A Structural Isomers Enumeration and Display System (SIEDS)

YOSHIHIRO KUDO, YUJI HIROTA, SHOTARO AOKI, YOSHITO TAKADA, TOYOAKI TAJI, ICHIRO FUJIOKA, KAZUKO HIGASHINO, HISAYUKI FUJISHIMA, and SHIN-ICHI SASAKI*

JEOL, Tokyo, and Miyagi University of Education, Sendai, Japan

Received May 21, 1975

A Structural Isomers Enumeration and Display System (SIEDS) completely enumerates unique structures composing "informational homologues" by means of the connectivity stack, and displays each of them in the form of a structural diagram and the corresponding connection table. SIEDS is incorporated in the Instrument JAL-30XA, which is a lightpen-controlled system with a cathode ray tube (CRT) and a cassette tape, and all information and instructions are input via the lightpen at desired terms on the CRT. Matching procedures for a final check of candidate structures are considerably complicated because ranges (the maximum and minimum values of partial structures) are contained in the input information of SIEDS.

We have reported the concept of the connectivity stack.[1] Since the connectivity stack specifies the connectivities of an organic structure and easily undergoes many kinds of checks to eliminate inappropriate constructions by pruning the potential logical tree as early as possible, it is utilized for not only quick building-up of a correct structure but also for exhaustive enumeration of unique structures consistent with given structural information.[2-4] Conversion of the stack into the corresponding structural diagram is justified by preference of the latter over the former by the chemist, in spite of rather high cost,[5] and because of the possibilities of enumeration of very large numbers of structures. A Structural Isomers Enumeration and Display System (SIEDS) has been developed and is used to help a chemist who wants to deduce an organic structure from various types of structural information. The goal of the system is to enumerate and display all logically valid structures (called the "informational homologues"[6]) according to the input information, which consists, in this case, of a molecular formula and, if any, partial structures. In addition to displaying a structure, the system is designed to be easy to operate by chemists so that they can easily communicate with the system, from viewpoints of both hardware and software. SIEDS is incorporated in the Instrument JAL-30XA,[7] which is a lightpen-controlled system with a cathode ray tube (CRT) and a cassette tape (Figure 1), and all information and instructions are input by aiming the lightpen at the desired term on the CRT. Enumerated structures are stored in the cassette tape and are displayed on the CRT, one at a time, according to indication by the operator. The representation of a structure is performed in the form of a structural formula diagram whenever possible and, if impossible or desired, a connection table.

SIEDS is an integrated program composed of several subprograms. They are divided into two parts from the viewpoint of their functions, e.g., Input/Output and Enumeration.[3] The former part consists of (1) selection of execution, (2) input of a molecular formula, (3) input of partial structures, (7) indication of the structure to be displayed, and (8) display of structures, and the latter, (4) combination of component segments, (5) building-up candidate structures of the informational homologues, and (6) final check of them by matching with input data. The numbers more or less show the order called (Figures 1 and 2).

Table I. Terms on the CRT Required for Description of Input Partial Structures[a]

(A)	C	CH	CH2	CH3	O	OH	FWD
	-	=	(C)	(?)	(O)	RNG	END
(B)	F	CL	BR	I	(X)	(*)	FWD
	(=)	#	DBL	TRI	R1	R2	BCK
(C)	N3	NH	NH2	(N)	P3	PH	FWD
	PH2	(P)	S2	SH	(S)		BCK
(D)	S4	SH	SH2	SH3	(S)	N5	FWD
	NH	NH2	NH3	NH4	(N)		BCK
(E)	P5	PH	PH2	PH3	PH4	(P)	FWD
	S6	SH	SH2	(S)			BCK
(F)	RG3	RG4	RG5	RG6	B66	ERR	FWD
	F-R	H-R	B-R	S-C	END		BCK
NUMERALS							
	(0)	(1)	(2)	(3)	(4)	(5)	FWD
	(6)	(7)	(8)	(9)	=	END	BCK

[a] See the text. Some of the terms are unused as yet.

REPRESENTATION OF STRUCTURES

To describe all organic structures without duplication, partial structures are logically divided and the resultants are called components; for any component C_i,

$$\bigcup_i C_i = \text{whole structures}$$

$$C_i \cap C_j = 0 \ (i \neq j)$$

Of many possible sets under the two conditions, SIEDS selects a level of components in which hydrogen is always subordinated to other elements (Table I), taking effective description of structures into consideration. Table I shows all components in SIEDS, afferent natures [(X) type], and other terms required for description of input partial structures. Adoption of the components, "double [=]" and "triple [#]" (bond) lighten iterative checks for confirmation of all pairs between which multibonds are caught. Some of the terms are unused as yet in the up-to-date system. Besides the component which corresponds to a chemical element, there is another concept, SEGMENT, which corresponds to an atom. In the system it is considered that a structure may be described by the assemblage of the component segments and the connectivities between them. All components have their peculiar properties: constitution of atom(s), valence, and efferent nature.

For example: methyl (CH_3–), one of the monovalent components, consists of a carbon atom and three hydrogen atoms and its efferent nature is carbon: secondary amino (–NH–), one of the bivalent components, consists of a nitrogen atom and a hydrogen atom, and its efferent nature is nitrogen. Providing components with such properties means connecting chemistry with mathematics.

A (complete) structure consists of only segments and the connectivities between them; both ends of a bond are always defined with component segments. On the other hand, in a

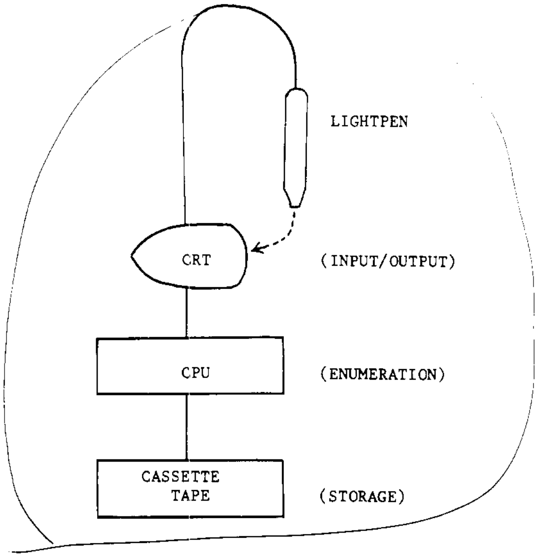


Figure 1. Hardware constitution (JAL-30XA).

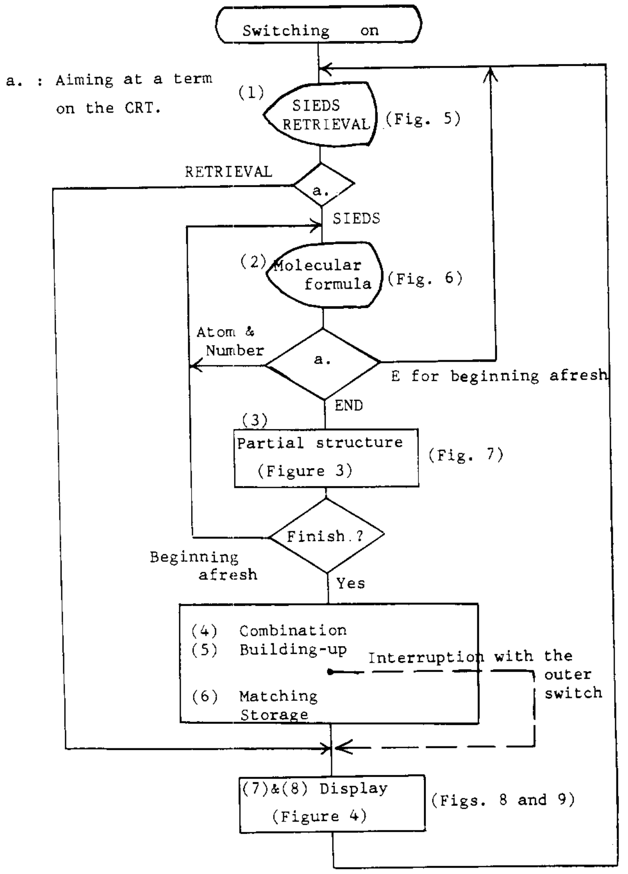
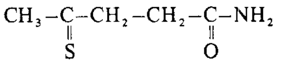
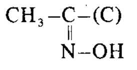
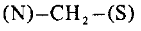


Figure 2. Outline of whole operation.

partial structure which requires a bond to be connected to another part of the structure, one of both ends of at least one bond is occupied by afferent nature (which is efferent nature of a neighbor segment). For instance, **1** is a structure, and **2** and **3** are partial structures:


1


2


3

Unintentionally they also suggest the essential method for description of structures in SIEDS. Namely, any structure is described not directly with atoms but with component segments adopted by SIEDS.

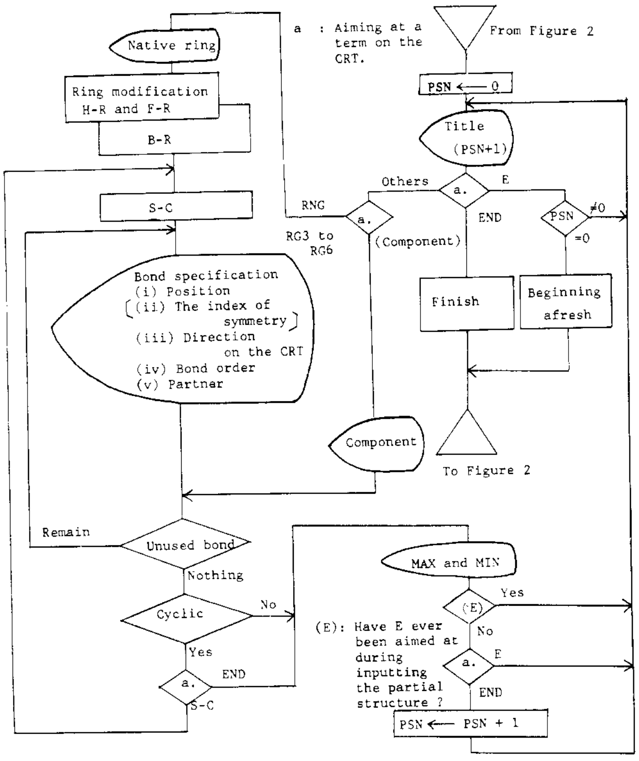


Figure 3. Outline of operation for input of partial structure.

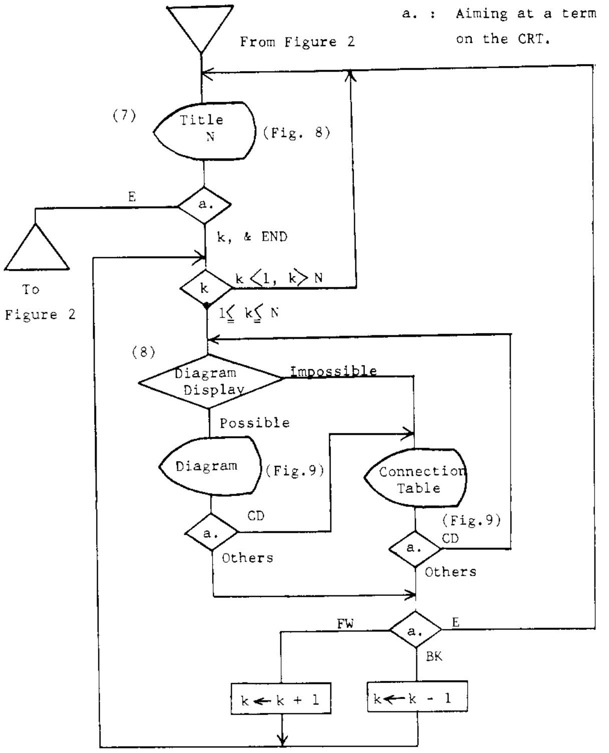


Figure 4. Outline of operation for display of desired structures.

A structure can be expressed equivalently in several ways, with different functions, namely, a structural diagram (input/output), a connectivity stack (enumeration), a connection matrix (matching), a connection table (output, and structure analysis for display), and a linear notation (structure analysis for display).

OPERATION

The outline of the whole operation in SIEDS is summarized in Figures 2, 3, and 4. Depicting a flow chart of the procedure cannot but be a little complicated because of the lightpen-

Table II. Examples of the Execution

	Input information			Required time, sec	
No.	Molecular formula	Partial structures	No.[a]	A[b]	B[c]
1	C_6H_6		217[4,7,8]	82	26
2	C_6H_6	$(CH)_6$	6[4,7]	13	<2
3 (Ex-4)	C_6H_8		159[4,8]	47	11
4	$C_6H_{10}O$		747[4,8]	155	28
C (Ex-1)	$C_6H_{10}O$	(Figure 7)	14	15	
6 (Ex-3)	$C_4H_7F_3$		14	14	<1
7 (Ex-2)	$C_5H_{10}O_3P^VS^{II}$	(Figure 7)	7	450	

[a] Number of informational homologues. [b] Enumeration plus storage. [c] Enumeration alone.

controlled system. Generally in the system, the terms E and END serve to indicate several levels of erasing and finishing input, respectively. Any numeral value is represented by proper numeral latters and an additional symbol, =, which corresponds to a decimal point.

Real Examination. Before describing the operation in detail, five examples (Ex-1 to Ex-5) of real execution are shown in Figures 6 to 10, divided into every step, and in Table II. The table shows (1) numbers of the informational homologues according to the input information, (2) required times in the up-to-date system for the enumeration plus the storage into the cassette tape, and (3) the enumeration alone. The latter was measured by cutting the route for the storages and suggests the required time when the storage will be extremely accelerated.

In Ex-1, molecular formula, $C_6H_{10}O$, and four kinds of partial structures are input information, as shown in Figures 6 and 7, respectively. Fourteen structures are enumerated (Figure 8), and they are displayed on the CRT in the form of a structural formula except one (no. 14), which is displayed as a connection table (Figure 9). The required time in the up-to-date system was 155 sec from the end of the information input to the time when pattern for indicating a number of structures (Figure 7) appeared (Table II, no. 5); if the time for storage of structures into the cassette tape becomes too small to be measured, the overall time would be decreased to 28 sec.

In Ex-2, atoms other than carbon, hydrogen, and oxygen, are also contained in the molecular formula, in this case phosphorus and sulfur. The condition, molecular formula ($C_6H_{15}O_3P^VS^{II}$) and two kinds of partial structures, resulted in seven informational homologues, one of them, no. 4, represented with a connection table.

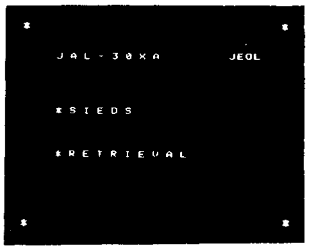
Ex-3 shows enumeration of all structural isomers because the input information consists of only molecular formula, $C_4H_7F_3$. Of the resulting 14 isomers, the last 3 do not contain the methyl group; then if the methyl group is denied, all structures except three would be eliminated.

Ex-4 also shows enumeration of all structural isomers. But of 159[8] possible structural isomers from molecular formula, C_6H_8, all 22 acyclic structures and 4 of 137 cyclic ones are illustrated.

Ex-5 is one in which a cyclic partial structure is input and drawn in various ways for all of the enumerated structures.

Job Selection. First of all, after switching on the system, a title pattern (Figure 5) appears. It shows two instructions, SIEDS and RETRIEVAL, which are keys to the enumerated and display of informational homologues, and the retrieval of structures previously stored in the cassette tape, respectively. Indication of SIEDS leads the operator to a pattern for input of a molecular formula (Figure 6).

Molecular Formula Input. To input a molecular formula, the number of atoms of all elements in the formula must be defined. The system watches and waits for the input. When aimed at, an atomic symbol will brighten up. Next, the number of each atom is registered by the use of a numerical value. By aiming at END, the molecular formula input finishes, and a new pattern (Figure 7) for input of a partial structure appears.



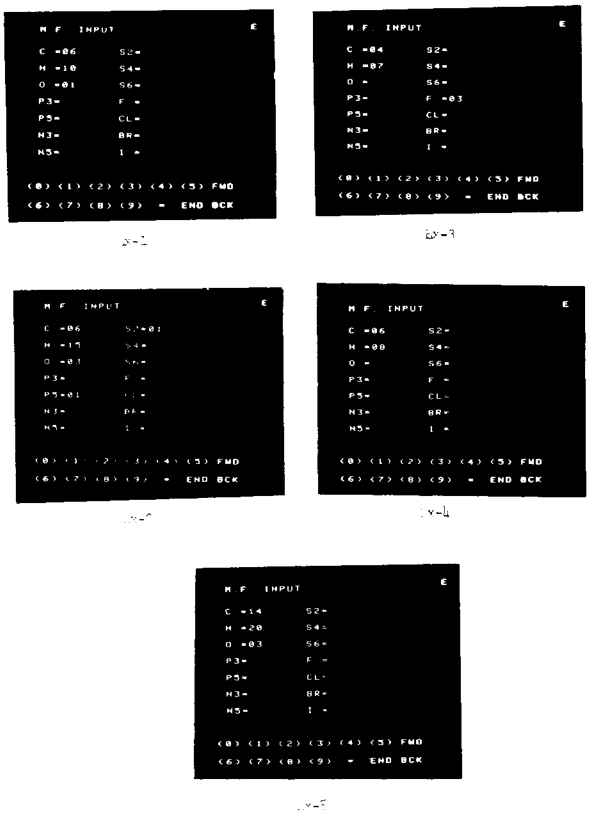
```
*                              *
   J A L - 3 0 X A       JEOL

   * S I E D S

   * R E T R I E V A L

*                              *
```

Figure 5. A title pattern.



```
M F   INPUT                  E
C =06       S2=
H =10       S4=
O =01       S6=
P3=         F =
P5=         CL=
N3=         BR=
N5=         I =

(0) (1) (2) (3) (4) (5) FWD
(6) (7) (8) (9)  =  END BCK
```

x-1

```
M.F.  INPUT                  E
C =04       S2=
H =07       S4=
O =         S6=
P3=         F =03
P5=         CL=
N3=         BR=
N5=         I =

(0) (1) (2) (3) (4) (5) FWD
(6) (7) (8) (9)  =  END BCK
```

Ex-3

```
M F   INPUT                  E
C =06       S2=01
H =15       S4=
O =03       S6=
P3=         F =
P5=01       CL=
N3=         BR=
N5=         I =

(0) (1) (2) (3) (4) (5) FWD
(6) (7) (8) (9)  =  END BCK
```

x-2

```
M F.  INPUT                  E
C =06       S2=
H =08       S4=
O =         S6=
P3=         F =
P5=         CL=
N3=         BR=
N5=         I =

(0) (1) (2) (3) (4) (5) FWD
(6) (7) (8) (9)  =  END BCK
```

x-4

```
M.F   INPUT                  E
C =14       S2=
H =20       S4=
O =03       S6=
P3=         F =
P5=         CL=
N3=         BR=
N5=         I =

(0) (1) (2) (3) (4) (5) FWD
(6) (7) (8) (9)  =  END BCK
```

x-5

Figure 6. Pattern for input of a molecular formula.

Partial Structure Input. In the pattern a partial structure number to be input and all necessary terms to describe it are shown. The terms are too many to be exhibited all at once on the CRT, so they are divided into six groups (cf. Table I), and the six patterns are exchangeable with each other by indication of FWD (forward) and BCK (back). The outline of the procedure is shown in Figure 3.

An acyclic partial structure is specified with components of segments, the connectivities between segments, at least one afferent nature, and the allowed range, MAX (the possible maximum number) and MIN (the necessary minimum number). Input of the range makes the partial structure input finish. To input an acyclic partial structure, first a component is aimed at, and then necessary bond(s), component(s), and afferent nature(s) are aimed at. Defining all ends of all bonds means complete specification of the partial structure. At that time, characters, MAX= and MIN=, automatically appear.

In the case of a ring system input, aspects are more complicated than those of acyclic moieties. Moreover, the principle of SIEDS is partially changed for simplification of operation. Namely, although a ring structure is input, the content is not specified except its composition and symmetric characters of efferent bonds, so duplication can occur when

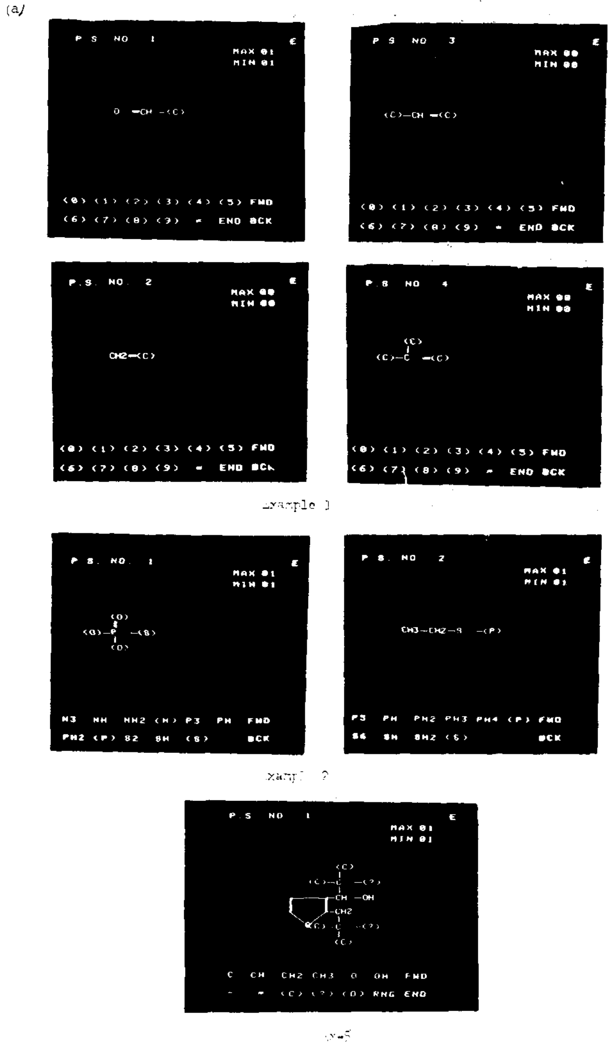


Figure 7. Pattern for input of partial structures.

the rest-segments unfortunately form an identical cyclic structure with the input one. The symmetric character is indicated with the index of symmetry by the operator. The index itself has no meaning, but whether two bonds have the same index or not determines whether they are equivalent or not. After aiming at RNG, a term, indicating the ring size (RG3 to RG6), is aimed at. A depicted native ring means an alicyclic hydrocarbon (cyclopropane to cyclohexane). To change a ring into a partial structure, at least one implied hydrogen atom has to be substituted for a bond by aiming at S-C (side chain) and a proper term [-, =, or # (a triple bond)]. Just at that time, the index of symmetry is assigned. Before using the S-C, all other ring modifications, H-R (for heteroatom), F-R (for ring fusion), and B-R (for conversion to multibond), must be finished. The former two will not function after aiming at B-R. Unlike in the case of acyclic partial structures, the definition of all ends of all bonds does not mean complete specification of a desired partial structure, because no hydrogen may be implied. Therefore the end point has to be indicated with END by the operator. After this, MAX and MIN appear and their values are input. During the operation, the operator can cancel the partial structure under input by aiming at E. Aiming at END after inputting MAX and MIN means accomplishment of the input of the partial structure.

Enumeration and Storage of the Informational Homologues. On finishing the information input, the system starts the enumeration. Every time when three new structures are built up, they are sent to the cassette tape. To write them into the tape, 2 sec is required. The up-to-date system may store about 2800 structures on one tape. The enumeration is automatically carried out, as will be described in sections after the next one.

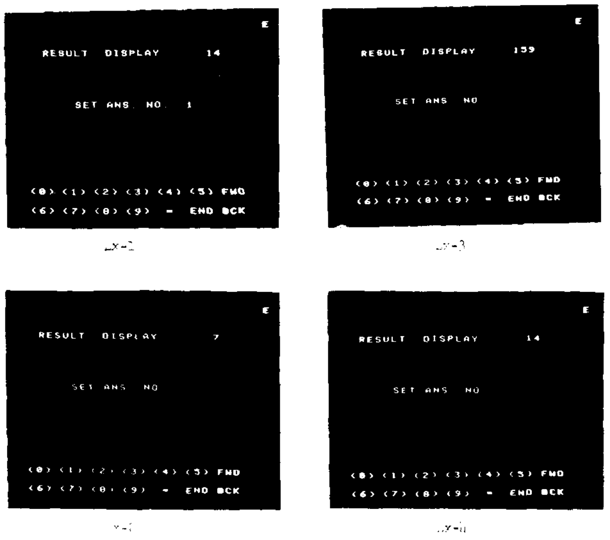


Figure 8. Number indication pattern.

Structure Display. A pattern for indication of number of desired structures, the number indication pattern (Figure 8), appears (1) when the enumeration of the informational homologues is finished, (2) when the enumeration is interrupted with the outer switch by the operator, (3) when in the first title pattern (Figure 5), RETRIEVAL instead of SIEDS is aimed at (Figure 2), or (4) when a too small or too large number is indicated as the structure number (Figure 4). The outline of the operation is shown in Figure 4. The number indication pattern is used to input a desired structural number, k. The pattern is accompanied by the total number of the enumerated structures, N, which defines the upper limit of k. Of course, the lower limit is one. A k beyond the limit is not accepted. In Figure 8, Ex-1 and the others show the patterns after and before inputting the k, respectively. If the k is reasonable, the kth structure is retrieved from the cassette tape and displayed. BK (back) and FW (forward) are keys to indicate structures with a lower by one and the next numbers, respectively. The character, E, plays its role in changing back the condition to the number indication pattern.

The display is usually performed in the form of a structural formula diagram (Figures 9 and 10). Although the system fails to put the diagram in the central area of the CRT or when any margin of the diagram protrudes beyond the CRT, the operator can easily shift the diagram to any direction on the CRT. A term, CD, is used for translation between a structural diagram and a corresponding connection table. When a structural diagram is predicted to be too complicated to display on the CRT, only a connection table is denoted. For example, a too large ring size (more than six), or a too high degree of ring fusion (more than four, like pentacene and hexacene) is beyond the power of the up-to-date SIEDS (to be expanded).

ENUMERATION OF THE INFORMATIONAL HOMOLOGUES

Combination of Component Segments. Homologues are the logical product of each input. First, a molecular formula defines the range of the homologues. It is realized by solving the set of equalities and inequalities

$$(CV)(CM) = (MF) \tag{1}$$

All elements of (CV) are zero or natural numbers where (CV) stands for component vector, which is a set of numbers of segments for all components, and (MF) for the vector

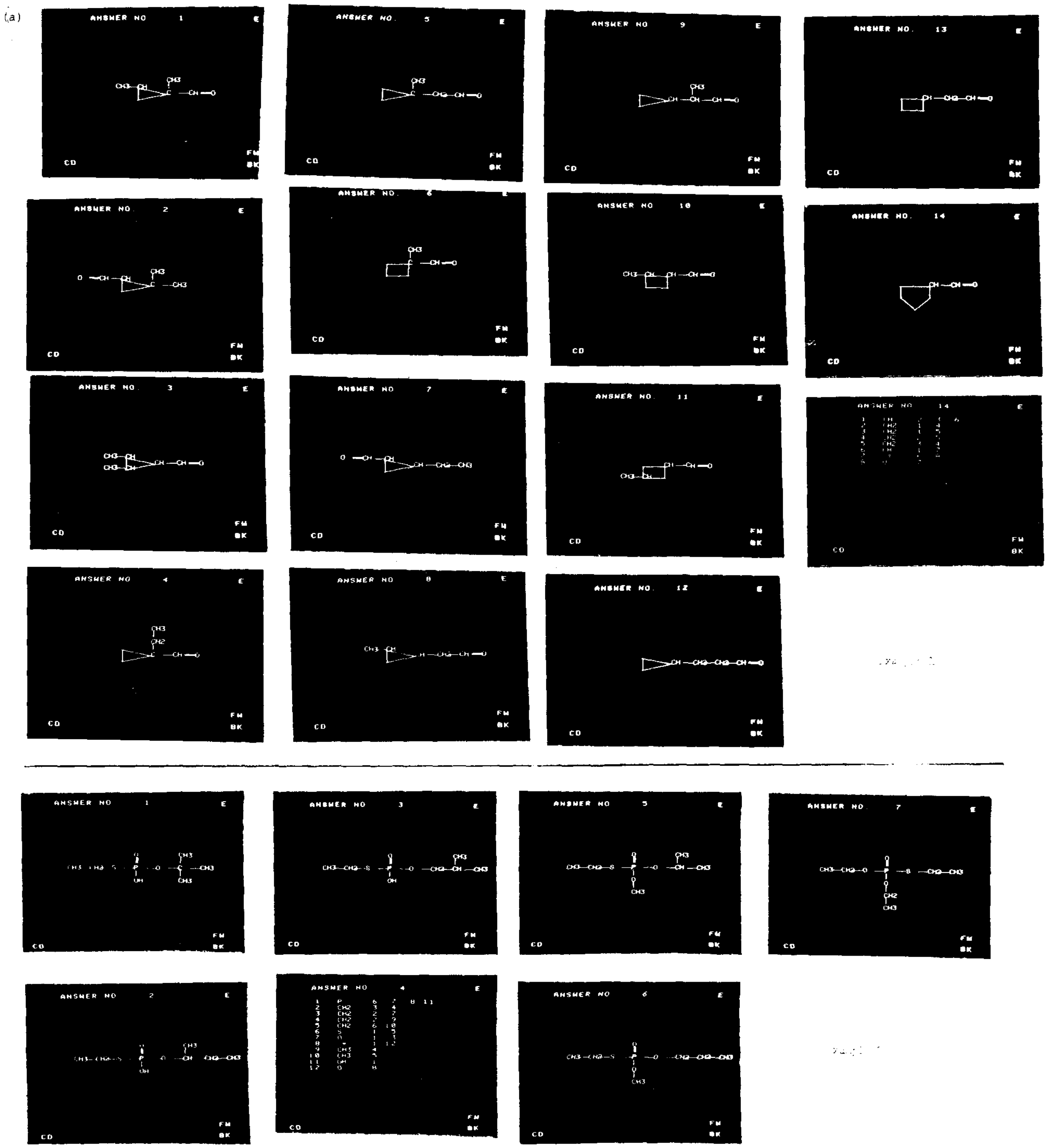


Figure 9. Informational homologues.

representation of a molecular formula. Let a molecular formula be C_2H_6O, with components, C, CH, CH_2, CH_3, O, OH, and others containing heteroatoms other than oxygen. The last components are not involved during calculation in this case because the molecular formula consists of only carbon, hydrogen, and oxygen. Since

$$(\mathrm{C\ CH\ CH_2\ CH_3\ O\ OH\ others})\begin{array}{c}\text{C H O others}\\ \begin{pmatrix}1&0&0&0\\1&1&0&0\\1&2&0&0\\1&3&0&0\\0&0&1&0\\0&1&1&0\\ *&*&*&*\end{pmatrix}\end{array} = \begin{array}{c}\text{C H O others}\\ (2\ 6\ 1\ 0)\end{array}$$

(C CH CH_2 CH_3 O OH others) equals (0 0 1 1 0 1 0) and (0 0 0 2 1 0 0). These two solutions are clearly ethanol and dimethyl ether, respectively. In the algorithm, (CM) is a set of certain constant values, and (MF) is contained in the input information. After assuming a possible (CV), the set of equations and inequalities (1) is examined for whether being true or not. Of course, in the algorithm, the system is designed to predict useless calculations as early as possible.

Building-Up of Candidates of the Informational Homologues. Building-up is carried out by means of the connectivity stack. As stated above the connectivity stack is created as a tool for quick building-up of a correct structure and exhaustive enumeration of unique structures consistent with given structural information, and its purpose has been attained. The aspects are depicted in our preceding article.[4]

Examination by Matching. Each candidate structure is examined by matching with the input information. The input consists of (1) the number, *M*, of kinds of partially structures,

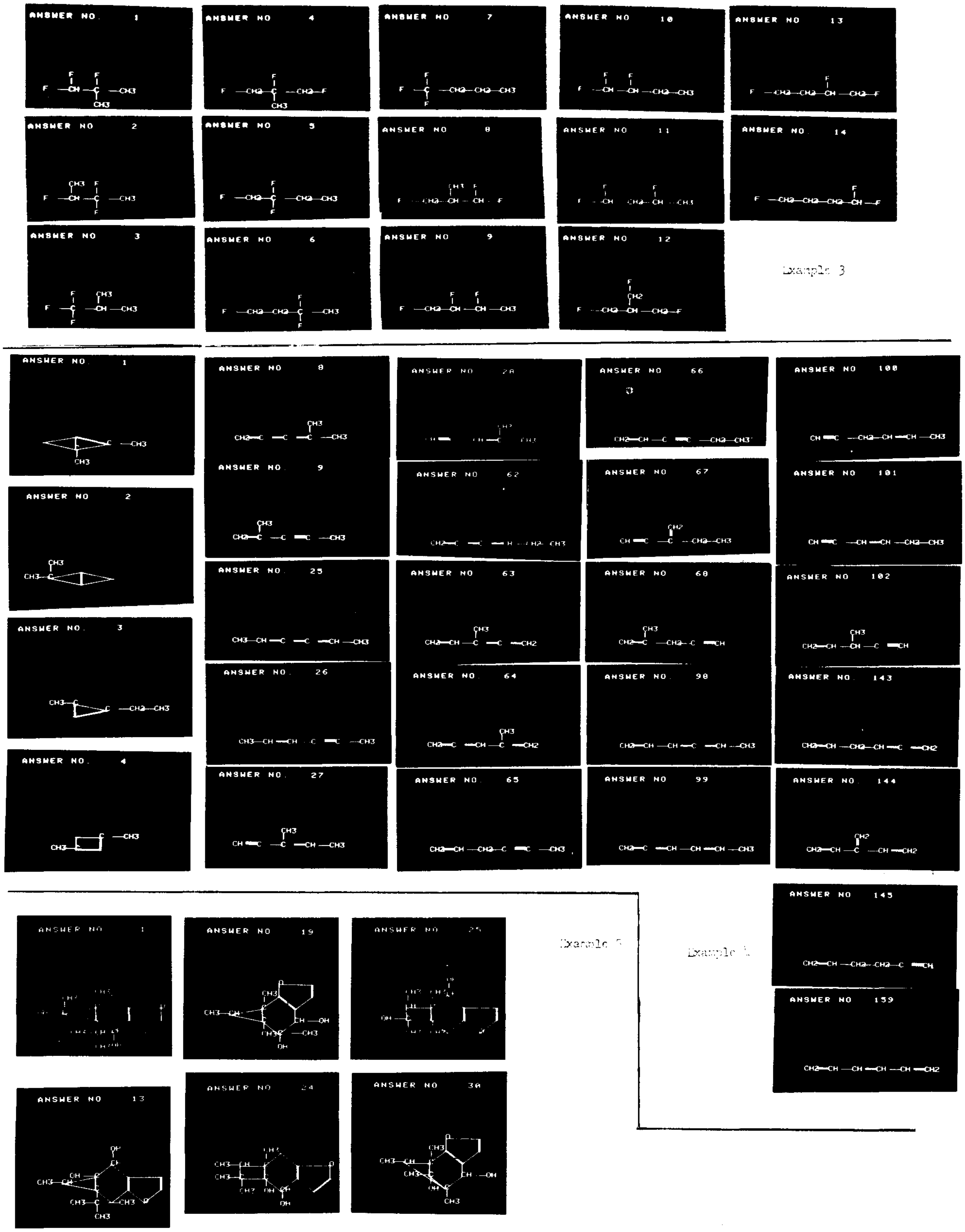


Figure 10. Informational homologues.

(2) the content of each partial structures, PS*i*, specified with terms on the CRT, and (3) its range of the possible numbers, MAX*i* and MIN*i*. The input forms a set of inequalities:

$MIN_1 \leqq$ a number of $PS_1 \leqq MAX_1$

$MIN_2 \leqq$ a number of $PS_2 \leqq MAX_2$

$MIN_M \leqq$ a number of $PS_M \leqq MAX_M$

like

$3 \leqq (PS_1) \leqq 4 \quad (PS_1) = 3, 4$

$0 \leqq (PS_2) \leqq 0 \quad (PS_2) = 0$

$0 \leqq (PS_3) \leqq 2 \quad (PS_3) = 0, 1, 2$

This set means six possibilities: (PS_1 PS_2 PS_3) = (4 0 2), (4 0 1), (4 0 0), (3 0 2), (3 0 1) and (3 0 0). Generally the

Table III. Examples of the Examination by the Matching

	The Input Information							
CH_3CH_2(?)	0-0	2-1	3-2	-	-	2-1	2-1	1-0
(?)CH_2CH_2(?)	-	-	-	1-0	3-2	1-1	1-0	0-0
Candidate								
C–C–OH · C	0	X	X	0	X	X	X	0
C–C–C–OH	X	0	X	0	X	X	0	0
C · C–C–C–OH · C	0	X	X	0	X	X	X	0
C · C–C–C–OH · C	X	0	X	0	X	X	0	0
C–C–C–OH · · C C	0	X	X	0	X	X	X	0
C–C–C–C–OH · C	0	X	X	0	X	X	X	X
C–C–C–C–OH · C	X	0	X	0	X	X	0	0
C–C–C–OH · C–C	X	0	0	0	X	X	0	X
C–C–C–C–OH · C	X	0	X	0	X	X	0	0
C–C–C–C–C–OH	X	0	X	0	0	0	0	X

[a] 0: registered as an informational homologue. X: rejected.

number of the possibilities is

$$\prod_{i=1}^{M}(\mathrm{MAX}_i - \mathrm{MIN}_i + 1)$$

If the candidate satisfies any of the possibilities, it is registered as an informational homologue; otherwise it is rejected. In the algorithm, homologues of a partial structure on segments and the connectivities between them, considering afferent segments, are extracted from a candidate structure according to a number between MAX and MIN. This operation continues until any set of the inequalities on numbers of partial structures becomes true. If there is no way to make it true, the candidate is rejected. Aspects of the matching procedure are not always simple. For example, *n*-butane (CH_3CH_2-CH_2CH_3) can survive, though all structures which contain any ethyl group are denied, if an ethylene group (CH_2CH_2) is permitted. In other words, extracting an ethylene group hides the existence of ethyl. Table III shows several examples of the matching process.

CONCLUSIONS

SIEDS enumerates all structures of the informational homologues and displays them in the form of structural diagrams. By means of the system, chemists may easily and quickly recognize an extent at which they arrive in the elucidation of organic structures, and, in cases that identification or determination of structures is not accomplished, they can consider how to obtain further structural information.

REFERENCES AND NOTES

(1) Y. Kudo and S. Sasaki, *J. Chem. Doc.*, **14**, 200 (1974).

(2) (a) S. Sasaki, Y. Kudo, S. Ochiai, and H. Abe, *Mikrochim. Acta*, 726 (1971); (b) S. Sasaki, Y. Kudo, S. Ochiai, and I. Fujioka, *Jpn. Anal.*, **22**, 25 (1973); (c) CHEMICS (Combined Handling of Elucidation Methods for Interpretable Chemical Structures) system; e.g., CHEMICS-F will be reported in the near future.

(3) Y. Kudo, *Kagaku no Ryoiki Zokan*, **98**, 115 (1972).

(4) Y. Kudo and S. Sasaki, *J. Chem. Inf. Comput. Sci.*, preceding paper in this issue.

(5) M. F. Lynch, J. M. Harrison, W. G. Town, and J. E. Ash, "Computer Handling of Chemical Structure Information", Macdonald, London, and American Elsevier, New York, 1971, p 36.

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

(7) This is a improved instrument of JAL-30X, which is a system controlled by a console typewriter with a joystick: Y. Kudo, and S. Ochiai, *Anal. Instrum.*, (*Bunseki-Kiki*), **11**, 654 (1973).

(8) L. M. Masinter, N. S. Sridharan, J. Lederberg, and D. H. Smith, *J. Am. Chem. Soc.*, **96**, 7702 (1974).

LETTERS TO THE EDITOR

Dear Sir:

I read with interest the paper, "Automatic Abstracting Research at CAS", by Pollack and Zamora in the *Journal of Chemical Information and Computer Sciences*, Vol. 15, No. 4, 1975.

On page 229, Table V is entitled, "Abbreviations and Symbols used in ACS Publications". In this table a period is used after all abbreviations and upper-case letters for the symbol for nuclear magnetic resonance. This is, of course, contrary to the information published by the American Chemical Society in "Handbook for Authors of Papers in the Journals of the American Chemical Society", 1st ed, pp 44 and 45.

It appears to me the authors became a little bit too enthusiastic in the title of Table V, especially since CAS deviates from the "Handbook...". It would seem to me that CAS should either conform to the recommendations in the "Handbook..." or the title of Table V should have been "Abbreviations and Symbols used in CAS Publications".

The authors are also inconsistent in the use of the hyphen, viz., p 228: in the title of Table IV the word, "Non-Substantive", is used, which is obviously incorrect; however, in the text in line 3 under "Final Editing" and also in the heading of the next paragraph the correct form, "nonsubstantive", is used.

The Division of Chemical Information is primarily concerned with communication of correct information, yet we continually allow inconsistent, nonconforming, and inaccurate statements to be published. This is very confusing, and, in this case, it makes the job of editing manuscripts in-house for subsequent publication in ACS journals very difficult and at times impossible when two ACS publications print two entirely different versions of a table.

Alan R. McGarvey, Manager
Technical Information Services
Armstrong Cork Company
Lancaster, Pennsylvania

Received December 18, 1975

Dear Sir:

Dr. Pollock and Mr. Zamora have asked me to respond to Mr. McGarvey's letter of 15 December 1975 concerning their article, "Automatic Abstracting Research at CAS", published