 35 codes, which have been previously prepared, is automatically canonicalized in a computer. The notation given by the user is standardized according to the rules for rearranging the codes into a dictionary order. The connectivity stack is estimated for each of the standard notations and its permuted derivatives. The notation whose stack is the largest amount is decided to be canonical. This notation method will be widely applicable in the field of structure manipulation because of its extreme simplicity.

Several methods for the representation of organic structures have been investigated for computer-aided storage and retrieval of the structures.[1] Linear notations and connection table methods are two major techniques for the topologically unambiguous and unique representation of chemical structures.[2] The connection table descriptions specify all the atoms of a molecule (hydrogen is often suppressed) and may explicitly describe the connectivity of each atom. On the other hand, one of the features of the linear notation method is that chemical structure can be expressed more compactly by the use of letters, numerals, and some symbols. The number of letters, numerals, and symbols used to represent a structure is, in general, much fewer than the number of atoms included in the structure. According to such compactness, the linear notation method seems to be more preferable than the connection table method for compilation of a vast number of structures to be treated in a computer. However, the procedure for canonicalization of a linear notation is generally so tedious and complicated that users hesitate to adopt the methods.

In this paper, we present a new notation system on the basis of a "connectivity stack", which has been published by Y.K. and S.S.[3] The notation system, CANOST (autoCANOnicalization system for organic STructures), has the following features. (1) The notation given arbitrarily by the user through rather simple procedures described later is automatically canonicalized in a computer. (2) Thirty-five symbols expressing atoms, atomic groups, ionic charges, and others as listed in Table I are used to make the arbitrary notation. Two or three hours is normally sufficient to learn how to encode chemical structures for even a beginner in chemistry. (3) Though most of structures are expressed with the 35 items, any other symbols consisting of up to four letters may be added if necessary. (4) The notation can be easily converted into a corresponding connection table. The latter, in some cases, is more usable and convenient than the linear notation for computer-aided manipulation of structures.

Table I. Code of Substructure in CANOST[a]

no.	substructure	code	no.	substructure	code
1	—C≡	T	16	=C=O	VD
2	HC≡	T1	17	—O—	Q
3	=C=	DD	18	—OH	Q1
4		DS	19	=O	QD
			20	—F	LF
5	—CH=	D1	21	—Cl	LC
6	H_2C=	D2	22	—Br	LB
			23	—I	LJ
7		C	24[d]	single bond	SG
			25	cation	+
8		C1	26	anion	−
			27	radical	·
9	—CH_2—	C2	28	chelation	/
10	—CH_3	C3	29	other atom	X
11[a]		Y	30[e]		XR
12[b]		Y1	31		XW
13[c]		YT	32	=X	XD
			33	=X=	XX
			34	≡X	XT
14		V	35	XH_p	XP
15	—CHO	V1			

[a] Aromatic carbon without hydrogen. [b] Aromatic carbon with hydrogen. [c] –C(OH)–C(O)– in troponoid. [d] Prepared for connecting D1 to clearly express conjugated double bond (see Figure 3). [e] Non-carbon atom in aromatic structure.

GENERAL ENCODING PROCEDURES

The following describes how to encode a chemical structure into CANOST notation.

Step 1. Select proper symbols from Table I for the atoms and atomic groups in a structure concerned. If two or more alternative encodings are possible for the structure, the one

† Present address: Faculty of Engineering, Yamagata University, Yonezawa, Yamagata, Japan 992.
‡ Present address: Mitsui Petrochemical Industry Ltd. Co., Iwakuni, Yamaguchi, Japan 740.
§ Present address: Asahi Research Center Co. Ltd., Uchisaiwai-cho, Chiyoda, Tokyo, Japan 100.

0095-2338/84/1624-0212$01.50/0

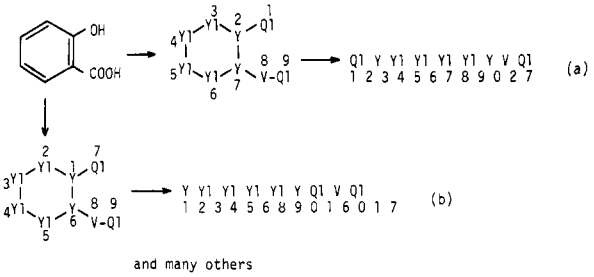


Figure 1. Arbitrary and noncanonical notation of salicyclic acid. Bar and hyphen are used only for graphical representation for user's convenience and should not be confused with anion (−) symbol in Table I or Figure 4. This applies corresponding to the graphical notation in Figures 2 and 3.

consisting of the smallest number of symbols should be selected.

Step 2. Numerals (1, 2, 3, ...) are allocated consecutively to every symbol. Since there is no particular rule for the numbering, it is allowed to start at any symbol.

Step 3. Place all symbols in a line according to the order of the numbers assigned above.

Step 4. To express connections of symbols, arrange the corresponding numerals in a line. There are two rules for arranging the numerals; one is that continuous arrangement of the numerals indicates the connection of the corresponding symbols, and the other is that zero is used when the connection is suspended.

Figure 1 illustrates the process of arbitrary and noncanonical notation of a structure with salicyclic acid as an example. In case (a) in the figure, first, six aromatic carbons, two hydroxyls, and one carbonyl of the structure are replaced by the corresponding symbols, four Y1's, two Y's, two Q1's, and one V, respectively, selected from the symbols listed in Table I. Then, numerals (1, 2, 3, ..., 9) are assigned arbitrarily to all the symbols. Second, the symbols are placed in a line, from left to right. To express connection of the symbols, the numerals are placed as shown in the figure. The connection of Q1, Y, Y1, Y1, Y1, Y1, Y, V, and Q1 is indicated by continuous arrangement of numerals, 1–9. After 9, 0 (zero) comes to indicate that the connection is suspended here. Then, again, 2 and 7 are placed to show two Y's numbered 2 and 7 are connected to each other. Another arbitrary notation (b) is also given in Figure 1, and many other notations, are, of course, possible for the structure, according to arbitrariness of an individual user. Any notation can be transformed into a canonical form along the algorithm mentioned later.

ADDENDUM RULES FOR STEP 1

Although these procedures for noncanonical notation of salicyclic acid are easily applied to all other structures, it will be necessary to add some more general rules in addition to the above-mentioned rules for the notation.

(1) Symbols for the atoms and atomic groups containing hetero elements other than oxygen and halogens are expressed by symbols 29–35 in Table I. Users can make the proper symbols as they like by substituting X with actual elemental symbols N, S, P, or other hetero elements. For example, nitro group and amino group are represented as NW and N2, respectively, by substituting X in symbols 31 and 35 with N. Another examples are shown in Figure 2.

(2) When a structure contains conjugated double bonds, the structure is expressed in either way with or without SG (24) as shown in Figure 3. SG is prepared to help a user out of his confusion of the encoding, especially, in the case of highly conjugated structure, e.g., carotenoid. Since this is a dummy, this is neglected in the process of canonicalization and does not appear in the final canonical notation.

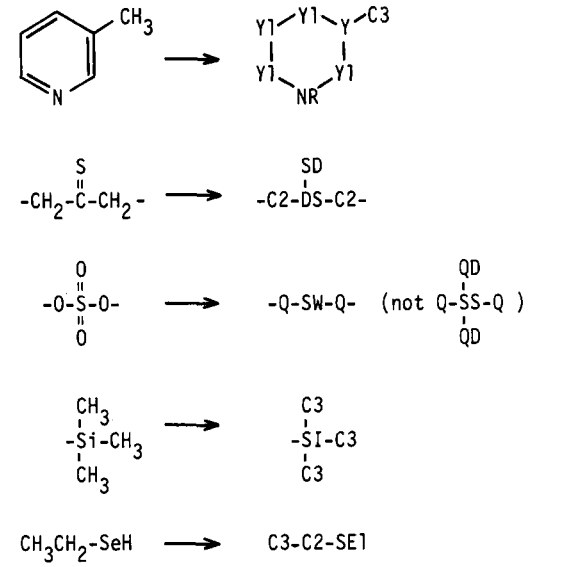


Figure 2. Encoding of structure with heteroatom.

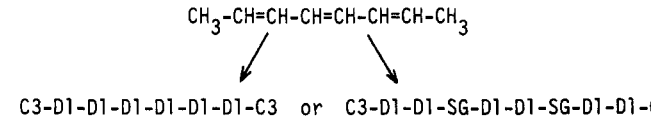


Figure 3. Encoding of structure with conjugation.

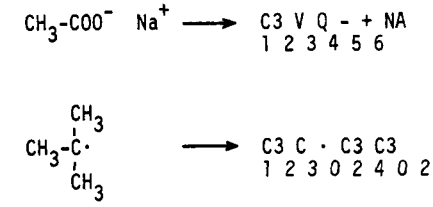


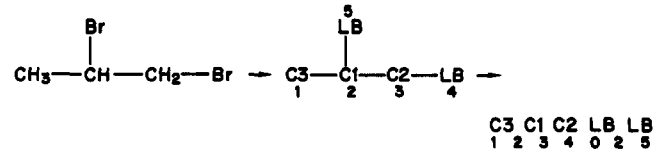
Figure 4. CANOST notation of salt and radical.

(3) Salt and radical are expressed as shown in Figure 4.

ALGORITHM OF CANONICALIZATION

The formal algorithm for the canonicalization has already been published by Y.K. and S.S.[3b] Thus, the routes to reach canonical form from noncanonical form are described with the example 1,2-dibromopropane. Before that, it is still necessary to mention two important principles for the canonicalization. The first is that CANOST codes consisting of a noncanonical notation should be rearranged in the order of Z Y X ... N M L ... C B A 9 8 ... 3 2 1 0 / ● − + □ (□ stands for a space). The second is that the notation whose connectivity stack has the largest binary value among the possible notations for a certain structure expressed by CANOST codes is defined to be canonical form.

Now, let us consider the case of 1,2-dibromopropane, as a simpler example. One of the noncanonical notations of the structure might be expressed by a user by the procedure previously mentioned as



The above notation is rearranged according to the first principle, and simultaneously, the code number is modified by reallocation of numerals 1–5; the result is

LB LB C3 C2 C1
1 2 3 4 5

1,2-Dibromopropane is expressed by two possible connectivities

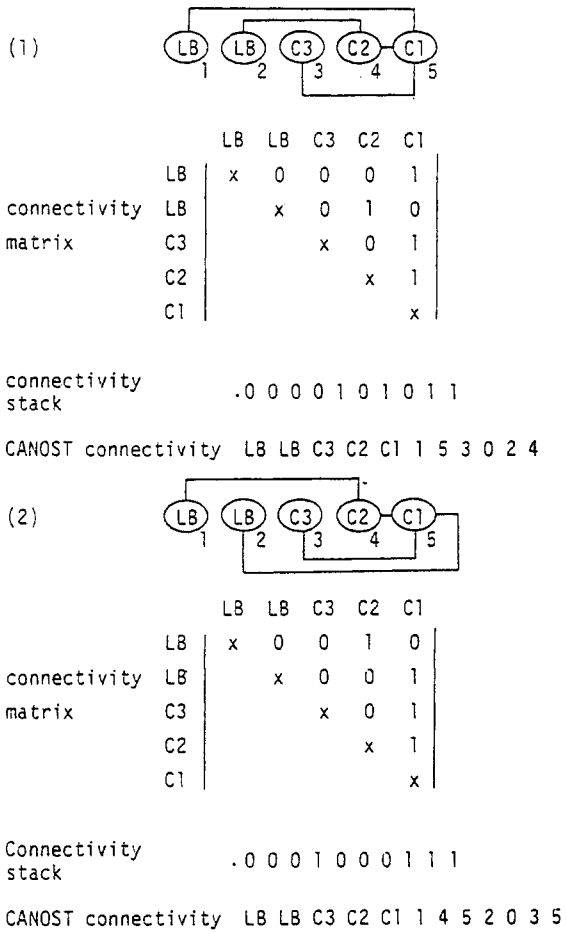
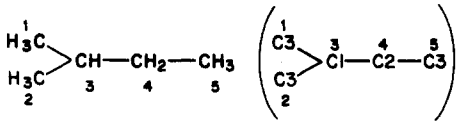


Figure 5. Two potential connectivities of CANOST codes of 2,3-dibromobutane.

of these codes as shown with (1) and (2) in Figure 5. According to the second principle, it may be decided whether (1) or (2) is canonical by comparing their connectivity stacks. As connectivity stack is defined to be a numeral string derived by arranging the off-diagonal upper triangle of a connectivity matrix in row by row manner, .0000101011 and .0001000111 are obtained for the connectivity matrixes of (1) and (2), respectively, as shown in Figure 5. Thus, the latter (2) is decided for the canonical form of the structure, which is expressed as LB LB C3 C2 C1 1 4 5 2 0 3 5. The rule to make a line of numerals, 1 4 5 2 0 3 5, indicating connectivity of codes of (2) is that LB (1) connecting to C2 (4) makes 1 4 and C2 (4) connecting to C1 (5) makes 4 5. C1 (5) connects to LB (2) and C3 (3). In such a case, the smaller numeral, 2, comes after 5. So far, 1 4 5 2 is made, and here, the linear connection is suspended by placing a zero (0) next. C3 (3) and C1 (5) makes 3 5, as the smaller numeral is always first. Thus, 1 4 5 2 0 3 5 is made to express the connection of all the codes, LB, LB, C3, C2, and C1. To have the canonicalization processed in a computer, the codes used to express structure are converted into eight-digit numerals by refering to Table II.

Let us show the processes of canonicalization performed by a computer. The first example is 2-methylbutane. Noncanonical notation, C3 C3 C1 C2 C3 1 3 4 5 0 2 3, originated from



is sent to a computer, where the following things are carried out.

Table II. Numerals Given for Elements of CANOST Codes

symbol	numeral	symbol	numeral	symbol	numeral
□	32	9	57	N	78
+	43	A	65	O	79
−	45	B	66	P	80
.	46	C	67	Q	81
/	47	D	68	R	82
0	48	E	69	S	83
1	49	F	70	T	84
2	50	G	71	U	85
3	51	H	72	V	86
4	52	I	73	W	87
5	53	J	74	X	88
6	54	K	75	Y	89
7	55	L	76	Z	90
8	56	M	77		

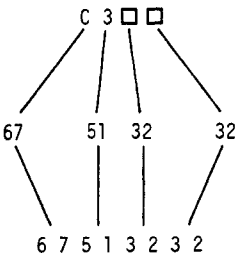


Figure 6. Numeralization of code C3. Two spaces always follow after code, which are replaced by 3 2 3 2.

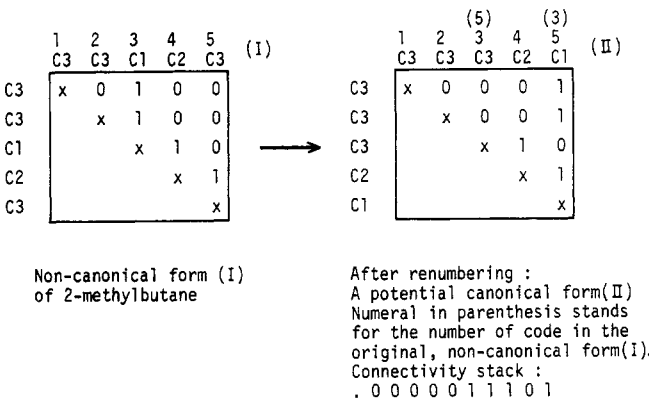
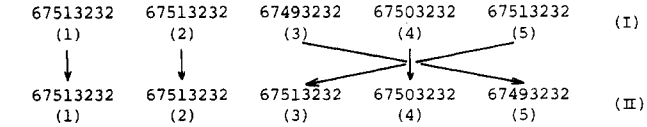


Figure 7. Generation of a possible canonical form by renumbering of code number.

(1) Each code of the input noncanonical notation inputted is converted into the corresponding numeral by applying Table II. For instance, C3 is replaced by 6 7 5 1 3 2 3 2 as shown in Figure 6. Thus, the noncanonical notation of the structure is finally represented like 67513232 67513232 67493232 67503232 67513232 1345023. These numeralized codes are rearranged in descending order in accordance with their amounts:



By this renumbering, the original connectivity matrix based on the connection of I is converted into a new connectivity matrix based on II, which corresponds to one of the potential canonical forms (Figure 7).

(2) The potential canonical form provided by the renumbering is one of the candidate canonical notations and its connectivity stack is .0000011101 as shown in Figure 7. If there are more than one particular CANOST code present in a structure, exhaustive permutation of their order should be done to look for the connectivity matrix with the largest

Table III. Six Possible Permutations of Three C3's That Are Labeled with 1, 2, and 3

	II	IV	V	VI	III	VII
C3	1	1	2	2	3	3
C3	2	3	1	3	1	2
C3	3	2	3	1	2	1
C2	4	4	4	4	4	4

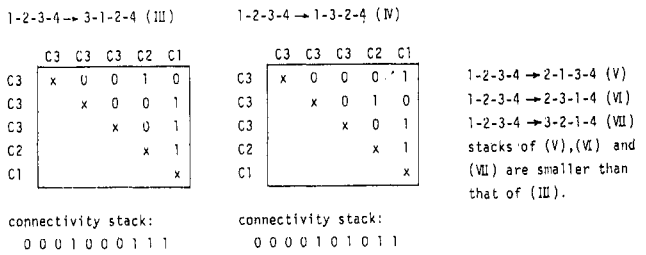


Figure 8. Comparison of connectivity stacks resulted by permutation of three C3's.

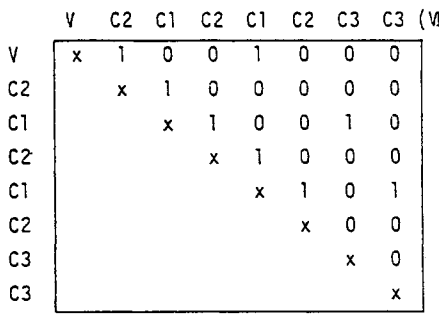


Figure 9. Connectivity matrix of a noncanonical notation of 3,5-dimethylcyclohexanone.

connectivity stack. In this example, three C3's are present; therefore, six operations, 1 2 3 4 (II), 1 2 3 4 → 1 3 2 4 (IV), 1 2 3 4 → 2 1 3 4 (V), 1 2 3 4 → 2 3 1 4 (VI), 1 2 3 4 → 3 1 2 4 (III), and 1 2 3 4 → 3 2 1 4 (VII), are necessary for the complete permutation (Table III).

(3) The permuted results are shown in Figure 8. When the numbers of codes in II (1 2 3 4) are rearranged into 3 1 2 4 (III), the connectivity stack of III is found to be the largest among those of II (Figure 4), IV, V, VI, and VII. Thus, III is decided to be canonical for 2-methylbutane.

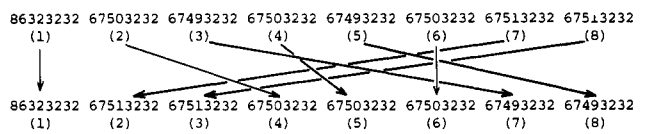
(4) Matrix III is then converted into canonical CANOST connectivity according to the procedure shown in Chart I.

(5) The output is the canonical CANOST notation of 2-methylbutane, C3 C3 C3 C2 C1 1 4 5 2 0 3 5. Example 2 is 3,5-dimethylcyclohexanone. The noncanonical notation given arbitrarily for this structure, V C2 C1 C2 C1 C2 C3 C3 1 2 3 4 5 6 1 0 3 7 0 5 8, is sent to the computer.

(6) All the codes are replaced by eight-digit numerals by referring to Table II: V C2 C1 C2 C1 C2 C3 C3, 86323232 67503232 67493232 67503232 67493232 67503232 67513232 67513232.

(7) The above noncanonical notation given arbitrarily is transformed into the corresponding connectivity matrix (VIII) (Figure 9).

(8) Numeralized codes are rearranged in descending order in accordance with their amounts:



Here, renumbering of codes, steps 2–8, is carried out.

(9) By the above renumbering, the original matrix (V) generated at step 7 is transformed into a new matrix (IX) in

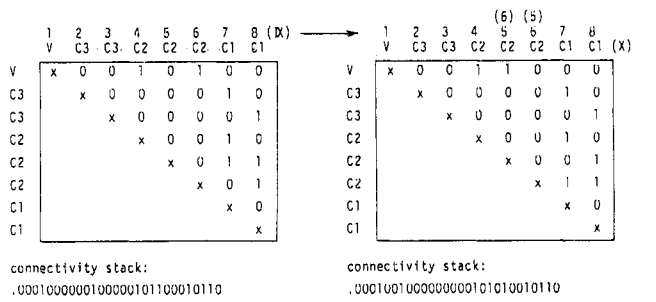


Figure 10. Comparison of the connectivity stacks resulted by permutation of identical nodes. Only two results are shown. Stack of X is the largest among those of 24 connectivity matrixes. Numerals in parentheses indicate the number of code in IX.

Chart I

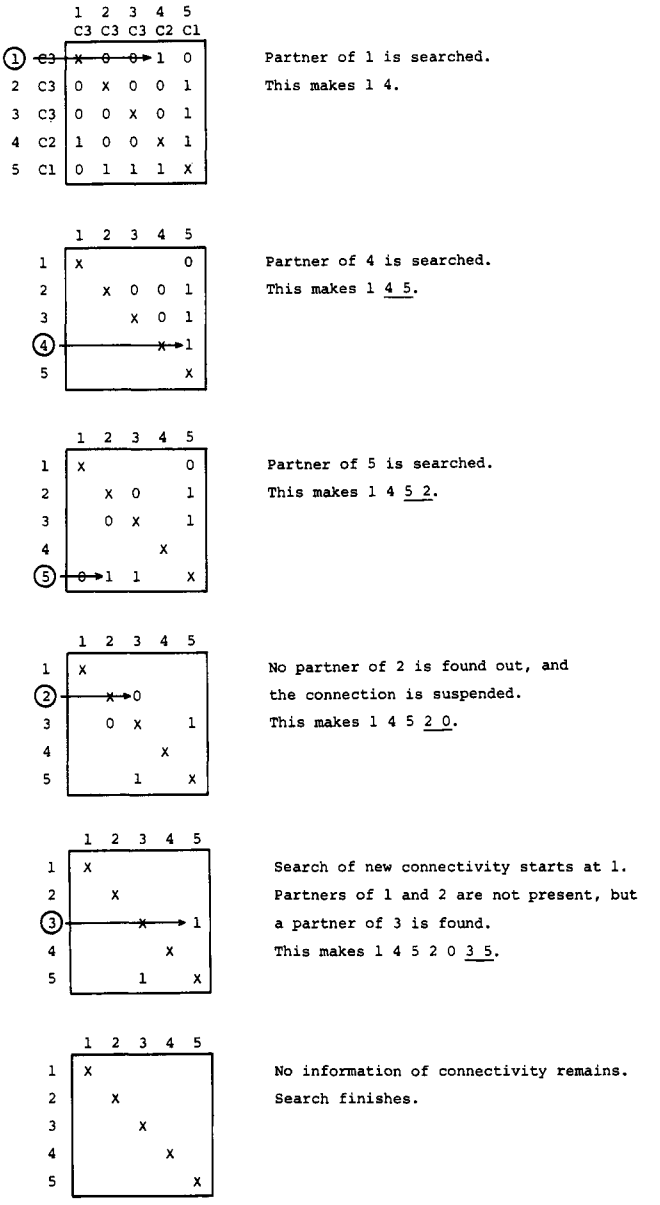


Figure 10 that becomes a standard to look for the canonical notation with the largest connectivity stack.

(10) As similarly in example 1, the same types of codes (two C3's, three C2's, and two C1's in the example) are exhaustively permuted. The computer found that the largest connectivity stack is obtained when the connection of 1 2 3 4 5 6 7 8 of the standard (IX) is rearranged like 1 2 3 4 6 5 7 8 (X) (Figure 10).

(11) V C3 C3 C2 C2 C2 C1 C1 1 4 7 2 0 1 5 8 3 0 6 7 0 6 8 is uniquely presented as the canonical notation of 3,5-dimethylcyclohexanone.

APPLICATION OF THE CANOST NOTATION SYSTEM

The present notation system is an undoubtedly convenient tool for a variety of data bases in which manipulation of structural formulas is required. Actually, the system has been applied to the system SPIRES (SPectral Information REtrieval System), in which a ^{13}C NMR data base system is contained. The data base system, an interactive retrieval system of structure (substructure)–spectral information, has already been reported briefly,[4] and more detail will be presented in the following paper.[5]

A function to represent stereochemical structure has not yet been included with the system. The problem, however, will be solved without much difficulty by adding new symbols indicating stereochemistry such as, for example, E and Z for geometrical and R and S for configurational isomers and by slightly modifying the program.

REFERENCES AND NOTES

(1) Zupan, J.; Heller, S. R.; Milne, G. W. A.; Miller, J. A. *Anal. Chim. Acta* **1978**, *103*, 141. Bremser, W. *Anal. Chim. Acta* **1978**, *103*, 355. Wipke, W. T.; Heller, S. R.; Feldmann, R. J.; Hyde, E., Ed. In "Computer Representation and Manipulation of Chemical Information"; Wiley-Interscience: New York, 1974.
(2) Lynch, M. F.; Harrison, J. M.; Town, W. G.; Ash, J. E. "Computer Handling of Chemical Structure Information"; Macdonald: London, 1971. Davis, C. H.; Rush, J. E. "Information Retrieval and Documentation in Chemistry"; Greenwood Press: Westport, CT, 1974.
(3) (a) Kudo, Y.; Yamasaki, T.; Sasaki, S. *J. Chem. Doc.* **1973**, *13*, 225. (b) Kudo, Y.; Sasaki, S. *J. Chem. Doc.* **1974**, *14*, 200. (c) Kudo, Y.; Sasaki, S. *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 43. (d) Sasaki, S.; Abe, H.; Hirota, Y.; Ishida, Y.; Kudo, Y.; Ochiai, S.; Saito, K.; Yamasaki, T. *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 211.
(4) Abe, H.; Sasaki, S.; Tanaka, K.; Osada, H. *CODATA Bull.* **1981**, *No. 40*, 31.
(5) Abe, H.; Hayasaka, H.; Miyashita, Y.; Sasaki, S. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 216-219.

Generation of Stereoisomeric Structures Using Topological Information Alone

HIDETSUGU ABE, HIROSHI HAYASAKA, YOSHIKATSU MIYASHITA, and SHIN-ICHI SASAKI*

Laboratory for Chemical Information Science, Toyohashi University of Technology, Toyohashi, 440 Japan

Received August 16, 1983

An algorithm for enumeration of stereoisomers due to asymmetric carbon, C=C double bond, and so on has been developed. By use of this algorithm, all the possible stereochemical structures for a molecule may be generated on the basis of its topological representation. The identification of each distinct stereoisomeric structure is performed by SEMA notation.

An algorithm has been developed to enumerate all the possible stereoisomers due to asymmetric carbons, C=C double bonds, and so on on the basis of topological data of chemical structure, for instance, the connection table, or the connectivity matrix, which is without stereochemical information.

Problems on computer-assisted enumeration of all the possible isomeric structures consistent with the molecular formula and/or spectral data of a certain compound have been studied, coupled with the studies of automated structure elucidation of organic compounds. The basic concept of the structure elucidation system is to infer the chemical structures that are not contradictory to such structural information of a sample compound. Up to now, such computer program systems as DENDRAL,[1] CHEMICS,[2] and CASE[3] have been developed for that purpose. The method of structure generation in these systems is the method that utilizes some of the distinctive features of computer, high-speed calculation and exhaustive enumeration and provides the key technique in the systems of structure elucidation as well as automated analyses of spectral data. Nourse and his co-workers[4] extended the isomer generation task from ordinary constitutional isomer level to stereoisomer level. Their work suggests that consideration for stereoisomers should be prerequisite in construction of the structure elucidation system in the future.

This study is concerned with a method to make exhaustive and unoverlapped enumeration of stereoisomers caused by asymmetric carbon atoms, C=C double bonds, etc., only from topological structural information such as connection table or connectivity matrix, which includes no stereochemical information. In practice, SEMA (Stereochemically Extended Morgan Algorithm) notation[5] is employed for identification of individual stereoisomeric structures. SEMA is an extension of Morgan method, which Wipke and Dyott have developed for recognition of stereochemical configurations in the organic synthetic design system (SECS).[6]

CONCEPT OF GENERATION OF STEREOISOMERIC STRUCTURES

Stereocenter. The most important things to be considered in generation of stereoisomeric structures are how to recognize the atoms and bonds responsible for stereoisomerism and how to represent the information of their configurations. In accordance with the proposal of Wipke and Dyott,[5] atoms and bonds responsible for the stereoisomerism are called here "stereocenters", which will be defined as follows.

An atom i is called a stereocenter if positional change of two attachments (substituents, atomic groups) bonded to the atom i will result in an alternative three-dimensional structure. As shown in Figure 1, exchange of groups between l and k in formula 1a changes the R configuration into an S configuration (1b). Thus, carbon atom i can be called a stereocenter. As usual, one stereocenter in a molecule results in two stereoisomers, and therefore, n stereocenters may produce 2^n stereoisomers. Of course, there may be cases where identical stereostructures are found in 2^n isomers, because of symmetry in the structure of a compound like tartaric acid. In this study, we will deal with tertiary carbons, quaternary carbons, and C=C double bonds as potential stereocenters but will not consider heteroatoms and double bonds connecting heteroatoms.

Configuration Mold. As one of the characteristics of the present method is that topological information alone is used for stereoisomeric structure generation, configuration mold

0095-2338/84/1624-0216$01.50/0