Algorithms for Unique and Unambiguous Coding and Symmetry Perception of Molecular Structure Diagrams. 5. Unique Coding by the Method of "Orbit Graphs"

MASAHIRO UCHINO

Laboratory of Resources Utilization, Tokyo Institute of Technology, 4259 Nagatsudacho, Midoriku, Yokohama, 227 Japan

Received August 1, 1981

The "orbit graph" proposed in this paper contains all information about symmetry which is necessary and sufficient for unique coding of a chemical graph, and provides a simple and efficient coding procedure when used in combination with the concept of the "fitness" of the node for the number i. For instance, the dodecahedron, often cited as the worst case of unique coding, can be coded even without a computer (i.e., executing the proposed algorithm manually). The validity of the procedure is proved mathematically, and several principles which are useful for improvement of other algorithms are included.

INTRODUCTION

So far many canonical codes of chemical graphs have been proposed.[1-18] However, the mathematical clarity of the method used for the determination of the canonical code, how it treats symmetry of graphs, and how it behaves under the worst case seem to be more important than the canonical code.

In this paper we have further developed the previous theoretical approach[15-18] and clarified the role of symmetry in unique coding. On the basis of the mathematical analysis of the coding problem, we have introduced two new concepts, i.e., "orbit graph" and "fitness" of the node for the number i.

The "orbit graph" used in the present approach is a directed graph which can be easily derived from the chemical graph to be coded (i.e., no difficult cases arise for the determination) and contains less information than the symmetry group itself (i.e., for instance, we cannot know all orbits of the symmetry group directly from the orbit graph), but it contains all information about symmetry which is necessary for coding and can be used to suppress completely the duplicated numberings due to symmetry of a graph.

Another concept, the "fitness" $f(i, v)$ of node v for the number i is a list of integers which is assigned, by some rule f, to each unnumbered node v, depending on the previous assignment of the numbers 1, 2, ..., $i-1$, and this is used for the choice of nodes receiving the number i.

The canonical numbering is defined in terms of the "fitness", and the numbering (i.e., the node sequence $(v_1, v_2, ..., v_n)$) which provides the lexicographically greatest sequence of the fitness, $(f(1, v_1), f(2, v_2), ..., f(n, v_n))$ is the canonical numbering. With the choice of an appropriate rule f, the code associated with the canonical numbering is identical with the unique code previously reported.[16]

Before presenting details of the mathematical treatment, we first explain the method intuitively.

I. INTUITIVE EXPLANATION OF THE METHOD

"Orbit Graph". Suppose a simple graph A shown in Figure 1. (1) Choose arbitrarily a node w and draw arrows from the node w toward the nodes x, y, and z which can be mapped onto the chosen node w by a symmetry operation of the graph (see graph A_1). (2) Choose arbitrarily a noide z among unchosen nodes and examine whether there exist other nodes than z which can be mapped onto z by a symmetry operation fixing previously chosen node w. Since such a node is not found, no arrow is drawn from the node z. (3) Choose arbitrarily a node y among the unchosen nodes and draw an arrow from the node y toward the node x which is mapped onto y by a symmetry operation fixing previously chosen nodes w and z. Thus, we have a directed graph, A_2, which represents the orbit graph of the chemical graph A.

For a more precise definition of orbit graph, suppose the node sequence $(v_1, v_2, ..., v_n)$ according to the order in which we have chosen the nodes and let G be the set of symmetry operations f (i.e., by operation f node v is mapped to node $f(v)$).[19] Then, the orbit graph is defined as a directed graph obtained by the following procedure. (1) Draw an arrow from the node v_1 toward the node x if and only if there exists an operation f in G such that $f(v_1) = x \neq v_1$. (2) For $i = 2, 3, ..., n-1$, draw an arrow from the node v_i toward the node x if and only if there exists an operation f in G such that $f(v_1) = v_1, f(v_2) = v_2, ..., f(v_{i-1}) = v_{i-1}$ and $f(v_i) = x \neq v_i$.

For a given chemical graph, several orbit graphs can be produced according to the order in which the nodes are chosen, but these are equally useful for coding of the chemical graph. The essential idea of the orbit graph is based on the fact that we can produce all representations of a chemical graph without duplication by assigning an earlier number to the node u than the node v for each arc uv of its orbit graph, as will be shown later.

Several examples of the orbit graph are shown in Figure 2 and briefly described below.

Graph B. There exists no node which can be mapped onto the node a, and no arrow from the node a is drawn. When we fix the node a, the nodes c and d can be mapped onto the node b. Hence, arrows from the node b toward the nodes c and d are drawn. When we fix the nodes a and b, only the node d can be mapped to the node c. Hence, an arrow is drawn from the node c toward node d. If we fix the nodes a-c, all nodes of the graph are fixed, and no more arrows are drawn.

Graph C. All nodes of a dodecahedron are equivalent, and arrows are drawn from a chosen node, say node a, toward all other nodes. When we fix node a, only the nodes c and d can be mapped onto the node b. Hence, arrows are drawn from the node b toward the nodes c and d. When we fix the nodes a and b, the node d can be mapped to the node c, and an arrow is drawn from the node c toward the node d. If we fix the nodes a-c, all nodes are fixed.

Graph D. From arbitrarily chosen node a, arrows are drawn toward all other nodes which are equivalent to the node a. And if we choose the node e and fix the node a, only the node d can be mapped to the node e. Hence an arrow is drawn from the node e toward the node d. When we fix the nodes a and e, all nodes are fixed and no more arrows are drawn.

"Fitness" and Procedure for Canonical Numbering. Assume we have a rule f which assigns to each unnumbered node v a list $f(i, v)$ depending on the previous assignment of the numbers 1, 2, ..., $i-1$. We use the term pseudocanonical numbering

0095-2338/82/1622-0201$01.25/0

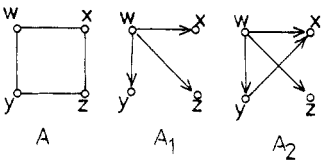

Figure 1. Determination of "orbit graph".

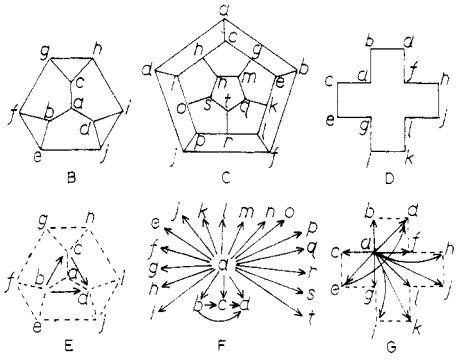

Figure 2. Some examples of graphs and the associated "orbit graphs". The orbit graphs E–G correspond to the graphs B–D, respectively, and the dotted lines in E and G show the bonds in the chemical graphs.

to denote a numbering obtained by choosing a node with the lexicographically maximum list for each $i = 1, 2, ..., n$, and if the rule f satisfies the following conditions, each list $f(i, v)$ will be called the "fitness" of node v for the number i.

Condition 1. If the representations (i.e., pairs of bond-matrix and atomic vectors) based on pseudocanonical numberings are identical, then the sequences of the list are identical.

Condition 2. If the sequences of the list for pseudocanonical numberings are identical, then the representations based on these numberings are identical.

The fitness can be introduced in various ways by using the invariant number $s(v)$ of node v and the bond values $m(v', v)$ representing the bond multiplicities between the nodes v' and v. The invariant number $s(v)$ can be given by the combination of atomic value $a(v)$ and another property, $h(v)$, of node v (i.e., for instance, the number computed by an appropriate illumination function[16]).

$$s(v) = ba(v) + h(v) \quad b = 1 + \max\{h(v)|v \in V\} \quad (1)$$

Alternatively,

$$s(v) = bh(v) + a(v) \quad b = 1 + \max\{a(v)|v \in V\} \quad (2)$$

where V is the set of nodes.

We give two examples of the fitness described in terms of the invariant numbers $s(v)$ and the bond values $m(v, v')$ as follows.

Example 1:

$$f(1, v) = (s(v))(v \in V)$$

$$f(i, v) = (s(v), m(v_1, v), m(v_2, v), ..., m(v_{i-1}, v)) \quad (v \in V - \{v_1, v_2, ..., v_{i-1}\}, i \in \{2, 3, ..., n\}) \quad (3)$$

where each v_j denotes the node with the number j.

The canonical numbering by this fitness provides the lower triangular maximum code.[16]

Example 2:

$$f(1, v) = (s(v))(v \in V)$$

$$f(i, v) = (-t(i), m(v_{t(i)}, v), s(v), m(v_{t(i)+1}, v), m(v_{t(i)+2}, v), ..., m(v_{i-1}, v)) \; (v \in V - \{v_1, v_2, ..., v_{i-1}\}, i \in \{2, 3, ..., n\}) \quad (4)$$

where $v_{t(i)}$ is the node with the smallest number among the nodes which are adjacent to unnumbered nodes; i.e., $t(i) = \min\{j|j \in \{1, 2, ..., i-1\}, m(v_j, w) \neq 0$ for a node $w \in V - \{v_1, v_2, ..., v_{i-1}\}\}$ $(i \in \{2, 3, ..., n\})$.

The pseudocanonical numbering by this fitness proceeds from a center to the outer spheres, and the sequence of $t(i)$ values ($i = 2, 3, ..., n$) provides a spanning tree of the graph (i.e., the edges $(v_{t(i)}, v_i)$ ($i = 2, 3, ..., n$) define a spanning tree of the graph).

The fitness permits us to construct the pseudocanonical numbering simply by choosing a node among unnumbered nodes which has the maximum fitness for the number i ($i = 1, 2, ..., n$). Usually the number of the pseudocanonical numberings is very small for the unsymmetrical graph, although there exist $n!$ pseudocanonical numberings for the complete graph of n nodes.

The orbit graph can be used to reduce the number of pseudocanonical numberings to be examined, and we need to examine only representatives of pseudocanonical numberings, defined by the orbit graph.

If we have a node sequence $(v_1, v_2, ..., v_n)$ obtained by any pseudocanonical numbering, the pseudocanonical numbering will be called representative if and only if there exists in its orbit graph no arrow directed from any $v \in V - \{v_1, v_2, ..., v_{i-1}\}$ toward each node v_i for $i = 1, 2, ..., n$.

The number of representatives of pseudocanonical numbering is very small in general, and, for instance, there exists only one representative for pseudocanonical numbering of the complete graph.

In the present approach, the canonical numbering is selected from the representatives of pseudocanonical numbering, and we give descriptively a simple procedure to determine it.

Procedure. For $i = 1, 2, ..., n$, choose among unnumbered nodes $V_i = V - \{v_1, v_2, ..., v_{i-1}\}$ a node which has the maximum fitness for the number i and receives, in the orbit graph, no arrow from any $v \in V_i$. If such a node is not found, change the previous assignment of the number to another possible assignment. Examining as above the representatives of pseudocanonical numbering, choose one numbering which provides the lexicographically maximum sequence of the fitness, i.e., $(f(1, v_1), f(2, v_2), ..., f(n, v_n))$. This chosen numbering is the canonical numbering.

For instance, in graph A of Figure 1 all nodes w–z have the same fitness (in the sequel, we assume the fitness of example 1) for the number 1, but only the node w has no arrow toward it. Thus, the node w receives the number 1. Among the remaining nodes, the nodes x and y have the maximum fitness for the number 2, but an arrow is directed from the node y to the node x. Hence, the number 2 is assigned to the node y. The number 3 is assigned to the node x since this node has greater fitness for the number 3 than the node z and has no arrow directed toward it from the node z. Thus, we have the canonical sequence of nodes, i.e., w, y, x, z.

The canonical numberings for the graphs shown in Figure 2 can be easily determined in a similar manner and are described below.

Graph B. The invariant numbers computed for each node by the method used in a previous paper are 3 for the node a, 2 for the nodes b–d, and 1 for the nodes e–j. Thus the node having the maximum fitness for the number 1 receives the number 1. The nodes b–d have the same maximum fitness for the number 2, but the node c receives an arrow from the node b, and the node d receives an arrow from the node c. Hence, the number 2 is assigned to the node b. In the same way, number 3 is assigned to node c, and number 4 is assigned to node d. Among the remaining nodes, the nodes e and f have the maximum fitness for the number 5, i.e., 1, 0, 1, 0, 0. If we assign number 5 to the node f, then numbers 6–10 must be assigned respectively to nodes e, g, h, j, and i. Alternatively, if we assign number 5 to node e, we must assign number 6 to

Table I. Canonical Numberings

graph[a]	assignment of numbers
B	numbers: 1 2 3 4 5 6 7 8 9 10
	nodes: $a \rightarrow b \rightarrow c \rightarrow d \rightarrow f \rightarrow e \rightarrow g \rightarrow h \rightarrow j \rightarrow i$
	$\rightarrow e \rightarrow f \rightarrow g$
C	numbers: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
	nodes: $a \rightarrow b \rightarrow c \rightarrow d \rightarrow e \rightarrow f \rightarrow g \rightarrow h \rightarrow j \rightarrow i \rightarrow k \rightarrow l \rightarrow m \rightarrow n \rightarrow p \rightarrow o \rightarrow q \rightarrow r \rightarrow s \rightarrow t$
	$\rightarrow f \rightarrow e \rightarrow g$
D	numbers: 1 2 3 4 5 6 7 8 9 10 11 12
	nodes: $a \rightarrow b \rightarrow c$
	$\rightarrow c \rightarrow b \rightarrow e \rightarrow d \rightarrow g \rightarrow f \rightarrow i \rightarrow h \rightarrow k \rightarrow j \rightarrow l$

[a] See Figure 2.

the node *f*, number 7 to node *g*, and so on. In the latter assignment, the maximum fitness for number 7 is lexicographically smaller than that of the former assignment, and hence, the former assignment is the canonical numbering.

Graph C. The assignment of numbers 1–4, respectively, to nodes *a*–*d* is evident from the orbit graph. The assignment of number 5 is at the point of choice, since nodes *e* and *f* have the same maximum fitness for number 5 (i.e., 1, 0, 1, 0, 0). If we assign number 5 to node *e*, we must assign numbers 6–20, respectively, to nodes *f*, *g*, *h*, *j*, *i*, *k*, *l*, *m*, *n*, *p*, *o*, *q*, *r*, *s*, and *t*. Alternatively, if we assign number 5 to node *f*, we must assign number 6 to node *e*, number 7 to node *g*, and so on. In the latter assignment, the maximum fitness for number 6 is equal to that of the former assignment, but the maximum "fitness" for number 7 is smaller than that of the former. Hence, the former assignment is the canonical numbering.

Graph D. The node *a* having no arrow toward it receives number 1. The assignment of number 2 is at the point of choice since nodes *b* and *c* have the same maximum fitness for number 2. If we assign number 2 to node *b*, number 3 must be assigned to node *c*, but there is no node to which number 4 can be assigned since node *d* having the maximum fitness for number 4 receives an arrow from node *e*. Therefore, we go back to the latest point of choice, assign number 2 to node *c*, and continue to assign numbers 3–12, respectively, to nodes *b*, *e*, *d*, *g*, *f*, *i*, *h*, *k*, *j*, and *l*.

Table I summarizes the process for the canonical numberings.

II. MATHEMATICAL TREATMENT

In this section we present the mathematical treatment of the coding procedure described in the preceding section. Several mathematical theorems which support the validity of the coding procedure are presented and proved rigorously. The notation used here is different from that of our previous papers[15-17] and is more standard.

Application of Idea of Coset to the Coding Problem. Any one-to-one mapping of $\{1, 2, ..., n\}$ onto $\{1, 2, ..., n\}$ is known as a permutation, and the set S_n of all permutations of $\{1, 2, ..., n\}$ is usually called the symmetric group of degree *n*. For permutations τ and σ, the product $\tau\sigma$ is defined as in eq 5.

$$\tau\sigma = \begin{pmatrix} 1 & 2 & \dots & n \\ \tau(\sigma(1)) & \tau(\sigma(2)) & \dots & \tau(\sigma(n)) \end{pmatrix} \quad (5)$$

Since permutation is an abstract mathematical idea, there are two possible interpretations of permutation σ as the renumbering of the nodes in a chemical graph.

Interpretation A. We assign the number $\sigma(i)$ to the node having the old number *i*. In this interpretation, the bond matrix $\mathbf{B} = [b_{ij}]$ and the atomic vector $\boldsymbol{a} = [a_i]$ which are based on the initial numbering of nodes are transformed respectively to $\mathbf{B}' = [b'_{ij}]$ and $\boldsymbol{a}' = [a'_i]$ as given in eq 6.

$$b'_{ij} = b_{\sigma^{-1}(i)\sigma^{-1}(j)} \quad a'_i = a_{\sigma^{-1}(i)} \text{ for } i, j \in \{1, 2, ..., n\} \quad (6)$$

Table II. Coset Partitioning by Bond-Atom Symmetry[a]

left cosets	right cosets	bond-atom symmetry
(1 2 3)	(1 2 3)	(1 2 3)
(3 2 1)	(3 2 1)	(3 2 1)
(1 3 2)	(1 3 2)	graph
(2 3 1)	(3 1 2)	1 2 3
		x-y-x
(2 1 3)	(2 1 3)	
(3 1 2)	(2 3 1)	

[a] Permutations are shown in the form of the permutation vector and the separation of cosets is shown by lines.

Interpretation B. We assign the new number *i* to the node having the old number $\sigma(i)$. In this case, the transformed bond matrix $\mathbf{B}'$ and atomic vector $\boldsymbol{a}'$ are given by eq 7.

$$b'_{ij} = b_{\sigma(i)\sigma(j)} \quad a' = a_{\sigma(i)} \text{ for } i, j \in \{1, 2, ..., n\} \quad (7)$$

We adopt the interpretation B, and we understand the renumbering σ to mean the renumbering by permutation σ, which assigns the number *i* to each node having the old number $\sigma(i)$, for $i = 1, 2, ..., n$. We denote the operation associated with renumbering σ as applied to the bond matrix, atomic vector, and node sequence by P_σ (i.e., $\mathrm{P}_\sigma\mathbf{B} = \mathbf{B}'$, $\mathrm{P}_\sigma\boldsymbol{a} = \boldsymbol{a}'$, and $\mathrm{P}_\sigma(v_1, v_2, ..., v_n) = (v_{\sigma(1)}, v_{\sigma(2)}, ..., v_{\sigma(n)})$). Then, we have eq 8 and 9.

$$(\mathrm{P}_\sigma)^{-1} = \mathrm{P}_{\sigma^{-1}} \quad (8)$$

$$\mathrm{P}_{\tau\sigma} = \mathrm{P}_\sigma\mathrm{P}_\tau \quad (9)$$

The symmetry of a chemical graph is defined by the subgroups $G^{(\mathbf{B})}$ and $G^{(\boldsymbol{a})}$ of S_n (eq 10–12).

$$\text{bond symmetry: } G^{(\mathbf{B})} = \{\sigma | \sigma \in S_n, \mathrm{P}_\sigma\mathbf{B} = \mathbf{B}\} \quad (10)$$

$$\text{atomic symmetry: } G^{(\boldsymbol{a})} = \{\sigma | \sigma \in S_n, \mathrm{P}_\sigma\boldsymbol{a} = \boldsymbol{a}\} \quad (11)$$

$$\text{bond-atom symmetry: } G^{(\mathbf{B})} \cap G^{(\boldsymbol{a})} \quad (12)$$

A right coset of bond-atom symmetry in S_n is a subset of the form $(G^{(\mathbf{B})} \cap G^{(\boldsymbol{a})})\sigma = \{\lambda | \lambda = \tau\sigma, \tau \in G^{(\mathbf{B})} \cap G^{(\boldsymbol{a})}\}$ for some σ in S_n, and a left coset of bond-atom symmetry is a subset of the form $\sigma(G^{(\mathbf{B})} \cap G^{(\boldsymbol{a})}) = \{\lambda | \lambda = \sigma\tau, \tau \in G^{(\mathbf{B})} \cap G^{(\boldsymbol{a})}\}$. The right (left) cosets of bond-atom symmetry in S_n form a partition of bond-atom symmetry in S_n, i.e., the union of all the right (left) cosets is S_n itself, and any pair of distinct cosets has an empty intersection. An example is illustrated in Table II to show that the group S_3 is the disjoint union of cosets of bond-atom symmetry. We can observe that permutations in the same right coset provide the same bond matrix and atomic vector after renumbering by them.

Now we define the right (left) coset representatives of bond-atom symmetry in S_n to be a subset R_0 (L_0) of S_n obtained by choosing one element from each of the cosets of bond-atom symmetry in S_n, without choosing two elements

from the same coset. Then, $S_n = (G^{(B)} \cap G^{(a)})R_0 = L_0(G^{(B)} \cap G^{(a)})$.

Suppose a sequence of subgroups of bond-atom symmetry defined by $G_1 = (G^{(B)} \cap G^{(a)})$ and $G_i = \{\sigma | \sigma \in G_1, \sigma(j) = j, j = 1, 2, ..., i-1\}$ $(2 \le i \le n)$. Evidently $G_1 \supset G_2 \supset ..., \supset G_n$, and G_n consists only of the identity permutation.

Since G_{i+1} is a subgroup of G_i, G_i can be given by the union of cosets of G_{i+1} in G_i, and we define right (left) coset representatives of G_{i+1} in G_i to be a subset R_i (L_i) of G_i obtained by choosing one element from each coset of G_{i+1} in G_i. Then, we have the following relations.

$G_1 = G_2R_1$	$G_1 = L_1G_2$
$G_2 = G_3R_2$	$G_2 = L_2G_3$
.....	.....
$G_i = G_{i+1}R_i$	$G_i = L_iG_{i+1}$
.....	.....
$G_{n-1} = G_nR_{n-1} = R_{n-1}$	$G_{n-1} = L_{n-1}G_n = L_{n-1}$

Hence

$$G^{(B)} \cap G^{(a)} = G_1 = L_1L_2 ... L_{n-1} = R_{n-1}R_{n-2} ... R_1 \quad (13)$$

Accordingly, if we determine a sequence of sets of representatives, i.e., $L_1, L_2, ..., L_{n-1}$, the bond-atom symmetry of a chemical graph can be given by their product. This result is extremely important for computing the bond-atom symmetry, since the determination of $L_1, L_2, ..., L_i, ...$ is much easier than the direct determination of the bond-atom symmetry.

Coding of a chemical graph does not require all the information of its bond-atom symmetry, and we need only the orbit O_i of the subgroup G_i, which contains the number i, i.e., $O_i = \{\sigma(i) | \sigma \in G_i\}$ $(i \in \{1, 2, ..., n\})$, since all representations of a chemical graph (i.e., all distinct pairs of bond matrix and atomic vector) are determined without duplication by R_0, and R_0 is determined by O_i $(i = 1, 2, ..., n)$.

Lemma I. Suppose bond matrix $\mathbf{B}$, atomic vector $\boldsymbol{a}$, and the bond-atom symmetry, based on an arbitrary numbering of nodes, and let $R_0 = \{\sigma, \sigma', ..., \sigma''\}$ be all right coset representatives of bond-atom symmetry in S_n. Then, all representations of the chemical graph are produced without duplication by the renumberings of nodes with permutations σ, σ', ..., σ''.

Proof. First we show that there is no duplication. Assume that for τ, $\sigma \in R_0$, $\mathbf{P}_\tau\mathbf{B} = \mathbf{P}_\sigma\mathbf{B}$, and $\mathbf{P}_\tau\boldsymbol{a} = \mathbf{P}_\sigma\boldsymbol{a}$. Then, $\mathbf{B} = (\mathbf{P}_\tau)^{-1}\mathbf{P}_\sigma\mathbf{B} = (\mathbf{P}_{\sigma\tau^{-1}})\mathbf{B}$ and $\boldsymbol{a} = (\mathbf{P}_{\sigma\tau^{-1}})\boldsymbol{a}$. Hence $\sigma\tau^{-1}$ is a member of bond-atom symmetry, and this implies that $\sigma \in (G^{(B)} \cap G^{(a)})\tau$. Since τ and σ are right coset representatives, we reached a contradiction. Thus, there is no duplication.

Now we show that all representations are produced. Suppose a permutation $\tau \in S_n$ such that $\tau \notin R_0$. Since R_0 is the set of all representatives, there exists σ in R_0 such that $\tau \in (G^{(B)} \cap G^{(a)})\sigma$. Hence, we can write $\tau = \lambda\sigma$ by choosing an approxiate $\lambda \in G^{(B)} \cap G^{(a)}$. Hence, $\mathbf{P}_\tau = \mathbf{P}_{\lambda\sigma} = \mathbf{P}_\sigma\mathbf{P}_\lambda$. Consequently, $\mathbf{P}_\tau\mathbf{B} = \mathbf{P}_\sigma\mathbf{P}_\lambda\mathbf{B} = \mathbf{P}_\sigma\mathbf{B}$ and $\mathbf{P}_\tau\boldsymbol{a} = \mathbf{P}_\sigma\boldsymbol{a}$. Hence, the bond matrix $\mathbf{P}_\tau\mathbf{B}$ and the atomic vector $\mathbf{P}_\tau\boldsymbol{a}$ are equivalent to those produced by a member of R_0, and hence we showed that all representations are produced.

Directly from lemma I, we can say there exist $|S_n|/|G^{(B)} \cap G^{(a)}|$ representations for a chemical graph of n nodes. For instance, there exist three possible representations (i.e., 4!/8) for cyclobutane having a symmetry of order 8.

If $(v_1, v_2, ..., v_n)$ is the node sequence of the initial numbering used to define bond-atom symmetry, the orbit graph defined in section I can be obtained simply by drawing an arrow from the node v_i to the node v_y for each pair (i, y) such that $y \in O_i, y \ne i, i \in \{1, 2, ..., n-1\}$. The numberings of nodes which assign an earlier number to the node u than the node v for each arc uv of the orbit graph are associated with the permutations of renumberings which provide such numberings. If σ is any such permutation, for each arc v_iv_j the new number $\sigma^{-1}(i)$ assigned to the node v_i is smaller than the number $\sigma^{-1}(j)$ assigned to the node v_j by σ, and all numberings which obey the orbit graph can be given by the renumberings $\{\sigma | \sigma \in S_n, \sigma^{-1}(i) < \sigma^{-1}(y)$ for each $y \in O_i$ such that $y \ne i$, and for $i = 1, 2, ..., n-1\}$.

Lemma II. The set $R = \{\sigma | \sigma \in S_n, \sigma^{-1}(i) < \sigma^{-1}(y)$, for each $y \in O_i$ such that $y \ne i$ and for $i = 1, 2, ..., n-1\}$ is a set of right coset representatives of bond-atom symmetry in S_n; i.e., the set R satisfies the condition of R_0.[20]

Proof. We define the order among the permutations of S_n by the lexicographic order of the permutation vector, i.e., $\sigma(1)$, $\sigma(2)$, ..., $\sigma(n)$. And we say permutation τ is smaller than permutation σ if and only if $\tau(1) < \sigma(1)$ or $\tau(j) = \sigma(j)$ for $j = 1, 2, ..., k$ and $\tau(k+1) < \sigma(k+1)$. If permutation τ is the smallest permutation of a set X, then we write $\tau \ll X$.

(1) First we show that if $\sigma \in R$, then $\sigma^{-1} \ll \sigma^{-1}(G^{(B)} \cap G^{(a)})$. Suppose any $\tau \in G^{(B)} \cap G^{(a)}$ such that $\sigma^{-1}\tau \ne \sigma^{-1}$. Then, for a certain integer s, $\sigma^{-1}\tau(j) = \sigma^{-1}(j)$, $1 \le j < s$, and $\sigma^{-1}\tau(s) \ne \sigma^{-1}(s)$. Hence, $\tau(j) = j$, $1 \le j < s$, and $\tau(s) \ne s$. By the definition of $G_1, G_2, ..., G_n$, such a τ must be a member of G_s, and hence, $\tau(s) \in O_s$. Since σ is a member of R, we reached $\sigma^{-1}(s) < \sigma^{-1}(\tau(s))$. Thus, $\sigma^{-1} \ll \sigma^{-1}(G^{(B)} \cap G^{(a)})$.

(2) Now we show that if $\sigma^{-1} \ll \sigma^{-1}(G^{(B)} \cap G^{(a)})$, then σ is a member of R. For any $y \in O_i$ and $i \in \{1, 2, ..., n-1\}$, we can find a permutation $\tau \in G^{(B)} \cap G^{(a)}$ such that $\tau(j) = j$, $1 \le j < i$, and $\tau(i) = y$, and evidently σ^{-1} is smaller than or equal to $\sigma^{-1}\tau$. Since $\tau(j) = j$ for $1 \le j < i$, this implies $\sigma^{-1}(i) < \sigma^{-1}(y)$ for $y \ne i$. Hence, σ is a member of R.

(3) From parts 1 and 2, σ^{-1} for $\sigma \in R$ is the smallest permutation in a left coset of bond-atom symmetry, and $\{\sigma^{-1} | \sigma \in R\}$ defines a set of left coset representatives. Hence, R is a set of right coset representatives.

Theorem I. All representations of a chemical graph are produced without duplication by the numberings of nodes which obey its orbit graph and the number of representations is given by $n!/(1 + \rho(v))(1 + \rho(v')) ... (1 + \rho(v''))$, where each $\rho(v)$ denotes the out degree of node v in the orbit graph.

Proof. By lemma II, the permutations which provide the numberings obeying the orbit graph are right coset representatives of bond-atom symmetry in S_n. And hence, the numberings obeying the orbit graph produce all representations without duplication (lemma I). If $(v_1, v_2, ..., v_n)$ is the initial numbering, evidently $|O_i| = 1 + \rho(v_i)$, $i = 1, 2, ..., n$. Since the order of bond-atom symmetry is given by $|L_1||L_2| ... |L_n|$ and it is easily shown that $|L_i| = |O_i|$ $(i = 1, 2, ..., n)$, the number of representations is given by $n!/(1 + \rho(v))(1 + \rho(v')) ... (1 + \rho(v''))$.

Although the orbit graph can be obtained by computing a sequence of orbits $O_1, O_2, ..., O_{n-1}$, the computation of bond-atom symmetry requires the determination of a sequence of sets of left coset representatives, i.e., $L_1, L_2, ..., L_{n-1}$, and the following theorem is useful for the computation of bond-atom symmetry.

Theorem II. If X is a subset of G_i $(i \in \{1, 2, ..., n-1\})$ such that $O_i = \{\sigma(i) | \sigma \in X\}$ and for any $\sigma, \tau \in X$ $(\sigma \ne \tau)$, $\sigma(i) \ne \tau(i)$, then the set X is a set of left coset representatives of G_{i+1} in G_i, i.e., L_i.

Proof. For any $\tau \in G_i$ such that $\tau(i) = x$, we can find a permutation $\sigma \in X$ such that $\sigma(i) = x$, and for such a σ, $\sigma^{-1}\tau(j) = j, j = 1, 2, ..., i$. Hence, $\sigma^{-1}\tau \in G_{i+1}$ and $\tau \in \sigma G_{i+1}$. Suppose $\sigma, \sigma' \in X$, and $\sigma \ne \sigma'$ and assume there exists $\tau \in \sigma G_{i+1} \cap \sigma' G_{i+1}$. Then, $\tau = \sigma\lambda = \sigma'\lambda'$ for some $\lambda, \lambda' \in G_{i+1}$. Hence, $\sigma = \sigma'\lambda'\lambda^{-1}$. Since $\lambda'\lambda^{-1}$ is a member of G_{i+1}, we reached $\sigma(i) = \sigma'(i)$, a contradiction. Thus X is a set of left coset representatives, and G_i is given by XG_{i+1}.

By the above theorem, the number of permutations to be generated for perception of bond-atom symmetry is greatly reduced, and we need to construct only, $1 + \sum(|O_i| - 1)$ permutations. For instance, the bond-atom symmetry of the

complete graph with six nodes consists of 720 permutations, but we need to generate only 16 permutations as illustrated below for a node sequence (*a*, *b*, *c*, *d*, *e*, *f*).

1. *abcdef*	9. *-dbcef*
2. *----fe*	10. *-ebcdf*
3. *--edf*	11. *-fbcde*
4. *---fde*	12. *bacdef*
5. *--dcef*	13. *cabdef*
6. *--ecdf*	14. *dabcef*
7. *--fcde*	15. *eabcdf*
8. *-cbdef*	16. *fabcde*

By the equivalence of nodes perceived during the permutation generation, the construction of the first 12 permutations is sufficient for the determination of the orbit graph.

For a node sequence (*w*, *y*, *x*, *z*) of graph A in Figure 1, five permutations (or numberings) are required for the determination of a sequence of sets of left coset representatives, i.e., $L_1, L_2, \ldots, L_{n-1}$, and the first three permutations are sufficient for the orbit graph. The eight permutations of bond-atom symmetry can be given by the product L_1L_2 as below.

1. *wyxz*	L_1 = {(1, 2, 3, 4), (2, 1, 4, 3), (3, 1, 4, 2), (4, 2, 3, 1)}
2. *-xyz*	L_2 = {(1, 2, 3, 4), (1, 3, 2, 4)}
3. *ywzx*	bond-atom symmetry = {(1, 2, 3, 4), (1, 3, 2, 4), (2, 1, 4, 3),
4. *xwzy*	(2, 4, 1, 3), (3, 1, 4, 2), (3, 4, 1, 2), (4, 2, 3, 1), (4, 3,
5. *zyxw*	2, 1)}

"Fitness" for Pseudocanonical Numberings. The use of the fitness permits us to construct the pseudocanonical numberings by a simple rule of choosing for each number a node with the maximum fitness among unchosen nodes, and it enables us to give the general description of procedures for orbit graph, bond-atom symmetry, and canonical numbering. The number of the pseudocanonical numberings depends on the choice of fitness, but it is generally small for the unsymmetrical graph if an invariant number of great discrimination is used and the fitness is defined so that the preceding assignment of numbers to nodes restricts the next assignment.

The pseudocanonical numberings obeying an orbit graph provide a small number of representations without duplication by conditions 1 and 2 of the fitness (section I) and theorem I, and the number of such representations, i.e., the number of representative pseudocanonical numberings, is given by (total number of pseudocanonical numberings)/((1 + $\rho(v)$)(1 + $\rho(v')$) ... (1 + $\rho(v'')$)), where each $\rho(v)$ represents the out-degree of node v in the orbit graph. One numbering among representative pseudocanonical numberings which provides the maximum sequence of the fitness is selected as the canonical numbering.

The pseudocanonical numbering is important not only in construction of the canonical numbering but also in determination of the bond-atom symmetry and of the orbit graph, since the permutations which provide the pseudocanonical numberings of the same sequence of the fitness constitute a coset of the bond-atom symmetry in S_n, and in generation of permutations of the coset the pseudocanonical numbering prevents many unwanted cosets. The fitness given as example 2 in section I is particularly suited for determination of the orbit graph and the bond-atom symmetry since the pseudocanonical numbering with this fitness defines a tree, and determination of $L_1, L_2, \ldots, L_{n-1}$ by theorem II can be carried out most easily for such numberings.

III. PRECISE PRESENTATION OF ALGORITHMS

Imagine n points on a line, denoted $p_1, p_2, \ldots, p_n$ from left to right. Suppose that each point p_i is equipped with a bucket W_i which stores the candidates of nodes to receive the number i. We assign number i to a node in the bucket at each point p_i, moving from point to point.

To start we go to the first point and determine the candidates for nodes to receive number 1 and put them into the bucket W_1. Whenever we go to points p_i ($2 \le i \le n$) from left to right, we determine the candidates for nodes which receive number i and put them into the bucket W_i. If there is no candidate, bucket W_i is left empty, and we move to the left if $i \neq 1$ and terminate the process if $i = 1$. If W_i is not empty, we take out a node from the bucket and assign number i to it.

Principal differences in the algorithms of orbit graph and canonical numbering are the determination of the candidates and the technique used for accelerating the execution.

Algorithm for the Orbit Graph. Step 1. Assign arbitrarily initial numbers $\alpha(v)$ for each node $v \in V$. Set $i = 0$, *flag* = 0, and $z = n + 1$ (i denotes the number to be assigned, *flag* indicates completion of a pseudocanonical numbering, and z is a variable used to accelerate execution of the algorithm.)

Step 2. If $i = n$, go to step 10. Determine the set X of nodes having the maximum fitness for the number $i + 1$; i.e., set FMAX = l-max$\{f(i + 1, v) | v \in V - \{v_1, v_2, \ldots, v_i\}\}$ and $X = \{v | v \in V - \{v_1, v_2, \ldots, v_i\}, f(i + 1, v) = \text{FMAX}\}$. Set *flagx* = 1. If *flag* = 1, compare FMAX with $f(i + 1, w_{i+1})$ of the first pseudocanonical node sequence $(w_1, w_2, \ldots, w_n)$, and if FMAX $\neq f(i + 1, w_{i+1})$, set *flagx* = 0. [The node sequence $(v_1, v_2, \ldots, v_i)$ is a current partial node sequence, and the set X represents the candidates for the number $i + 1$ if *flagx* = 1.]

Step 3. If *flagx* = 0, go to step 5. (According to flag$x$ we decide whether we store X into bucket W_{i+1} or if we examine another possibility for the number i.)

Step 4. Set $i = i + 1$ and $W_i = X$.

Step 5. If $i = 0$, go to step 11. If i is greater than z or W_i is empty, go to step 6. (We must find out the nearest point p_i such that $i \le z$.) Take out a node y with the smallest initial number and set $W_i = W_i - \{y\}$. Set $z = n + 1$. (We abolish the indication of acceleration.) Go to step 7.

Step 6. Set $i = i - 1$, and go to step 5.

Step 7. Set *flag7* = 0. If *flag* = 0 or $i \neq 1$, go to step 8. If a member of O_1 is equivalent to the node y, i.e., if $x \in O_1$ and $x, y \in E_j$ for some j, set *flag7* = 1 and go to step 8. If the node y is equivalent to a node x having a smaller initial number than y, i.e., if $x, y \in E_j$ for some j, $\alpha(x) < \alpha(y)$, and $x \notin O_1$, go to step 5 (*flag7* indicates whether the node y must be added as a member of O_1 or not).

Step 8. Assign the number i to the node y, i.e., set $v_i = y$. If *flag7* = 0, go to step 2.

Step 9. Add v_1 to O_1; i.e., set $O_1 = O_1 \cup \{v_1\}$. Draw an arrow from the node w_1 toward v_1. Go to step 5.

Step 10. If *flag* = 0, set $E_j = \{v_j\}$ and $w_j = v_j$ for $j = 1, 2, \ldots, n$, set $O_1 = \{v_1\}$, set *flag* = 1, and go to step 5. If *flag* = 1, add v_1 to O_1, determine the minimum m such that $w_m \neq v_m$, draw an arrow from w_m toward v_m, set $z = m$, remember the equivalence of the nodes v_j and w_j for $j = 1, 2, \ldots, n$, i.e., for each $j \in \{1, 2, \ldots, n\}$, find classes E_t and E_s such that $v_j \in E_t$ and $w_j \in E_s$, and set $E_t = E_t \cup E_s$ and $E_s = 0$ if $t \neq s$. Go to step 5.

Step 11. End of the algorithm. The directed graph obtained is the orbit graph of a chemical graph.

Algorithm for Bond-Atom Symmetry. The algorithm described for the orbit graph can be adapted to computation of bond-atom symmetry, and the modifications required are as follows.

Step 7. If *flag* = 0 or $i \neq 1$, go to step 8. If the node y is equivalent to a node x having smaller initial number than y and x is not a member of O_1, go to step 5.

Step 8. Assign the number i to the node y; i.e., set $v_i = y$. Go to step 2.

Step 9. Deleted.

Step 10. If *flag* = 0, set $E_j = \{v_j\}$ and $w_j = v_j$ for $j = 1, 2, ..., n$, set $k = 0$, set $O_1 = \{v_1\}$, set *flag* = 1, and go to step 5. If *flag* = 1, add v_1 to O_1, determine the minimum m such that $w_m \neq v_m$, set $z = m$, set $k = k + 1$, set $M_k = m$, set $V_k = (v_1, v_2, ..., v_n)$, for each j find classes E_t and E_s such that $v_j \in E_t$ and $w_j \in E_s$, set $E_t = E_t \cup E_s$ and $E_s = 0$ if $t \neq s$, and go to step 5. (k is a counter for the pseudocanonical numberings, and M_k stores the value of m for the numbering V_k.)

Step 11. End of the algorithm. The node sequence $w_1, w_2, ..., w_n$ and the sequences $V_1, V_2, ..., V_k$ are the pseudocanonical numberings which determine a sequence of the coset representatives $L_1, L_2, ..., L_{n-1}$ in theorem II.

If we change the nodes in the generated node sequences to the numbers according to the first node sequence $(w_1, w_2, ..., w_n)$, i.e., each node w_i is changed to the number i, the sequences $V_1, V_2, ..., V_k$ which provide permutation vectors $V_1', V_2', ..., V_k'$ and $L_1, L_2, ..., L_{n-1}$ of the bond-atom symmetry based on the numbering $w_1, w_2, ..., w_n$ can be given in the form of a permutation vector as follows: $L_i = \{V_j' | M_j = i, j = 1, 2, ..., k\} \cup \{\text{identity}\}$ $(i = 1, 2, ..., n - 1)$.

Algorithm for Canonical Numbering. Step 1. Determine the orbit graph. Set $i = 0$, $i^* = n + 1$, and *flag* = 0.

Step 2. If $i = n$, go to step 8. Determine the set X of nodes having the maximum fitness for the number $i + 1$; i.e., set FMAX = l-max $\{f(i + 1, v) | v \in V - \{v_1, v_2, ..., v_i\}\}$ and $X = \{v | v \in V - \{v_1, v_2, ..., v_i\}, f(i + 1, v) = \text{FMAX}\}$. Eliminate from X the nodes having entering arrows from members of the unnumbered nodes, i.e., $V - \{v_1, v_2, ..., v_i\}$. Set *flagx* = 1. If X is empty, set *flagx* = 0 and go to step 3. If *flag* = 0 or i^* is less than $i + 1$, go to step 3. If *flag* = 1 and $i^* \geq i + 1$, set $i^* = n + 1$ and compare FMAX with $f(i + 1, w_{i+1})$ of the best node sequence $(w_1, w_2, ..., w_n)$. If FMAX $< f(i + 1, w_{i+1})$, set *flagx* = 0, and set $i^* = i + 1$ if FMAX $> f(i + 1, w_{i+1})$. [When FMAX is lexicographically greater than $f(i + 1, w_{i+1})$, we must not compare the maximum fitness with that of the best node sequence in determination of the candidates at the points to the right of p_{i+1}, and i^* remembers this point.]

Step 3. If *flagx* = 0, go to step 5.

Step 4. Set $i = i + 1$ and $W_i = X$.

Step 5. If $i = 0$, go to step 9. If W_i is empty, go to step 6. Take out a node y from bucket W_i and set $S_i = W_i - \{y\}$. Go to step 7.

Step 6. Set $i = i - 1$ and go to step 5.

Step 7. Assign the number i to the node y; i.e., set $v_i = y$. Go to step 2.

Step 8. Store the node sequence $(v_1, v_2, ..., v_n)$ as the best node sequence $(w_1, w_2, ..., w_n)$; i.e., set $w_j = v_j$ for $j = 1, 2, ..., n$. Set *flag* = 1 and $i^* = n + 1$.

Step 9. End of the algorithm. The node sequence $w_1, w_2, ..., w_n$ is the canonical node sequence.

IV. GENERAL APPLICABILITY OF THE METHOD

The method described in this paper is based on the particular choice of the invariant number and of the "fitness", but this does not deny the general applicability of the method. By changing the invariant number and the fitness used for unique coding, we provide an efficient procedure to obtain many other canonical codes. The usefulness of the orbit graph is essentially independent of the definition of the canonical code, although a particular choice of fitness is desirable for an efficient coding method. In any coding method, the information carried by the orbit graph is required to avoid the duplicated numberings due to symmetry, and the orbit-graph can suppress completely such numberings.

If the list $f(i, v)$ assigned to each unnumbered node v does not satisfy condition 2 of the fitness, i.e., if $f(i, v)$ is the "partial fitness", then several representative pseudocanonical numberings of the same sequence of the lists $f(1, v_1), f(2, v_2), ..., f(n, v_n)$ may be generated. Even in this case, the method of the orbit graph can be applicable if the name associated with the numbering is used for the final selection of the canonical numbering. For instance, if we employ the orbit graph, Morgan's name of the dodecahedron can be easily determined.

ACKNOWLEDGMENT

I express my gratitude to Professor M. Iri (Department of Computer Sciences, University of Tokyo) for his understanding, valuable criticism about my previous notation, and encouragement, to Dr. C. A. Shelley (Eastman Kodak) for his encouragement, and to Mr. A. Zamora (Chemical Abstracts Service) and Professor W. T. Wipke (University of California) for valuable suggestions in revising the paper.

REFERENCES AND NOTES

(1) For a bibliography on unique coding and graph isomorphism see: Rush, J. E. "Status of Notation and Topological Systems and Potential Future Trends". *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 202. Read, R. C.; Corneil, D. G. "The graph Isomorphism Disease". *J. Graph Theory* **1977**, *1*, 339. Gati, G. "Further Annotated Bibliography on the Isomorphism Disease". *Ibid.* **1979**, *3*, 95.
(2) Enomoto, H.; Katayama, T.; Yonezaki, N. "On the Retrieval of Isomorphic Graphs by Partitioning Algorithm". *Johoshori* **1977**, *18*, 1209 (in Japanese).
(3) Lynch, M. F.; Harrison, J. M.; Town, W. G.; Ash, J. E., Eds. "Computer Handling of Chemical Structure Information"; Macdonald: London, 1971. Wipke, W. T.; Heller, S. R.; Feldmann, R. J.; Hyde, E., Eds.; "Computer Representation and Manipulation of Chemical Information"; Wiley: New York, 1974. Ash, J. E.; Hyde, E., Eds. "Chemical Information Systems"; Ellis Horwood Ltd.: Chichester, England, 1975.
(4) Morgan, H. L. *J. Chem. Doc.* **1965**, *5*, 107.
(5) Dubois, J. E.; Viellard, H. *Bull. Soc. Chim. Fr.* **1968**, 900, 905, 913; **1971**, 839.
(6) Yoneda, Y. "CHEMOGRAM"; Maruzen: Tokyo, 1972.
(7) Kudo, Y.; Sasaki, S. *J. Chem. Doc.* **1974**, *14*, 200.
(8) Kudo, Y.; Sasaki, S. *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 43.
(9) Wipke, W. T.; Dyott, T. M. *J. Am. Chem. Soc.* **1974**, *96*, 4834.
(10) Shelley, C. A.; Munk, M. E. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 110; **1979**, *19*, 247.
(11) Randic, M. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 105; **1977**, *17*, 171. Randic, M.; Brissey, G. M.; Wilkins, C. L. *Ibid.* **1981**, *21*, 52.
(12) Jochum, C.; Gasteiger, J. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 113; **1979**, *19*, 49.
(13) Schubert, W.; Ugi, I. *J. Am. Chem. Soc.* **1978**, *100*, 37.
(14) Walentowski, R. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 181.
(15) Uchino, M. *Kagaku Sosetsu* **1978**, *18*, 47.
(16) Uchino, M. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 116, 121, 124.
(17) Uchino, M. "Algorithms for Unique and Unambiguous Coding and Symmetry Perception of Molecular Structure Diagrams. 4. A Simple Algorithm for Determining Morgan-Type Name of Chemical Graphs", in press.
(18) Uchino, M. In "Processing of Number, Formula, and Sentence"; Iri, M., Ed.; Iwanami: Tokyo, 1981.
(19) The set G of symmetry operations is defined as follows. Let V be the set of nodes, $m(v,v')$ be the bond value between the nodes v and v', and $a(v')$ be the atomic values of the nodes v and v'. Then, $G = \{f | \text{bijection } f: V \rightarrow V, m(v, v') = m(f(v), f(v')), a(v) = a(f(v)), a(v') = a(f(v')), \text{for all } v, v' \in V\}$.
(20) See also the lemma due to Sims in: Brown, H.; Masinter, L. M.; Hjelmeland, L. "Constructive Graph Labeling Using Double Cosets"; Report STAN-CS-72-318; Computer Science Department, Stanford University: Stanford, CA, Oct 1972.