Algorithms for Unique and Unambiguous Coding and Symmetry Perception of Molecular Structure Diagram. II. Basic Algorithm for Unique Coding and Computation of Symmetry Group†

MASAHIRO UCHINO

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

Received June 4, 1979; Revised Manuscript Received January 21, 1980

A procedure for unique and unambiguous coding of a chemical graph is described on a group theoretical basis. During the procedure, a chemical graph is transformed to a smaller graph by the use of logical atoms, and the invariant vector under the bond symmetry group of the derived graph is computed. The invariant vector is used to define the normalized representation from which a unique code is generated together with the associated symmetry group. A new efficient method was devised for the generation of the permutation vectors and is implemented by an extremely compact APL program.

INTRODUCTION

This paper is concerned with the results obtained by the application of the elementary group theory[1] to the problem of the unique and unambiguous coding of the chemical graph. Except for few papers,[2-4] the procedures to represent molecular structures uniquely and unambiguously have been proposed without stating the mathematical proofs. Rush[5] provides numerous references to work in this and related areas.

In this paper, we present a mathematical method which permits a clearer understanding of unique coding and related problems. In section I, a list of logical atoms is given to provide a method for a chemical graph representation of molecular structures. Next we describe the normalized representation of a chemical graph on the basis of an invariant vector under a bond-atom symmetry group (section II). Fundamental theorems of unique coding and symmetry perception are presented in that section, and in section III a simple but efficient method to determine unique code and symmetry group is described. In the final section, APL implementation of the present approach is presented.

I. LOGICAL ATOMS FOR CHEMICAL GRAPH REPRESENTATION OF MOLECULAR STRUCTURES

To describe a molecular structure by a smaller chemical graph, usually hydrogen atoms in an organic molecule are neglected. We similarly simplify in a systematic way: (1) fuse hydrogen to adjacent carbon, oxygen, and nitrogen; (2) fuse methyl groups to adjacent carbon, oxygen, and nitrogen; (3) fuse unsaturated oxygen to adjacent carbon, oxygen, and nitrogen. In this way, we can simplify a graph consisting of physical atoms to that of logical atoms (see Figure 1). A set of logical atoms derived in this manner is given in Table I together with molecules of a single logical atom. A physical atom which is not a part of the logical atoms listed is treated as itself. Since only one physical atom in a logical atom possesses free valences, no difficulty arises in graph construction, and the bond value between two logical atoms is defined by that of two physical atoms in them. When we discriminate between double bonds in cis and trans configurations and in aromatic rings, special bond values may be employed.

II. FUNDAMENTAL THEOREMS OF UNIQUE CODING AND SYMMETRY PERCEPTION

A chemical graph consisting of logical atoms and physical atoms can be represented by a bond matrix $\mathbf{B}'$ and an atomic vector $\boldsymbol{a}'$ on the basis of an arbitrary numbering of nodes, and, by one of the methods described in the preceding paper,[6] an invariant vector under bond symmetry $G^{(\mathbf{B}')}$ can be computed. For such vector $\boldsymbol{h}'$, we define an invariant vector $\boldsymbol{s}'$ under intersection group $G^{(\mathbf{B}')} \cap G^{(\boldsymbol{a}')}$ by the following:

$$\boldsymbol{s}' = b\boldsymbol{h}' + \boldsymbol{a}' \qquad (1)$$

where the components of the vector $\boldsymbol{a}'$ are atomic numbers (physical or logical) and b = 1 + the maximum component of $\boldsymbol{a}'$.

If the components of the vector $\boldsymbol{s}'$ thus defined are in descending order, the atomic vector $\boldsymbol{a}'$ and the bond matrix $\mathbf{B}'$ will be said to be normalized. If we consider the group F of the n-element permutation vectors for the graph of n nodes, we can always find a vector $\boldsymbol{f}$ in F such that the components of the permuted vector $\boldsymbol{s} = \boldsymbol{s}'_{\circ}\boldsymbol{f}$ are in descending order, and we have a normalized representation by:

$$\boldsymbol{a} = \boldsymbol{a}'_{\circ}\boldsymbol{f}$$

$$\mathbf{B} = \mathbf{B}'_{\circ}0(\boldsymbol{f};\boldsymbol{f})$$

Table I. Logical Atoms of Hydrogen, Carbon, Oxygen, and Nitrogen

(a) H, CH_3, and =O which are bonded to atoms other than C, H, O, and N

1, -H; 110, $-CH_3$; 111, =O

(b) Logical carbons which can be bonded to physical or logical atoms other than -H, $-CH_3$, and =O

112, $-\overset{|}{C}-$; 113, $-\overset{|}{C}=$; 114, $-C\equiv$; 115, $=C=$; 116, $-\overset{|}{C}H-$; 117, $-CH=$; 118, $CH\equiv$; 119, $-CH_2-$; 120, $=CH_2$; 121, $-\overset{|}{C}CH_3$; 122, $\overset{|}{C}CH_3$; 123, $-\overset{|}{C}HCH_3$; 124, $=CHCH_3$; 125, $-CH_2CH_3$; 126, $-\overset{|}{C}(CH_3)_2$; 127, $=C(CH_3)_2$; 128, $-CH(CH_3)_2$; 129, $-C(CH_3)_3$; 130, $-\overset{|}{C}=O$; 131, -CHO; 132, $-COCH_3$; 133, =CO

(c) Logical oxygens which can be bonded to physical or logical atoms other than -H, $-CH_3$, and =O

134, -O-; 135, -OH; 136, $-OCH_3$

(d) Logical nitrogens which can be bonded to physical or logical atoms other than -H, $-CH_3$, and =O

137, $-\overset{|}{N}-$; 138, $-N=$; 139, $N\equiv$; 140, -NH-; 141, =NH; 142, $-NH_2$; 143, $-\overset{|}{N}CH_3$; 144, $=NCH_3$; 145, $-NHCH_3$; 146, $-N(CH_3)_2$; 147, -NO; 148, $-NO_2$

(e) Single atom molecules

148, H_2; 149, CH_4; 150, H_2O; 151, C_2H_6; 152, $C(CH_3)_4$; 153, $CH(CH_3)_3$; 154, $CH_2(CH_3)_2$; 155, O_2; 156, ozone; 157, CO_2; 158, CH_3OCH_3; 159, CH_3OH; 160, HCHO; 161, CH_3CHO; 162, CH_3COCH_3; 163, NH_3; 164, NH_2CH_3; 165, $NH(CH_3)_2$; 166, $N(CH_3)_3$; 167, ONH; 168, $ONCH_3$

† Presented at the Division of Chemical Information, ACS/CSJ Congress, Honolulu, Hawaii, April 3, 1979.

0095-2338/80/1620-0121$01.00/0

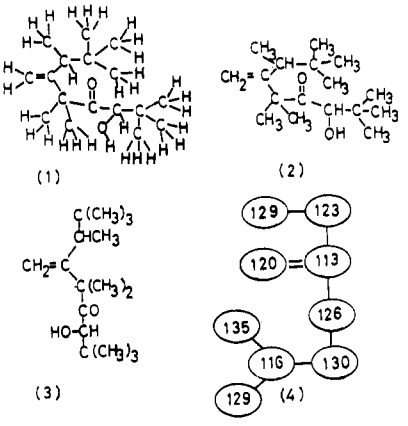

Figure 1. Graph of physical atoms and that of logical atoms: (1) graph of physical atoms; (2) fusion of hydrogen; (3) fusion of methyl group and of unsaturated oxygen; (4) final graph used for representation of chemical structure.

where the circle notation implies that the ith component of $\boldsymbol{a}$ is the f_ith component of $\boldsymbol{a}'$ and the (i,j)th component of $\mathbf{B}$ is the (f_i,f_j)th component of $\mathbf{B}'$.

Now suppose other initial numbering of the same graph and the associated atomic vector $\boldsymbol{a}''$, bond matrix $\mathbf{B}''$, and permutation vector $\boldsymbol{f}''$ of normalization. Then,

$$\boldsymbol{s} = \boldsymbol{s}'\circ\boldsymbol{f}' = \boldsymbol{s}''\circ\boldsymbol{f}''$$

and

$$\boldsymbol{a} = \boldsymbol{a}'\circ\boldsymbol{f}' = \boldsymbol{a}''\circ\boldsymbol{f}''$$

since $\boldsymbol{a}'\circ\boldsymbol{f}' \neq \boldsymbol{a}''\circ\boldsymbol{f}''$ implies that $\boldsymbol{s}'\circ\boldsymbol{f}' \neq \boldsymbol{s}''\circ\boldsymbol{f}''$. Thus, the vectors $\boldsymbol{s}$ and $\boldsymbol{a}$ are independent of the initial numbering of the nodes. Consequently, the atomic symmetry[6] $G^{(\mathbf{a})}$ of the normalized vector $\boldsymbol{a}$ and the following group $G^{(\mathbf{s})}$ are also independent of the initial numbering.

$$G^{(\mathbf{s})} = \{\boldsymbol{f} \in F;\ \boldsymbol{s}_0\boldsymbol{f} = \boldsymbol{s}\} \tag{2}$$

Unfortunately the normalized bond matrix depends on the initial numbering and $\mathbf{B}'\circ(\boldsymbol{f}';\boldsymbol{f}')$ is not necessarily equal to $\mathbf{B}''\circ(\boldsymbol{f}'';\boldsymbol{f}'')$.

For a normalized bond matrix $\mathbf{B}$, we define a set M of matrices by the following:

$$M = \{\mathbf{B}\circ(g;g);\ g \in G^{(\mathbf{s})}\} \tag{3}$$

and we state theorem I.

Theorem I. A bond matrix which is isomorphic to $\mathbf{B}$ is normalized if and only if it is a member of set M.

Proof. This is obvious since $G^{(\mathbf{s})}$ consists of all $\boldsymbol{f} \in F$ such that $\boldsymbol{s}\circ\boldsymbol{f} = \boldsymbol{s}$.

By theorem I, the set M is a numbering independent set of all normalized bond matrices, and if we select one matrix $\mathbf{B}^*$ from M by a certain criterion, $\mathbf{B}^*$ and $\boldsymbol{a}^*$ $(=\boldsymbol{a})$ provide a unique and unambiguous representation of a chemical graph. To select t($\boldsymbol{g}$,$\mathbf{B}$) bond matrix, we define the following vector t($\boldsymbol{g}$,$\mathbf{B}$) for $\mathbf{B}$ and $\boldsymbol{g} \in G^{(\mathbf{s})}$:

$$\mathrm{t}(\boldsymbol{g},\mathbf{B}) = (\mathbf{B}_{g_2g_1}, \mathbf{B}_{g_3g_1}, \mathbf{B}_{g_3g_2}, \ldots, \mathbf{B}_{g_ig_1}, \mathbf{B}_{g_ig_2}, \ldots, \mathbf{B}_{g_ig_{i-1}}, \ldots, \mathbf{B}_{g_ng_1}, \mathbf{B}_{g_ng_2}, \ldots, \mathbf{B}_{g_ng_{n-1}}) \tag{4}$$

The vector t(g,$\mathbf{B}$) corresponds to the lower triangular part of matrix $\mathbf{B}\circ(\boldsymbol{g};\boldsymbol{g})$, and we refer to the lexicographically maximum t($\boldsymbol{g}$,$\mathbf{B}$), i.e., $\ell\text{-max}_{\boldsymbol{g}\in G^{(\mathbf{s})}}\mathrm{t}(\boldsymbol{g},\mathbf{B})$ as the maximum lower triangular code of $\mathbf{B}$, and this can be used together with atomic vector $\boldsymbol{a}$ as a unique and unambiguous code of the chemical graph. If t($\boldsymbol{g}^*$,$\mathbf{B}$) for $\boldsymbol{g}^*$ in $G^{(\mathbf{s})}$ is such a maximum vector, this determines a subset S of the group $G^{(\mathbf{s})}$:

$$S = \{\boldsymbol{g} \in G^{(\mathbf{s})};\ \mathrm{t}(\boldsymbol{g},\mathbf{B}) = \mathrm{t}(\boldsymbol{g}^*,\mathbf{B})\} \tag{5}$$

The set S is not necessarily the group and we state theorem II to relate the set S to symmetry group $G^{(\mathbf{B})} \cap G^{(\mathbf{a})}$.

Theorem II. For any $\boldsymbol{g}$ and $\boldsymbol{g}'$ of $G^{(\mathbf{s})}$, $\mathbf{B}\circ(\boldsymbol{g};\boldsymbol{g}) = \mathbf{B}\circ(\boldsymbol{g}';\boldsymbol{g}')$ if and only if $\boldsymbol{g}' \in (G^{(\mathbf{B})} \cap G^{(\mathbf{a})})\boldsymbol{g}$.

Proof. To simplify the notation, we write $\mathbf{B}\circ\boldsymbol{g}$ for $\mathbf{B}\circ(\boldsymbol{g};\boldsymbol{g})$ and G for $G^{(\mathbf{B})} \cap G^{(\mathbf{a})}$, and suppress the circle notation for group elements. If $\boldsymbol{g}' \in G\boldsymbol{g}$, then there exists $\boldsymbol{f} \in G$ such that $\boldsymbol{g}' = \boldsymbol{f}\boldsymbol{g}$, and $\mathbf{B}\circ\boldsymbol{g}' = \mathbf{B}\circ\boldsymbol{f}\boldsymbol{g}$. This implies $\mathbf{B}\circ\boldsymbol{g}' = \mathbf{B}\circ\boldsymbol{g}$ since $\boldsymbol{f} \in G$ and $\mathbf{B}\circ\boldsymbol{f} = \mathbf{B}$. Assume $\mathbf{B}\circ\boldsymbol{g}' = \mathbf{B}\circ\boldsymbol{g}$; then (i) $\mathbf{B}\circ\boldsymbol{g}'\boldsymbol{g}^{-1} = \mathbf{B}$. Since $G^{(\mathbf{s})}$ is a subgroup of $G^{(\mathbf{a})}$ and G is a subgroup of $G^{(\mathbf{s})}$, $G = G^{(\mathbf{s})} \cap G^{(\mathbf{B})}$. Consequently, (i) implies $\boldsymbol{g}'\boldsymbol{g}^{-1} \in G$. Hence, $\boldsymbol{g}' \in G\boldsymbol{g}$.

By theorem II, there is one-to-one correspondence between the right coset of $G^{(\mathbf{B})} \cap G^{(\mathbf{a})}$ in $G^{(\mathbf{s})}$ and the normalized bond matrix. And hence, we can write the set S of (5) as:

$$S = (G^{(\mathbf{B})} \cap G^{(\mathbf{a})})\boldsymbol{g}^* \tag{6}$$

where $\boldsymbol{g}^* \in S$.

The bond-atom symmetry $G^{(\mathbf{B}^*)} \cap G^{(\mathbf{a}^*)}$, associated with the unique matrix $\mathbf{B}^* = \mathbf{B}\circ(\boldsymbol{g}^*;\boldsymbol{g}^*)$ and $\boldsymbol{a}^* = \boldsymbol{a}\circ\boldsymbol{g}^*$ $(=\boldsymbol{a})$ can be computed by theorem III.

Theorem III. $G^{(\mathbf{B}^*)} \cap G^{(\mathbf{a}^*)} = \boldsymbol{g}^{*-1}(G^{(\mathbf{B})} \cap G^{(\mathbf{a})})\boldsymbol{g}^*$.

Proof. We employ simplified notation as in the proof of theorem II. (i) Transformation $\boldsymbol{g}^{*-1}\boldsymbol{f}\boldsymbol{g}^*$ is one to one; i.e., for any $\boldsymbol{f}$ and $\boldsymbol{f}' \in F$, $\boldsymbol{g}^{*-1}\boldsymbol{f}\boldsymbol{g}^* = \boldsymbol{g}^{*-1}\boldsymbol{f}'\boldsymbol{g}^*$ if and only if $\boldsymbol{f} = \boldsymbol{f}'$. (ii) For any $\boldsymbol{f} \in G$, $\mathbf{B}\circ\boldsymbol{f}^{-1} = \mathbf{B}$, and hence $\mathbf{B} = \mathbf{B}^*\circ\boldsymbol{g}^{*-1}\boldsymbol{f}$; consequently $\mathbf{B}^* = \mathbf{B}^*\circ\boldsymbol{g}^{*-1}\boldsymbol{f}\boldsymbol{g}^*$. Hence, $\boldsymbol{g}^{*-1}\boldsymbol{f}\boldsymbol{g}^* \in G^{(\mathbf{B}^*)} \cap G^{(\mathbf{a}^*)}$. By the same way, for any $f \in G^{(\mathbf{B}^*)} \cap G^{(\mathbf{a}^*)}$ $\boldsymbol{g}^*\boldsymbol{f}\boldsymbol{g}^{*-1} \in G$.

By theorem III and (6), the bond-atom symmetry $G^{(\mathbf{B}^*)} \cap G^{(\mathbf{a}^*)}$ of the unique numbering is given as:

$$G^{(\mathbf{B}^*)} \cap G^{(\mathbf{a}^*)} = \boldsymbol{g}^{*-1}S \tag{7}$$

If $\mathbf{B}'$ is the bond matrix of initial numbering and $\boldsymbol{f}'$ is the permutation vector used for normalization, then $\mathbf{B}^* = \mathbf{B}'\circ(\boldsymbol{f}'\boldsymbol{g}^*;\boldsymbol{f}'\boldsymbol{g}^*)$. Consequently $\mathbf{B}^*$ is obtained by assigning the ith component of $\boldsymbol{g}^{*-1}\boldsymbol{f}'^{-1}$ to the ith node of the initial numbering.

III. PROCEDURE TO DETERMINE THE SUBSET S OF $G^{(\mathbf{s})}$

To determine the subset S of (5), t($\boldsymbol{g}^*$,$\mathbf{B}$), and bond-atom symmetry, a generator of permutation vectors is required. The permutation generators described by others[7,8] use a backtrack procedure and generate the permutation vectors one by one. The efficiency of the generator used for the determination of the set S depends on the initial numbering of nodes. A new method for permutation generation is devised and used for the determination of S.

Before presenting this material, some definitions are given.

The components of the vector s defining $G^{(\mathbf{s})}$ are separated into several intervals and in each interval the components are same. Let q be the number of such intervals and define q-element vector $\boldsymbol{c}$ whose ith component is the number of elements in the ith interval. We define vectors $\boldsymbol{d}$, $\boldsymbol{u}^1$, $\boldsymbol{u}^2$, . . ., $\boldsymbol{u}^q$ by the following:[6]

$$\boldsymbol{d}_1 = 0,\ \boldsymbol{d}_j = \sum_{i=1}^{j-1}\boldsymbol{c}_i,\ j = 2, 3, \ldots, q$$

$$\boldsymbol{u}_i^j = \boldsymbol{d}_j + i,\ i = 1, 2, \ldots, \boldsymbol{c}_j,\ j = 1, 2, \ldots, q$$

Now we present a basic algorithm to determine the subset S of $G^{(\mathbf{s})}$.

Basic Algorithm.[10] *Step 1.* By any component x of $\boldsymbol{u}^1$, make partially constructed vector $\boldsymbol{x}$ of length 1 for $G^{(\mathbf{s})}$ and initialize a set S' by all such vectors: $m \leftarrow 0$, $b \leftarrow 1$ + maximum

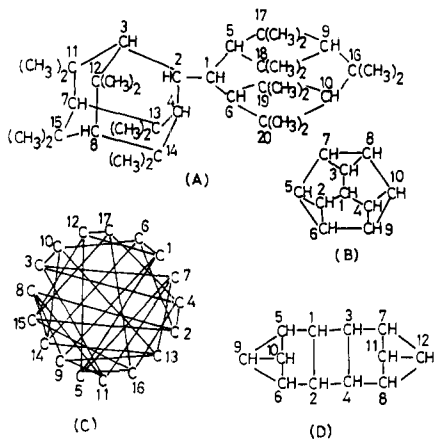

Figure 2. Examples of unique numbering. For partitioning of graph nodes, the invariant vector produced by the function PARTITION1[6] was used.

Table II. Maximum Lower Triangular Code and Symmetry Group

unique code:	1 1 0 1 0 0 0 1 0 0 0 1 0 0 1 0 0 1 0 1 0 0 0 0 1
	0 0 0 1 0 0 0 1 0 1 0 0 0 0 0 1 0 0 0 1 1
bond-atom symmetry:	1 2 3 4 5 6 7 8 9 10
	1 2 4 3 6 5 9 10 7 8
	1 3 2 4 7 8 5 6 10 9
	1 3 4 2 8 7 10 9 5 6
	1 4 2 3 9 10 6 5 8 7
	1 4 3 2 10 9 8 7 6 5

Table III. The Number of Partial Vectors Processed

struct	no. of vectors processed at each stage
A	1→1→2→2→2→2→4→4→4→4→8→8→8→8→16→16→16→ 16→64→16→48→16→32→16→16→16→16→16→16→ 16→64→16→48→16→32→32→32→32
B	3→3→6→6→6→6→36→12→60→12→48→6→18→6→12→ 6→6→6[a]
C	1
D	12→8→16→8→8→8→32→8→24→8→16→8→8→8→32→8→ 24→8→16→16→16→16

[a] The maximum number observed for this case is exceptionally large, i.e., 10 times the order of symmetry group. This number can be reduced to 12 by a more advanced algorithm.[10]

component of **B**. $n \leftarrow$ the number of nodes.

Step 2. $m \leftarrow m + 1$. If $m > n$, stop. Otherwise, determine k such that $d_k < m \leq d_{k+1}$. For each g in S', determine the elements x of u^k which are not the components of g, and "grow" the vector g to $g_1, g_2, \ldots, g_{m-1}, x$. Replace the members of S' by the grown vectors.

Step 3. For each g in S', compute $f(g) = \sum_{i=1}^{m-1} b^{m-(i+1)} \mathbf{B}_{g_m g_i}$ and remove from S' the vector g such that $f(g) \neq \max_{g' \in S'} f(g')$, and go to step 2.

When the algorithm stops, the set S' is equivalent to the set S of (5).

IV. EXAMPLES

Examples of the unique numbering determined by the present algorithm are shown in Figure 2. Method I of the preceding paper[6] was used for the computation of the invariant vector under bond symmetry. Structure A of Figure 2 consists of 127 physical atoms, and by the use of the logical atoms of Table I, this structure is represented by a chemical graph of 20 nodes. The normalized bond matrix obtained by the use of method I[6] and (1), was processed by the basic algorithm of section III. An example of the unique code and of the associated bond-atom symmetry $G^{(\mathbf{B}^*)} \cap G^{(a^*)}$ is given in Table II for structure B of Figure 2.

The hypothetical structure C is related to the graph described by Mackay,[9] but in the present approach this can be coded even without the generation of permutation vectors since the bond-atom symmetry of this structure consists of identity elements and method I[6] provides the automorphism partitioning for this structure.

The change of the number of partial vectors processed during the procedure is given in Table III for the structures of Figure 2. The maximum number of the vectors processed was usually not more than twice the order of symmetry group.

```
      ∇NORMALISATION[⎕]∇
    ∇ NORMALISATION HL;MAX;SET;LT;SHIFT;M;B1;HN;T
[1]   ΔA←A
[2]   MAX←⌈/A←AL⍋A]⍳A
[3]   MAX←1+MAX⌈⌈/HL
[4]   ΔOD←⍋HN←MAX⊥(2,ΔN)⍴HL,A
[5]   B1←HN[ΔOD]
[6]   ΔE←⍴SET←SETSELECT B1
[7]   B1←B1⍳SET
[8]   M←⌈/ΔC←,((1↓B1),ΔN+1)-B1
[9]   ΔD←0,(-1)↓T←+\ΔC
[10]  LT←(~LT),[2] LT←(ΔE,M)⍴0
[11]  SHIFT←M-ΔC
[12]  LT←(ΔE,M)↑LT←SHIFT⊖[2] LT
[13]  ΔK←(,LT)/,⍳(M,ΔE)⍴⍳ΔE
[14]  B←B[ΔOD;ΔOD]
[15]  ΔA←ΔA[ΔOD]
    ∇
```

Figure 3. APL function for normalization.

```
      ∇GROW[⎕]∇
    ∇ GROW;S;S1;S2;K;REP;ID
[1]   K←⍳/S←⍴ΔR
[2]   S1←S[1]
[3]   S2←S[2]
[4]   REP←,⍳(S2,S1)⍴⍳S1
[5]   ΔS←ΔS[REP;]
[6]   ΔS←ΔS,[2](K,1)⍴ΔR
[7]   ID←(S2×S2)⍴1
[8]   ID[(⍳S2)+0,S2×⍳S2-1]←S2⍴0
[9]   ID←(K×S2)⍴ID
[10]  ΔR←ΔR[REP;]
[11]  ΔR←(K,S2-1)⍴ID/,ΔR
    ∇
```

Figure 4. APL function for "growth" of partial vectors.

Table IV. Arrays Used in the Programs

program	array	explanation
LTRMAXCODE	A	atomic vector a' of arbitrary numbering
	B	bond matrix $\mathbf{B}'$ of arbitrary numbering
	HN	invariant vector under $G^{(\mathbf{B}')}$, computed by PARTITION1.
	ΔK	vector whose component is the integer k in step 2 of the basic algorithm.
	ΔS	set S' in basic algorithm; at the final step, this represents bond-atom symmetry
	CODE	$t(g^*, \mathbf{B})$ at the final step
	ΔG	g^*
NORMALISATION	ΔE	number of intervals, i.e., q in section III
	ΔOD	vector f' of normalization
	ΔC	vector c in section III
	ΔD	vector d in section III
	ΔA	normalized atomic vector, i.e., a $(=a^*)$
GROW	ΔR	array of the elements of u^k which are not the components of the partially constructed vector

V. APPENDIX. PROGRAMS

The theory and algorithms described in this paper have been implemented on an IBM 5110 and a MELCOM COSMO 700II UTS/VS (Mitsubishi Electric Corp.) in APL. For the computation of the invariant vector under bond symmetry, the function PARTITION1 of the preceding paper[6] was used. By the use of this invariant vector, the program Normalization (see Figure 3) computes the vector f of the normalization, the vectors c and d in section III, and a vector whose mth component represents the interval number associated with the mth component of the permutation vectors, i.e., the integer k in

```
      ∇LTRMAXCODE[□]∇
     ∇ CODE←A LTRMAXCODE B;M;K;M1;GOSU;X;HN
[1]    ΔS← 1 0 ρCODE←⍳M←K←0
[2]    AN←1↑ρB
[3]    AD←⌈/(,C←B×⍳0)/,B
[4]    HN←PARTITION1 B
[5]    NORMALISATION HN
[6]    B←,B
[7]  L1:→(ΔN<M←M+1)/EXIT
[8]    M1←ΔK[M]
[9]    →(K=M1)/NORM
[10]   GOSU←1↑ρΔS
[11]   ΔR←(GOSU,ΔCCM1])ρΔDEM1]←⍳ΔCCM1]
[12] NORM:GROW
[13]   K←M1
[14]   →(M=1)/L1
[15]   GOSU←1↑ρΔS
[16]   X←(((M-1),GOSU)ρ((ΔSC;M]-1)×ΔN))+⍳(0,¯1)↓ΔS
[17]   V←V=⌈/V←(ΔR+1)↓B[X]
[18]   ΔS←V/[1] ΔS
[19]   ΔR←V/[1] ΔR
[20]   CODE←CODE,B[X[;V⍳1]]
[21]   →L1
[22] EXIT:ΔS←(¢ΔC←ΔSC1;])[ΔS]
      ∇
```

Figure 5. APL function for unique coding and symmetry perception.

step 2 of the basic algorithm. The program GROW (see Figure 4) "grows" the partially constructed permutation vectors. PARTITION1, NORMALIZATION, and GROW are used by LTRMAXCODE (see Figure 5) which provides the maximum lower triangular code t(g^*,**B**), normalized atomic vector a ($=a^*$), and bond-atom symmetry $G^{(B^*)} \cap G^{(a^*)}$. The function SETSELECT used by NORMALIZATION is given in the preceding paper.[6]

The principal arrays used in the programs are described in Table IV.

REFERENCES AND NOTES

(1) B. Baumslag and B. Chandler, "Group Theory", McGraw-Hill, New York, 1968.
(2) W. T. Wipke and T. M. Dyott, *J. Am. Chem. Soc.*, **96**, 4834 (1974).
(3) C. Jochum and J. Gasteiger, *J. Chem. Inf. Comput. Sci.*, **17**, 113 (1977).
(4) M. Uchino, *Kagakusosetsu*, **18**, 47 (1978).
(5) J. E. Rush, *J. Chem. Inf. Comput. Sci.*, **16**, 202 (1976).
(6) M. Uchino, preceding paper in this issue.
(7) Y. Kudo and S. Sasaki, *J. Chem. Doc.*, **14**, 200 (1974).
(8) A. Nijenhuis and H. S. Wilf, "Combinatorial Algorithms", Academic Press, New York, 1975.
(9) A. L. Mackay, *J. Theor. Biol.*, **54**, 399 (1975); *J. Chem. Phys.*, **62**, 309 (1975).
(10) For more advanced algorithm, see following paper in this issue.

Algorithms for Unique and Unambiguous Coding and Symmetry Perception of Molecular Structure Diagram. III. Method of Subregion Analysis for Unique Coding and Symmetry Perception†

MASAHIRO UCHINO

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

Received June 4, 1979; Revised Manuscript Received January 21, 1980

A new efficient method to provide unique codes and symmetry groups is proposed. In this method, a normalized bond matrix (or adjacency matrix) is separated into several subregions, and each subregion is analyzed to avoid unnecessary generation of the permutation vectors for determination of unique codes and symmetry groups. The number of permutation vectors to be examined can be greatly reduced by this analysis of subregions and is usually smaller than (or equal to) the order of symmetry group.

INTRODUCTION

In the basic algorithm[1] to determine the right coset of the symmetry group $G^{(B)} \cap G^{(a)}$ in $G^{(s)}$, we "grow" the permutation vectors from the first interval of the permutation to the last interval (i.e., from left to right), testing at each stage to see if the partially constructed vectors can be extended to the vectors which are members of the right coset $(G^{(B)} \cap G^{(a)})g^*$. In this paper, we describe a more advanced technique to determine the right coset satisfying the criterion of "the maximum lower triangular code".[1]

In the new method, we "grow" the segments of the permutation vectors in the intervals of permutation in a certain optimal way and not mechanically from left to right. The decision for the interval in which we "grow" the segments of the permutation vectors can be made by separating the permuted bond matrix into several subregions. And for this purpose, we introduce in section I two vectors and two matrixes as functions of partially constructed vector and interval indexes. The analysis of the subregions by these functions provides six possible cases, and for each case, the action to be taken is described in section II. In section III, examples of the proposed method are presented.

I. FUNCTIONS OF PARTIALLY CONSTRUCTED PERMUTATION VECTOR

In the basic algorithm,[1] the partially constructed vector at the mth stage consists of the first m components of the permutation vector. In this paper, the partially constructed vector $\boldsymbol{g}$ consists of the constructed segments $\boldsymbol{g}(1)$, $\boldsymbol{g}(2)$, . . ., $\boldsymbol{g}(i)$ of lengths p_1, p_2, . . ., p_i which are associated with the permutation intervals of lengths c_1, c_2, . . ., c_i. When the number of permutation intervals is q, we define q-element vector $\boldsymbol{d}$ such that[1]

$$d_1 = 0, \qquad d_j = \sum_{i=1}^{j-1} c_i, \qquad j = 2, 3, \ldots, q$$

and vector $\boldsymbol{u}^j$ for $j = 1, 2, \ldots, q$ as

$$u^j{}_i = d_j + i, \qquad i = 1, 2, \ldots, c_j$$

and let $r(\boldsymbol{g}(j))$ denote the vector obtained from $\boldsymbol{u}^j$, suppressing the elements of $\boldsymbol{g}(j)$.

† Presented at the Division of Chemical Information, ACS/CSJ Congress, Honolulu, Hawaii, April 3, 1979.

0095-2338/80/1620-0124$01.00/0