CHEMICS-F: A Computer Program System for Structure Elucidation of Organic Compounds

SHIN-ICHI SASAKI,* HIDETSUGU ABE, YUJI HIROTA, YOSHIAKI ISHIDA, YOSHIHIRO KUDO, SHUKICHI OCHIAI, KEIJI SAITO, and TOHRU YAMASAKI

Miyagi University of Education, Aoba, Sendai 980 Japan

Received November 3, 1977

A computer program system CHEMICS-F for the structure elucidation of organic molecules containing C, H, and O is described in detail. CHEMICS-F involves the software used to analyze spectra, to convert the spectral information into "components" (defined substructures), and to construct a molecular structure based on the components by means of a newly developed method. Multiple structural formulas as final output are often constructed based on both desirable and undesirable components designated by the spectral information. To minimize the multiplicity of answers, a file handling procedure has been integrated in which the spectra of NMR, IR, and MS are stored in compressed and concentrated shapes capable of identification of compounds.

Many studies have been done for structure elucidation and identification of organic compounds with the aid of computers. The methodologies and the techniques are classified into two categories. One is the retrieval method in which the identification is carried out by refining the most likely structure from a data bank by comparing data, for instance, chemical spectra, of an unknown with those of organic compounds stored there. The other is a structure generation method; that is, the most probably structure is generated by the automated analysis of data (also, for instance, chemical spectra) of an unknown using empirical and theoretical rules.

Many studies using the retrieval method have already been reported and some are now working as practical systems.[1] However, this method is always limited by rather serious weak points: (1) when the number of stored members are not sufficient, the system will scarcely function as a good tool for the characterization of compounds, even if it is built in an excellent way; (2) it will become more difficult to extract one and only one compound without noise (incorrect answers) when the searching is executed in a larger bank in which many spectra are stored; and (3) it is impossible to collect the spectral data of all existing organic compounds, whose number amounts to millions, with new compounds being produced every day.

For these reasons, the second way, the structure generation method through which the generation of a reasonable structure is performed by the analysis of spectral data and other properties of an unknown, has been investigated by several workers.[2-10] This method, which is at the opposite side of the retrieval method, seems to work well when the class of compounds is limited rather narrowly; otherwise a large number of structures is produced in response to the given structural information.

Thus, the first and the second methods were integrated to reduce the unnecessary answers; that is, the system CHEMICS-F was designed so as to be endowed with these two functions (see Figure 1).

CHEMICS-F plays its role by deducing all logically valid structures from a set of input information concerned with structure, on the basis of empirical and/or theoretical rules. These valid structures, equivalent under the given structural information, form a definite class, and each individual in the class is called an "informational homologue" which is a structure to be taken into consideration from a logical viewpoint, i.e., regardless of a chemist's expectation. The system is designed to elucidate structures of organic compounds with C, H, and O by narrowing informational homologues by the examination of the molecular formula and the IR, NMR, and MS[16] analyses of the sample compound. The system also works so as to make a rational integration of two kinds of approaches—deductive interpretation of chemical data and comparison of the chemical spectra of the sample compound with the data file. To execute this strategy the system, four major programs and more than 30 subprograms were written. Four major programs are, as shown in Figure 1, INPUT CONTROL, DATA ANALYSIS, STRUCTURE GENERATOR, DATA COMPARISON. In addition, there are utility routines for compiling new spectral data to be packed into the data file. Each role of the major programs is as follows: INPUT CONTROL receives a set of input data (molecular formula, IR and NMR spectra and, if available, MS) and sends them to DATA ANALYSIS. Here, spectral data are also converted into certain fixed formats for sending to DATA COMPARISON. DATA ANALYSIS sifts out all the partial structures (components) contradictable to the input information. STRUCTURE GENERATOR first selects all the possible sets of components satisfying the molecular formula. After these processes, the "informational homologues" consistent with the input information are enumerated. DATA COMPARISON determines the plausibility of each generated structure by searching the spectral data file with the structure as a keyword.

COMPONENT

"Components", which mean partial structures of organic molecules in the present paper, play the most important role in the system. They are used not only as fragments to construct structures but also as carriers of the spectral information and the elucidated results in terms of various parameters. It is desirable for simplifying the structure construction procedure that the valence bonds of a component are equivalent to each other. On the other hand, as much information as possible obtained from spectral data analysis should be kept in the components.

A simple partial structure, for example, –CO–O–, can be easily determined by analyzing spectral data, and a component with that structure satisfies the latter condition. However, from the standpoint of the former condition, the structure is not as preferable. If an imaginary component, A, which has that structure is employed, a representation X–A–Y means two different structures X–CO–O–Y and Y–CO–O–X; this makes the structure construction procedure complicated. In order to prevent these cases, the partial structure is divided into two components –CO– and –O–, both of which are symmetrical with respect to their bonds, but still contain as much information as possible by limiting their partners. This is a basic principle for defining the components in the CHEMICS system. Another condition is shown below:

$$\bigcup_i C_i = \text{all whole structures, and } C_i \cap C_j = \emptyset \ (i \neq j)$$

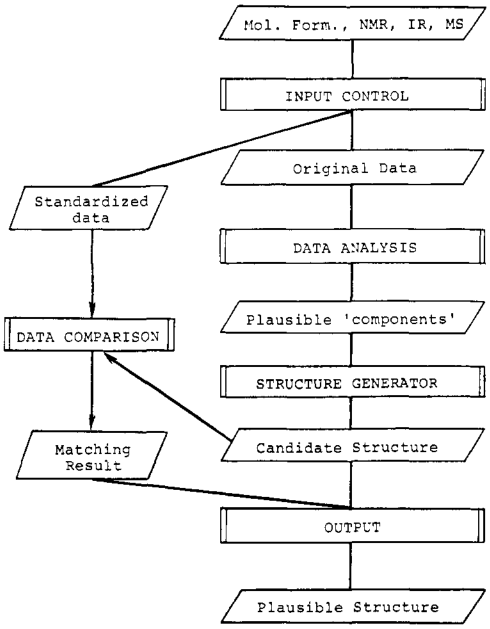
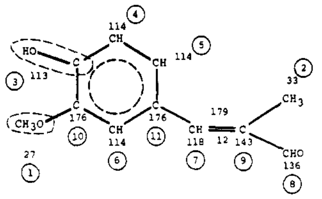


Figure 1. Block diagram of CHEMICS-F system.



```
component #   --- 27 --- 33 --- 113, 114 --- 118 --- 136 --- 143 --- 176 --- 179
   number     ---  1 ---  1 ---   1 ,  3 ---   1 ---   1 ---   1 ---   2 ---   1
```

component set	1*	2	3	4	5	6	7	8	9	10	11	12
	27	33	113	114	114	114	118	136	143	176	176	179

connectivities 1 - 10, 2 - 9, 3 - 4, 3 - 10, 4 - 5, 5 - 11, 6 - 10, 6 - 11, 7 - 11, 7 - 12, 8 - 9, 9 - 12

* This is not the actual numbering, refer to the text.

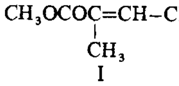
Figure 2. Example of structure representation using "components".

where C_i and C_j mean the *i*th and the *j*th components, respectively. Namely, any structure in the universe of discourse could be constructed with an appropriate set of components, and any pair of components should have no overlapping part. Now 179 components are arbitrarily defined for the system CHEMICS-F based upon our experience (Table I). To represent an organic structure with the components, a set of components and a set of connectivities are adopted in the system as shown in Figure 2.

Each component has its own attributes to specify it from all the other components. What the attributes mean are elemental composition and several parameters indicating efferent nature, afferent nature, the number of other components to be connected to the said component, and the number of bonds with the efferent nature. These parameters are used in the structure construction (see later section).

DATA ANALYSIS

Using methyl tiglate (I) as an example, the input data for



I

it are shown in Figure 3. They are sent to the next step, DATA ANALYSIS. As mentioned previously, the function of this block is to pick out all the components that do not contradict the input information.

Molecular Formula $C_6H_{10}O_2$

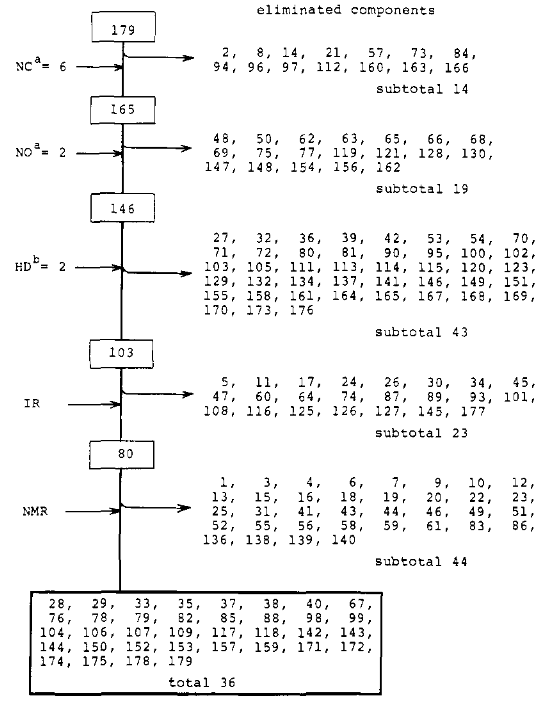
IR DATA

NO.	POSI.	INT.
1	2996	81
2	2960	250
3	1724	471
4	1658	285
5	1440	343
6	1386	207
7	1349	200
8	1266	436
9	1194	286
10	1142	407
11	1085	344
12	1037	156
13	990	128
14	923	130
15	869	83
16	814	132
17	740	313
18	661	176

NMR DATA

NO.	POSI.	AREA	HEIGHT
1	410.0	214.	16.
2	408.2	99.	16.
3	406.0	74.	11.
4	402.6	214.	15.
5	400.6	241.	16.
6	221.2	2646.	369.
7	111.2	2554.	202.
8	109.6	1608.	204.
9	107.8	236.	52.
10	105.2	1005.	97.
11	104.2	268.	63.

Figure 3. Input data for compound I.



eliminated components

2, 8, 14, 21, 57, 73, 84,
94, 96, 97, 112, 160, 163, 166
subtotal 14

48, 50, 62, 63, 65, 66, 68,
69, 75, 77, 119, 121, 128, 130,
147, 148, 154, 156, 162
subtotal 19

27, 32, 36, 39, 42, 53, 54, 70,
71, 72, 80, 81, 90, 95, 100, 102,
103, 105, 111, 113, 114, 115, 120, 123,
129, 132, 134, 137, 141, 146, 149, 151,
155, 158, 161, 164, 165, 167, 168, 169,
170, 173, 176
subtotal 43

5, 11, 17, 24, 26, 30, 34, 45,
47, 60, 64, 74, 87, 89, 93, 101,
108, 116, 125, 126, 127, 145, 177
subtotal 23

1, 3, 4, 6, 7, 9, 10, 12,
13, 15, 16, 18, 19, 20, 22, 23,
25, 31, 41, 43, 44, 46, 49, 51,
52, 55, 56, 58, 59, 61, 83, 86,
136, 138, 139, 140
subtotal 44

28, 29, 33, 35, 37, 38, 40, 67,
76, 78, 79, 82, 85, 88, 98, 99,
104, 106, 107, 109, 117, 118, 142, 143,
144, 150, 152, 153, 157, 159, 171, 172,
174, 175, 178, 179
total 36

a: NC and NO mean number of carbons and hydrogens in the molecule.

b: HD is the index of hydrogen deficiency which is calculated as : HD = NC + 1 - NH/2.

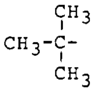
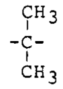
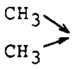
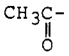
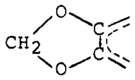
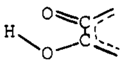
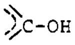
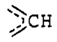
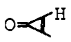
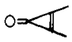
Figure 4. Selection of plausible "components".

The molecular formula of an unknown works as an effective filter in the first step of DATA ANALYSIS. Comparison of the composition of the molecular formula with those of all the components is carried out. The components which exceed the molecular formula in any compositions are discarded as inappropriate. Since compound I has six carbons and two oxygens, those components which require more than six carbons and/or more than two oxygens for their existence in a molecule are eliminated as shown in Figure 4. In the same manner, those components which require more than two as an index of hydrogen deficiency are eliminated.

Here, 103 components (179 - 14 - 19 - 43) survived as the components not contradictable with the molecular formula, $C_6H_{10}O_2$. After examination of the molecular formula, the analysis of the IR spectrum is carried out using the procedure stated below.

To analyze the existence of carbonyl and hydroxy groups via the IR spectrum, two parameters ICO and IOH are determined according to the relative intensity of three strongest

Table I

#	Structure	adjacent groups[a]
1		O
2		Y
3		K
4		D
5		T
6		C
7		O
8		Y
9		K
10		D
11		T
12		C
13	CH_3-C-	O
14		Y
15		K
16		D
17		T
18		C
19[b]		#82
20		#83
21		#84
22		#85
23		#86
24		#87
25		#88
26	CH_3O-	O
27		Y
28		K
29		D
30		T
31		C
32	CH_3-	Y
33		D
34		T
35		O
36		Y
37		K
38		D
39		T
40		C
41	CH_3-	#104
42		#105
43		#106
44		#107
45		#108
46		#109
47		#74
48		#75
49		#76
50		#77
51		#78
52		#79
53		#80
54		#81
55		#82
56		#83
57		#84
58		#85
59		#86
60		#87
61		#88
62	CH-	OOO
63		OOA
64		OOP
65		OAA
66		OAP
67		OPP
68		AAA
69		AAP
70		APP
71		QQP
72		QTT
73		TTT
74		COO
75		CAO
76		COP
77		CAA
78		CAP
79		CQQ
80		CQT
81		CTT
82		CCO
83		CCA
84		CCY
85		CCK
86		CCD
87		CCT
88		CCC
89	$-CH_2-$	OO
90		OY
91		OK
92		OD
93		OT
94		YY
95		YK
96		YD
97		YT
98		KK
99		KD
100		KT
101		DD
102		DT
103		TT
104		CO
105		CY
106		CK
107		CD
108		CT
109		CC
110[c]	CH_2=	
111[d]		
112[d]		
113[d]		
114[d]		
115		
116	CH≡C-	
117[e]	-CH=	
118[c]	-CH=	
119	HCOO-	O
120		Y
121		K
122		D
123		T
124		C
125	-OH	O
126		D
127		C
128	-COOH	O
129		Y
130		K
131		D
132		T
133		C
134	-CHO	Y
135		K
136		D
137		T
138		-C-
139		-CH-
140		$-CH_2-$
141		
142[e]	>C=	
143[c]	>C=	
144[e]	=C=	
145[c]	=C=	
146	-O-	
147		KK
148		KO
149		KY
150		KD
151		KT
152		KC
153		
154	-CO-	OO
155		OY
156		OK
157		OD
158		OT
159		OC
160		YY
161		YK
162		KK
163		DY
164		DK
165		DD
166		TY
167		TK
168		TD
169		TT
170		CY
171		CK
172		CD
173		CT
174		CC
175	O=C=	
176[d]		
177	-C≡C-	
178	-C-	
179[c]	[D]	

[a] The symbols O, Y, K, D, T, C, and A indicate oxygen, aromatic carbon, carbonyl carbon, SP^2 carbon, SP carbon, and acyl oxygen (O-CO-), respectively. The two other symbols P and Q indicate four (Y, K, D, and T) and three (Y, K, and T) kinds of adjacent groups, respectively.

[b] Pairs of methyl groups which will compose isopropyl groups.

[c] These components are used to build up various olefinic structures.

[d] For aromatic structures, the term 'aromatic' is defined arbitrarily a structure composed from n of there components, where n is 4, 6, 8, 10, ··· .

[e] These are for ketenic structures.

Table II. Analysis of IR Data

ICO	+1			0			-1		
IOH	+1	0	-1	+1	0	-1	+1	0	-1
ICOX	NO-1	NO	NO	NO-1	NO	NO	0	0	0
ICON	1	1	1	0	0	0	0	0	0
IOHX	NO-1	NO-1	0	NO	NO	0	NO	NO	0
IOHN	1	0	0	1	0	0	1	0	0
IOX	NO-2	NO-1	NO-1	NO-1	NO	NO	NO-1	NO	NO
ION	0	0	0	0	0	0	0	0	NO

bands appearing at the regions of 3700–3200 cm^{-1} (IS1), 1899–1700 cm^{-1} (IS2), and 1699–1500 cm^{-1} (IS3). The values IS1, IS2, and IS3 are calculated by the equation

$$\mathrm{IS}i = 512(1 - (T_i/100)) \qquad (1)$$

where, T_i is the percent transmittance of the strongest absorption band of the corresponding spectral region. Depending on the scores of IS1, IS2, and IS3, one of the numerals, 1, 0, and −1, is given for ICO, and IOH follows:

$$\begin{array}{lll} \mathrm{IOH} = 1 & \text{when} & \mathrm{IS1} > 270 \\ \quad 0 & & 270 \geqslant \mathrm{IS1} > 0 \\ \quad -1 & & \mathrm{IS1} = 0 \\ \mathrm{ICO} = 1 & \text{when} & \mathrm{IS2} \geqslant 390 \\ \quad 0 & & 270 \leqslant \mathrm{IS2} < 390 \text{ and/or } 270 \leqslant \mathrm{IS3} \\ \quad -1 & & \mathrm{IS2} < 270 \text{ and } \mathrm{IS3} < 270 \end{array}$$

These threshold values are arbitrarily determined based upon our experience. Numerals assigned to ICO and IOH indicate the presence of the corresponding functional groups; i.e., +1, 0, and −1 mean presence, uncertainty, and absence, respectively. Next, three pairs of parameters, ICOX/ICON, IOHX/IOHN, and IOX/ION, which indicate maximum and minimum numbers of CO, OH, and ethereal oxygen in a sample compound are determined by using ICO, IOH, and molecular formula, respectively (Table II). Only four parameters (ICOX, ICON, IOHX, and IOX) among these six parameters are used in the IR analysis step and the remainder in the later step (STRUCTURE GENERATOR). Five values of ICOX, IOHX, IOX, HD–ICON, and NO–ICON obtained from the IR spectrum of a sample dealt with the above-mentioned processes are compared with the intrinsic parameters of each component (refer to Table VI). If anyone of the parameters of the sample spectrum is smaller than those of the corresponding component, the component is eliminated as an inappropriate one.

The IR spectrum of compound I was analyzed by the above-mentioned procedures. Then 80 components were selected out of 103 components which survived on examination of the molecular formula alone (refer to Figure 4).

The third step of DATA ANALYSIS is analysis of the NMR spectrum. Since the methodology of NMR analysis has already been reported by the authors,[7] only a brief explanation is made here.

Grouping of the signals every 20 Hz or more is the first process of NMR analysis. The number of hydrogens allocated to each group is determined based on the area intensity of each signal and molecular formula of the sample compound. Referring the approved chemical shift range for each component (Table VI) to the position and number of hydrogens allocated to each signal group of the unknown, probable components and their maximum (MAX) and minimum (MIN) number possibly present in the unknown are computed in this process.

Analysis of spectral pattern, i.e., A_m and AX_n (m = 3, 6, and 9, and n = 3 and 6), is also performed. The results are represented with three sets of parameters (MAX(i), MIN(i) for each survived component which contains hydrogens), (LY(j), j = 1, JN), and NM matrix where JN is the number of signal groups. The LY(j) and NM matrix are prepared for use in the next step, i.e., STRUCTURE GENERATOR. The LY(j) represents allocated hydrogen numbers for the jth signal group. Each row of the NM matrix corresponds to the survived component and each column to each signal group. Each element of the matrix, i.e., AOC$_{ij}$, indicates the maximum number of the component assigned to the corresponding signal group.

The analysis of the NMR of compound I was executed by the above-mentioned method and 35 components (refer to Figure 4) survived through DATA ANALYSIS. Twenty-two hydrogen-containing components of them were given in the form of an NM matrix. Then, the vector for all the survived components and the matrix for hydrogen-containing components in the survived components were sent from DATA ANALYSIS to STRUCTURE GENERATOR as the data for the generation of informational homologues.

STRUCTURE GENERATOR

The structure generation procedure is executed in three steps as follows:

1. Selecting all the possible sets which satisfy the given molecular formula from the components picked out in DATA ANALYSIS. Actually, the selection is made within a range of the MIN and the MAX.

2. Making a check of whether the contents of each set is consistent with the inputted NMR spectrum; if not, the set is discarded. This procedure uses the NM matrix.

3. Using all possible combinations of components in the survived set to construct informational homologues.

1. Combination of Numbers of Components. Theoretically, all possible combinations which satisfy a given molecular formula can be generated by varying every element of a 178 dimensional vector one by one from MIN(i) to MAX(i), where i = 1 to 178,[17] as shown in (I), but this is very wasteful and

$$\begin{array}{llllllllll} i & 1 & 2 & 3 & 4 \text{---------} & 174 & 175 & 176 & 177 & 178 \\ \mathrm{MIN} & 0 & 0 & 0 & 0 \text{---------} & 0 & 0 & 0 & 0 & 0 \\ & 1 & 0 & 0 & 0 \text{---------} & 0 & 0 & 0 & 0 & 0 \\ & \cdot & \cdot & \cdot & \cdot \cdots\cdots & \cdot & \cdot & \cdot & \cdot & \cdot \\ & \cdot & \cdot & \cdot & \cdot \cdots\cdots & \cdot & \cdot & \cdot & \cdot & \cdot \\ \mathrm{MAX} & a & b & c & d \cdots\cdots & v & w & x & y & z \end{array} \qquad (\mathrm{I})$$

time-consuming for the computer. To improve the efficiency of the process, 179 components are hierarchically summarized into two kinds of component groups, secondary and the primary components; there are 37 secondaries and 7 primaries as shown in Table III. The original 179 components are called tertiary, and the chemical elements, carbon, hydrogen, and oxygen, are called elementary. All the sets of components used to construct structures are represented in the form of a vector, in terms of which the procedure of the combination of the number of components may be described.

There are four kinds of vectors in the system, namely, elementary, primary, secondary, and tertiary, and they form a series of parents and daughters in turn; e.g., an elementary vector is a parent of primary ones.

The relation is expressed by the general equation:

$$(\mathrm{DCV})(\mathrm{DCM}) = (\mathrm{PCV}) \qquad (2)$$

where DCV and PCV are the daughter and parent vectors, respectively, and DCM stands for a component matrix, which defines the daughter components in terms of the parent components. Each row of the matrix corresponds to each component of the parent vector and each column to that of the daughter vector. For calculation of the vectors i.e., boundary conditions, upper and lower limits of the elements of vectors are required. They are also prepared in the form

Table III. Relations between Three Groups of Components

		number of primary components						
tertiary	secondary	OH	O	CH_3	CH_2	CH	C	HD*
1 - 6	22	0	0	3	0	0	1	0
7 - 12	23	0	0	2	0	0	1	0
13 - 18	27	0	0	1	0	0	1	0
19 - 25	24	0	0	2	0	0	0	0
26 - 31	20	0	1	1	0	0	0	0
32 - 34	21	0	0	1	0	0	0	0
35 - 40	19	0	1	1	0	0	1	2
41 - 46	26	0	0	1	0	0	0	0
47 - 61	25	0	0	1	0	0	0	0
62 - 88	18	0	0	0	0	1	0	0
89 - 109	13	0	0	0	1	0	0	0
110	10	0	0	0	1	0	0	1
111	3	0	2	0	1	0	2	4
112	1	1	1	0	0	0	2	3
113	2	1	0	0	0	0	1	1
114	5	0	0	0	0	1	0	1
115	15	0	1	0	0	1	2	6
116	17	0	0	0	0	1	1	4
117	7	0	0	0	0	1	0	1
118	11	0	0	0	0	1	0	1
119 - 124	14	0	2	0	0	1	0	2
125 - 127	29	1	0	0	0	0	0	0
128 - 133	28	1	1	0	0	0	1	2
134 - 140	16	0	1	0	0	1	0	2
141	30	0	1	0	0	0	3	6
142	9	0	0	0	0	0	1	1
143	12	0	0	0	0	0	1	1
144	34	0	0	0	0	0	1	2
145	35	0	0	0	0	0	1	2
146	4	0	1	0	0	0	0	2
147 - 153	32	0	1	0	0	0	0	0
154 - 174	31	0	1	0	0	0	1	2
175	8	0	1	0	0	0	1	3
176	6	0	0	0	0	0	1	1
177	33	0	0	0	0	0	2	4
178	36	0	0	0	0	0	1	0
179	37	0	0	0	0	0	0	0

* The value of HD is the twice value of the actual index of hydrogen defficiency for simplifying.

of vectors, LPN and LPX, for minimum and maximum number of the primary vector elements and they are obtained from LSN and LSX which mean minimum and maximum number of secondary vector elements, respectively. These LSN and LSX are detemined from MIN and MAX which are afforded by the DATA ANALYSIS.

Elementary Component Vector. The chemical elements constitute the elementary components and their number (molecular formula) form the elementary vector [NC NH NO], where NC, NH, and NO correspond to the number of carbon, hydrogen, and oxygen atoms, respectively.

Primary Component Vector. The primary component vector [LP] is composed of seven elements, but the seventh is separated from the other six because of its nature. Its number is equal to the HD, the index of hydrogen deficiency.

On the basis of eq 2, [LP]'s are obtained from an elementary component vector. For example, from an elementary vector [2 6 1] which corresponds to a molecular formula C_2H_6O, DCV (primary vector in this case) is given to solve eq 3.

$$\underset{[\mathrm{LP}]}{\overset{\mathrm{OH}\ \ \mathrm{O}\ \ \mathrm{CH_3}\ \mathrm{CH_2}\ \mathrm{CH}\ \ \mathrm{C}}{[\ X_1\ X_2\ \ X_3\ \ \ X_4\ \ \ X_5\ X_6]}}\ \underset{[\mathrm{LPM}]}{\begin{bmatrix} 0\,1\,1 \\ 0\,0\,1 \\ 1\,3\,0 \\ 1\,2\,0 \\ 1\,1\,0 \\ 1\,0\,0 \end{bmatrix}} = \underset{[\mathrm{DCV}]}{\overset{\mathrm{C\ H\ O}}{[2\ 6\ 1\ ]}} \tag{3}$$

Boundary condition LPN = [0 0 0 0 0 0]; LPX = [1 1 2 2 2 2]

Unfortunately, there is no general procedure to solve this equation because only three equations are derived for six variables. This is the usual case in the system; therefore, to get an appropriate answer for the vector DCV, each X_i is substituted one by one with values LPN(i) through LPX(i) as follows:

$$\overset{\mathrm{LPN}}{[0\ 0\ 0\ 0\ 0\ 0]} \rightarrow [0\ 0\ 0\ 0\ 0\ 1] \rightarrow \underline{[0\ 1\ 2\ 0\ 0\ 0]} \rightarrow$$

$$[0\ 1\ 2\ 0\ 0\ 1] \rightarrow \underline{[1\ 0\ 1\ 1\ 0\ 0]} \rightarrow$$

$$[1\ 0\ 1\ 1\ 0\ 1] \rightarrow [1\ 1\ 2\ 2\ 2\ 1] \rightarrow \underset{\mathrm{LPX}}{[1\ 1\ 2\ 2\ 2\ 2]}$$

Among all the possibilities, only the two vectors underlined satisfy eq 3; they mean [O, $(CH_3)_2$] and [CH_3, CH_2, OH], respectively.

Secondary and Tertiary Component Vector. The procedure to obtain the secondary vector [LS] from [LP] and [LSM] is the same as for the primary ones. The tertiary component vector [LT] is derived from parent [LS] and [LTM] in the same manner.

2. NMR Consistency Check. As described previously, the NMR signal groups are treated as if they are independent of each other and the component which can be assigned to at least one signal group survives without any further examination at the DATA ANALYSIS. However, it is necessary to examine whether the set [LT] is consistent with a given NMR spectrum or not; in other words, all the components in the set that are properly assigned to all signal groups without any excess or any deficiency should be confirmed.

3. Generation of the Informational Homologues. The number of double bonds is calculated as

$$\mathrm{LT}(179) = [\mathrm{LT}(110) + \mathrm{LT}(118) + \mathrm{LT}(143)]/2 + \mathrm{LT}(145) \tag{4}$$

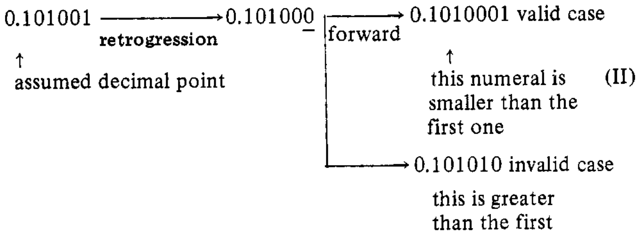
The total number of components of an LT is represented by a parameter NAT. The NAT components are numbered from one to NAT according to the hierarchical order settled for empirically for improving the efficiency of the structure generation. If the number of components of the highest order is n, they are numbered from 1 to n. Components of the next highest, whose number is m, are numbered from $n + 1$ to $n + m$, and so on.

The connectivities between the constituents are described in the form of a connectivity stack,[11] which is a series of local connectivities: a connectivity between the ith and the jth components forms the $[i + (j - 1)(j - 2)/2]$th position of the stack. The double bond in olefins, allenes, and cumulenes is set as component 179. All other kinds of double bonds and triple bonds are implied in other components, e.g., 119, 128, and 154 for double bond and 116 and 177 for triple bond. The elements of a stack should be represented with 0 or 1.

First of all, the second and the first members of the set ([LT]) are picked up and examined whether they can make a valid connectivity or not. Generally, the jth and ith ($i = 1$ to $j - 1$) components are picked up, and their connectivity validity is examined. This operation means an examination of the $[i + (j - 1)(j - 2)/2]$th element of the stack. If not, the element is set to zero and the next element is examined. The creation of the stack with (NAT − 1)NAT/2 elements means the creation of one of the informational homologues. At the time of creation, the stack cannot be extended, so it retrogrades.

The principle of retrogression is that the nonzero element which will appear after retrogression should be at the later position than the position occupied by the latest element which was converted to zero; in other words, assuming the stack to be a decimal numeral, it should become smaller after retrogression as shown in (II). With this rather simple rule, fast and exhaustive enumeration of structural isomers (informational homologues) is performed without any duplication and omission.[12]

The retrogression continues until the last nonzero element is converted to zero. Sooner or later, nonzero elements are



gradually shifted to later positions, and the stack vanishes; this is the end point of the enumeration. Whether two components may be connected with each other or not is examined by using several parameters defined for every component.

Five informational homologues were totally generated for the input data (Figure 3) of compound I. The structures of the informational homologues are shown in Figure 5 (the underlined one is the structure of compound I).

DATA COMPARISON

Data comparison has the following functions: (a) retrieves the data corresponding to each informational homologue from the data file, (b) compares the data with those of the sample, and (c) shows the results in the form of comments.

1. Data Conversion for Comparison. NMR.[13] The spectrum is expressed with a set of two values, G and S, which are the center of gravity of the whole spectrum and the standard deviation of each signal to the center, respectively, according to the equations

$$G = \sum_{i=1}^{n} (\omega_i A_i) / \sum_{i=1}^{n} A_i \tag{5}$$

$$S = \left(\sum_{i=1}^{n} (\omega_i - G)^2 A_i / \sum_{i=1}^{n} A_i \right)^{1/2} \tag{6}$$

where A_i stands for the intensity of the ith signal and ω_i for the position of the ith signal (ref. Me_4Si).

IR. At first, the spectrum is divided into 18 blocks, at 3200, 2800, 2300, 2000, 1900, 1800, 1700, 1600, 1500, 1400, 1300, 1200, 1100, 1000, 900, 800, and 700 cm^{-1}. The ith block is expressed with a value $100I_i + P_i$. The values of I_i and P_i indicate the intensity and position of the highest peak in the block, respectively. The intensity parameter, I_i, is determined according to the relative absorbance, A_i, which is a ratio of absorbance,

$$A_i = -\log (T_i/100) = 2 - \log T_i, \text{ to the largest } A$$

$I = 0$ for $A_i = 0.0$, $I = 1$ for $0.0 < A_i \leq 0.3$, $I = 2$ for $0.3 < A_i \leq 0.7$, and $I = 3$ for $A_i > 0.7$. The P_i is obtained as the subdivision number (1 to 10) which indicates the position of the strongest absorption in the block. The value 0 for P_i means that there is no peak in the corresponding block.

Mass. The spectrum is divided in a set of 14 ms from m/e 6, and the two most intense peaks are picked up from every block.[14] They are arranged in order of ms together with the intensities which are expressed with parameters I_i's: $I = 1$ for the so-called relative intensity less than 35%, $I = 2$ for the intensity between 35 to 75%, and $I = 3$ for the intensity over 75%.

2. Data Retrieval. The data of each informational homologue are retrieved by means of three kinds of keywords, i.e., molecular formula, a set of components, and connectivities between the components. The keyword works properly stepwise according to the three elements of representation, and this makes the efficiency of retrieval much better. In each step, a particular list of structures is prepared, and in the next step, only the resulting list is scanned.

3. Data Comparison. Whether two spectra, say X and Y, clearly mismatch or not is examined.

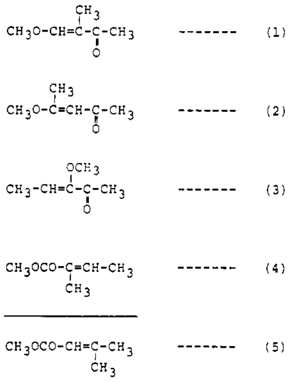


Figure 5. Resulted structures.

NMR. For a set of two values, G and S, if $|G_X - G_Y|$ is greater than 0.03, and/or if $|S_X - S_Y|$ is greater than 0.04, the two are determined clearly mismatching.

IR. A value, E, which is defined by eq 7 is used to carry

$$E = \sum_{i}^{18} |I_{xi} - I_{yi}| \tag{7}$$

out the sorting of IR spectra. Here the suffixes x and y mean the spectra X and Y, respectively. Another value, M, is also defined to represent the number of blocks where the P_{xi} and P_{yi} are not identical. If M is greater than 6, and/or if E is greater than 20, the two are determinced clearly mismatching.

MS. If, for every unregistered m/e value, its $I_{m/e}$ is assigned to zero for simplification of the explanation, the spectra would have the form [$I_{m/e}$: (m/e 6, 7, . . .)]. After this, two parameters, P and D, are set to zero. (a) When $(I_{m/e})_X$ and $(I_{m/e})_Y$ are both unregistered, or (b) both are registered and equal to each other, P and D are not changed. (c) When one $(I_{m/e})_Y$ is registered but unequal to the other, D is increased by one. If P is greater than 7, and/or if D is greater than 5, the two spectra are determined clearly mismatching.

4. Results of the Comparison. The results are not directly used for elimination of "unplausible" structures, because the structure representation is kept at the level of a structural isomer (i.e., not stereoisomers), and more than one stereoisomer belonging to the same class of the structural isomer often gives quite different features of the spectra. CHEMICS neither can control the conditions nor deal with checking them, and even if their difference can be detected, it would be difficult to predict the effect of the difference in all cases.

Therefore, the result of the data comparison is shown in the form of comments. The comments are given for the following three cases: (i) a structure of the informational homologue is not registered in the file, so the data comparison cannot be carried out; (ii) the corresponding data are found in the file and do not clearly mismatch with the data of the sample; and (iii) the two data clearly mismatch. For the case of (i), no comment is given, for (ii) the structure is most probable, and for (iii) the structure may be wrong.

Each of five structures (1, 2, 3, 4, and 5 in Figure 5) generated by CHEMICS for compound I was compared with the storage of file (F). Structures 1 to 3 were not found in the file; thus no comparison was carried out. Structures 4 and 5 were found in the file and the filed data for the former were well-matched for the input data, but the data for the latter were not. The conclusion afforded by the structure generation (CHEMICS) followed by the file searching (F) for compound I is that structure 4 is the most probable, 5 is the most unprobable, and the rest (1, 2, and 3) still survive as candidates because no comparison of data has been carried out. Tables IV and V summarize the results of the structure elucidation computed by using the CHEMICS and CHEMICS-F systems,

Table IV. The Amounts of Informational Homologues for a Variety of Compounds by CHEMICS

Compounds	molecular formula			Amount of Informational Homologues
	C	H	O	
2,4-Dimethyl-1,3-dioxane	6	12	2	72
2-Methylpentane	6	14		3
3-Methylpentane	6	14		3
2,3-Dimethylbutane	6	14		4
Anisole	7	8	1	5
o-Cresol	7	8	1	62
o-Xylene	8	10		40
m-Xylene	8	10		40
p-Xylene	8	10		40
Ethylbenzene	8	10		5
2-Phenoxyethanol	8	10	2	143
Cyclohexyl methyl ketone	8	14	1	30
α-Methylstyrene	9	10		110
Phenyl allyl ether	9	10	1	181
Propiophenone	9	10	1	28
p-Cymene	10	14		305
Borneol	10	18	1	353
n-Decanol	10	22	1	50
β-Ethylnaphthalene	12	12		79

Table V. The Amounts of Informational Homologues for Variety of Compounds by CHEMICS-F

Compounds	molecular formula			generated structures	results of file retrieval		
	C	H	O		matched	mismatched	not found
Methyl tiglate (I)	6	10	2	5	1	1	3
2-Methyl-5-hexanone	7	14	1	7	1	1	5
2,4-Dimethyl-3-pentanone	7	14	1	1	1	0	0
2-Heptanone	7	14	1	1	1	0	0
3-Heptanone	7	14	1	2	1	1	0
4-Heptanone	7	14	1	2	1	1	0
2-Heptanone	7	16	1	16	1	2	13
3-Heptanone	7	16	1	11	1	2	8
4-Heptanone	7	16	1	3	1	1	1
1,1-Diethylpropanol	7	16	1	6	1	1	4
1,1-Dimethylpentanol	7	16	1	3	1	1	1
2-Octanone	8	16	1	2	1	0	1
3-Octanone	8	16	1	12	1	0	11

respectively, for a variety of compounds which were likened to unknowns.

CONCLUSION

Among several automated structure elucidation systems, the strategies for searching substructures used as building blocks can be classified into two major categories. The first one is that the presence of the predefined substructures are determined by an automatic interpreter as actualized in the CHEMICS-F system, and the other is that the chemists (users) should provide selected substructures for the system; CASE[9] and CONGEN[10] have employed this strategy.

The reliability of answers by using the system which employs the latter strategy strongly depends upon the user's experience in chemistry, because a wrong input results in a wrong answer. On the contrary, the system with the former strategy does not require the user to have any chemical experience. The reliability depends upon solely how properly the substructures are defined.

The system can cover the structure elucidation of any organic compound with C, H, and O. Especially, the structure construction works perfectly to enumerate all the structures corresponding to the molecular formula and the partial structures (components) afforded by chemical spectra of an unknown. Should the analyses of the chemical spectra give good partial structures in quality and quantity, one can expect to obtain one correct structure in a moment. Also the function of file search is proven to be excellent in giving three ranking marks to the candidate structures.

Presently the analyses of spectra are not performed deeply and accurately enough. Only poor information—the presence of OH, CO, and ethereal oxygen—is given by IR analysis, and the use of MS has not been realized. Further, though the file search functions excellently, the storage—only 600 structures and their MS, NMR, and IR—is too few to make the system practical and pragmatic.

Thus we are endeavoring to raise the accuracy of IR analysis and plan to introduce the MS analysis as another powerful weapon in the system. In addition to that, C-13 NMR analysis as a new information source was added to the system and some of the new results have been reported by ACS[15] and in other meetings. The system which is rather hard to operate easily should be remodeled into an interactive one by which any kind of information can be input at any step of the structure elucidation procedure. Building up any structure freely as one pleases by inputting partial structures, such as nonyl, octyl, phenyl, benzyl, and so on, through man-machine conversation is one of the important functions to be added to the present system.

APPENDIX

1. Parameters for Components. As shown in Table VI, several parameters are defined for all components and they are used in the DATA ANALYSIS and the STRUCTURE GENERATOR. The first three parameters, NC, NO, and HD, express the necessary conditions for the existence of a component in a molecule. They mean the number of carbons and oxygens, and the value of index of hydrogen deficiency, respectively. The fourth, IQG, expresses the efferent nature of a component. The numerals 1 through 6 correspond to oxygen (O), aromatic (Y), carbonyl (K), olefinic (D), acetylenic (T), and saturated (C) carbons, respectively. The value 8 requires special treatment in the program; 9 and 10 are prepared for the components which compose olefinic and aromatic structures but have no additional bonds with other components which have IQG's 4 (D) or 2 (Y).

On the other hand, the afferent nature, IQT the fifth, is rather complicated because most components have plural bonds, and they are required to combine with different components of a different nature. Explanations for them are summarized in Table VII. The next two parameters IBF and IBG correspond to the number of bonds and the number of bonds with the efferent nature, respectively. The following five parameters are for the IR data analysis and the last two are for NMR data analysis.

2. Implementation of NMR Data Analysis. After the grouping of signals, intensities (INTC) of signals whose positions are within the approval range of chemical shift of a component are summarized for a signal group. Then, the amount of *i* component (AOC_{ij}) for the *j*th signal group is calculated by using the value of summarized INTC and the signal intensity for unit proton (INTUP) which is evaluated by molecular formula and summation of intensities of the entire signals. Finally, MAX is expressed by the integer form of AOC for each signal group. This procedure is expressed by eq 8, 9, and 10.

$$\mathrm{INTUP} = \Sigma\mathrm{INT}/\mathrm{NH} \qquad (8)$$

Here INT and NH represent intensity of each signal and the number of hydrogens in the molecule, respectively.

$$\mathrm{AOC}_{ij} = \Sigma\mathrm{INTC}/(\mathrm{INTUP} \times \mathrm{PAR}) \qquad (9)$$

Here PAR is the proton number of component *i*.

$$\mathrm{MAX}_i = \Sigma\ \mathrm{AOC}_{ij} \qquad (10)$$

For compound I, 11 signals of the NMR data (Figure 3) were grouped into three, i.e., signals 1 through 5 as first group, number

	COMPONENT	MAX	MIN	NM matrix		
1	28	1	0	0	1	0
2	29	1	0	0	1	0
3	33	2	0	0	0	2
4	35	1	0	0	0	1
5	37	1	0	0	0	1
6	38	1	0	0	0	1
7	40	1	0	0	0	1
8	67	4	0	0	4	0
9	76	4	0	0	4	0
10	78	1	0	1	0	0
11	79	4	0	0	4	0
12	82	4	0	0	4	0
13	85	6	0	0	4	3
14	88	6	0	0	4	6
15	98	2	0	0	2	0
16	99	2	0	0	2	0
17	104	2	0	0	2	0
18	106	3	0	0	0	3
19	107	4	0	0	0	4
20	109	4	0	0	0	4
21	117	1	0	1	0	0
22	118	1	0	1	0	0
			LY	[1	3	6]

Figure 6. NM matrix and vectors MAX and MIN.

6 as second, and 7 through 11 as third group. Then, 1, 3 and 6 hydrogens were allocated for these three groups, respectively. So the vector **LY** is [1, 3, 6] and the value of JN is 3.

Secondly, 44 components of which approved chemical shift ranges did not accord with any of the signals were eliminated as shown in Figure 4. Out of the 35 components survived, 22 hydrogen-containing components were given in the forms of MAX vector and NM matrix according to the eq 9 and 10.

As an example, consider component 28 which survived through this stage. Since the chemical shift range of this component has been evaluated as 246.0 to 210.0 Hz (refer to Table VI), only the second signal group which contains one signal, the 6th, is considered as a candidate to be assigned to the component. Therefore, the subscripts i and j in eq 9 are 1 and 2, respectively.

The value of INTUP is calculated as 915.9 (9159/10) by eq 8 and that of $AOC_{1,2}$ is equal to 1.07 [2646/(915.9 × 2.7)] by eq 9 where summarized INTC for the second signal group is directly placed as 2646 and PAR has the value 2.7, reduced by approximately 10% from the ideal value, 3.0.

Since any other value of $AOC_{1,j}$ (j = 1 or 3) is zero, the value of MAX_1 is determined as one by eq 10. In this manner, all other elements of the vector and the matrix were determined as shown in Figure 6; they were sent to STRUCTURE GENERATOR with an additional vector MIN which indicates the minimum number of components. Each element of the vector is usually set to zero.

3. Generation of LT Vectors. As described before, the first step in the structure generation is the derivation of all the possible component sets represented by the vector form LT's, which satisfy the given molecular formula and conditions MAX and MIN which were given at previous step. During the formation of [LT], MA(i) and ME(i), which are the number of bonds with the ith afferent and efferent natures, respectively, are counted out. The afferent nature of a component is implied in a value of IQT; e.g., the IQT whose value, 274 (=256 × 1 + 16 × 1 + 2) means two [O]'s and one [Y]. Therefore, each element of the MA(i) (i = 1, 6) is summed up by decoding the IQT of the every member of a [LT].

On the other hand, the elements of ME(i) (i = 1, 6) are obtained by applying eq 11 to 16 from a parent [LS], because the values of

$$ME(1) = LS(14) + LS(20) + LS(29) + LS(32) \times 2 \quad (11)$$

$$ME(2) = LS(6) + LS(15) + LS(30) \times 2 \quad (12)$$

$$ME(3) = LS(16) + LS(28) + LS(31) \times 2 \quad (13)$$

$$ME(4) = LS(7) + LS(9) \times 2 + LS(11) + LS(12) \times 2 \quad (14)$$

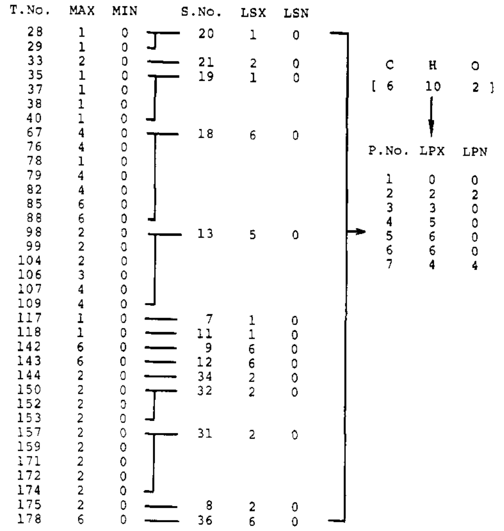
$$ME(5) = LS(17) + LS(33) \times 2 \quad (15)$$

$$ME(6) = LS(13) \times 2 + LS(18) \times 3 + LS(22) + LS(23) \times 2 + LS(24) \times 2 + LS(25) + LS(26) + LS(27) \times 3 + LS(36) \times 4 \quad (16)$$

the IQG for the tertiary components combined in one secondary component are the same (refer to Table III). These two new vectors, [MA] and [ME], are used for checking whether a generated [LT] could afford any structures. That is, every MA(i) should not be greater than the corresponding ME(i) as follows.

(a) A set composed of two components, 1[t-Bu(O)] and 2[t-Bu(Y)], is denied because [ME] and [MA] are [0 0 0 0 0 2] and [1 1 0 0 0], respectively, where ME(1) and ME(2) are smaller than MA(1) and MA(2), respectively.

(b) A set composed of two components, 6[t-Bu(C)] and 125 [(O)OH], should be denied, but their contradicting character for connection cannot be detected yet at this step because [ME] and [MA] are both [1 0 0 0 0 1]. The potential contradiction will be disclosed later.



T.No.	MAX	MIN	S.No.	LSX	LSN
28	1	0	20	1	0
29	1	0			
33	2	0	21	2	0
35	1	0	19	1	0
37	1	0			
38	1	0			
40	1	0			
67	4	0	18	6	0
76	4	0			
78	1	0			
79	4	0			
82	4	0			
85	6	0			
88	6	0			
98	2	0	13	5	0
99	2	0			
104	2	0			
106	3	0			
107	4	0			
109	4	0			
117	1	0	7	1	0
118	1	0	11	1	0
142	6	0	9	6	0
143	6	0	12	6	0
144	2	0	34	2	0
150	2	0	32	2	0
152	2	0			
153	2	0			
157	2	0	31	2	0
159	2	0			
171	2	0			
172	2	0			
174	2	0			
175	2	0	8	2	0
178	6	0	36	6	0

C	H	O
[6	10	2]

P.No.	LPX	LPN
1	0	0
2	2	2
3	3	0
4	5	0
5	6	0
6	6	0
7	4	4

Figure 7. Derivation of LSN, LSX, LPN, and LPX.

$$[\,x_1\ x_2\ x_3\ x_4\ x_5\ x_6\,] \begin{array}{ccc} C & H & O \\ \left[\begin{array}{ccc} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 3 & 0 \\ 1 & 2 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \end{array}\right] \end{array} = [\,6 \quad 10 \quad 2\,]$$

```
boundary condition ─┬─ LPX = [ 0    2    3    5    6    6 ]
                    │          X1   X2   X3   X4   X5   X6
                    └─ LPN = [ 0    2    0    0    0    0 ]
```

LP 1 =	0	2	0	4	2	0	4
2 =	0	2	0	5	0	1	4
3 =	0	2	1	2	3	0	4
4 =	0	2	1	3	1	1	4
5 =	0	2	2	0	4	0	4
6 =	0	2	2	1	2	1	4
7 =	0	2	2	2	0	2	4
8 =	0	2	3	0	1	2	4

Figure 8. Generation of LP's.

The following example illustrates the above-mentioned procedures for compound I. The parametric vectors MIN and MAX given for this compound are shown in Figure 6. Figure 7 shows the process of derivation of constant vectors [LSN], [LSX], [LPN], and [LPX], and the resulted vectors. LSX(13) was determined as 5 with reference to MAX(98) through MAX(109) and the given molecular formula, and so on. Figure 8 shows the equation for obtaining primary component vectors and resulted vectors; as indicated in this figure, a total of 8 [LP]'s were generated for the compound. They are all the possible combinations of primary components which are consistent with given molecular formula under the conditions restricted with [LPN] and [LPX]. From those 8 [LP]'s, 64 [LS]'s were generated and 197 [LT]'s were derived from the [LS]'s. For example, from the 8th [LP], 14 [LS]'s were generated and a total of 11 [LT]'s were derived from the [LS]'s as shown in Figure 9. The [LP] consists of 2[O], 3[CH_3], 1[CH], and 2[C], and the sum totals of carbon, hydrogen, and oxygen atoms are, of course, consistent with the molecular formula. Among the 14 [LS]'s the first (LS51), the second (LS52), and the 8th (LS58) to the 14th (LS64) vectors also afford several candidate [LT]'s but all of them were discarded by the ME and MA check. The fourth [LS] in the figure consists of 1 [-CH=], 1 [C=], 1 [CH_3O-], 2 [CH_3(U)]'s (the symbol U include Y, K, D, and T), and 1 [CO]. Then the [LT] consisting of [CH_3O(K)], [CH_3(D)], [-CH=], [C=], and [(O)CO(D)] is derived from the [LS]. The [ME] and the [MA] vectors of this [LT] are [1 0 1 2 0 1] and [1 0 1 2 0 0], respectively, and these values satisfy the condition of efferent and afferent nature mentioned before. Other ten [LT]'s also pass this check and are sent to the next step.

Table VI. Intrinsic Parameters for Component

#	NC	HD	NO	IQG	IQT	IBF	IBG	IOHX	IOX	NO-ICON	ICOX	HD-ICON	chemical shift(H_z)	
1	4	0	1	6	1	1	1	0	0	1	0	0	84.0	60.0
2	8	3	0	6	2	1	1	0	0	0	0	3	96.0	72.0
3	5	1	1	6	3	1	1	0	0	0	1	0	90.0	60.0
4	6	1	0	6	4	1	1	0	0	0	0	1	90.0	54.0
5	6	2	0	6	5	1	1	0	0	0	0	2	90.0	54.0
6	5	0	0	6	6	1	1	0	0	0	0	0	66.0	36.0
7	4	0	1	6	129	2	2	0	0	1	0	0	84.0	48.0
8	8	3	0	6	130	2	2	0	0	0	0	3	84.0	66.0
9	5	1	1	6	131	2	2	0	0	0	1	0	90.0	54.0
10	6	1	0	6	132	2	2	0	0	0	0	1	90.0	48.0
11	6	2	0	6	133	2	2	0	0	0	0	2	90.0	48.0
12	5	0	0	6	134	2	2	0	0	0	0	0	89.4	36.0
13	2	0	1	6	129	3	3	0	0	1	0	0	90.0	48.0
14	8	3	0	6	130	3	3	0	0	0	0	3	108.0	60.0
15	5	1	1	6	131	3	3	0	0	0	1	0	84.0	42.0
16	6	1	0	6	132	3	3	0	0	0	0	1	84.0	42.0
17	6	2	0	6	133	3	3	0	0	0	0	2	84.0	42.0
18	5	0	0	6	134	3	3	0	0	0	0	0	90.0	42.0
19	3	0	1	6	4097	1	2	0	0	1	0	0	84.0	54.0
20	4	1	2	6	4103	1	2	0	1	1	1	0	90.0	54.0
21	7	3	0	6	4098	1	2	0	0	0	0	3	90.0	54.0
22	4	1	1	6	4099	1	2	0	0	0	1	0	90.0	48.0
23	5	1	0	6	4100	1	2	0	0	0	0	1	90.0	48.0
24	5	2	0	6	4101	1	2	0	0	0	0	2	90.0	48.0
25	4	0	0	6	4102	1	2	0	0	0	0	0	84.0	30.0
26	1	0	2	1	1	1	1	0	1	2	0	0	210.0	186.0
27	5	3	1	1	2	1	1	0	1	1	0	3	246.0	210.0
28	2	1	2	1	3	1	1	0	1	1	1	0	246.0	210.0
29	3	1	1	1	4	1	1	0	1	1	0	1	246.0	210.0
30	3	2	1	1	5	1	1	0	1	1	0	2	246.0	210.0
31	2	0	1	1	6	1	1	0	1	1	0	0	210.0	186.0
32	5	3	0	8	2	1	1	0	0	0	0	3	168.0	120.0
33	3	1	0	8	4	1	1	0	0	0	0	1	144.0	90.0
34	3	2	0	8	5	1	1	0	0	0	0	2	132.0	108.0
35	2	1	2	3	1	1	1	0	1	1	1	0	150.0	108.0
36	6	4	1	3	2	1	1	0	0	0	1	3	150.0	108.0
37	3	2	2	3	3	1	1	0	0	0	2	0	150.0	108.0
38	4	2	1	3	4	1	1	0	0	0	1	1	150.0	108.0
39	4	3	1	3	5	1	1	0	0	0	1	2	150.0	108.0
40	3	1	1	3	6	1	1	0	0	0	1	0	150.0	108.0
41	2	0	1	6	4097	1	1	0	0	1	0	0	89.0	54.0
42	6	3	0	6	4098	1	1	0	0	0	0	3	90.0	54.0
43	3	1	1	6	4099	1	1	0	0	0	1	0	90.0	48.0
44	4	1	0	6	4100	1	1	0	0	0	0	1	90.0	48.0
45	4	2	0	6	4101	1	1	0	0	0	0	2	90.0	48.0
46	3	0	0	6	4102	1	1	0	0	0	0	0	84.0	30.0
47	2	0	2	6	4097	2	1	0	0	2	0	0	90.0	54.0
48	3	1	3	6	4119	2	1	0	1	2	1	0	90.0	30.0
49	3	1	1	6	4121	2	1	0	0	1	0	0	90.0	30.0
50	4	2	4	6	4103	2	1	0	2	2	2	0	90.0	30.0
51	4	2	2	6	4217	2	1	0	1	1	1	0	90.0	30.0
52	4	2	0	6	4232	2	1	0	0	0	0	0	90.0	30.0
53	5	3	0	6	4229	2	1	0	0	0	0	2	90.0	30.0
54	6	4	0	6	4101	2	1	0	0	0	0	4	90.0	54.0
55	3	0	1	6	4118	2	1	0	0	1	0	0	84.0	54.0
56	4	1	2	6	4214	2	1	0	1	1	1	0	90.0	30.0
57	7	3	0	6	4134	2	1	0	0	0	0	3	90.0	30.0
58	4	1	1	6	4150	2	1	0	0	0	1	0	90.0	30.0
59	5	1	0	6	4166	2	1	0	0	0	0	1	90.0	30.0
60	5	2	0	6	4182	2	1	0	0	0	0	2	90.0	30.0
61	4	0	0	6	4102	2	1	0	0	0	0	0	90.0	30.0
62	1	0	3	6	1	3	3	0	0	3	0	0	420.0	210.0
63	2	1	4	6	279	3	3	0	1	3	1	0	432.0	330.0
64	2	1	2	6	281	3	3	0	0	2	0	0	390.0	210.0
65	3	2	5	6	375	3	3	0	2	3	2	0	480.0	390.0
66	3	2	3	6	377	3	3	0	1	2	1	0	432.0	324.0
67	3	2	1	6	409	3	3	0	0	1	0	0	360.0	210.0
68	4	3	6	6	7	3	3	0	3	3	3	0	492.0	390.0
69	4	3	4	6	1913	3	3	0	2	2	2	0	480.0	390.0
70	4	3	2	6	1945	3	3	0	1	1	1	0	432.0	344.0
71	4	3	0	6	2185	3	3	0	0	0	0	0	360.0	210.0
72	6	5	0	6	2133	3	3	0	0	0	0	4	360.0	120.0
73	7	6	0	6	5	3	3	0	0	0	0	6	156.0	60.0
74	2	0	2	6	278	3	3	0	0	2	0	0	390.0	210.0
75	3	1	3	6	374	3	3	0	1	2	1	0	432.0	330.0
76	3	1	1	6	406	3	3	0	0	1	0	0	360.0	210.0
77	4	2	4	6	1910	3	3	0	2	2	2	0	480.0	390.0
78	4	2	2	6	1942	3	3	0	1	1	1	0	420.0	324.0
79	4	2	0	6	2182	3	3	0	0	0	0	0	360.0	210.0
80	5	3	0	6	2134	3	3	0	0	0	0	2	360.0	120.0
81	6	4	0	6	1366	3	3	0	0	0	0	4	156.0	60.0
82	3	0	1	6	358	3	3	0	0	1	0	0	327.7	144.0
83	4	1	2	6	1894	3	3	0	1	1	1	0	372.0	270.0
84	7	3	0	6	614	3	3	0	0	0	0	3	240.0	138.0
85	4	1	1	6	870	3	3	0	0	0	1	0	240.0	110.0
86	5	1	0	6	1126	3	3	0	0	0	0	1	204.0	120.0
87	5	2	0	6	1382	3	3	0	0	0	0	2	156.0	60.0
88	4	0	0	6	6	3	3	0	0	0	0	0	222.0	0.0
89	1	0	2	6	1	2	2	0	0	2	0	0	300.0	252.0
90	5	3	1	6	18	2	2	0	0	1	0	3	320.0	252.0
91	2	1	2	6	19	2	2	0	0	1	1	0	324.0	240.0
92	3	1	1	6	20	2	2	0	0	1	0	1	318.0	240.0
93	3	2	1	6	21	2	2	0	0	1	0	2	312.0	228.0
94	9	6	0	6	2	2	2	0	0	0	0	6	252.0	210.0
95	6	4	1	6	35	2	2	0	0	0	1	3	252.0	192.0
96	7	4	0	6	36	2	2	0	0	0	0	4	246.0	192.0
97	7	5	0	6	37	2	2	0	0	0	0	5	246.0	192.0
98	3	2	2	6	3	2	2	0	0	0	2	0	240.0	162.0
99	4	2	1	6	52	2	2	0	0	0	1	1	240.0	150.0
100	4	3	1	6	53	2	2	0	0	0	1	2	264.0	192.0
101	5	2	0	6	4	2	2	0	0	0	0	2	216.0	150.0
102	5	3	0	6	69	2	2	0	0	0	0	3	264.0	192.0
103	5	4	0	6	5	2	2	0	0	0	0	4	264.0	192.0
104	2	0	1	6	22	2	2	0	0	1	0	0	282.0	120.0
105	6	3	0	6	38	2	2	0	0	0	0	3	222.0	144.0
106	3	1	1	6	54	2	2	0	0	0	1	0	220.0	108.0
107	4	1	0	6	70	2	2	0	0	0	0	1	162.0	102.0
108	4	2	0	6	86	2	2	0	0	0	0	2	180.0	114.0
109	3	0	0	6	6	2	2	0	0	0	0	0	144.0	0.0
110	1	1	0	9	126	1	0	0	0	0	0	1	396.0	264.0
111	5	4	2	10	5000	2	0	0	2	0	0	4	378.0	330.0
112	7	5	2	10	5000	2	0	1	0	1	1	4	840.0	540.0
113	4	3	1	10	5000	2	0	1	0	1	0	3	840.0	240.0
114	4	3	0	10	5000	2	0	0	0	0	0	3	540.0	372.0
115	3	3	1	2	11	1	1	0	0	0	1	2	480.0	390.0
116	2	2	0	5	11	1	1	0	0	0	0	2	192.0	120.0
117	2	2	1	4	11	2	1	0	0	0	1	1	480.0	228.0
118	2	1	0	4	126	2	1	0	0	0	0	1	480.0	228.0
119	1	1	3	1	1	1	1	0	1	2	1	0	526.2	469.8
120	5	4	2	1	2	1	1	0	1	1	1	3	535.2	448.8
121	2	2	3	1	3	1	1	0	1	1	2	0	526.2	469.8
122	3	2	2	1	4	1	1	0	1	1	1	1	526.2	469.8
123	3	3	2	1	5	1	1	0	1	1	1	2	526.2	469.8
124	2	1	2	1	6	1	1	0	1	1	1	0	526.2	469.8
125	0	0	2	1	1	1	1	1	1	2	0	0	600.0	120.0
126	2	1	1	1	4	1	1	1	0	1	0	1	600.0	120.0
127	1	0	1	1	6	1	1	1	0	1	0	0	600.0	105.0
128	1	1	3	3	1	1	1	0	1	2	1	0	833.4	300.0
129	5	4	2	3	2	1	1	0	1	1	1	3	838.8	398.4
130	2	2	3	3	3	1	1	0	1	1	2	0	833.4	300.0
131	3	2	2	3	4	1	1	0	1	1	1	1	830.4	497.4
132	3	3	2	3	5	1	1	0	1	1	1	2	833.4	300.0
133	2	1	2	3	6	1	1	0	1	1	1	0	823.8	268.8
134	5	4	1	3	2	1	1	0	0	0	1	3	664.2	563.4
135	2	2	2	3	3	1	1	0	0	0	2	0	696.0	507.0
136	3	2	1	3	4	1	1	0	0	0	1	1	649.2	483.6
137	3	3	1	3	5	1	1	0	0	0	1	2	696.0	507.0
138	2	1	1	3	3254	1	1	0	0	0	1	0	630.6	501.6
139	2	1	1	3	3254	1	1	0	0	0	1	0	630.6	501.6
140	2	1	1	3	3254	1	1	0	0	0	1	0	630.6	501.6
141	3	3	1	2	11	2	2	0	0	0	1	2		
142	2	1	1	4	128	3	2	0	0	0	1	1		
143	2	1	0	4	126	3	2	0	0	0	0	1		
144	2	2	1	8	128	2	0	0	0	0	1	1		
145	3	2	0	9	126	2	0	0	0	0	0	2		
146	4	3	1	10	5000	2	0	0	1	1	0	3		
147	2	2	3	1	3	2	2	0	1	1	2	0		
148	1	1	3	1	19	2	2	0	1	2	1	0		
149	5	4	2	1	35	2	2	0	1	1	1	3		
150	3	2	2	1	52	2	2	0	1	1	1	1		
151	3	3	2	1	53	2	2	0	1	1	1	2		
152	2	1	2	1	54	2	2	0	1	1	1	0		
153	0	0	1	1	254	2	2	0	1	1	0	0		
154	1	1	3	3	1	2	2	0	2	2	1	0		
155	5	4	2	3	18	2	2	0	1	1	1	3		
156	2	2	3	3	19	2	2	0	1	1	2	0		
157	4	2	2	3	20	2	2	0	1	1	1	1		
158	3	3	2	3	21	2	2	0	1	1	1	2		
159	2	1	2	3	22	2	2	0	1	1	1	0		
160	9	7	1	3	2	2	2	0	0	0	1	6		
161	6	5	2	3	35	2	2	0	0	0	2	3		
162	3	3	3	3	3	2	2	0	0	0	3	0		
163	7	5	1	3	36	2	2	0	0	0	1	4		
164	4	3	2	3	52	2	2	0	0	0	2	1		
165	5	3	1	3	4	2	2	0	0	0	1	2		
166	7	6	1	3	37	2	2	0	0	0	1	5		
167	4	4	2	3	53	2	2	0	0	0	2	2		
168	5	4	1	3	69	2	2	0	0	0	1	3		
169	5	5	1	3	5	2	2	0	0	0	1	4		
170	6	4	1	3	38	2	2	0	0	0	1	3		
171	3	2	2	3	54	2	2	0	0	0	2	0		
172	4	2	1	3	70	2	2	0	0	0	1	1		
173	4	3	1	3	86	2	2	0	0	0	1	2		
174	3	1	1	3	6	2	2	0	0	0	1	0		
175	2	2	1	8	128	1	0	0	0	0	1	1		
176	4	3	0	2	5000	3	1	0	0	0	0	3		
177	2	2	0	5	11	2	2	0	0	0	0	2		
178	1	0	0	6	11	4	4	0	0	0	0	0		

4. NMR Consistency Check. As mentioned before, it is necessary to examine whether all the components in the set are properly assigned to all signal groups without any excess or any deficiency. To make this examination, the selective NM matrix is made for each [LT] by placing selected components at every row out of the NM matrix previously given and the elements of this matrix are to be converted into numbers of hydrogens of the assigned components. This matrix is called a modified NM matrix (Figure 10). Then, each element of the modified matrix is substituted with variables (x_{ij}) to yield new matrix **X**. A set of simultaneous equations (eq 17 and 18) is made from the **X** and two vectors **H** and **LY**. The vectors present the number of hydrogens of the components and the allocated hydrogens to the signal groups, respectively.

$$\sum_{j=1}^{\mathrm{JN}} x_{ij} = h_i \tag{17}$$

$$\sum_{i=1}^{\mathrm{IN}} x_{ij} = \mathrm{ly}_j \tag{18}$$

Here, x_{ij}, h_i and ly_j represent the elements of **X**, **H**, and **LY**, respectively, and IN stands for the size of the **H** vector.

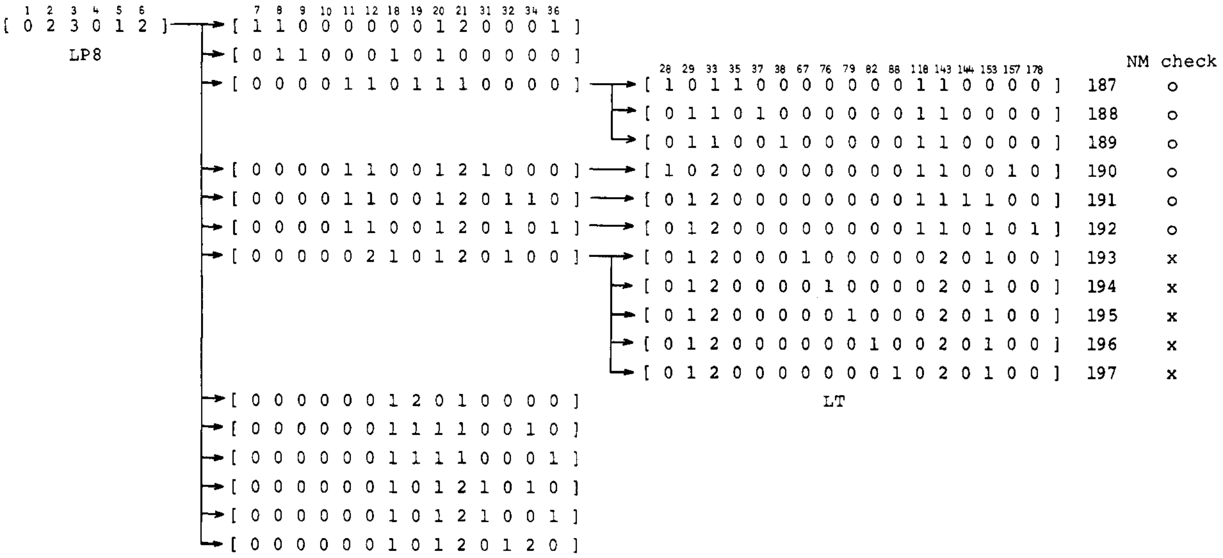


Figure 9. Generation of LS's and LT's from LP8.

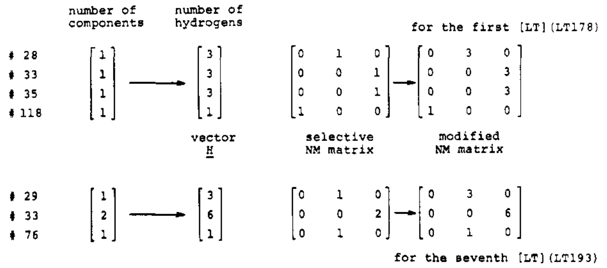


Figure 10. Pretreatment for NMR consistency check.

Table VII. Values of IQT's and Their Meanings

Value of IQT	Meanings
1 - 7	indicating the efferent nature of the partner directly.
11*	indicating no limitation for partner
18 - 86	=16α+β, α<β, and α,β= 1 - 7; indicating components with IQG values α and β as partners.
126	indicating olefinic nature
127	for component #179
129 - 134	=128+α, the value α indicates the highest hierarchical order** of efferent natures of the component.
254*	for component #153
278 - 2185	=256α+16β+γ, α≤β<γ or α<β≤γ and α, β, γ= 1 - 7 indicating components with IQG values α, β and γ as partners.
3254*	indicating formyl group.
4097 - 4103	=4096+α, indicating a pair of methyls or a methyl which should be connected to a methine or a methylene with afferent nature α.
4118 - 4214	=4096+16α+β, indicating a methyl which should be connected to a methine with afferent natures α and β.
5000	indicating aromatic nature

* Those components which have these IQT values require some exceptional treatments in STRUCTURE GENERATOR.

** The order is O>Y>K>D>T>C.

The number of equations is equal to that of the components in the set plus the signal groups. The equations are placed under two restraints; namely, the variable x_{ij} should not exceed the range between zero and the value of the corresponding modified NM matrix element. To solve these simultaneous equations is the major function of this step. When no solution is obtained, the set is judged to be an inappropriate one; when a solution is given, the set is sent to the following step. All the [LT]'s are examined in this manner.

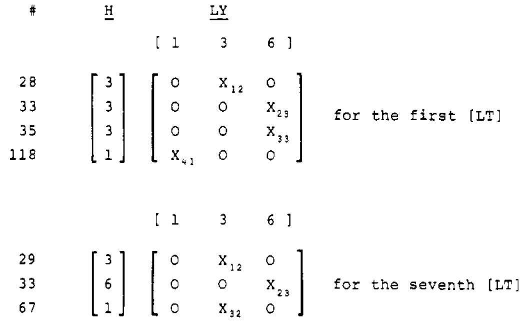


Figure 11. NMR consistency check.

Among the 11 [LT]'s in Figure 9, the first 6 [LT]'s gave solutions, but the remaining five did not afford any solution. As examples, the matrix X's and the vectors H's and LY's for the first and seventh [LT]'s are shown in Figure 11. In these cases the matrices are small and the answer will be obtained at a glance; i.e., for the first case, the solution of x_{12}, x_{23}, x_{33}, and x_{41} will be 3, 3, 3, and 1, respectively. On the other hand, all the elements of the first column of the second matrix are zero, and this means no solution will be given for the second case.

5. Structure Construction Procedure. Some pretreatments of [LT]'s are required before construction of structures.

If any of methyl groups, #41 to #46, #47 to #61, and #19 to #25, are there, they are changed to ethyl, monomethylmethine, and isopropyl groups simply by counterbalancing the corresponding methylene (#104 to #109), and methine (#74 to #88, and #82 to #88) components.

The NAT components are numbered from one to NAT according to the order of the parameter LTG's (Table VIII). The order is settled empirically and it represents the hierarchical order for the construction; it reflects the efferent nature of components and improves the efficiency of the function of the parameter IQT. If the number of components of the highest order is n, they are numbered from 1 to n. Components of the next highest, whose number is m, are numbered from $n + 1$ to $n + m$, and so on. The parameter INJ in Table VIII is the maximum number of remaining bonds for connection with the later order of component.

Since the principle of structure construction from an LT is described in the text, the actual procedure for the forth LT (LT190 in Figure 9) will be described here.

In this case, the components whose values are not zero in the [LT] are #28, #33, #118, #143 and #157. The value of LT(179) was calculated as 1 according to eq 4. Thus NAT is 7, and therefore the stack length is 21, (NAT-1)NAT/2.

Table VIII. Exchanging the Order between Components for the Construction

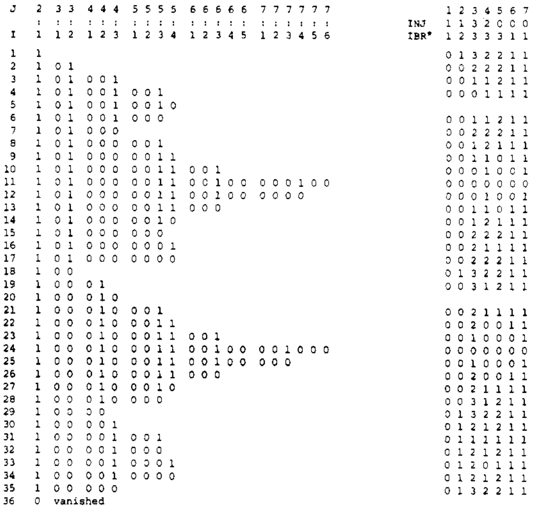
LTG	original #	INJ	LTG	original #	INJ	LTG	original #	INJ
1	153	2	61	171	2	121	24	0
2	148	2	62	172	2	122	5	0
3	149	2	63	173	2	123	39	0
4	147	2	64	174	2	124	132	0
5	150	2	65	37	1	125	137	0
6	151	2	66	40	1	126	30	0
7	28	2	67	130	1	127	123	0
8	31	2	68	133	1	128	74	1
9	121	2	69	135	0	129	75	1
10	127	2	70	138	1	130	76	1
11	26	0	71	139	1	131	77	1
12	119	0	72	140	1	132	78	1
13	125	0	73	143	3	133	79	1
14	152	2	74	118	2	134	80	1
15	124	1	75	145	2	135	81	1
16	176	3	76	110	1	136	91	0
17	113	2	77	179	0	137	95	0
18	114	2	78	142	3	138	98	0
19	111	2	79	117	2	139	99	0
20	112	0	80	144	1	140	100	0
21	146	0	81	175	0	141	82	2
22	141	2	82	163	0	142	83	2
23	115	1	83	165	0	143	84	2
24	160	0	84	92	0	144	85	2
25	94	0	85	96	0	145	86	2
26	32	0	86	101	0	146	87	2
27	42	0	87	52	0	147	13	2
28	21	0	88	33	0	148	14	2
29	2	0	89	44	0	149	15	2
30	129	0	90	23	0	150	16	2
31	36	0	91	4	0	151	17	2
32	134	0	92	38	0	152	104	1
33	27	0	93	131	0	153	105	1
34	120	0	94	136	0	154	106	1
35	154	0	95	29	0	155	107	1
36	155	0	96	122	0	156	108	1
37	156	1	97	126	0	157	7	1
38	157	1	98	177	2	158	8	1
39	158	1	99	116	1	159	9	1
40	159	1	100	64	0	160	10	1
41	35	0	101	66	0	161	11	1
42	128	0	102	67	0	162	55	1
43	62	0	103	69	0	163	56	1
44	63	0	104	70	0	164	57	1
45	65	0	105	71	0	165	58	1
46	68	0	106	72	0	166	59	1
47	89	0	107	73	0	167	60	1
48	90	0	108	93	0	168	43	0
49	47	0	109	97	0	169	22	0
50	48	0	110	102	0	170	3	0
51	50	0	111	103	0	171	178	4
52	41	0	112	49	0	172	88	3
53	19	0	113	51	0	173	18	3
54	1	0	114	53	0	174	109	2
55	20	0	115	54	0	175	61	2
56	161	1	116	166	0	176	12	2
57	170	1	117	168	0	177	6	1
58	162	2	118	169	0	178	25	1
59	164	2	119	34	0	179	46	0
60	167	2	120	45	0			

Table IX. Conditions for Combination of Every Pair of Components

NO		#	1	2	3	4	5	6	7	IQT	IQG	IBR	INJ
1	CH_3O-(K)	28	\	o	x	x	x	x	x	3	1	1	1
2	(O+CO-(D)	157	o	\	Δ	Δ	x	x	x	20	3	2	1
3	C=	143	x	Δ	\	Δ	o	o	o	126	4	3	3
4	-CH=	118	x	Δ	Δ	\	o	o	o	126	4	2	2
5	-D-	179	x	x	o	o	\	x	x	127	-	2	0
6	CH_3-(D)	33	x	x	o	o	x	\	x	4	8	1	0
7	CH_3-(D)	33	x	x	o	o	x	x	\	4	8	1	0

o: valid connection
x: invalid connection
Δ: connection is validated under special condition in the text

Table IX shows the correlations between every pair of components in the set, i.e., whether they can be connected to each other. The parameters IQT and IQG of the first component are 3 and 1, respectively. Thus, this can be connected only to the second one (#157) which has the parameters, IQG = *3* and IQT = 20 = 16 × *1* + 4.



```
J    2  3 3  4 4 4  5 5 5 5  6 6 6 6 6  7 7 7 7 7 7             1 2 3 4 5 6 7
     :  : :  : : :  : : : :  : : : : :  : : : : : :        INJ  1 1 3 2 0 0 0
I    1  1 2  1 2 3  1 2 3 4  1 2 3 4 5  1 2 3 4 5 6        IBR* 1 2 3 3 3 1 1

 1   1                                                          0 1 3 2 2 1 1
 2   1  0 1                                                     0 0 2 2 2 1 1
 3   1  0 1  0 0 1                                              0 0 1 1 2 1 1
 4   1  0 1  0 0 1  0 0 1                                       0 0 0 1 1 1 1
 5   1  0 1  0 0 1  0 0 1 0
 6   1  0 1  0 0 1  0 0 0                                       0 0 1 1 2 1 1
 7   1  0 1  0 0 0                                              0 0 2 2 2 1 1
 8   1  0 1  0 0 0  0 0 1                                       0 0 1 2 1 1 1
 9   1  0 1  0 0 0  0 0 1 1                                     0 0 1 1 0 1 1
10   1  0 1  0 0 0  0 0 1 1  0 0 1                              0 0 0 1 0 0 1
11   1  0 1  0 0 0  0 0 1 1  0 0 1 0 0  0 0 0 1 0 0             0 0 0 0 0 0 0
12   1  0 1  0 0 0  0 0 1 1  0 0 1 0 0  0 0 0 0                 0 0 0 1 0 0 1
13   1  0 1  0 0 0  0 0 1 1  0 0 0                              0 0 1 1 0 1 1
14   1  0 1  0 0 0  0 0 1 0                                     0 0 1 2 1 1 1
15   1  0 1  0 0 0  0 0 0                                       0 0 2 2 2 1 1
16   1  0 1  0 0 0  0 0 0 1                                     0 0 2 1 1 1 1
17   1  0 1  0 0 0  0 0 0 0                                     0 0 2 2 2 1 1
18   1  0 0                                                     0 1 3 2 2 1 1
19   1  0 0  0 1                                                0 0 3 1 2 1 1
20   1  0 0  0 1 0
21   1  0 0  0 1 0  0 0 1                                       0 0 2 1 1 1 1
22   1  0 0  0 1 0  0 0 1 1                                     0 0 2 0 0 1 1
23   1  0 0  0 1 0  0 0 1 1  0 0 1                              0 0 1 0 0 0 1
24   1  0 0  0 1 0  0 0 1 1  0 0 1 0 0  0 0 1 0 0 0             0 0 0 0 0 0 0
25   1  0 0  0 1 0  0 0 1 1  0 0 1 0 0  0 0 0                   0 0 1 0 0 0 1
26   1  0 0  0 1 0  0 0 1 1  0 0 0                              0 0 2 0 0 1 1
27   1  0 0  0 1 0  0 0 1 0                                     0 0 2 1 1 1 1
28   1  0 0  0 1 0  0 0 0                                       0 0 3 1 2 1 1
29   1  0 0  0 0                                                0 1 3 2 2 1 1
30   1  0 0  0 0 1                                              0 1 2 1 2 1 1
31   1  0 0  0 0 1  0 0 1                                       0 1 1 1 1 1 1
32   1  0 0  0 0 1  0 0 0                                       0 1 2 1 2 1 1
33   1  0 0  0 0 1  0 0 0 1                                     0 1 2 0 1 1 1
34   1  0 0  0 0 1  0 0 0 0                                     0 1 2 1 2 1 1
35   1  0 0  0 0 0                                              0 1 3 2 2 1 1
36   0  vanished
```

* The parameter IBR indicates the number of the remaining bonds, (whose initial value is equal to IBG in Table VI). Its value is frequently increased or decreased with the cleavage or the construction of the component - component linkage, respectively. If IBR(j) × IBR(i) is zero, components, j and i, cannot be connected with each other.

Figure 12. Generation of "informational homologues".

As for the second component (#157), the first one is the valid partner as mentioned above, but the third and the fourth ones require some additional conditions to be connected to the second one.

The third component (#143) has a special nature; it should be connected to component #179 (the fifth one in this case) with a single bond, and should not be connected to any components which already have been connected to the same #179 component. This condition is fully applied for the fourth one. The fifth one can only be connected to the third and/or the fourth components in this case, because of the double bond as mentioned previously. The sixth and the seventh (#33) have two components (the third and the fourth) as valid partners in this case.

A part of the construction process by using the stack, s(*i*), is illustrated in Figure 12, and a brief explanation for the contents follows:

lines 1 – 4	The stack is growing.
line 5	At this point, IBR(5) is still 1 but INJ(5) is 0; it means one remaining bond of the fifth one has no partner hereafter.
line 6	Then, retrogression begins.
line 7	Retrogression succeeds.
lines 8 – 10	Forward.
line 11	The first informational homologue (structure 4 in Figure 5) is generated.
line 12	Retrogression begins.
line 13	Retrogression succeeds because the sixth and the seventh are the same kind, and the total number of remaining valence bonds of components earlier than them is only two, so there is no possibility of generating another informational homologue in forwarding from this point.
lines 14 – 17	Retrogression succeeds.
line 18	A bond between the second and the third is cleft.
lines 19 – 23	The stack is growing toward another possibility.
line 24	The second informational homologues (structure 5 in Figure 5) is generated.
line 25	Retrogression begins.
lines 26 – 29	Retrogression succeeds.
lines 30, 31	Forward.
line 32	Retrogression.

Table X. Retrieved Data for Structures 4 and 5, and Inputted Data for Compound I

	NMR data		IR data					
	G	S						
Structure 4	168.74	90.11	0 102	0	0	0	0 404 202	0 202 102 404 204 203 102 102 202 101
Compound I	168.74	90.11	0 102	0	0	0	0 404 202	0 202 102 404 204 203 102 102 202 101
Structure 5	176.76	72.73	0 103	0	0	0	0 404 202	0 203 102 405 404 102 101 101 101 0

line 33 Forward.
lines 34, 35 Retrogression.
line 36 The stack vanished because the first component has no other partner except for the second one in this case.

Other three informational homologues (structures 1 through 3 in Figure 5) were also generated from the LT 189 in the same manner as above.

6. Data Comparison. As mentioned before, the generated structures 1 through 3 for compound I were not found in the file; thus no comparison was carried out. Structures 4 and 5 found in the file and their NMR and IR spectra are observed as shown in Table X.

G and *S* values for the NMR of compound I are exactly equal to those of structure 4, and the differences between I and structure 5 in *G* and *S* values exceed the extent of the allowance described in the text ($\Delta G \leq 0.03$ and $\Delta S \leq 0.04$). Similarly there is no difference between lines of numerals expressing IR and I and structure 4, while a significant difference between I and 5 in position and in intensity is observed as $M = 8$ and $E = 5$, respectively (cf. $M < 6$ and $E < 20$). Therefore the structure of compound I is suggested to be 4 as far as the present filed data are used.

REFERENCES AND NOTES

(1) S. R. Heller, G. W. A. Milne, and R. J. Feldmann, *J. Chem. Inf. Comput. Sci.*, **16**, 232 (1976).

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

(3) G. Beech, R. T. Jones, and K. Miller, *Anal. Chem.*, **46**, 714 (1974).

(4) N. A. B. Gray, *Anal. Chem.*, **47**, 2426 (1975).

(5) L. A. Gribov and M. E. Elyashberg, *J. Mol. Struct.*, **9**, 357 (1971).

(6) H. Abe and P. C. Jurs, *Anal. Chem.*, **47**, 1829 (1975).

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

(8) S. Sasaki, "Automated Chemical Structure Analysis Systems" in "Determination of Organic Structure by Physical Methods", Vol. 5, F. C. Nachod and J. J. Zuckerman, Ed., Academic Press, New York, N.Y., 1973.

(9) (a) C. A. Shelley, H. B. Woodruff, C. R. Snelling, and M. E. Munk, "Interactive Structure Elucidation" in ACS Symposium Series, No. 54, "Computer-Assisted Structure Elucidation", D. H. Smith, Ed., American Chemical Society, Washington, D.C., 1977, p 92. (b) H. B. Woodruff and M. E. Munk, *J. Org. Chem.*, **42**, 1761 (1977).

(10) R. E. Carhart, D. H. Smith, H. Brown, and C. Djerassi, *J. Am. Chem. Soc.*, **97**, 5755 (1975).

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

(12) Y. Kudo and S. Sasaki, *J. Chem. Inf. Comput. Sci.*, **16**, 43 (1976).

(13) S. Sasaki, Y. Yotsui, and S. Ochiai, *Bunseki Kagaku*, **24**, 213 (1975).

(14) B. A. Knock, I. C. Smith, D. E. Wright, R. G. Ridley, and W. Kelley, *Anal. Chem.*, **42**, 1516 (1970).

(15) S. Sasaki, H. Abe, Y. Kudo, and T. Yamasaki, "CHEMICS: A Computer Program System for Structure Elucidation of Organic Compounds" in ref 9a, p 108.

(16) MS analysis will be introduced in the near future.

(17) The value of 179th element is calculated after completing the 178th dimensional vector (cf. eq 4).

A Compact and Efficient File Structure for Searching Large Generic-Keyed Databases. An Application to Mass Spectral Data

R. GEOFF DROMEY*

Research School of Chemistry, Australian National University, Canberra, A.C.T. 2600, Australia

Received July 1, 1977

Conventional file structures do not satisfactorily handle large generic-keyed databases. Seemingly a compromise must always be made between storage requirements and retrieval efficiency. A new inverted bitmap file structure that does not involve a high storage cost is suggested as a viable alternative to existing systems. It requires less storage and is faster for retrieval than other systems that are currently being used. Implementation and performance evaluation for a mass spectral database are given.

INTRODUCTION

Key-based information storage and retrieval systems usually fall into one of three categories. The simplest of these is the single-key file, where each record possesses just one key which may or may not be unique. A chemical name file and a molecular formula file would fit into this category. Efficient methods for handling this type of system are readily available.[1,2] The second category is typified by a bibliographic file, where each record is characterized by perhaps an author, title, and several keywords. Methods for handling these systems with their relatively few keywords per record are also straightforward.[3] It is in the third category, where each record may consist of many generic keys, that the real difficulties arise. Application of existing methods to this problem has always resulted in a tradeoff between storage and processing efficiency.

The present paper represents an attempt to design an efficient system for handling large databases in the third category. The aim has been to present a file structure which is simple, easy to construct, and yet at the same time is highly economical on both storage requirements and retrieval times.

Lefkovitz[4] has discussed the characteristics of generic-key files and suggested that a hybrid inverted list-bitmap file structure is a practical way of handling these systems. This hybrid file design is tailored to cope with the Zipfian-like[5] distribution among keys that almost invariably exists for large generic-keyed databases. The small proportion of keys occurring at high frequency is stored in fixed-length inverted bitmaps while the majority of keys, which occur only infrequently, are stored in inverted lists. This approach certainly economizes on storage but at the cost of introducing very inefficient Boolean operations between the two data structures (bits and lists). For any given system there is a degree of uncertainty as to what is the most useful mix between bitmaps and lists. The hybrid file approach, although clearly superior to other available systems, still must trade efficiency for storage. A desirable goal would therefore seem to be to devise a data structure that could exploit the processing efficiency of bitmap systems without incurring the excessive storage costs that they conventionally entail.

Zatocoding[4,6] has been suggested as a possible bitmap-oriented solution to the large database dilemma. Its drawbacks are that it involves a sequential rather than inverted search and that the super-imposed code can produce a significant risk of false retrievals. Lefkovitz[4] gives a detailed account of why zatocoding would appear to be unsatisfactory for very large systems.

Efficiency demands that some type of inverted bitmap be used for large databases. The storage problem still remains.

* Address correspondence to author at Department of Computing Science, University of Wollongong, Wollongong, N.S.W. 2500, Australia.