A New Scheme for Assignment of a Canonical Connection Table

Chang-Yu Hu and Lu Xu*

Applied Spectroscopy Laboratory, Changchun Institute of Applied Chemistry, Academia Sinica,
Changchun 130022, Jilin, People's Republic of China

Received November 2, 1993®

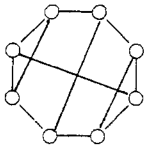
A new algorithm for deriving canonical numbering of atoms in a molecular graph has been developed. Some graph invariants, such as node properties, degree (connectivity), topological path, the smallest node ring index, etc., are encoded together to partition the atoms in a molecule. A canonical numbering of atoms is generated after the atom partitioning, and thus a unique connection table is constructed.

INTRODUCTION

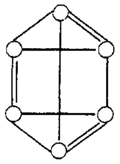
Unique coding of chemical structures for manipulation by computers has become increasingly important. Documentation and information retrieval systems, computer-aided synthetic design, computer-assisted structure elucidation, and research on structure–activity correlations are some of the most prominent fields of applications. Chemical structures can be represented in many ways, such as symbolic representation, e.g., the well-known Wiswesser linear notation,[1] connection matrix,[2] connection stack,[3] connection table,[4,5] etc. The connection table is a major technique for the topologically unambiguous and unique representation of chemical structures and is widely used in many fields.

The extensive works going on in our laboratory in developing computer programs for data base systems,[6,7] computer-assisted structure elucidation,[8,9] and studies on structure–activity correlations[10] are based on the connection table. A unique connection table is necessary to avoid multiple storage of the same chemical structures in our CIAC (Changchun Institute of Applied Chemistry) Comprehensive Information System of Rare Earths[6] and CIAC ^{13}C-NMR Database System.[7] An interactive program, ESESOC (Expert System for the Elucidation of the Structures of Organic Compounds) has been developed to reduce comprehensive spectroscopic information to its structural implications.[8] As an integral part of this system, the structure generator builds all the candidate structures consistent with the known information.[9] To remove the duplicate candidates, a unique connection table must be assigned to each isomer.

Various algorithms have been developed to produce a unique description for molecular structures.[11–23] Morgan's scheme[11] is based on the extended connectivity, which is derived by repeatedly summing the connectivity values of the nearest neighbors. The maximum differentiation of atoms is not achieved by this technique. Randic's algorithm associates the numbering atoms with the smallest binary code of the connectivity matrix.[12,13] Another approach suggested by Randic is based on the eigenvector associated with the largest eigenvalue of a molecule.[14] In Gasteiger's algorithm,[15] a new graph invariant, NOON (number of the outmost occupied neighbor sphere), is used. Though NOON is a useful graph invariant, this approach fails to detect topologically nonequivalent atoms in structure I. Similar to Ugi's algorithm,[16] Munk's scheme[17] partitions the atoms of a molecular structure into equivalence classes by utilizing the class membership of nearest neighbors. This algorithm completely partitions the nodes into equivalence classes for most molecular structures but fails to partition the nodes of structures I and II and some structures in Figure 1. Therefore, an algorithm using the permutation method was developed.[18] This algorithm is rigorous, but it is time-consuming to construct the $\prod(C_i)!$ permutations, where C_i is the number of nodes in class i (partitioned by other algorithms). Wipke,[19] Uchino,[20–22] and Elk[23] all have their own viewpoints. In this paper, we propose a new algorithm to partition the atoms in a molecule and assign a unique connection table for the molecule.



I



II

CONCEPTS

A molecular structure can be described by a chemical graph. In our scheme, four graph invariants are applied: node properties, degree, topological path, and the smallest node ring index.

(I) Node Properties. A node of a chemical graph stands for an atom in the molecule. In our system, the node properties include: (1) element type defining a specific non-hydrogen atom; (2) the number of attached hydrogen atoms; and (3) the partial bonds by which it can join to another node. The partial bonds may be single, double, or triple (aromatic carbon atom treated as >C= and —CH=).

All the nodes containing elements, such as C, N, O, P, S, F, Cl, Br, I, ..., were generated exhaustively and stored in the node library. An index number is assigned to each node in the node library. Table 1 shows several nodes of the node library.

(II) Degree. The degree (or connectivity) of a node is the number of other nodes attaching to it. According to Wipke,[17] Morgan's extended connectivity is a measure of how centrally involved a node is within a molecular graph.

(III) Topological Path. Gasteiger's NOON[15] is the minimum number of neighbor spheres necessary to accommodate all atoms of a molecule starting at that atom. The atoms with the minimum NOON value are defined as the center of the molecule. The nodes of the center form the first layer. The outer neighbors connecting immediately to the nodes of the first layer form the second layer, and then the

[®] Abstract published in *Advance ACS Abstracts,* April 1, 1994.

0095-2338/94/1634-0840$04.50/0

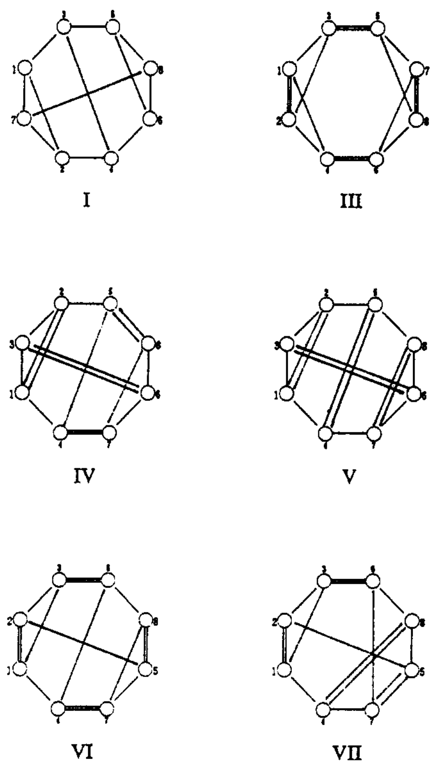


Figure 1. Some selected plocyclic structures.

Table 1. Several Nodes of the Node Library

no.	symbol	no.	symbol	no.	symbol
1	CH_3—	14	NH_2—	27	—S—
2	—CH_2—	15	—NH—	28	S=
3	CH_2=	16	NH=	29	>S=
4	>CH—	17	>N—	30	=S=
5	—CH=	18	—N=	31	>S(=)=
6	CH≡	19	N≡	32	H_2P—
7	>C<	20	>N(=)—	33	—PH—
8	>C=	21	—N(=)=	34	>P—
9	—C≡	22	=N≡	35	=P<—
10	=C=	23	N^-=	36	F—
11	HO—	24	=N^+=	37	Cl—
12	—O—	25	—N^+≡	38	Br—
13	O=	26	HS—	39	I—

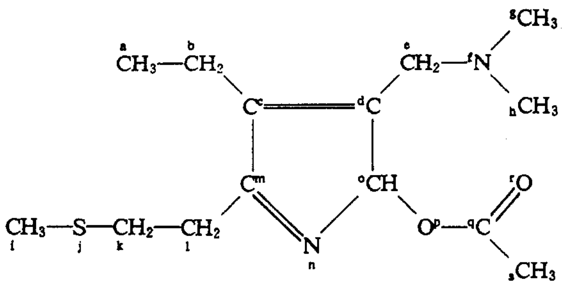
third layer, and so on. The layers are counted from the outmost to the center, the number of the layer is defined as the topological path (TP).

TP can be calculated by Gasteiger's NOON,

$$\mathrm{TP}[i] = \max_{i=1,n} \mathrm{NOON}[i] - \mathrm{NOON}[i] + 1$$

Where TP[i] stands for the topological path of the ith node, n is the total number of the nodes in a molecule, and NOON[i] is the NOON of the ith node.

As an example, the TP and NOON of structure **VIII** are listed in Table 2.



VIII

Table 2. TP Values of the Nodes in Structure VIII

node	TP	NOON	node	TP	NOON
a	3	7	k	3	7
b	4	6	l	4	6
c	5	5	m	5	5
d	4	6	n	5	5
e	3	7	o	4	6
f	2	8	p	3	7
g	1	9	q	2	8
h	1	9	r	1	9
i	1	9	s	1	9
j	2	8			

Node	NRI		SNRI	PI
c	21121 12112	⟶	12112	2
d	21211 11212	⟶	11212	1
m	21121 12112	⟶	12112	2
n	21211 11212	⟶	11212	1
o	12121 12121	⟶	12121	3

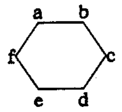
Figure 2. Partitioning by using the SNRI applied to the ring nodes c, d, m, n, and o in structure **VIII**.

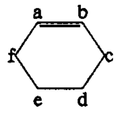
(IV) The Smallest Node Ring Index. For the nodes in a ring system, the smallest node ring index (SNRI) is introduced, with 1 for single bond, 2 for double bond, and 3 for triple bond. From a node in a ring, the codes of the bonds in the ring are put together sequentially; thus, a number is constructed which is defined as the node ring index (NRI).

From a node, along different directions (clockwise or reverse) or different ring paths, different NRIs will be obtained. The smallest of those NRIs is defined as the smallest node ring index (SNRI). The NRIs and SNRIs of nodes c, d, m, n, and o in structure **VIII** are shown in Figure 2.

The nodes in rings can be partitioned by the SNRI, and the partitioning identifier (PI) is assigned to each node in a molecular graph. The number of different SNRI values (NSNRI) are counted, and integers between one and NSNRI are assigned to the PI of each node in the ring system. The PI of the nodes with the largest SNRI becomes 1, the PI of the nodes with the smallest SNRI becomes NSNRI, and the nodes with the same SNRI have the same PI value. For the nodes not in ring system, PI is equal to 0. Figure 2 traces the implementation of partitioning of the ring nodes c, d, m, n, and o in structure **VIII** by using the SNRI.

The SNRI can mark the positions of the multiple bonds in a ring path in a molecular graph. For example, all the six bonds in the ring of cyclohexane are single bonds, so the SNRIs of each node are equivalent (111111). For cyclohexene, SNRIs





of node a, b, c, d, e, and f are 111112, 111112, 111121, 111211, 111211, and 111121, respectively. They are different ac-

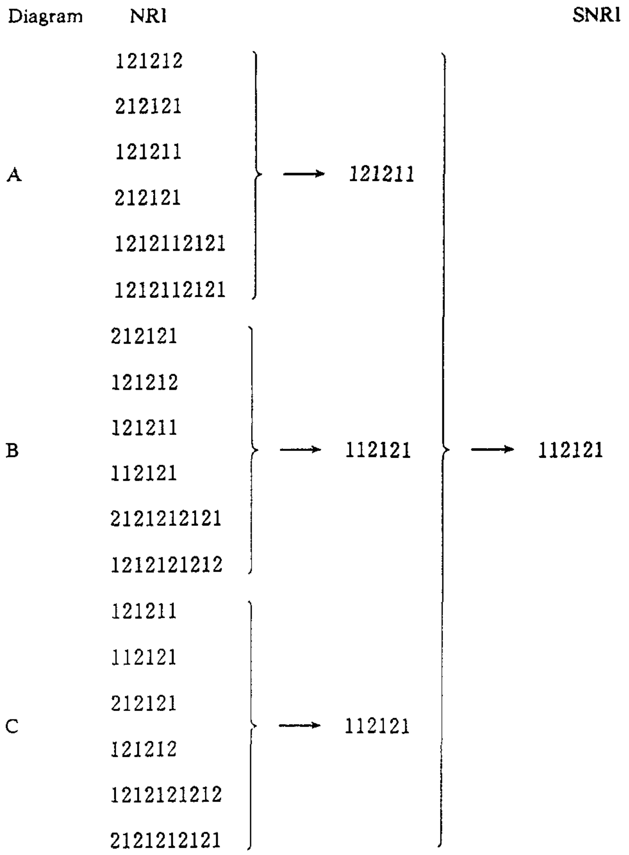
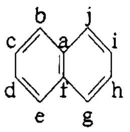


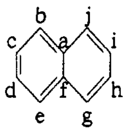
Figure 3. Calculation of the SNRI of node a in naphthalene.

cording to the positions of the node and double bond, so nodes a and b, c and f, and d and e are divided into different classes, respectively.

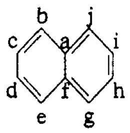
For aromatic compounds, a structure may have different diagrams. For example, naphthalene has three diagrams, A, B, and C, based on the positions of the double bonds. The calculation of the SNRI of node a in naphthalene is described in Figure 3.



A



B



C

By repeating the processes illustrated in Figure 3, the SNRIs of nodes b, e, c, d, e, f, g, h, and i can be obtained. The SNRI of the node f is 112121, the SNRI of the nodes b, e, g, and i is 111212, and the SNRI of the nodes c, d, h, and i is 121112. So the nodes a and f belong to a category, the nodes b, e, g, and j belong to a category, and the nodes c, d, i, and h belong to another category; thus, the SNRI can partition the three kinds of the nodes of naphthalene correctly.

ALGORITHM FOR THE PERCEPTION OF TOPOLOGICAL SYMMETRY

The algorithm in this study for perception of topological symmetry possesses some of the characteristics of the algorithms developed by Morgan,[11] Munk,[16] and Ugi.[18] It extends node environments throughout the graph, but some new graph invariants mentioned above are used to partition the atoms of a molecule into topological equivalent sets.

For perception of topological symmetry, in this system, the weight of each node in the molecular graph is calculated by the following function, which specifies the four graph invariants, i.e., node properties, degree, topological path, and partition of the nodes in ring system by the smallest node ring index:

$$W[i] = TP[i] \times 1000000 + D[i] \times 10000 + PI[i] \times 100 + N[i]$$

where W[i] stands for the weight of the ith node, TP[i] for the topological path of the ith node, D[i] for the degree of the ith node, PI[i] for the partition identifier of the ith node by using SNRI, and N[i] for the node index of the ith node in the node library.

An integer termed the class identifier (CI) is set for each node according to its weight value, and then the topological symmetry is represented by the CI. The algorithm for perception of topological symmetry (CI) is described as the following steps.

(1) Count the number of different weight values (NW), and assign a class identifier (CI) between 1 and NW to each node. Thus, the CI of the nodes with the largest weight value becomes NW, the CI of the nodes with the smallest weight value is 1, and the nodes with the same weight values possess the same CI. The number of the class identifier (NCI) is equal to NW.

(2) If the NCI is equal to the total number of nodes, then go to END, else, assign a trial weight (TW) to each node by utilizing the class membership of the nearest neighbors. The format consists of five three-digit integers. The leftmost field contains the CI of the node itself. The next four fields contain an ordered ascending list of nearest neighbor nodes (CI × 10 + b), where b is the code of the bond between the node and its neighbor. The format of the TW provides for up to four nearest neighbors. If there are less than four nearest neighbors, then the ordered ascending list is right justified in the four available fields and the leading digits are zero filled. The TW describes class membership of all nearest neighboring nodes and the previous class membership nodes of the given node.

(3) Count the number of different TW values (NTW). If NTW is not greater than NCI, then go to 4, else, set the W of each node to its TW, go to step 1.

(4) END; EXIT.

Table 3 traces the algorithm for structure **VIII**.

ASSIGNMENT OF UNIQUE CONNECTION TABLE

The sequence numbers of each node is determined according to its CI values by using the following algorithm.

(1) Choose the node with the highest CI value as the current node, and give it the sequence number 1.

(2) If there are any attachments to the current node which have not been assigned sequence numbers, then assign the unnumbered attachment with the highest CI value the next sequence number and repeat this step, else, go to step 3.

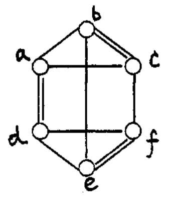
After the implementation of the algorithm for the perception of topological symmetry, the topological symmetry is represented by CI. If the nodes have same CI value, then they have topological symmetry. After being assigned, the topological equivalence of the unnumbered nodes may be changed in the case of their different attachments to numbered nodes. If more than one unnumbered attachment of the current node possess the same highest CI values, then the topological equivalence must be recalculated before resolving choices between attachments with the same CI values by following procedure: (i) set the CI values of the numbered nodes to 2n minus their sequence number, where n is the total number of the nodes in the molecular graph, and (ii) implement the algorithm for the perception of topological symmetry again.

Table 3. Results of the Algorithm Applied to Structure VIII

node	TP	D	PI	N	W	CI[a]	TW[b]	CI[c]	TW[d]
a	3	1	0	1	03010001	6	060000000000091	8	008000000000121
b	4	2	0	2	04020002	9	009000000061131	12	012000000081171
c	5	3	2	8	05030208	13	013000091112131	17	017000121152181
d	4	3	3	8	04030308	11	011000071101132	15	015000101141172
e	3	2	0	2	03020002	7	007000000051111	10	010000000071151
f	2	3	0	17	02030017	5	005000011011071	7	007000031031101
g	1	1	0	1	01010001	1	001000000000051	3	003000000000071
h	1	1	0	1	01010001	1	001000000000051	3	003000000000071
i	1	1	0	1	01010001	1	001000000000031	1	001000000000051
j	2	2	0	27	02020027	3	003000000011071	5	005000000011091
k	3	2	0	2	03020002	7	007000000031091	9	009000000051131
l	4	2	0	2	04020002	9	009000000071131	13	013000000091181
m	5	3	2	8	05030208	13	013000091122131	18	018000131162171
n	5	2	3	18	05020318	12	012000000101132	16	016000000141182
o	4	3	1	4	04030104	10	010000081111121	14	014000111151161
p	3	2	0	12	03020012	8	008000000041101	11	011000000061141
q	2	3	0	8	02030008	4	004000011022081	6	006000021042111
r	1	1	0	13	01010013	2	002000000000042	4	004000000000062
s	1	1	0	1	01010001	1	001000000000041	2	002000000000061

[a] NCI = 13 [b] NTW = 18 [c] NCI = 18 [d] NTW = 18

Table 4. Recalculation of the Topological Equivalence of Structure II after Assignment of Node b



II

node	CI	TW	CI[a]
a	1	001000012021111	2
b	11	011000011021022	6
c	2	002000011021112	4
d	1	001000012021021	1
e	2	002000011022111	5
f	2	002000011021022	3

[a] NCI = 6 (= N)

Table 5. Unique Connection Table of Structure VIII

sequence no.	node	connections		
1	>C=	2(1)	3(2)	4(1)
2	>C=	1(1)	5(2)	6(1)
3	—N=	1(2)	7(1)	
4	—CH₂—	1(1)	8(1)	
5	>C=	2(2)	7(1)	9(1)
6	—CH₂—	2(1)	10(1)	
7	>CH—	3(1)	5(1)	11(1)
8	—CH₂—	4(1)	12(1)	
9	—CH₂—	5(1)	13(1)	
10	CH₃—	6(1)		
11	—O—	7(1)	14(1)	
12	—S—	8(1)	15(1)	
13	>N—	9(1)	16(1)	17(1)
14	>C=	11(1)	18(2)	19(1)
15	CH₃—	12(1)		
16	CH₃—	13(1)		
17	CH₃—	13(1)		
18	O=	14(2)		
19	CH₃—	14(1)		

For example, in structure II (nodes a, b, ... are marked in Table 4), after the implementation of the algorithm for perception of topological symmetry, the CI value of the nodes a and d is equal to 1, and the CI value of the nodes b, c, e, and f is 2, so the nodes a and d and the nodes b, c, e, and f have symmetry, respectively. When the sequence number 1 is assigned to the node b, the attachments of node b are a, c, and e, and the nodes c and e have the same highest CI value, 2; therefore, the topological equivalence must be recalculated before the sequence number 2 is assigned. Let CI of node b be equal to 2 (6) – 1 = 11. After the implementation of the algorithm for topological symmetry perception again, the CI values of the nodes a, b, c, d, e, and f are 2, 6, 4, 1, 5, and 3, respectively, i.e., after node b is assigned, the nodes a, c, d, e, and f are all nonequivalent. Table 4 traces the process mentioned above.

(3) If the structure is completely numbered, then go to step 4, else the node with sequence number equal to that of the current node plus 1 becomes the current node, and go to step 2.

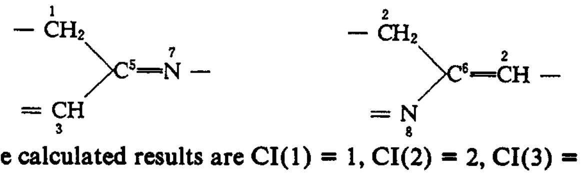
(4) Done, the sequence numbers have been assigned.

Structure VIII would be numbered as m → 1, c → 2, n → 3, l → 4, d → 5, b → 6, o → 7, k → 8, e → 9, a → 10, p → 11, j → 12, f → 13, q → 14, i → 15, g → 16, h → 17, r → 18, s → 19, and the unique connection table of structure VIII used in CIAC ^{13}C-NMR Database System and ESESOC-I is described in the Table 5.

CONCLUSIONS

This research has created a scheme capable of identifying all the equivalent atoms of a molecule and generating a unique connection table for each and every constitutionally isomer, independent of the original number of the molecular structure. We believe the topological symmetry algorithm to be rigorous in identification of equivalence classes. For rigorous tests of the method, some complex polycyclic graphs in Figure 1 were examined, and the results are listed in Table 6. From Figure 1 and Table 6, it can be seen that topological equivalence was correctly determined.

The algorithm for perception of topological equivalence can also be applied to partially assembled structures. For example, in the structure generation process of ESESOC-I, it is necessary to detect the topological equivalence of bonding sites 1 and 2 for the following partially assembled structure,



The calculated results are CI(1) = 1, CI(2) = 2, CI(3) = 3,

Table 6. Results of Topologically Equivalent Classes for the Selected Structures in Figure 1

structure	sets of equivalence nodes	CI
I	(1,2,5,6)	1
	(3,4)	3
	(7,8)	2
III	(1,2,7,8)	2
	(3,4,5,6)	1
IV	(1,2)	1
	(3)	3
	(4,5)	5
	(6)	4
	(7,8)	2
V	(1,2,7,8)	1
	(3,6)	2
	(4,5)	3
VI	(1,8)	2
	(2,5)	3
	(3,7)	1
	(4,6)	4
VII	(1,2)	5
	(3)	4
	(4,5)	3
	(6)	1
	(7,8)	2

CI(4) = 4, CI(5) = 7, CI(6) = 8, CI(7) = 6, CI(8) = 5. Therefore, bonding sites 1 and 2 are nonequivalent.

This algorithm is programmed in FORTRAN and runs on a Micro VAX II computer. It has been applied to ESESOC-I successfully.

ACKNOWLEDGMENT

The authors acknowedge the financial support of the National Natural Science Foundation of China.

REFERENCES AND NOTES

(1) Smith, E. G., *The Wiswesser Line-formula Chemical Notation*; New York: McGraw-Hill, 1986.

(2) Spialter, L. The Atom Connectivity Matrix (ACM) and its Characteristic Polynomial (ACMCP). *J. Chem.* Doc. **1974**, *14*, 200.

(3) Kudo, Y.; Sasaki, S. The Connectivity Stack, A New Format for Representation of Chemical Structures. *J. Chem. Doc.* **1974**, *14*, 200.

(4) Ray, L. C.; Kirsch, A. Finding Chemical Records by Digital computers. *Science* **1975**, *126*, 814.

(5) Goodson, A. L. Graph-Based Chemical Nomenclature. 1. Historical Background and Discussion. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 167.

(6) Xu, L.; Li, G. Q.; Wang, S. Y.; Lu, H.; Wang, H. Y.; Hu, C. Y.; Xiao, Y. H.; Xiao, Y. D.; Jiang, X. H.; Lu, X. Y. Comprehensive Information System of Rare Earths. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 375.

(7) Xu, L.; Li, G. Q.; Wang, S. Y.; Zhao, K. D.; Zhang, J. G.; Sun, J. CIAC-Carbon-13 NMR Search System. *Chin. J. Anal. Chem.*, **1986**, *14*, 431.

(8) Hu, C. Y.; Xu, L. The Computer Automatic Structure Elucidation Expert System ESESOC. I. Chin. J. Anal. Chem. **1991**, *20*, 643.

(9) Hu, C. Y.; Xu, L. An Expert System for the Elucidation of the Structures of Organic Compounds Containing C, H, O Elements. Chin. *J. Org. Chem.* **1993**, *13*, 129.

(10) Yao, Y. Y.; Xu, L., Yang, Y. Q.; Yuan X. S. Studies on Structure–Activity Relationships of Organic Compounds—Three Topological Indexes and Their Applications *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 590.

(11) Morgan, L. The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107.

(12) Randic, M. On the Recognition of Identical Graphs Representing Molecular Topology. *J. Chem. Phys.* **1974**, *60*, 3920.

(13) Randic, M. On Unique Numbering of Atoms and Unique Codes for Molecular Graph. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 105.

(14) Randic, M. On Canonical Numbering of Atoms in a Molecule and Graph Isomorphism. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 171.

(15) Jochum, C.; Gasteiger, J. Canonical Numbering and Constitutional Symmetry. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 113.

(16) Schubert, W.; Ugi, I. Constitutional Symmetry and Unique descriptors of Molecules. *J. Am. Chem. Soc.* **1978**, *100*, 37.

(17) Shelley, C. A.; Munk, M. E. Computer Perception of Topological Symmetry. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 110.

(18) Shelley, C. A.; Munk, M. E. An Approach to the Assignment of Canonical Connection Tables and Topological Symmetry Perception. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 247.

(19) Wipke, W. T.; Dyott, M. T. Stereochemically Unique Algorithm. *J. Am. Chem. Soc.* **1974**, *96*, 4834.

(20) Uchino, M. Algorithms for Unique and Unambiguous Coding and Symmetry Perception of Molecular Structure Diagram. I. Vectors Functions for Automorphism Partitioning. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 116.

(21) Uchino, M. Algorithms for Unique and Unambiguous Coding and Symmetry Perception of Molecular Structure Diagram. II. Basic Algorithm for Unique Coding and Computation of Symmetry Group. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 121.

(22) Uchino, M. Algorithms for Unique and Unambiguous Coding and Symmetry Perception of Molecular Structure Diagram. III. Method of Subregion Analysis for Unique Coding and Symmetry Perception. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 124.

(23) Elk, S. B. Graph–Theoretical Algorithm to Canonically Name the Isomers of the Regular Polyhedranes. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 14.