

A Computer Program for Generation of Constitutionally Isomeric Structural Formulas

HIDETSUGU ABE, TOHRU OKUYAMA, IWA0 FUJIWARA, and SHIN-ICHI SASAKI*

Department of Materials Science, Toyohashi University of Technology, Hibarigaoka, Tempaku, Toyohashi, 440 Japan

Received December 28, 1983

A computer program for generation of constitutionally isomeric structural formulas has been developed. The program can generate all the possible isomeric structural formulas for a given molecular formula, which is any one of the combinations of C, H, N, O, S, and halogens. The structure construction algorithm employed in this study is the connectivity stack method, which has been developed by S.S. Because this is a rather simple combinatorial algorithm, the efficiency is excellent. For example, the time required for the generation of 355 isomeric structural formulas of $C_{12}H_{26}$ is only 3.5 s on a Hitachi M-200H computer. Moreover, there are no redundancies and/or deficiencies in the generated isomeric formulas for tested molecular formulas.

It may be an interesting and important problem for any organic chemist to deduce the number and the structures of all the constitutional isomers consistent with a given molecular formula.¹ Any automated structure elucidation system for organic compounds²⁻⁴ also requires a structure generator that can generate all the relevant chemical structures from minimum information, a molecular formula alone.

In this paper a method for generation of all the possible structures from a given molecular formula will be described in connection with the study of the computer-assisted structure elucidation system CHEMICS.⁴ The structures to be generated are constitutional isomers of organic compounds consisting of any combination of C, H, N, O, S, and halogens. These nine elements have been selected because they are frequently found in most organic compounds and constitute the greater part of organic compounds so far known.

BASIC COMPOSITION OF STRUCTURE GENERATOR

The structure generator has developed the following sequence. (1) Determination of a set of augmented atoms (so-called components) as media for information transfer. (2) Development of algorithms and programs for generating all the possible sets of components that satisfy a given molecular formula. (3) Development of programs for generating relevant structures from a given set of components by use of the connectivity stack method.^{5,6}

Figure 1 shows the fundamental composition of the generator, which consists of two basic units, corresponding to the second and the third stages of development described above. The unit for the second stage is a set generator that generates a set of all the component combinations that satisfy a molecular formula. The unit for the third stage is a structure generator where the connectivity stack method plays an important role.

DETERMINATION OF COMPONENTS

A component, here, is an organic partial structure that has the following three characteristics. (1) Any structural formula that contains any combinations of C, H, O, S, and halogens can be expressed as a combination of some of the components. (2) No pair of the components overlap each other, even partly. (3) A component is defined by four attributes, i.e., symbol, root element, composition, and bond, as shown in Table I.

The first and the second attributes, symbol and root element, stand for a symbol used for expressing a component and characteristic of an atom that has a bond(s) for another component(s), respectively. As to the latter, when two or more bonds are involved in a single component, the atom having these bonds is defined to have the same root element. The third attribute composition expresses the elemental composition of

a component. The last attribute bond denotes the characteristic of the bond. When any two components are to be bonded, the attribute of one should be matched to that of the other.

Table I shows 63 partial structures, which are defined as the entire set of components (called secondary components) that have three characteristics mentioned above. They have been defined so that N, O, and S atoms may not be subject to any influence even in the cases when the valencies N, O, and S deviate from normal valencies, 3, 2, and 2, respectively. The table does not contain ionic components, because no structure with ionic property is dealt with here.

The first component in Table I is defined to be a methyl group that does not connect to a saturated carbon. The reason is that, as for bonding it with the saturated carbon, other components having methyl group are separately resulted, such as ethyl (no. 18) and isopropyl group (no. 19), and this causes serious duplication of final structures. Components from no. 49 to no. 58 are specially prepared to deal with resonance structures of aromatic compounds. Two alternative formulas may be written for either benzene or *o*-xylene, as shown in Figures 2 or 3, respectively. On Figure 3, although structure a is identical with structure b in a chemical sense, they cannot be said to be identical from the graph theoretical point of view, because a single bond is interposed between two methyls in (a) while a double bond is interposed between them in (b). Therefore, structures a and b are recognized to be different by their graphical difference, leading to generation of two independent structures. To prevent these sort of undesirable duplications, components no. 49-58 in Table I are prepared, which have root elements and bond attributes implying aromatic features. However, even the arrangement so far made cannot make comprehensive and proper handling of all the aromatic structures. There are still problems pertinent to non-benzenoid aromatics, such as tropolone and aromatic heterocycles. For proper solution of these problems, we should trace back to the definition of aromaticity itself. Thus, we have determined to have only the structures generated, in each of which the total number of π -electrons of the aromatic structures among the components involved in a component set will satisfy Hückel's law ($4n + 2$, $n = 1, 2, 3, \dots$). The number of π -electrons should be calculated by reference to the π -electron numbers listed in Table II. Also, we have decided that, in structure generation, any bonds of any aromatic components, except the bond of component no. 49, should be connected to aromatic components. The bond attribute of such bond is called "AB". The structure generator will output both the structures expressed by aromatic components and 1,3,4-cyclohexatriene for composition C_6H_6 . Tropolone structure is exceptionally displayed in the form of extreme structure only. Component no. 63 is a dummy that is indicative of a double

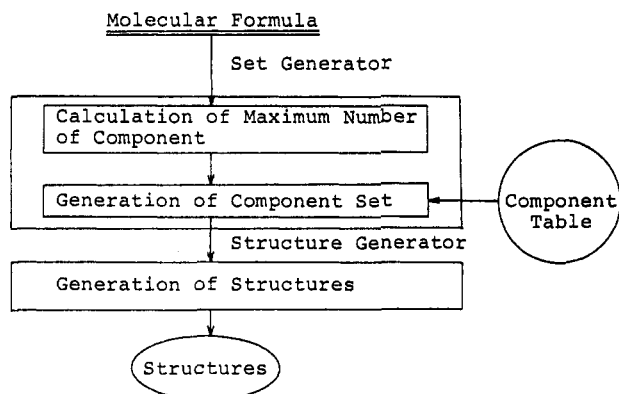


Figure 1. Fundamental construction of structure generator.

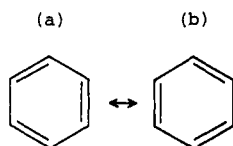


Figure 2. Limit formulas of benzene.

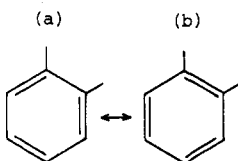


Figure 3. Limit formulas of o-xylene.

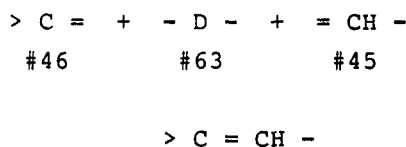


Figure 4. Double-bond representation used by double-bond dummy component.

bond. This dummy is related to characteristics of the structure generation algorithm. Since the algorithm does not allow one to register the bond multiplicity ($n = 2, 3$) on the connectivity stack, a double bond is represented by placing the dummy between two components containing the double-bond attribute (Figure 4). On the contrary, a triple bond is originally included in the components, as shown with components no. 5, 6, and 27.

PRINCIPLE OF STRUCTURE GENERATION

Hierarchization of Components. The first step for structure generation is to prepare an appropriate set of components that satisfy the elemental composition of a given molecular formula. To facilitate this procedure, we have proposed a set of new "augmented atoms", which are situated between atoms themselves and the 63 components give in Table I.

Table III shows the proposed augmented atoms. Hereafter, we call these 16 augmented atoms in the table primary components and also call the 63 components proposed before secondary components. The primary components are more basic atomic groups than the secondary components. Root element and bond attribute are not defined for the former. The generation of a set of components to satisfy a molecular formula is made hierarchically in the sequence molecular formula \rightarrow primary components \rightarrow secondary components, as shown in the block diagram in Figure 5.

Set Generator. According to the sequence of the block diagram in Figure 5, a secondary component vector is generated from a given molecular formula as shown in Procedure 1.

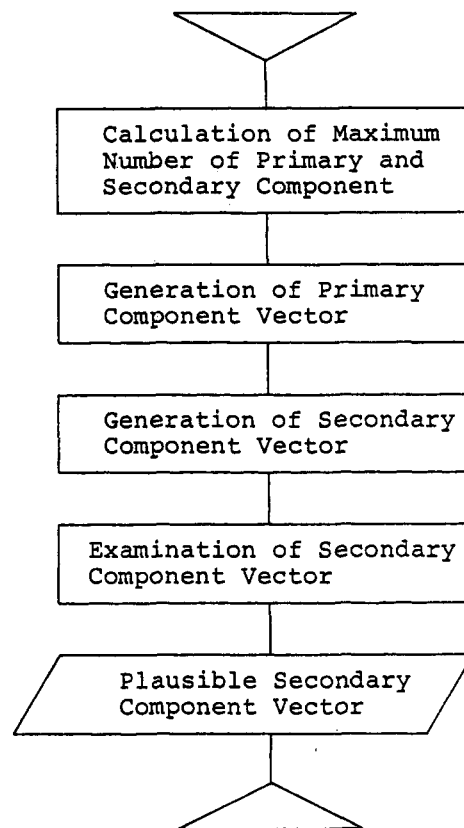


Figure 5. Block diagram of generation of secondary component vector.

Procedure 1. Generation and Verification of Secondary Component Vectors

- (1) To determine the maximum numbers of possible primary and secondary components from given molecular formula. The maximum numbers thus determined are stored in the maximum value vector for primary and secondary components, respectively.
- (2) To generate all the possible sets of primary components that satisfy the elemental composition of the molecular formula. The sets are called primary component vectors.
- (3) To generate all the possible sets of secondary components that satisfy each of the primary component vectors. The sets are called secondary component vectors.
- (4) To verify the secondary component vector from the viewpoint of graph theory and with root element and bond attributes.

To obtain a primary component, vector P is solved from the following equation

$$P \cdot C_p = M \quad (1)$$

where C_p is an atomic composition matrix of primary components and M is a molecular formula vector. Such an equation as this almost always gives plenty of solutions. But it is easy for a computer to generate all these solutions within few seconds. Actually, only the combinations are selected that are coincidental with a given molecular formula out of all the combinations that do not exceed the maximum value defined by a primary component vector.

Similarly, the secondary component vector S can be defined to be the solutions of the following equation

$$S \cdot C_s = P \quad (2)$$

where C_s is a primary component composition matrix for secondary components.

In the case of secondary components, however, relevant structures may not be generated from all the secondary components generated, because of limitations in the bond attribute. Therefore, the generated secondary component vector should be verified. The verification is made whether or not the vector

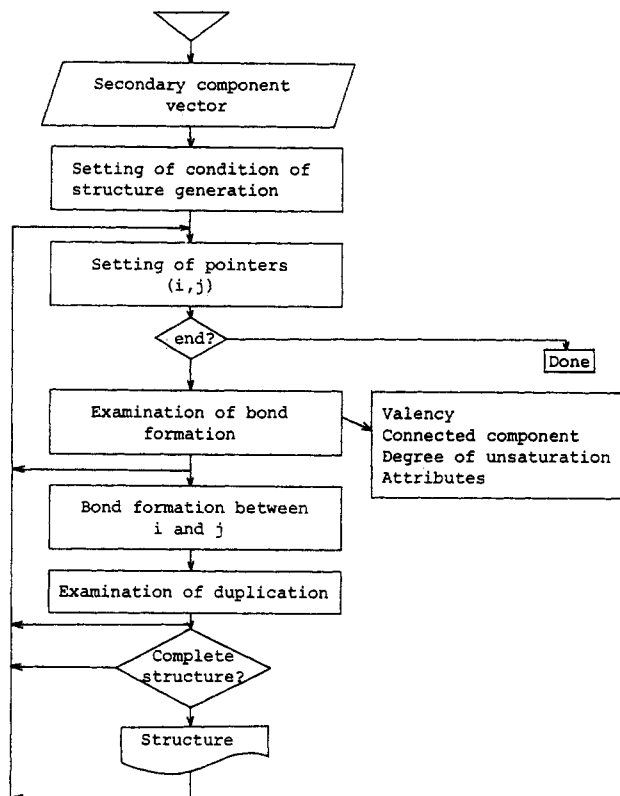


Figure 6. Flow diagram of structure generator.

satisfies the requirements of root element and bond attributes: (1) verification of the number of no. 1 components and a set of components with a saturated-carbon attribute; (2) verification of a set of aromatic components; (3) a set of double-bond components. In verification 1, the number of components is checked with eq 3. Here, C_1 is the number of no. 1 com-

$$C_1 \leq TD(v) - \sum_i V_i C_i - C_1 \quad (3)$$

ponents, V_i is the number of bonds of components with the i th saturated carbon attribute (see Table IV), and C_i is the number of components with the i th saturated carbon attribute. The component vectors that do not satisfy this condition are discarded. In verification 2, similarly, components in a vector are examined with the following conditions

$$\sum_i \pi_i C_i = k(4n + 2) \quad (n = 1, 2, 3, \dots; k = 1, 2, 3, \dots) \quad (4)$$

$$C_{49} \neq 0 \text{ if } \sum_i C_i \neq \sum_j C_j \quad (5)$$

where π_i is the number of π -electrons of the i th aromatic components (Table II shows π_i values), C_i is the number of i th aromatic components, and C_j the number of j th components in the component vectors. Component vectors are adopted if they satisfy these two conditions.

The last verification for double-bond components (see Table V) is made by Procedure 2. Step 4 in the procedure is based on the fact that the groups that could be connected to $=NH$ are limited to $-N=$ (no. 47) or $-N(O)=$ (no. 48), due to the presence of components $=C=NH$ (no. 42), $-CH=NH$ (no. 4), and $>C=NH$ (no. 26). Step 5 is a verification to avoid duplication between $-N_3$ (no. 17) and $-N=$ (no. 47) + $=N^+=N^-$ (no. 44). The component vectors that have passed the verification are sent to the stage of structure generation.

Structure Generator. In the second stage of structure generation, structures are generated from the secondary component vectors generated by the set generation. The

Procedure 2. Verification of Double-Bond Components

- (1) Double bond components are grouped into the following two groups.
 - (I) Components that have no bonds other than double bonds (no. 38–44).
 - (II) Components that have other bonds as well as double bonds (no. 45–48).
- (2) If the total number of double bonds is odd, then go to step 7.
- (3) Else, if $TDB^I < TDB^{II}$, then go to step 7. Here, TDB is the total number of double bonds involved in each group (i = double bond component each group):

$$TDB = \sum_i V_i^d C_i^d \quad (6)$$
- (4) If $C_{43}^d > C_{47}^d + C_{48}^d$, then go to step 7.
- (5) Else, if $C_{44}^d > C_{39}^d + C_{45}^d + C_{46}^d + C_{48}^d$, then go to step 7.
- (6) Else, the component vector is adopted, and the procedure ends.
- (7) The component vector is discarded, and the procedure ends.

Procedure 3. Examination of Bond Formation by RE and B Attributes

- (1) Make binary vectors, E^i and E^j , of the RE attributes for the j th and i th components, respectively, as well as binary vectors, B_1^i, B_2^i, \dots and B_1^j, B_2^j, \dots , of the B attributes of remaining bonds for the individual components.
- (2) Make $E^i \wedge E^j, E^i \wedge E_2^j, \dots$, the elements of which are logical products of the elements E^i and those of B_1^j, B_2^j, \dots vectors. Similarly, make $E^j \wedge B_1^i, E^j \wedge B_2^i, \dots$
- (3) Make logical sums of the elements of the product vectors.

$$L_1^i = (E^i \wedge B_1^j)_1 \vee (E^i \wedge B_2^j)_2 \vee \dots \vee (E^i \wedge B_m^j)_m$$

$$L_2^i = (E^i \wedge B_1^j)_1 \vee (E^i \wedge B_2^j)_2 \vee \dots \vee (E^i \wedge B_m^j)_m$$

$$\vdots$$

$$L_j^i = (E^i \wedge B_1^j)_1 \vee (E^i \wedge B_2^j)_2 \vee \dots \vee (E^i \wedge B_m^j)_m$$

$$L_j^j = (E^j \wedge B_2^i)_1 \vee (E^j \wedge B_2^i)_2 \vee \dots \vee (E^j \wedge B_2^i)_m$$

$$\vdots$$
- (4) $U^i = L_1^i \vee L_2^i \vee \dots$ and $U^j = L_1^j \vee L_2^j \vee \dots$. If $U^i \wedge U^j$ is true (1), a bond is formed. If it is false (0), no bond is formed.
- (5) Select L_k^i and L_l^j for true L_1^i, L_2^i, \dots and L_1^j, L_2^j, \dots values, respectively. And, remove the l th bond of the i th component and the k th bond of the j th component from the bond list.

Procedure 4. Analysis of Ring Structures

- (1) Define one of the two components to be bonded as an origin and another as a target.
- (2) Set the origin component at the level of $l = 1$.
- (3) $l = l + 1$.
- (4) Assuming the levels for the components adjacent to a components are the same, investigate whether the target component exists at the present level l or not. If there is, go to step 6. If all the components at the same level have been investigated and yet there remain some uninvestigated components, go back to step 3.
- (5) It is confirmed that no ring is formed. Thus, the procedure is finished.
- (6) Formation of ring structures has been confirmed. The level l in this step indicates the ring size.

connectivity stack method is employed for structure generation. Since registration of multiple bonds is not allowed in the stack, multiple bonds are implied in several components. Therefore, a connectivity stack is expressed as a sequence of 0 and 1 elements. In addition to the four kinds of verifications relevant to bond formation that are described above, verifications for root element (RE) and bond (B) attributes will be made in this step.

Figure 6 shows the flow diagram of the structure generator. The examination for RE and B attributes are made according to Procedure 3. A typical examination process for the attributes is also shown in Figure 7. Next, the number of rings must be examined because all unsaturations $U(V)$ should be consumed for formation of rings. The procedure for ring

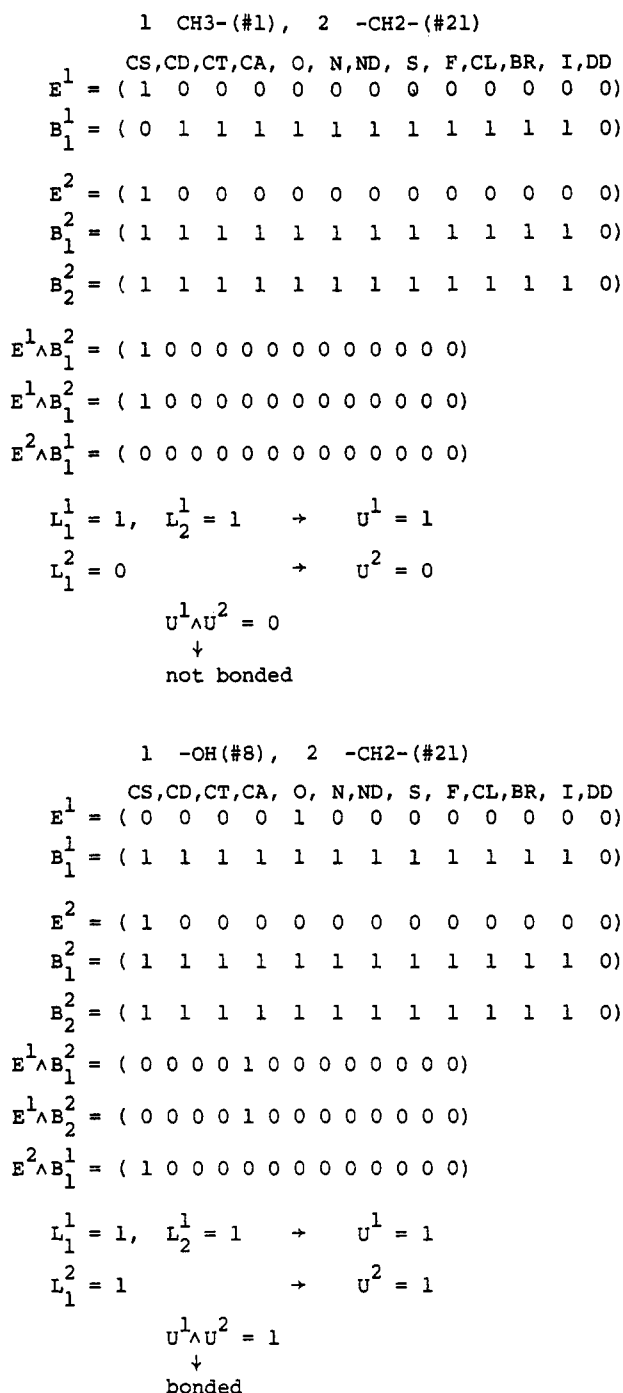


Figure 7. Examination of bond formation used by root element and bond attributes.

structure analysis in the CHEMICS-UBE system⁷ is adopted for this examination. Procedure 4 describes examination of ring structure formation, and Figure 8 illustrates its logic. In this procedure, care should be taken of the treatment of a dummy component between double bonds. In examination of ring formation, it is necessary to consider the dummy component. However, when it is needed to extract the ring size as necessary information, the dummy should be previously omitted. When the structure formation procedure is done for all the component vectors generated by a set generator, all the possible structures that satisfy a given molecular formula are to be generated.

RESULTS AND DISCUSSION

Generation of Component Vectors. Table VI shows the numbers of primary component vectors and Table VII indi-

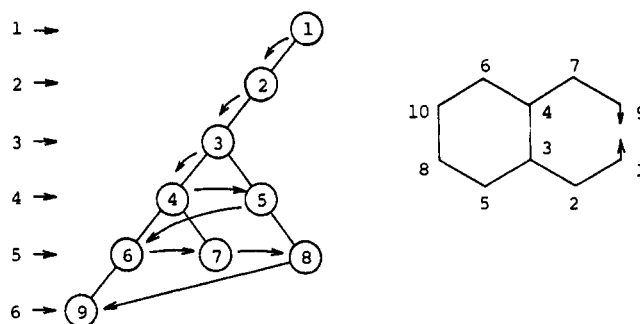


Figure 8. Example of examination of formation of ring structure.

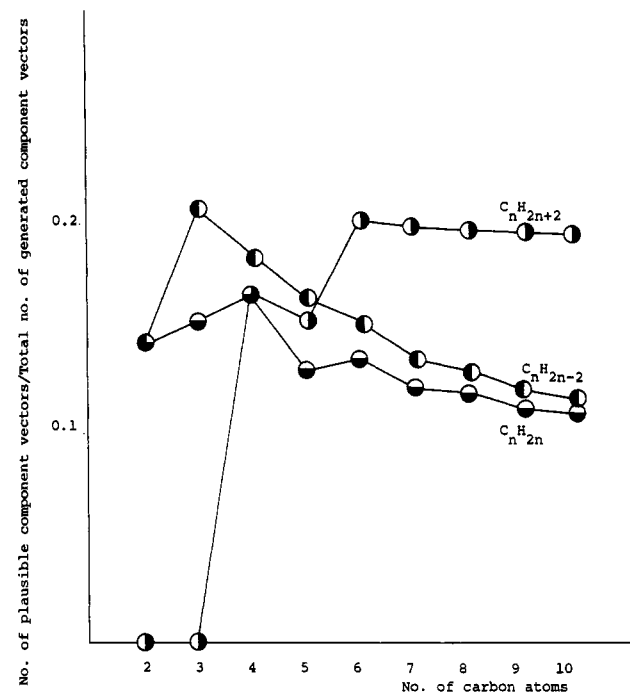
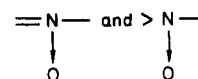


Figure 9. Efficiency of examination routines of secondary component vector generator.

cates the number of secondary component vectors generated for various molecular formula. S1 and S2 in Table VII are the numbers of vectors before and after the above examination procedures are performed, respectively. In Figure 9, the ratio between numbers of secondary component vectors before and after examination is plotted against the number of carbon atoms in saturated and unsaturated hydrocarbons.

For saturated hydrocarbons, nearly 20% of vectors pass the examination, while for unsaturated hydrocarbons with two or less indexes of hydrogen deficiency about 10% pass the examination. Thus, it can be understood that the examination is important, particularly for unsaturated hydrocarbons.

Generation of Chemical Structures. Table VIII shows the number of secondary component vectors from which structures are generated finally. Table IX shows the number of structures generated from a given molecular formula. The number 12 for $C_nH_{2n+1}N$ ($n = 3$) in the table differs from Gribov's result (=14).⁸ The latter will not be correct.⁹ In the cases of $C_nH_{2n+1}NO$ and $C_nH_{2n+3}NO$, the authors take into consideration the *N*-oxide components



therefore structures 23, 87, and 308 are afforded for $C_nH_{2n+1}NO$ ($n = 2, 3$, and 4), respectively. This is also not identical with Gribov's result. In Figure 10, the ratio of the number of secondary component vectors from which structures

Table I. List of Components (So-Called Secondary Components)









no.	symbol ^a	bond ^b	root ^c	composition									HD ^d
				C	H	O	N	S	F	Cl	Br	I	
1	—CH ₃	NSC	CS	1	3	0	0	0	0	0	0	0	0
2	—CH=O	NBC	CD	1	1	1	0	0	0	0	0	0	2
3	—CH=S	NBC	CD	1	1	0	0	1	0	0	0	0	2
4	—CH=NH	NBC	CD	1	2	0	1	0	0	0	0	0	2
5	—C#CH	NBC	CT	2	1	0	0	0	0	0	0	0	4
6	—CN	NBC	CT	1	0	0	1	0	0	0	0	0	4
7	—N=C	NBC	ND	1	0	0	1	0	0	0	0	0	4
8	—OH	NBC	O	0	1	1	0	0	0	0	0	0	0
9	—SH	NBC	S	0	1	0	0	1	0	0	0	0	0
10	—NH ₂	NBC	N	0	2	0	1	0	0	0	0	0	0
11	—N=O	NBC	ND	0	0	1	1	0	0	0	0	0	2
12	—N=O	NBC	ND	0	0	2	1	0	0	0	0	0	2
13		NBC	ND	0	0	0	1	1	0	0	0	0	2
14		NBC	ND	0	0	1	1	1	0	0	0	0	2
15		NBC	S	0	1	2	0	1	0	0	0	0	0
16		NBC	S	0	1	3	0	1	0	0	0	0	0
17	—N ₃	NBC	ND	0	0	0	3	0	0	0	0	0	4
18	—CH ₂ CH ₃	NFM	CS	2	5	0	0	0	0	0	0	0	0
19	—CH(CH ₃) ₂	NFM	CS	3	7	0	0	0	0	0	0	0	0
20	—C(CH ₃) ₃	NFM	CS	4	9	0	0	0	0	0	0	0	0
21	—CH ₂ —	NFM, NFM	CS	1	2	0	0	0	0	0	0	0	0
22	>CHCH ₃	NFM, NFM	CS	2	4	0	0	0	0	0	0	0	0
23	>C(CH ₃) ₂	NFM, NFM	CS	3	6	0	0	0	0	0	0	0	0
24	>C=O	NBC, NBC	CD	1	0	1	0	0	0	0	0	0	2
25	>C=S	NBC, NBC	CD	1	0	0	0	1	0	0	0	0	2
26	>C=NH	NBC, NBC	CD	1	1	0	1	0	0	0	0	0	2
27	—C#C—	NBC, NBC	CT	2	0	0	0	0	0	0	0	0	4
28	—O—	NBC, NBC	O	0	0	1	0	0	0	0	0	0	0
29	—S—	NBC, NBC	S	0	0	0	0	1	0	0	0	0	0
30	—NH—	NBC, NBC	N	0	1	0	1	0	0	0	0	0	0
31	—S—	NBC, NBC	S	0	0	1	0	1	0	0	0	0	0
32		NBC, NBC	S	0	0	2	0	1	0	0	0	0	0
33	>CH—	NFM, NFM, NFM	CS	1	1	0	0	0	0	0	0	0	0
34	>C(CH ₃)—	NFM, NFM, NFM	CS	2	3	0	0	0	0	0	0	0	0
35	>N—	NBC, NBC, NBC	N	0	0	0	1	0	0	0	0	0	0
36	>N—	NBC, NBC, NBC	N	0	0	1	1	0	0	0	0	0	0
37	>C<	NFM, NFM, NFM, NFM	CS	1	0	0	0	0	0	0	0	0	0
38	=CH ₂	DB	CD	1	2	0	0	0	0	0	0	0	1
39	=C=	DB, EXB	CD	1	0	0	0	0	0	0	0	0	2
40	=C=O	DB	CD	1	0	1	0	0	0	0	0	0	3
41	=C=S	DB	CD	1	0	0	0	1	0	0	0	0	3
42	=C=NH	DB	CD	1	1	0	1	0	0	0	0	0	3
43	=NH	EXB	ND	0	1	0	1	0	0	0	0	0	1
44	=N=N	EXB	ND	0	0	0	1	0	0	0	0	0	3
45	=CH—	EXB, EXB	CD	1	1	0	0	0	0	0	0	0	1
46	=C<	EXB, EXB, EXB	CD	1	0	0	0	0	0	0	0	0	1
47	=N—	EXB, EXB	ND	0	0	0	1	0	0	0	0	0	1
48	=N—	EXB, DB	ND	0	0	1	1	0	0	0	0	0	1
49		NBC, AB, AB	CA	1	0	0	0	0	0	0	0	0	1
50	—ACH—	AB, AB	CA	1	1	0	0	0	0	0	0	0	1
51	—MDO— ^h	AB, AB	CA	3	2	2	0	0	0	0	0	0	4
52	—AO—	AB, AB	OA	0	0	1	0	0	0	0	0	0	0
53	—AS—	AB, AB	SA	0	0	0	0	1	0	0	0	0	0
54	—AN—	AB, AB	NA	0	0	0	1	0	0	0	0	0	1
55	—ANH—	AB, AB	NA	0	1	0	1	0	0	0	0	0	0
56	—AN— ^e	AB, AB	NA	0	0	1	1	0	0	0	0	0	1
57		AB, AB	SA	0	0	1	0	1	0	0	0	0	0

Table I (Continued)

no.	symbol ^a	bond ^b	root ^c	composition									
				C	H	O	N	S	F	Cl	Br	I	HD ^d
58	—AS— ^f 	AB, AB	SA	0	0	2	0	1	0	0	0	0	0
59	—F	NBC	F	0	0	0	0	0	1	0	0	0	0
60	—Cl	NBC	CL	0	0	0	0	0	0	1	0	0	0
61	—Br	NBC	BR	0	0	0	0	0	0	0	1	0	0
62	—I	NBC	I	0	0	0	0	0	0	0	0	1	0
63	—D—	DBD, DBD	DD	0	0	0	0	0	0	0	0	0	0

^a #, triple bond; AC, aromatic carbon; ACH, aromatic carbon with hydrogen; AO, aromatic oxygen; AS, aromatic sulfur; —AN—, aromatic nitrogen; ANH, aromatic nitrogen with hydrogen. ^b NSC, not bonded to saturated carbon; NBC, no bonding constraint; NFM, not bonded to free methyl (no. 1); DB, only bonded to double bond (no. 63); EXB, exceptional bond (described in the text); AB, aromatic bond; DBD, dummy double bond. ^c CS, carbon with sp³ bond only; CD, carbon with sp² bond; CT, carbon with sp bond; O, oxygen; S, sulfur; N, nitrogen; ND, doubly bonded nitrogen; CA, aromatic carbon; SA, aromatic sulfur; NA, aromatic nitrogen; DD, dummy double bond. ^d Index of hydrogen deficiency. ^e Aromatic N-oxide (N is a member of an aromatic ring). ^f Aromatic sulfoxide (S is a member of an aromatic ring). ^g Aromatic sulfone (S is a member of an aromatic ring). ^h Denotes the following structure

Table II. Number of π Electrons

no.	symbol	no. of π electrons	no.	symbol	no. of π electrons
49	AC	1	54	AN	1
50	ACH	1	55	ANH	2
51	MDO	2	56	ANO	1
52	AO	2	57	ASO	2
53	AS	2	58	ASO2	2

are generated vs. that of secondary component vectors that passed the examination is plotted against the number of carbon atoms for the molecular formulas containing one oxygen atom. As shown in the figure, the present examination has a screening efficiency of 100% for saturated compounds. However, it is less effective for unsaturated compounds. Thus, it is required to devise another examination method that may utilize component attributes more effectively.

Examination of Output Results. For the estimation of the structure generation method described above, the examination was conducted with respect to the following items: (1) comparison of the number of structures generated to that previously reported; (2) examination of duplicate structures; (3) examination by the generation of specified structures from some kinds of molecular formulas. Table X shows the number of isomeric structures, previously reported,^{1,10} consistent with various molecular formulas. From its comparison to Table IX the same number of isomeric structures is found to have been generated from the structure generator. However, for molecular formulas containing both nitrogen and oxygen atoms, the generated structures are more than the reported ones

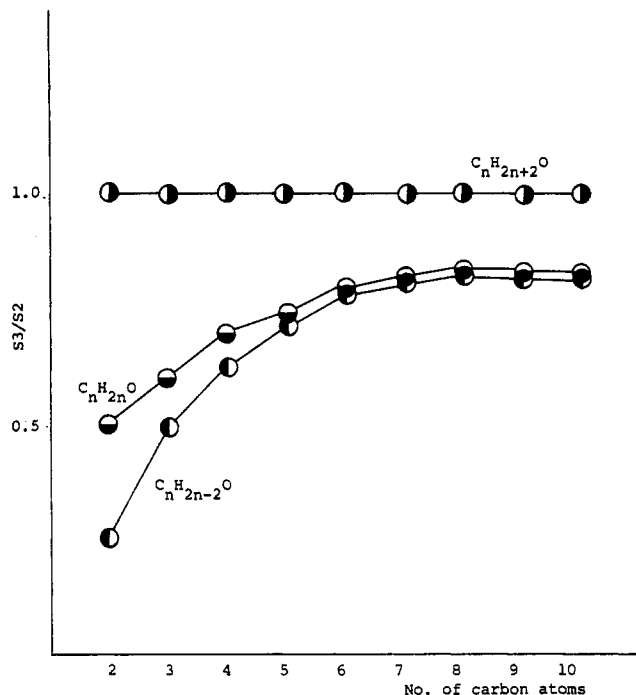


Figure 10. Plot of number of carbon atoms vs. ratio of number of secondary component vectors from which structures are generated (S3) and number of generated secondary component vectors (S2).

because previous works neglected amine oxide structures. Examination of duplicate structures will be conducted ac-

Table III. List of Augmented Atoms (So-Called Primary Components)

no.	symbol	composition									
		C	H	O	N	S	F	Cl	Br	I	HD ^a
1	CH3	1	3	0	0	0	0	0	0	0	0
2	CH2	1	2	0	0	0	0	0	0	0	0
3	CH	1	1	0	0	0	0	0	0	0	0
4	C	1	0	0	0	0	0	0	0	0	0
5	OH	0	1	1	0	0	0	0	0	0	0
6	O	0	0	1	0	0	0	0	0	0	0
7	NH2	0	2	0	1	0	0	0	0	0	0
8	NH	0	1	0	1	0	0	0	0	0	0
9	N	0	0	0	1	0	0	0	0	0	0
10	SH	0	1	0	0	1	0	0	0	0	0
11	S	0	0	0	0	1	0	0	0	0	0
12	F	0	0	0	0	0	1	0	0	0	0
13	Cl	0	0	0	0	0	0	1	0	0	0
14	Br	0	0	0	0	0	0	0	1	0	0
15	I	0	0	0	0	0	0	0	0	1	0
16	U ^b	0	0	0	0	0	0	0	0	0	2

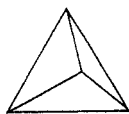
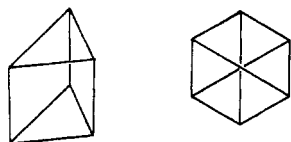
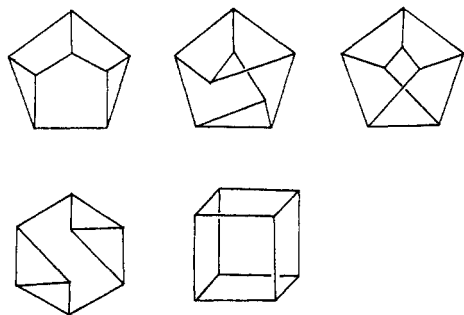
^a HD, hydrogen deficiency. ^b U, double bond.

Table IV. Saturated Carbon Components

component no. (i)	symbol	component no. (i)	symbol
18	CH ₃ CH ₂ -	23	(CH ₃) ₂ C<
19	(CH ₃) ₂ CH-	33	-CH<
20	(CH ₃) ₃ C-	34	-(CH ₃)C<
21	-CH ₂ -	37	>C<
22	CH ₃ CH<		

Table V. Double-Bond Components

component no. (i)	symbol	component no. (i)	symbol
38	CH ₂ =	44	=N ⁺ =N ⁻
39	=C=	45	-CH=
40	=C=O	46	>C=
41	=C=S	47	-N=
42	=C=NH	48	-N=
43	=NH		O

Figure 11. Generated trivalent regular graph ($n = 4$).Figure 12. Generated trivalent regular graphs ($n = 6$).Figure 13. Generated trivalent regular graphs ($n = 8$).

according to the following procedure. (a) By definition of components, structures generated from different component vectors are different from each other. Therefore, examination could be made within each set of structures generated from each component vector. (b) Canonical code (stack) is calculated for each of the structures generated from a component vector. Then, a check is made whether there exists the equivalents or not.

It has been found that there are no equivalents within a set of structures generated from a component vector. Finally, generation of specified structures from a certain molecular formula was examined. For example, the benzenoid isomeric structures of C₈H₁₀ are supposed to be *o*-, *m*-, and *p*-xylene and ethylbenzene. The generator has, surely, generated these four structures. Furthermore, regular trivalent graphs were selected as more general examples. From the viewpoint of chemical structures, regular trivalent graphs are considered to represent structures of hydrocarbon consisting entirely of methyne groups. Randić studied generation of regular trivalent graphs with the number of methyne groups (n) up to 10.¹¹ Figures 11–14 show regular trivalent graphs with n up to 10, generated by our structure generator. These graphs are all coincidental with the result of Randić. It has been reported

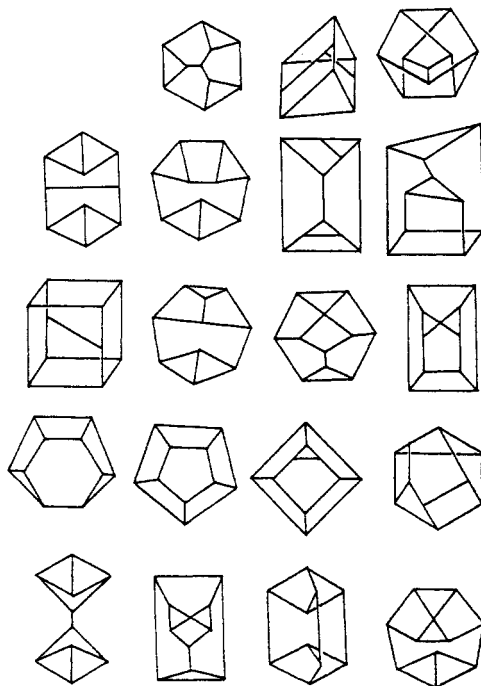
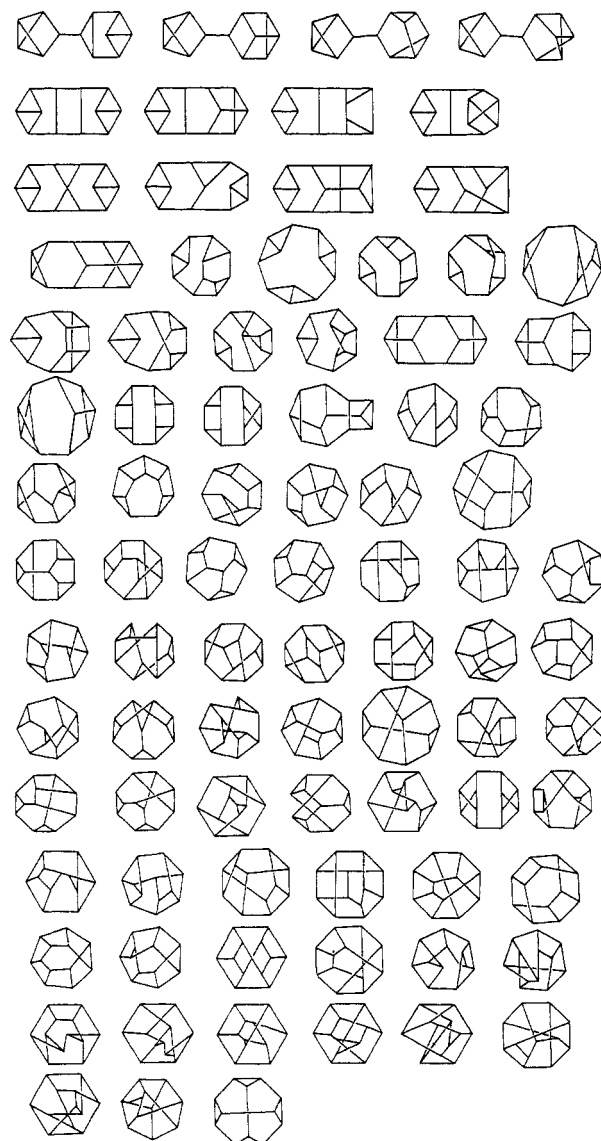
Figure 14. Generated trivalent regular graphs ($n = 10$).Figure 15. Generated trivalent regular graphs ($n = 12$).

Table VI. Result of Generation of Primary Component Vector

formula ^a	2	3	4	5	6	7	8	9	10
C _n H _{2n+2}	1	1	2	3	4	5	7	8	10
C _n H _{2n}	2	3	4	5	7	8	10	12	14
C _n H _{2n-2}	2	3	5	6	8	10	12	14	17
C _n H _{2n+2} O	2	3	5	7	9	12	15	18	22
C _n H _{2n} O	4	6	8	11	14	17	21	25	29
C _n H _{2n-2} O	3	6	9	12	16	20	24	29	34
C _n H _{2n+3} N	2	4	6	9	12	16	20	25	30
C _n H _{2n+1} N	5	8	11	15	19	24	29	35	b
C _n H _{2n+2} O ₂	4	6	9	12	16	b	b	b	b
C _n H _{2n} O ₂	6	9	13	17	22	b	b	b	b
C _n H _{2n+4} N ₂	5	8	13	18	25	b	b	b	b
C _n H _{2n+2} N ₂	10	17	26	35	46	b	b	b	b
C _n H _{2n+3} NO	6	15	21	28	b	b	b	b	b
C _n H _{2n+1} NO	11	17	24	32	41	b	b	b	b
C _n H _{2n+1} X	1	2	3	4	5	7	8	10	12
C _n H _{2n-1} X	2	3	4	6	7	9	11	b	b
C _n H _{2n} X ₂	2	3	4	5	7	8	10	12	14
C _n H _{2n-2} X ₂	2	3	5	6	8	b	b	b	b
C _n H _{2n} XY	2	3	4	5	8	b	b	b	b
C _n H _{2n-2} XY	2	3	5	6	8	b	b	b	b

^a X and Y = F, Cl, Br, and I. ^b Not calculated.

Table VII. Result of Generation of Secondary Component Vector

formula ^a		2	3	4	5	6	7	8	9	10
C _n H _{2n+2}	S1	1	2	6	13	25	46	82	134	218
	S2	0	0	1	2	5		16	26	42
C _n H _{2n}	S1	7	20	48	109	226	431	792	1381	2313
	S2	1	3	8	14	30	52	93	154	250
C _n H _{2n-2}	S1	14	44	121	283	626	1271	2423	4413	7696
	S2	2	9	22	46	92	172	309	528	878
C _n H _{2n+2} O	S1	4	9	23	46	87	158	272	444	714
	S2	2	3	7	13	24	41	68	107	165
C _n H _{2n} O	S1	26	72	174	389	790	1497	2721	4697	7815
	S2	6	15	33	69	134	239	414	684	1093
C _n H _{2n-2} O	S1	48	162	445	1062	2344	4740	9020	16358	28370
	S2	8	27	72	160	337	652	1186	2063	3448
C _n H _{2n+3} N	S1	4	10	25	52	101	185	323	536	866
	S2	2	4	8	16	31	55	94	153	241
C _n H _{2n+1} N	S1	41	116	279	618	1255	2387	4322	7476	b
	S2	8	21	47	99	193	351	612	1022	b
C _n H _{2n+2} O ₂	S1	10	23	53	104	195	b	b	b	b
	S2	6	13	27	50	89	b	b	b	b
C _n H _{2n} O ₂	S1	60	163	396	870	1755	b	b	b	b
	S2	14	34	79	164	317	b	b	b	b
C _n H _{2n+4} N ₂	S1	11	27	64	131	254	b	b	b	b
	S2	7	16	36	70	131	b	b	b	b
C _n H _{2n+2} N ₂	S1	370	1139	3000	6995	14966	b	b	b	b
	S2	53	108	369	825	1711	b	b	b	b
C _n H _{2n+3} NO	S1	16	93	188	361	b	b	b	b	b
	S2	9	46	89	164	b	b	b	b	b
C _n H _{2n+1} NO	S1	161	445	1066	2328	4677	b	b	b	b
	S2	32	82	189	395	765	b	b	b	b
C _n H _{2n+1} X	S1	2	5	11	20	37	66	108	176	278
	S2	1	2	4	7	12	20	32	49	74
C _n H _{2n-1} X	S1	11	29	71	154	305	573	1025	b	b
	S2	2	6	15	32	62	111	191	b	b
C _n H _{2n} X ₂	S1	3	7	13	25	46	76	127	204	314
	S2	3	6	11	20	34	54	85	129	190
C _n H _{2n-2} X ₂	S1	13	34	84	177	356	b	b	b	b
	S2	4	9	23	47	92	b	b	b	b
C _n H _{2n-2} XY	S1	13	34	84	177	356	b	b	b	b
	S2	4	9	23	47	92	b	b	b	b
C _n H _{2n} XY	S1	3	7	13	25	46	b	b	b	b
	S2	3	6	11	20	34	b	b	b	b

^a X and Y = F, Cl, Br, and I. ^b Not calculated.

that there are 85 regular trivalent graphs for $n = 12$. Our structure generator has succeeded in generating the identical number of distinctive graphs as shown in Figure 15.

Problems To Be Solved in the Structure Generator. The present structure generator has omission of such structures as shown in Table XI. They are not generated by one of the following reasons: (1) inorganic structures with no carbon atom (H₂O, NH₃, etc.) are not generated. (2) Structures

constructed by a carbon-containing component with a single bond and hydrogen are not generated. Examples are CH₄ (CH₃ + -H), HCHO (H- + -CHO), etc. The reason is that the hydrogen component (H-) has not been prepared. (3) Structures in which a methyl group is bonded to components with a saturated carbon as the root element attribute (CH₃-CH₃, CH₃CH₂CH₃, etc.) are not generated. This is because component no. 1, methyl group, cannot bond with

Table VIII. Number of Secondary Component Vectors from Which Structures Are Generated

formula ^a	2	3	4	5	6	7	8	9	10
C_nH_{2n+2}	0	0	1	2	5	9	16	26	42
C_nH_{2n}	0	2	5	10	22	43	81	143	238
C_nH_{2n-2}	0	3	9	24	59	126	251	454	783
$C_nH_{2n+2}O$	2	3	7	13	24	41	68	107	165
$C_nH_{2n}O$	3	9	23	51	106	196	346	574	917
$C_nH_{2n-2}O$	2	13	45	115	260	528	984	1719	2861
$C_nH_{2n+3}N$	2	4	8	16	31	55	94	153	241
$C_nH_{2n+1}N$	4	12	31	72	152	291	522	882	<i>b</i>
$C_nH_{2n+2}O_2$	5	10	20	36	63	<i>b</i>	<i>b</i>	<i>b</i>	<i>b</i>
$C_nH_{2n}O_2$	9	25	61	131	254	<i>b</i>	<i>b</i>	<i>b</i>	<i>b</i>
$C_nH_{2n+4}N_2$	6	13	28	53	98	<i>b</i>	<i>b</i>	<i>b</i>	<i>b</i>
$C_nH_{2n+2}N_2$	30	104	288	680	1449	<i>b</i>	<i>b</i>	<i>b</i>	<i>b</i>
$C_nH_{2n+3}NO$	8	36	68	124	<i>b</i>	<i>b</i>	<i>b</i>	<i>b</i>	<i>b</i>
$C_nH_{2n+1}NO$	21	59	145	315	631	<i>b</i>	<i>b</i>	<i>b</i>	<i>b</i>
$C_nH_{2n+1}X$	1	2	4	7	12	20	32	49	74
$C_nH_{2n-1}X$	1	4	10	23	46	86	149	<i>b</i>	<i>b</i>
$C_nH_{2n}X_2$	2	4	7	12	20	31	48	72	105
$C_nH_{2n-2}X_2$	2	5	14	31	60	<i>b</i>	<i>b</i>	<i>b</i>	<i>b</i>
$C_nH_{2n}XY$	2	4	7	12	20	<i>b</i>	<i>b</i>	<i>b</i>	<i>b</i>
$C_nH_{2n-2}XY$	2	5	14	31	60	<i>b</i>	<i>b</i>	<i>b</i>	<i>b</i>

^aX and Y = F, Cl, Br, and I. ^bNot calculated.**Table IX.** Number of Structures Generated by Authors' Method

formula ^a	2	3	4	5	6	7	8	9	10
C_nH_{2n+2}	<i>b</i>	<i>b</i>	<i>b</i>	<i>b</i>	5	9	18	35	75
C_nH_{2n}	<i>b</i>	2	5	10	25	56	139	388	852
C_nH_{2n-2}	<i>b</i>	3	9	26	77	222	654	1894	5495
$C_nH_{2n+2}O$	2	3	7	14	32	72	171	405	989
$C_nH_{2n}O$	3	9	26	74	211	596	1684	4745	13373
$C_nH_{2n-2}O$	<i>b</i>	13	55	205	745	2564	8608	28162	90769
$C_nH_{2n+3}N$	2	4	8	17	39	89	211	507	1238
$C_nH_{2n+1}N$	4	12	35	100	284	801	2258	6355	<i>c</i>
$C_nH_{2n+2}O_2$	5	11	28	69	179	<i>c</i>	<i>c</i>	<i>c</i>	<i>c</i>
$C_nH_{2n}O_2$	10	34	122	400	1313	<i>c</i>	<i>c</i>	<i>c</i>	<i>c</i>
$C_nH_{2n+4}N_2$	6	14	38	97	260	<i>c</i>	<i>c</i>	<i>c</i>	<i>c</i>
$C_nH_{2n+2}N_2$	32	145	652	2696	10612	<i>c</i>	<i>c</i>	<i>c</i>	<i>c</i>
$C_nH_{2n+3}NO$	8	57	152	405	<i>c</i>	<i>c</i>	<i>c</i>	<i>c</i>	<i>c</i>
$C_nH_{2n+1}NO$	23	87	308	1041	3418	<i>c</i>	<i>c</i>	<i>c</i>	<i>c</i>
$C_nH_{2n+1}X$	1	2	4	8	17	39	89	211	507
$C_nH_{2n-1}X$	1	4	12	35	100	284	801	<i>c</i>	<i>c</i>
$C_nH_{2n}X_2$	2	4	9	21	52	129	332	859	2261
$C_nH_{2n-2}X_2$	2	6	22	71	231	<i>c</i>	<i>c</i>	<i>c</i>	<i>c</i>
$C_nH_{2n}XY$	2	5	12	31	80	<i>c</i>	<i>c</i>	<i>c</i>	<i>c</i>
$C_nH_{2n-2}XY$	2	8	30	108	367	<i>c</i>	<i>c</i>	<i>c</i>	<i>c</i>

^aX and Y = F, Cl, Br, and I. ^bNot possible to calculate by ours (see Table XI). ^cNot calculated.**Table X.** Number of Structures Generated by Reed's¹ and Smith's⁸ Methods

formula ^a	ref	2	3	4	5	6	7	8	9	10
C_nH_{2n+2}	1	1	1	2	3	5	9	18	35	75
C_nH_{2n}	8	1	2	5	10	25	<i>b</i>	<i>b</i>	<i>b</i>	<i>b</i>
C_nH_{2n-2}	8	1	3	9	26	77	<i>b</i>	<i>b</i>	<i>b</i>	<i>b</i>
$C_nH_{2n+2}O$	8	2	3	7	14	<i>b</i>	<i>b</i>	<i>b</i>	<i>b</i>	<i>b</i>
$C_nH_{2n}O$	8	3	9	26	74	<i>b</i>	<i>b</i>	<i>b</i>	<i>b</i>	<i>b</i>
$C_nH_{2n-2}O$	8	3	13	55	205	<i>b</i>	<i>b</i>	<i>b</i>	<i>b</i>	<i>b</i>
$C_nH_{2n+3}N$	8	2	4	9	17	<i>b</i>	<i>b</i>	<i>b</i>	<i>b</i>	<i>b</i>
$C_nH_{2n+1}N$	8	4	12	35	100	<i>b</i>	<i>b</i>	<i>b</i>	<i>b</i>	<i>b</i>
$C_nH_{2n+2}O_2$	8	5	11	28	<i>b</i>	<i>b</i>	<i>b</i>	<i>b</i>	<i>b</i>	<i>b</i>
$C_nH_{2n+4}N_2$	8	6	14	36	<i>b</i>	<i>b</i>	<i>b</i>	<i>b</i>	<i>b</i>	<i>b</i>
$C_nH_{2n+1}X$	1	1	2	4	8	17	39	89	211	507
$C_nH_{2n}X_2$	1	2	4	9	21	52	129	332	859	2261
$C_nH_{2n}XY$	1	2	5	12	31	80	210	555	1479	3959

^aX and Y = F, Cl, Br, and I. ^bNot calculated.**Table XI.** Structures Not Possibly Generated by Authors' Method

type	structures
1	H ₂ O, HF, HCl, HBr, HI, H ₂ SO ₃ , NH ₃ , etc.
2	CH ₄ , HCHO, HCHS, HCHNH, HC≡CH, HCN, etc.
3	CH ₃ CH ₃ , CH ₃ CH ₂ CH ₃ , (CH ₃) ₂ CHCH ₃ , etc.
4	CH ₂ =CH ₂ , O=C=C=O, O=C=C=S, O=C=O, etc.

saturated carbon due to its bond attribute. (4) Structures constructed by bonding between the components only with double bonds (NH=C=NH, etc.) or similar structures

(O=C=O, etc.) are not generated.

These kinds of structures originated from molecular formulas with low molecular weights do not have many isomeric structures. Since they can be easily generated manually, it has been decided not to generate them for the sake of a simpler algorithm. Structures concerned with limitation 3 can be generated from the generator by using the primary components.

Comments on CPU Time. Table XII shows the average cpu time (HITAC M-200H) required for generation of one

Table XII. Time Required for Generation of One Structure

formula ^a	av CPU time (ms)	formula ^a	av CPU time (ms)
C _n H _{2n+2}	9.52	C _n H _{2n+3} N	4.51
C _n H _{2n}	8.13	C _n H _{2n+1} N	9.04
C _n H _{2n-2}	8.87	C _n H _{2n+1} X	3.76
C _n H _{2n+2} O	4.03	C _n H _{2n} X ₂	4.39
C _n H _{2n} O	9.88	C _n H _{2n} XY	4.25
C _n H _{2n-2} O	16.99		

^a X and Y = F, Cl, Br, and I.**Table XIII.** Time Required for Generation of Trivalent Regular Graphs

n	no. of generated graphs	total CPU time (s)	total CPU time/ no. of generated graphs
4	1	0.049	0.049
6	2	0.072	0.036
8	5	0.221	0.044
10	19	2.832	0.149
12	85	40.966	0.832

structure from various types of molecular formulas. In general, the more indexes of hydrogen deficiency, the more cpu time required to generate one structure. This may be due to the increase of secondary component vectors unnecessary for the generation of appropriate structures. The average cpu time was 8 ms for the generation of one structure.

Table XIII shows the cpu time required to generate regular trivalent graphs. The graphs are known for their difficulty in coding. For example, the coded form of dodecahedrone with $n = 20$ is often cited as one of the worst cases of coding. Regular trivalent graphs require much longer cpu time for their

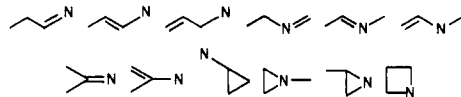
generation than normal chemical graphs. For example, only 3.5 s is required to generate 355 structures for a saturated hydrocarbon with 12 carbon atoms, while 41 s is required to generate 85 tertiary regular graphs for a compound with $n = 12$. This is due to the long time required to avoid giving rise to the same graphs, because n nodes are all equivalent in the regular graph. Hereafter, it is necessary to develop an algorithm for high-speed generation of these graphs containing many equivalent nodes.

ACKNOWLEDGMENT

We thank the Computer Center, Institute for Molecular Science, for affording computation facilities (Hitachi M-200 computer systems).

REFERENCES AND NOTES

- (1) Read, R. C. "The Enumeration of Acyclic Chemical Compounds". In "Chemical Applications of Graph Theory"; Balaban, A. T., Ed.; Academic Press: London, 1976.
- (2) Lederberg, J.; Sutherland, G. L.; Buchanan, B. G.; Feigenbaum, E. A.; Robertson, A. V.; Duffield, A. M.; Djerassi, C. *J. Am. Chem. Soc.* **1969**, *91*, 2973.
- (3) Shelley, C. A.; Hays, T. R.; Munk, M. E.; Roman, R. V. *Anal. Chim. Acta* **1978**, *103*, 121.
- (4) Abe, H.; Fujiwara, I.; Nishimura, T.; Okuyama, T.; Kida, T.; Sasaki, S. *Comput. Enhanced Spectrosc.* **1983**, *1*, 55.
- (5) Kudo, Y.; Sasaki, S. *J. Chem. Soc.* **1974**, *14*, 200.
- (6) Kudo, Y.; Sasaki, S. *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 43.
- (7) Oshima, T.; Ishida, Y.; Saito, K.; Sasaki, S. *Anal. Chim. Acta* **1980**, *122*, 95.
- (8) Serov, V. V.; Elyashberg, M. E.; Gribov, L. A. *J. Mol. Struct.* **1976**, *31*, 381.
- (9) Incorrect structures are


- (10) Smith, D. H. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 203.
- (11) Randić, M. *Acta Crystallogr., Sect. A* **1978**, *A34*, 275.