have not been fully refined with respect to the constraints and mutually contradictory sets of constraints. With an energy function, one can produce conformations that have minimal energy while truly satisfying all the constraints. The customary approach of adding constraint penalty terms to an energy function merely produces a compromise between partially satisfied constraints and partially optimal energy. With various other sorts of objective functions, AL can be used to more efficiently explore the range of allowed conformations. One can answer questions such as "What is the closest approach possible between these two atoms?", "How far can they be separated?", "How nearly trans can this bond become?", or "How dissimilar can I make this conformation compared to a standard structure?". In addition to handling such optimizations subject to very general geometric equality and inequality constraints, AL gives valuable insight concerning the interactions among constraints and between the constraints and the objective at the final conformation. We believe this last feature will be very important.

ACKNOWLEDGMENT

This work is based on some earlier efforts by R. Nambudripad and Y.-S. Sung. We acknowledge funding support by NSF Grant DMB-8705006 and NIH Grant GM37123.

REFERENCES AND NOTES

(1) Crippen, G. M. *Distance Geometry and Conformational Calculations*; Chemometrics Research Studies Series 1, Research Studies (Wiley): New York, 1981.

(2) Crippen, G. M.; Havel, T. F. *Distance Geometry and Molecular Conformation*; Research Studies (Wiley): Chichester, England, 1988.

(3) Bertsekas, D. P. *Constrained Optimization and Lagrange Multiplier Methods*; Academic: New York, 1982; pp 20, 96-156.

(4) Mosberg, H. I.; Subramanian, P.; Sobczyk, K.; Crippen, G. M.; Ramalingam, K.; Woodard, R. W. *Combined Use of Stereospecific Deuteration, NMR, and Distance Geometry for Conformational Analysis of* [D-*Pen*^{2},D-*Pen*^{5}]*Enkephalin*; presented at the 10th American Peptide Symposium, St. Louis, MO, May 1987.

Canonical Numbering and Coding of Imaginary Transition Structures. A Novel Approach to the Linear Coding of Individual Organic Reactions

SHINSAKU FUJITA

Research Laboratories, Ashigara, Fuji Photo Film Co., Ltd., Minami-Ashigara, Kanagawa, Japan 250-01

Received October 9, 1987

The canonical numbering and coding of an imaginary transition structure (ITS) are described. The nodes of an ITS are partitioned partially into (pseudo)equivalent classes in light of four kinds of extended connectivities. Each of the nodes of the highest class is selected as a root, to which possible spanning trees are constructed. The nominated sets of canonical numbering are obtained from the respective spanning trees. Then the canonical code is obtained by comparing newly defined lists based on the sets of numbering. The concept of a reduced ITS is proposed. The canonical numbering and coding of the reduced ITS are also discussed.

In previous papers,[1] we presented the concept of imaginary transition structures (ITS's) for the description of organic reactions. The ITS of a given reaction is a structure that has out- (—//—), in- (—o—), and par-bonds (—) in accord with structural changes during the reaction. This formulation provides an explicit method for describing an *individual organic reaction.*[1j] The ITS contains 15 kinds of imaginary bonds, each of which is a combination of the out-, in-, and/or par-bonds.[1a] Then the ITS is stored and manipulated in terms of an ITS connection table, in which the imaginary bonds are represented by complex bond numbers. In order to construct an effective computer system, the canonical numbering of the nodes (vertices) of the ITS and the canonical coding of the ITS are remaining problems to be solved.

Many methods were reported for canonizing organic compounds (molecular graphs).[2] One of the most familiar methods is Morgan's procedure, in which (1) the nodes of a molecular graph are partially partitioned by the iterative calculation of extended connectivities and then (2) numbered after the formation of a spanning tree rooted to each of the uppermost nodes, and finally (3) the best name is selected by comparison between nominated names.[3]

The present paper deals with the canonical numbering and coding of ITS's. The resulting canonical names of ITS's (CANITS) are the first unambiguous codes for the description of *individual organic reactions.* This method is an extension of Morgan's procedure, in which (1) four kinds of extended connectivities are introduced to partition the nodes of an ITS partially, and (2) the selection of the best name is based upon a newly defined linear code. In addition, we propose the concept of reduced imaginary transition structures and their canonical coding.

PARTIAL PARTITIONING BY MEANS OF FOUR KINDS OF EXTENDED CONNECTIVITIES

The ITS of a given reaction contains (1) *intra*string hydrogen atoms (hydrogen reaction centers), (2) implicit or explicit *extra*string hydrogen atoms (hydrogen atoms other than reaction centers), and (3) non-hydrogen atoms.[4] Among them, we consider (1) and (3) for the present coding of the ITS unless the description of stereochemistry requires the consideration of (2).

In the present method, four kinds of extended connectivities, EC1(i), EC2(i), EC3(i), and EC4(i), are computed and assigned to each node i of a given ITS. Then the nodes of the ITS are partitioned into (pseudo)equivalent classes in light of these extended connectivities.[5] Figure 1 shows the flow chart of the partial partitioning of the nodes to be examined.

Step 1. The initial values of the extended connectivities are calculated for each node i as follows: EC1(i), the number of neighboring reaction centers (hydrogen and non-hydrogen atoms) that are linked to the current node (i) by in- or out-bonds; EC2(i), the number of neighboring atoms (except extrastring hydrogen atoms) attached to the node i with any kind

0095-2338/88/1628-0128$01.50/0 © 1988 American Chemical Society

Table I. Initial Values of Extended Connectivities (EC1–EC4) of Each Node of ITS **1**[a]

node	EC1	EC2	EC3	EC4
1	2	2	1	3
2	2	2	1	2
3	2	2	2	2
4	2	2	1	3
5	0	2	2	3
6	0	2	2	2
7	0	2	2	2
8	0	2	2	3
9	2	3	3	4
10	2	3	2	4

[a] Iteration = 0; number of classes = 7.

Table II. First Trial Values of Extended Connectivities of Each Node of ITS **1**[a]

node	EC1	EC2	EC3	EC4
1	4	5	4	6
2	4	4	3	5
3	4	4	2	5
4	4	5	4	6
5	2	5	4	6
6	0	4	4	5
7	0	4	4	5
8	2	5	5	6
9	4	7	5	10
10	4	7	6	10

[a] Iteration = 1; number of classes = 8.

of imaginary bonds; EC3(i), the number of neighboring atoms (except extrastring hydrogen atoms) linked to the node i with one or more par-bonds; EC4(i), the number of second neighbors (except extrastring hydrogens) with respect to the node i.

Step 2. The order of the node i is denoted by SETNO(i). The initial SETNO(i) values are zero for all i.

Step 3. An array of EC(i) of 35 characters is defined as follows

$$\mathrm{EC}(i) = |^{1}999-\mathrm{SETNO}(i)|^{4}\mathrm{EC1}(i)|^{12}\mathrm{EC2}(i)|^{20}\mathrm{EC3}(i)|^{28}\mathrm{EC4}(i)^{35}|$$

wherein the three-character top section of the string is a value 999 − SETNO(i) and the four remaining sections of eight characters are set to values of EC1–EC4.[6]

Step 4. The EC(i) values for all i are then sorted in descending order. The order is in turn stored in the array SETNO(i) as a new value. The maximum number of SETNO(i) is defined as NEC, which is equal to the number of different EC(i) values (i.e., the number of classes divided) in the first partitioning step.[7]

Step 5. The four kinds of trial extended connectivities, TEC1(i), TEC2(i), TEC3(i), and TEC4(i), are calculated for each node i from EC1(i), EC2(i), EC3(i), and EC4(i), respectively. Each trial value is the sum of the values assigned to the nodes that adjoin the node i under consideration.

Step 6. An EC(i) is assigned to the node i as follows

$$\mathrm{EC}(i) = |999-\mathrm{SETNO}(i)|\mathrm{TEC1}(i)|\mathrm{TEC2}(i)|\mathrm{TEC3}(i)|\mathrm{TEC4}(i)|$$

wherein the value of SETNO(i) is that of the last trial.

Step 7. The EC(i) values for all i are then sorted in descending order. The order is stored in the array SETNO(i). The maximum number of SETNO(i) is NTEC.

Step 8. If NTEC is equal to NEC, go to step 10.[8]

Step 9. If NTEC is greater than NEC, assign NTEC to NEC, TEC1(i) to EC1(i), TEC2(i) to EC2(i), TEC3(i) to EC3(i), and TEC4(i) to EC4(i). Then go to step 5.

Step 10. Done. The SETNO(i) values show the final classes divided.

Table III. Second Trial Values of Extended Connectivities of Each Node of ITS **1**[a]

node	EC1	EC2	EC3	EC4
1	8	11	8	15
2	8	9	6	11
3	8	9	7	11
4	8	11	8	15
5	4	11	10	15
6	2	9	8	11
7	2	9	9	11
8	4	11	9	15
9	10	17	15	22
10	10	17	13	22

[a] Iteration = 2; number of classes = 9.

Table IV. Third Trial Values of Extended Connectivities of Each Node of ITS **1**[a]

node	EC1	EC2	EC3	EC4
1	18	26	21	33
2	16	20	15	26
3	16	20	14	26
4	18	26	20	33
5	12	26	21	33
6	6	20	19	26
7	6	20	17	26
8	12	26	24	33
9	22	39	30	52
10	22	39	33	52

[a] Iteration = 3; number of classes = 10.

Table V. Fourth Trial Values of Extended Connectivities of Each Node of ITS **1**[a]

node	EC1	EC2	EC3	EC4
1	38	59	45	78
2	34	46	35	59
3	34	46	35	59
4	38	59	47	78
5	28	59	52	78
6	18	46	38	59
7	18	46	43	59
8	28	59	47	78
9	52	91	78	118
10	52	91	71	118

[a] The number of divided classes is equal to that of the third iteration. End of iteration: iteration = 4; number of classes = 10.

Table VI. Final Values of EC Based on the Fourth Iteration (Table V) and the Resulting Partitioning into 10 Classes of the Nodes[a]

		EC				
node	class	999-SETNO	EC1	EC2	EC3	EC4
1	4	995	38	59	45	78
2	6	993	34	46	35	59
3	3	996	34	46	35	59
4	5	994	38	59	47	78
5	8	991	28	59	52	78
6	10	989	18	46	38	59
7	9	990	18	46	43	59
8	7	992	28	59	47	78
9	1	998	52	91	78	118
10	2	997	52	91	71	118

[a] Number of classes = 10.

Let us illustrate the above partitioning process by using as an example the Claisen rearrangement of 1-(allyloxy)cyclohexene forming 2-allylcyclohexanone (Figure 2). This reaction is represented by ITS **1**, in which the initial numbering is given in an arbitrary fashion. Tables I–V show the iterative calculation of extended connectivities for this ITS. Since the number of classes divided by the fourth iteration (Table V) is the same value as that of the third one (Table IV), the iteration process is ended at this point. The final number of classes is 10, which indicates complete partitioning in this case.

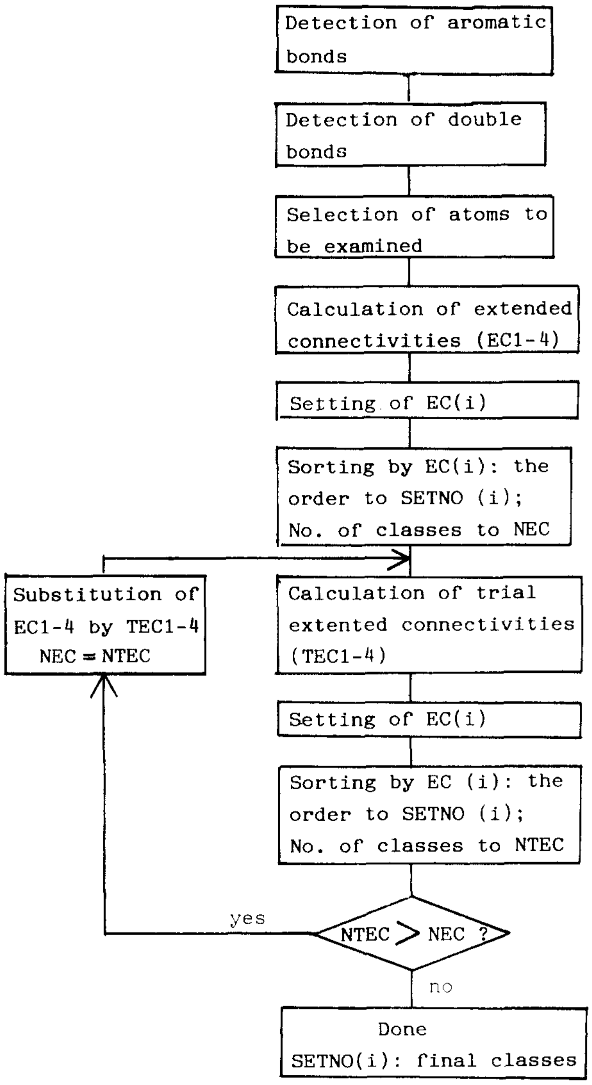

Detection of aromatic bonds
Detection of double bonds
Selection of atoms to be examined
Calculation of extended connectivities (EC1-4)
Setting of EC(i)
Sorting by EC(i): the order to SETNO (i); No. of classes to NEC
Substitution of EC1-4 by TEC1-4 NEC = NTEC
Calculation of trial extented connectivities (TEC1-4)
Setting of EC(i)
Sorting by EC (i): the order to SETNO (i); No. of classes to NTEC
yes
NTEC > NEC ?
no
Done SETNO(i): final classes

Figure 1. Flow chart of the subroutine for the calculation of four kinds of extended connectivities and the partial partitioning.

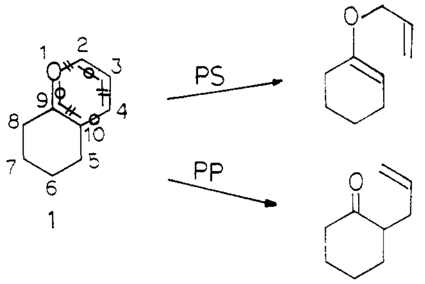


Figure 2. Imaginary transition structure of the Claisen rearrangement. The initial number of each node is given arbitrarily. The PS and PP operations afford the substrate and product, respectively.

Table VI collects EC(i) and SETNO(i) values of the final step of the iteration.

CANONICAL NUMBERING AND CODING BASED ON A SPANNING TREE

Construction of a Linear Code. Figure 3 illustrates a flow chart for the process of canonical numbering. The canonical numbering is based on a spanning tree rooted to an uppermost node that is selected in the preceding partial partitioning.

The processes of the canonical numbering and coding are as follows.

Step 1. (a) Select an uppermost node (i.e., SETNO(i) = 1) as the current atom and give it the sequence number 1. (b) In the case that two or more nodes have SETNO(i) of 1, select one of them as the node 1.

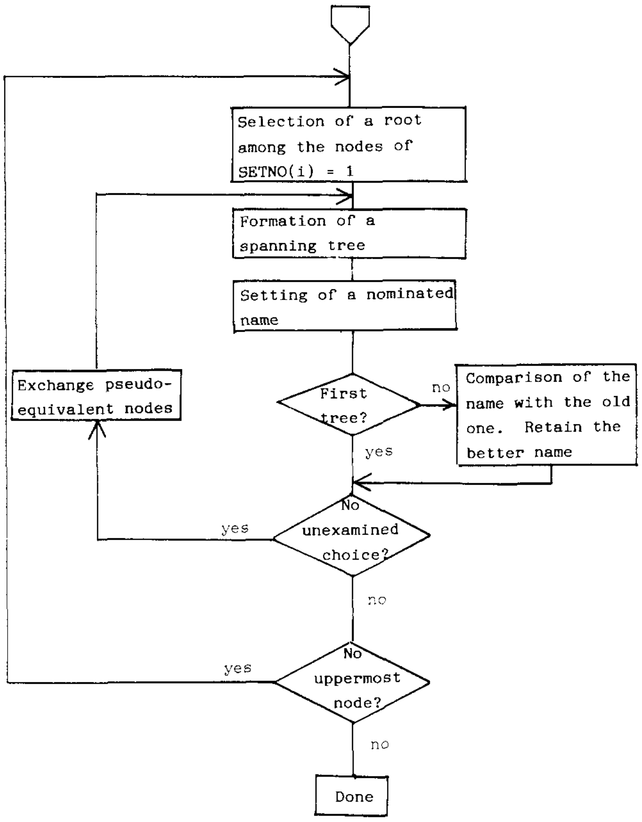

Selection of a root among the nodes of SETNO(i) = 1
Formation of a spanning tree
Setting of a nominated name
Exchange pseudo-equivalent nodes
First tree?
no
Comparison of the name with the old one. Retain the better name
yes
yes
No unexamined choice?
no
yes
No uppermost node?
no
Done

Figure 3. Flow chart of the subroutine for the canonical numbering and coding.

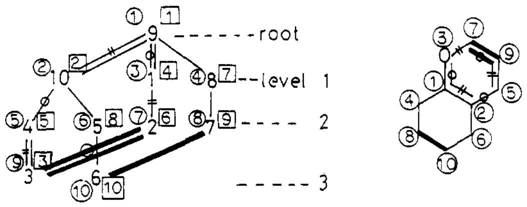


Figure 4. Spanning tree for ITS **1**. A number in a square indicates the class assigned to each node (i.e., SETNO(i) for node i). A circled number in the tree (left) or in the ITS (right) is the canonical number of each node. A bold-faced bond is a ring-closure bond.

Step 2. Construct a spanning tree rooted to the node 1. The maximum depth of the tree is max(d).

Step 3. d = 1, wherein the value (d) is a depth (or level) from the root.

Step 4. (a) The nodes i and j in the level d are numbered in this sequence if SETNO(LNK(i)) is smaller than SETNO(LNK(j)), wherein LNK(i) and LNK(j) are the nodes of level $d - 1$ linked to the node i and j, respectively. (b) If LNK(i) is equal to LNK(j) and SETNO(i) is smaller than SETNO(j), then the nodes i and j are numbered in this sequence. (c) If two or more nodes have the same SETNO and the same LNK, select an unexamined permutation of the nodes and number them in this sequence.

Step 5. If $d < \max(d)$, then $d = d + 1$ and go to step 4; otherwise, go to step 6.

Step 6. At the first run, construct and store an initial (and an old) name based on the numbering. (For the constitution of this name, see below.) At the other runs, construct a new nominated name, compare this with the old one, and then retain the better (i.e., lexicographically smaller) name.

Step 7. If there are other permutations, then go to step 3. If all the permutations have been examined, go to step 8.

Step 8. If all of the uppermost nodes have been examined, the process is completed. The retained name is the canonical name (CANITS). Otherwise, select the next node and go to step 1b.

Figure 4 shows a tree for ITS **1**. The numbers on the nodes (vertices) are the arbitrary sequential numbers given initially to the nodes (see Figure 2); each number in a square indicates the class assigned to the respective node (i.e., SETNO), and the circled numbers denote the canonical numbering of this ITS. The root of the tree is node 9, which has SETNO(*9*) = $\boxed{1}$. The depth of the tree is 3. The array defined in the above steps can be assigned easily. For instance, node 3 has SETNO(3) = $\boxed{3}$ (the number in a square) and LNK(3) = 4 (the initial numbering of the upper level), and then SETNO(LNK(3)) = SETNO(4) = $\boxed{5}$ (the number in a square at node 4) etc.

The node 9 is renumbered to be No. ①, since SETNO(9) = $\boxed{1}$. Since SETNO(10) = $\boxed{2}$, SETNO(1) = $\boxed{4}$, and SETNO(8) = $\boxed{7}$ in the level (depth) 1, the three nodes are ordered to be 10 > 1 > 8 and renumbered sequentially to be ②, ③, and ④. The nodes in the lower levels are renumbered similarly. The canonical numbering based on this tree is shown separately on the right ITS of Figure 4.

Each nominated name consists of the following lists that indicate the connectivity of the ITS based on the numbering examined. The canonical numbering and coding are obtained after comparison between all possible names as described above:

(**a**) Length of the canonical name (four characters).

(**b**) Number of nodes considered (three characters).

(**c**) Number of rings contained in the ITS (three characters).

(**d**) FROM list. The FROM list contains LNK(*i*) for all *i* ($i \geq 2$) in the ascending order of *i*, wherein *i* is the canonical number of each node and LNK(*i*) is renumbered in accord with the canonical numbering. Three characters are used for each node. For example, node 2 (corresponding to the initial node 10) of ITS **1** has the FROM list value of 001, which corresponds to the initial node 9 as shown in Figures 2 and 4.

(**e**) RING-CLOSURE list. The RING-CLOSURE list defines the remaining connectivity of the ITS that indicates the presence of ring structures (ITS rings). The pair of integers corresponding to the two terminal nodes of each ring-closure bond is collected in ascending order to form a six-character string. Then the six-character strings are listed in ascending order. In the case of ITS **1**, the ring-closure bonds are shown by heavy lines in Figure 4 and listed in ascending order, i.e., 007009008010.

(**f**) PAR-BOND list. The PAR-BOND list contains the number of par-bonds for each linkage, aligned in the order defined in FROM and RING-CLOSURE lists (one character for each linkage).

(**g**) IN-BOND list. In-bonds are listed in the same order as (**f**) in this list.

(**h**) OUT-BOND list. Out-bonds are listed in the same order as (**f**) in this list.

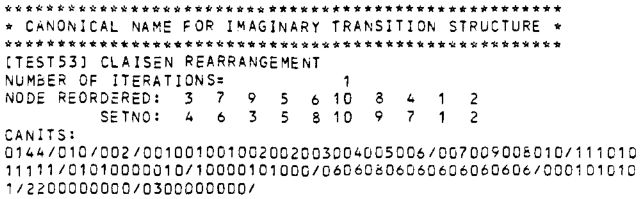
(**i**) ATOM list. The atomic numbers (two characters for each node) of the nodes are listed in the order of the canonical numbering in this list.

(**j**) INTACT NODE list. The INTACT NODE list contains 0 for a reaction center and 1 for a node other than the reaction center. These numbers are listed in the order of the canonical numbering.

(**k**) STARTING STEREO list. The STARTING and PRODUCT STEREO lists are concerned with the stereochemistry of the ITS (see below).

(**l**) PRODUCT STEREO list.

A canonical code for an ITS (CANITS) is the linear combination of the lists a–l in the order defined above. The INTACT NODE list (j) is necessary to indicate the reaction centers of cis–trans isomerizations of double bonds (see below). In addition, the presence of the list j is convenient to detect reaction centers and to examine their changes by referring to the lists i, k, and l. The heading lists a–c are omissible for compactness. However, the matching of the codes would be easier if the lists concerned with whole topological features were compared at an early stage of the matching process.

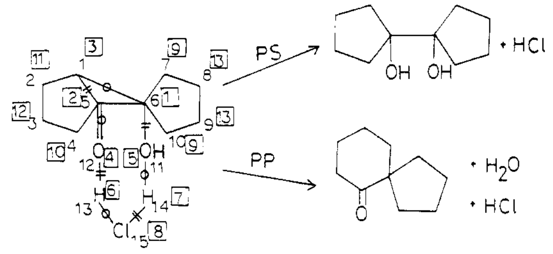


```
*****************************************************
* CANONICAL NAME FOR IMAGINARY TRANSITION STRUCTURE *
*****************************************************
[TEST53] CLAISEN REARRANGEMENT
NUMBER OF ITERATIONS=          1
NODE REORDERED:  3  7  9  5  6 10  8  4  1  2
         SETNO:  4  6  3  5  8 10  9  7  1  2
CANITS:
0144/010/002/001001001002002003004005006/007009008010/111010
11111/0101000001 0/1000010100 0/06060806060606060606/000101010
1/2200000000/0300000000/
```

Figure 5. Canonical code of ITS **1**. The corresponding canonical numbering is found in Figure 2.

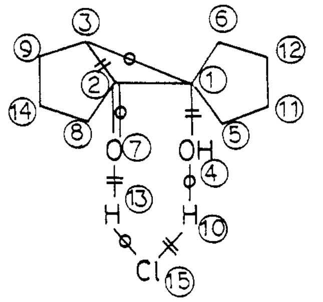
In comparison with the original Morgan name, the CANITS of the present work contains additional lists representing structural changes during a reaction, i.e., lists g, h, and j. The order of lists is decided by the descending priority of the categories of information, i.e., information on topological features (a–e), information on imaginary bonds (f–h), and information on nodes (i–l). Thus, the CANITS contains the ATOM list after the PAR-, IN-, and OUT-BOND lists, whereas the Morgan name prefers the ATOM list to the (PAR-)BOND list.[3a] The preference of the three BOND lists in the CANITS affords an effective representation, since an electron transfer during a reaction is suitably expressed by the BOND lists. This advantage is amplified when the CANITS is applied to the coding of reaction graphs, which will be discussed in the accompanying paper.

Figure 5 shows the CANITS of ITS **1**, which distinguishes lists a–l from each other with slashes (/). The "number of iterations" in Figure 5 is the number of trees constructed in the process of Figure 3. In this case, the tree is formed only once, since the nodes of ITS **1** are perfectly differentiated from each other by the values of SETNO. The "node reordered" in Figure 5 shows the new sequential numbers (the canonical numbers) in the order of the initial node numbers. This is illustrated by the comparison between the numbering of Figure 2 and that of the circled numbering of Figure 4. The SETNO collects the class values of the nodes in the initial order of node numbers. In the CANITS of Figure 5, the top three sections are header sections, in which 0144 is the length of this CANITS code, 010 is the number of nodes codified, and 002 is the number of rings (or ring-closure bonds). The fourth section shows the FROM list in which the three-digit groups of 001001001002... represent that the nodes 2, 3, and 4 are linked to node 1 (i.e., 001), node 5 is adjacent to node 2 (i.e., 002), and so on. The fifth section is the RING-CLOSURE list, where every six-digit number consists of a pair of three-digit numbers corresponding to the two terminal nodes of a ring-closure bond. Thus, the numbers 007009 indicate that the bond between nodes 7 and 9 is a ring-closure bond. The bond (C±C) between nodes 1 and 2 (the first bond derived from the FROM list), for example, is characterized by the values in the sixth to eighth sections, i.e., 1 (the first number of the PAR-BOND list), 0 (that of the IN-BOND list), and 1 (that of the OUT-BOND list). Similarly, the ring-closure bond (C±C) between nodes 7 and 9 (the first bond in the RING-CLOSURE list) is denoted by the values 1 (the 10th value of the PAR-BOND list), 1 (that of the IN-BOND list), and 0 (that of the OUT-BOND list). The ninth section of the CANITS contains atomic numbers in the order of the canonical numbering of the nodes, wherein 06 is for carbon and 08 is for oxygen. The 10th section indicates the reaction centers designated by the value 0. The stereochemical in-



2

Figure 6. ITS of the pinacol rearrangement. The initial numbering of the nodes is arbitrary. A number in a square indicates the class assigned to each node. This case has 13 classes of nodes.



```
CANITS:
0217/015/004/001001001001001002002003004005006007008010/0020
03009014011012013015/100111110110100110/010001001000000001/0
01000000001011000/060606080606080606010606010617/00001101101
1010/000000000000000/000000000000000/
```

Figure 7. Canonical numbering and code of ITS **2**.

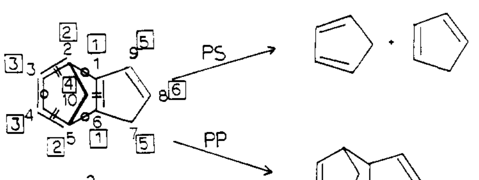
formation is collected in the 11th and 12th sections. For example, the second value (i.e., 3) in the 12th section of the CANITS indicates that node 2 of the product is a chiral center but the stereochemistry is unknown (see below).

Presence of Choice Points during the Numbering Process Based on a Tree. In the case that two or more pseudoequivalent nodes are present at the same level of the spanning tree (see step 4c in the preceding section), a choice must be made between the nodes. The corresponding names are generated and compared with each other. Then the numbering is selected to give the better (i.e., lexicographically smaller) name that contains the lists a–l. All permutations at each choice point should be compared to obtain the canonical numbering. However, the comparing step can be omitted if two or more nodes of the choice point are terminal and have the same atomic number and if LNK(i) is equal to LNK(j), wherein all i and all j are for all of the nodes. This terminal shortcut is exemplified later.

ITS **2**, which represents a pinacol rearrangement, provides 13 classes of nodes in the partial partitioning step (three iterations). A spanning tree rooted to node 6 has a choice point at level 1 (nodes 7 and 10), since SETNO(7) = SETNO(10) = $\boxed{9}$) as shown in Figure 6. Two names are generated at the choice point and compared with each other. In this case, the two names are the same. The result is shown in Figure 7.

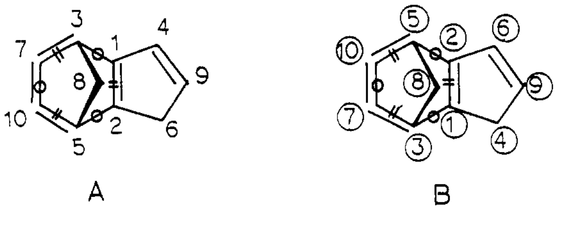
It is noted that there is another pair of equivalent nodes (nodes 8 and 9) in ITS **2**, i.e., SETNO(8) = SETNO(9) = $\boxed{13}$. However, this is no choice point, since the priority between them is decided automatically after the numbering of nodes 7 and 10 is established.

Presence of Two or More Pseudoequivalent Roots. Suppose that two or more nodes belong to the highest class (SETNO(i) = $\boxed{1}$). Then a spanning tree is constructed as above to be rooted to each of the highest nodes, and the best name is generated as a nominated name with respect to the spanning tree. Finally, the nominated names for the respective trees are compared with each other, and the final best name is selected as the canonical linear name.



3

Figure 8. ITS of the Diels–Alder addition (**3**). For the attached numbers, see the caption to Figure 6.



```
CANITS:
0153/010/003/001001001002002003003004005/005008006009007010/
101011111121/010100000001/100001001000/06060606060606060606/
0001010110/1110211012/1220211011/
```

Figure 9. Canonical numbering (B) as compared with a nominated numbering (A) and the canonical code of ITS **3** based on (B).

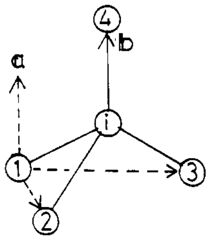


Figure 10. Calculation of parity around a chiral center.

Figures 8 and 9 exemplify this case. The nodes of ITS **3**, representing the Diels–Alder reaction, are divided into six classes after one iteration.

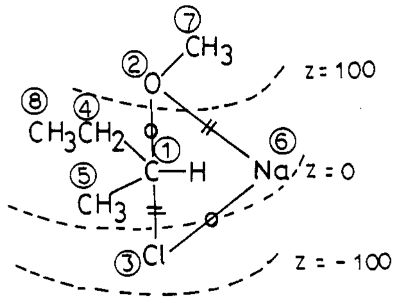
Nodes 1 and 6 belong to the highest class (SETNO(1) = SETNO(6) = $\boxed{1}$).[9] Hence, two nominated modes of numbering are obtained to be A and B. The numbering B is selected as the better one. This decision is accomplished in the PAR-BOND list, i.e., 10101112**1**111 for A and 10101111112**1** for B. The results are summarized in Figure 9.

Description of Stereochemical Changes. The stereochemistries of the starting and product stages are described in the STARTING and PRODUCT STEREO lists.[10] In the ITS approach,[1a] the starting stage is reproduced by PS operation (projection to the starting stage). This operation is the adoption of out- and par-bonds from the imaginary bonds. On the other hand, the product stage is obtained by PP operation (projection to the product stage), which corresponds to the adoption of in- and par-bonds from the imaginary bonds.

The stereo centers are denoted by z coordinates, which are given in an ITS connection table.[11] After the canonical numbering of the ITS, the vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ defined below are calculated at each stereo center. Suppose that the canonical numbers are given in ascending order, 1, 2, 3 and 4, as shown in Figure 10. The vector $\boldsymbol{a}$ is defined as the vector product $\overrightarrow{12} \times \overrightarrow{13}$. The vector $\boldsymbol{b}$ is defined as $\overrightarrow{i4}$, wherein node (i) is the stereo center. Then a parity value is defined at each of the stereo centers as follows

$$\begin{aligned} \text{parity} &= 1 \text{ for } \cos(\hat{\boldsymbol{a}\boldsymbol{b}}) > 0 \\ &= 0 \text{ for } \cos(\hat{\boldsymbol{a}\boldsymbol{b}}) = 0 \\ &= 2 \text{ for } \cos(\hat{\boldsymbol{a}\boldsymbol{b}}) < 0 \\ &= 3 \text{ not denoted} \end{aligned}$$

wherein $\hat{\boldsymbol{a}\boldsymbol{b}}$ represents the angle of the vector $\boldsymbol{a}$ with the vector $\boldsymbol{b}$.



4

CANITS:
0113/006/001/001001001001002002004/003006/00110110/10000001/
01001000/0608170606110606/00011011/10000000/20000000/

Figure 11. Canonical numbering and code of the Walden inversion reaction.

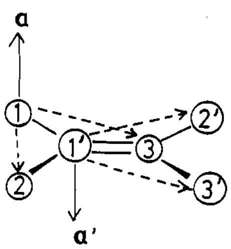


Figure 12. Calculation of parity for a double bond.

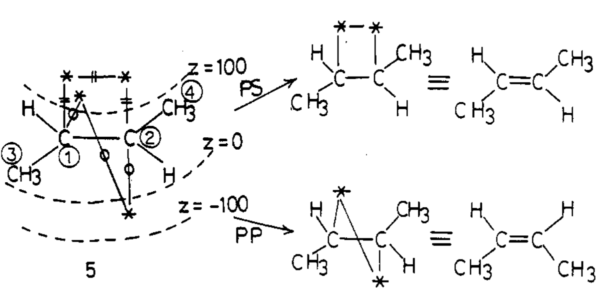
For example, the ITS **4**, representing the Walden inversion, provides the canonical name shown in Figure 11.[11] The starting and the product parities are 1 and 2, respectively, at the stereo center (node 1). This fact corresponds to the stereochemical change of this reaction.

The configurations around a double bond in the starting stage and the product stage are denoted also by the STARTING and PRODUCT STEREO lists. Suppose that one atom of a double bond is attached by three atoms that are canonically numbered in ascending order, 1, 2, and 3, and that the other atom of the double bond is attached by three atoms as 1′, 2′, and 3′ in ascending order (Figure 12). Then two vectors $\mathbf{a}$ and $\mathbf{a}'$ are defined as $\mathbf{a} = \overrightarrow{12} \times \overrightarrow{13}$ and $\mathbf{a}' = \overrightarrow{1'2'} \times \overrightarrow{1'3'}$, respectively. The parities of the double-bond atoms are defined as follows:

$$\begin{aligned} \text{parity} &= 1 \text{ for } \cos(\widehat{\mathbf{a}\mathbf{a}'}) > 0 \\ &= 0 \text{ for } \cos(\widehat{\mathbf{a}\mathbf{a}'}) = 0 \\ &= 2 \text{ for } \cos(\widehat{\mathbf{a}\mathbf{a}'}) < 0 \\ &= 3 \text{ not denoted} \end{aligned}$$

An example has been shown in the canonical name of ITS **3**. The STARTING STEREO list is 1110211012, though all the double bonds have cis configuration. Thus, difference between parities is not always ascribed to the difference between the configuration of the double bond. It is noted that the parities depend on the canonical numbering of the ITS, not on that of the starting molecules.

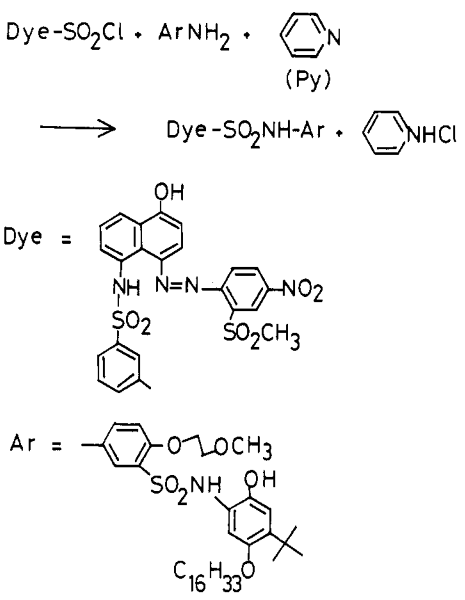
Cis–trans isomerization is treated by using a radical space that contains asterisks as shown in Figure 13.[12] The isomerization of *trans*- to *cis*-2-butene is represented by ITS **5**, in which asterisks provide information on the change of configuration. The calculation of parities around the double bonds are corrected by using the *z* coordinates of the asterisks. The canonical code of this reaction is computed and collected in Figure 11. Since the parity of the starting double bond is 2, the STARTING STEREO list results in 2200. The PRODUCT STEREO list is 1100, the value 1 of which is obtained by considering the *z* coordinates of the asterisks. The IN-BOND and OUT-BOND lists contain all zeros, which shows no change of the double bonds formally. However, the reaction centers are shown by the INTACT NODE list, which has a zero value for each of the reaction centers.



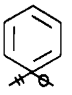
CANITS:
0060/004/000/001001002//211/000/000/06060606/0011/2200/1100/

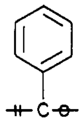
Figure 13. Canonical numbering and code of ITS **5**, which corresponds to cis–trans isomerization. The PS and PP operations show the change of the configuration.

Scheme I

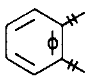


Treatment of Aromatic Rings. Aromatic bonds should be treated by distinguishing the following two cases in the present ITS approach: (a) The character of an aromatic ring is unchanged during a given reaction, e.g.





(b) An aromatic ring changes into a nonaromatic one and vice versa, e.g.

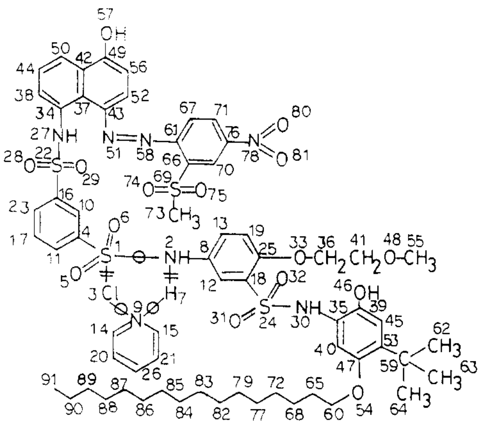


In the ITS approach, the former case (a) is characterized in the PAR-BOND list, in which both the formal single and double bonds of the aromatic ring are denoted by integer 4. The latter case (b) is treated in the exact sense that the ITS expresses.

The formation of a cyan dye releaser for instant color photography[13] (Scheme I) affords an example of treating aromatic rings in the ITS approach. The corresponding ITS **6** provides the canonical name as shown in Figure 14. The aromatic bonds (single or double) are characterized by integer 4 in the PAR-BOND list. In this process of canonization, the terminal shortcut described above is particularly effective. Iterations at the terminal groups such as $N(=O)_2$, $S(=O)_2$, and $C(CH_3)_3$ are omitted by this shortcut. Hence, the number of iterations is reduced from 1152 (=3!3!(2!)^5) to 12 (=3!2!) in this case.

Table VII. ITS and CANITS Integrating a Set of Natural Language Terms

ITS	CANITS	natural language terms
1	Figure 5	Claisen rearrangement (**r**) of 1-(allyloxy)cyclohexene (**s-1**) into 2-allylcyclohexanone (**s-2**)
2	Figure 7	pinacol rearrangement (**r**) of 1,1′-dihydroxy-1,1′-bicyclopentyl (**s-1**) into spiro[4.5]decan-2-one (**s-2**) with hydrochloric acid (**s-3**) as a catalyst
3	Figure 9	Diels–Alder-type dimerization (**r**) of cyclopentadiene (**s-1**) to afford dicyclopentadienyl (**s-2**)
4	Figure 11	nucleophilic substitution (**r-1**) of optically active 2-chlorobutane (**s-1**) with sodium methoxide (**s-2**) into 2-methoxybutane (**s-3**) having the inverse configuration (**r-2**)
5	Figure 13	isomerization (**r**) of *trans*- (**s-1**) to *cis*-2-butene (**s-2**)
6	Figure 14	condensation (**r**) between 5-[[[3-(chlorosulfonyl)phenyl]sulfonyl]amino]-4-[2-methylsulfonyl]-4-nitrophenylazo)-1-naphthol (**s-1**) and 2-[[[5-amino-2-(2-methoxyethoxy)phenyl]sulfonyl]amino]-5-*tert*-butyl-4-(hexadecyloxy)phenol (**s-2**) to form *N*-[3-[*N*-(4-*tert*-butyl-5-(hexadecyloxy)-2-hydroxyphenyl)sulfamoyl]-4-(2-methoxyethoxy)phenyl]-*N*′-[5-hydroxy-8-[2-(methylsulfonyl)-4-nitrophenylazo)-1-naphthyl]-1,3-benzenedisulfonamide (**s-3**) by using pyridine (**s-4**) as a basic reagent


57 OH
50 42 49
44 56
38 52
34 37 43
27NH N=N 61 67 71 76 N 78 O 80 O 81
28 22 O=S=O 29 51 58 66 70
16 10 74 O=S=O 75 69
23 73 CH3 13
17 O6 2 8 19 25 33 36 41 48 55
4 S 1 NH O-CH2CH2O-CH3
11 5 O 3 Cl H 7 12 18 O 32 46 OH
9 N 15 S-NH 35 39
14 31 O 24 30 45 62
20 21 40 O 53 C CH3
91 89 87 26 85 83 79 72 65 47 59 CH3 63
90 88 86 84 82 77 68 60 54 64 CH3

6

```
CANITS:
1089/091/008/00100100100100100200200300400400800800900901001
101201301401501601601801802002202202202402402402502703003303
403403503503603703703803903904004104204204304304504704804904
905105305405805905905906006106106506606606706806906906907007
2076077078078079082083084085086087088089090/0070090170230190
25021026044050047053052056071076/001220104444444444441414412
212211114444144441414414411412111111144114411224111221111111
11104444444/1000000100000000000000000000000000000000000000000
000000000000000000000000000000000000000000100000000/010001000
000000000000000000000000000000000000000000000000000000000000
00000000000000000000000000000/160717060808010607060606060606
060606060606160616060607080807080808060606060606060606060606
080608060607060608060608070606060606060606061606060606060808
06060706080806060606060606060606/000111010111111111111111111
111111111111111111111111111111111111111111111111111111111111
1111/0000000000000000000000000000000000000000000000000000000
000000000000000000000000000000000000/00000000000000000000000
000000000000000000000000000000000000000000000000000000000000
00000000
```

Figure 14. Canonical numbering and code of ITS **6** representing a sulfonamide formation.

CANITS AS THE UNITARY CODE OF AN INDIVIDUAL ORGANIC REACTION

Unique and Unambiguous Representation of an ITS That Unambiguously Represents an Individual Organic Reaction.[14] Organic chemists used to discuss an individual organic reaction by using a reaction diagram in which starting materials and products are combined with an arrow.[1j,15] We express the reaction with a combination of natural language terms (Table VII). For example, the reaction shown in Figure 2 is represented by the Claisen rearrangement (**r**) of 1-(allyloxy)cyclohexene (**s-1**) into 2-allylcyclohexanone (**s-2**). This representation consists of three terms, **r** being concerned with a reaction type and **s-1** and **s-2** with structures. The other representations collected in Table VII also consist of several natural language terms.

Most computer systems have succeeded the methodology that treats a reaction as a set of such terms of different categories but not per se. Hence, the corresponding codes for reaction types and for structures are used separately. The integration of such codes as belong to the different categories

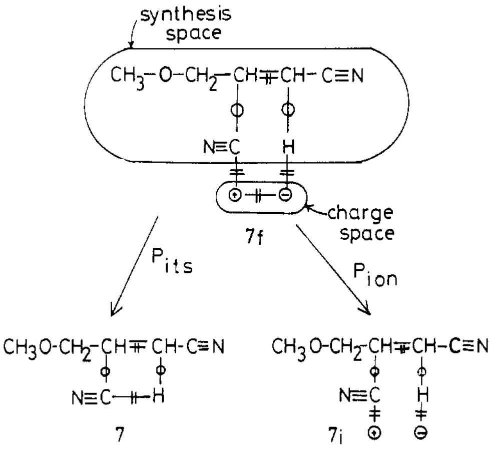
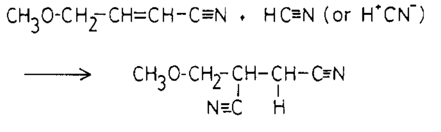


Figure 15. Full ITS for the addition of Scheme I.

Scheme II



would require a complex algorithm if we remain in the conventional methodology.

On the other hand, ITS's have all pieces of information on structural changes and represent individual organic reactions per se. For example, the PS operation (deletion of in-bonds) on ITS **6** affords the molecules in the starting stage, i.e., dye–SO_2Cl, Ar–NH_2, and pyridine. The PP operation (deletion of out-bonds) produces the products, i.e., dye–SO_2NH–Ar and pyridine hydrochloride. These items are sufficient to regenerate Scheme I. Moreover, the formation of an S–N bond and the cleavage of an S–Cl bond are easily characterized by examining in- and out-bonds of the ITS. These pieces of information are not contained explicitly in the conventional diagram (Scheme I).

The canonical codes (CANITS's) presented in this paper correspond uniquely and unambiguously to the ITS's. The present formulation thereby provides the unambiguous description of individual organic reactions. The CANITS integrates natural language terms or other conventional representations for describing individual organic reactions (Table VII). It should be noted that since an individual organic reaction may correspond to two or more ITS's, the selection of the best ITS is necessary to give unique and unambiguous description of an individual organic reaction (next section).

Covalent and Ionic Representations. We discuss here the ionic and covalent representations of a bond. We must overcome this duality in order to afford a unique and unambiguous name to an individual organic reaction. In the ITS approach, the ionic and covalent representations are integrated in the form of an ITS with charges (a full ITS).[1h] For ex-

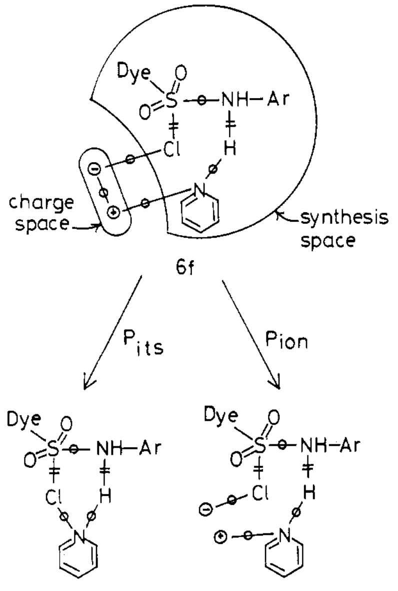


Figure 16. Full ITS for the sulfonamide formation of Scheme II.

ample, the addition of hydrogen cyanide to a double bond (Scheme II) is represented by the full ITS **7f** that possesses a synthesis space and a charge space (Figure 15). The projection to ITS (P_{ITS}) is the operation in which the charges and the interspace bonds of a full ITS are deleted and the intraspace bonds in the charge space are mapped onto the synthesis space. The usual ITS **7** thus obtained is based on the covalent representation of bonds. The present ITS approach adopts this covalent form for the purpose of canonical coding. However, since the ITS connection table contains the connectivities based on the full ITS, the corresponding ionic representation (**7i**) can be generated if necessary.[1h]

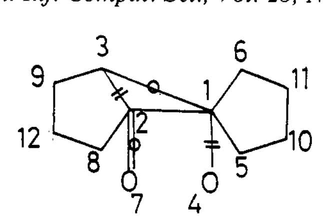
Figure 16 shows the full ITS **6f** corresponding to the ITS **6**. This indicates that the quinquevalent nitrogen in ITS **6** is the mapping from the chargen space of **6f**.[16]

The methodology to overcome the duality of bond character differentiates the ITS approach from Vladutz's method reported independently.[17] Vladutz's representation permits superimposed reaction graphs (covalent forms, e.g., the counterpart of **7**) as well as the corresponding completely legitimate alternative forms (ionic forms, e.g., the counterpart of **7i**). The permission of the two types of representations that are not integrated makes the canonical naming difficult. As a result, his system would encounter some difficulties such as multiple registration. Moreover, the ionic form produces H^+ and CN^- as independent species in his methodology. Since his method lacks the counterpart of the full ITS, such a combination as H^+CN^- requires an implicit presumption that stems from chemists' intelligence and is not inherent in a computer. His ionic form (e.g., $|^-CN$) contains the dual representation of a minus charge and an unshared electron pair. This succeeds to the conventional representation of formal charges but would be cumbersome to computer manipulation.

Representation of Organic Structure and That of ITS. It is worthwhile to mention manipulation of organic compounds by comparison with that of organic reactions. An *individual* organic compound is represented by the corresponding structure, which is in turn manipulated in terms of a canonical name. A compound type (e.g., a ketone, an alcohol, or an ether) is discussed on the basis of a substructure (e.g., C=O, OH, or –O–). The relationships between these concepts are represented in the following schemes:

(a) organic compound ↔ structure ↔ canonical code

(b) compound type ↔ substructure ↔ code



2r

```
0175/012/003/001001001001001002002003005006006/0020030090120
10011/10011111111011/01000100000000/00100000000100/060606050
606080606060606/000011011111/00000000000000/0000000000000/
```

Figure 17. Structure of the reduced ITS **2r** and its canonical code.

The relationships between the counterparts of the ITS approach are as follows:

(a′) organic reaction ↔ ITS ↔ canonical code

(b′) reaction type ↔ substructure of ITS ↔ canonical code

The correspondence of (a) to (a′) and that of (b) to (b′) provides broad prospects to the representation of organic reactions. Thus, concepts and techniques for compound manipulation would be applicable to the field of reaction manipulation with simple modification or with appropriate extension.

Comparison between the schemes reveals that the conventional methodology lacks the viewpoint concerned with the schemes a′ and b′, that the reaction types are simple "from–to" combinations of functional groups represented by the scheme b, and that the conventional methods represent an individual organic reaction on the basis of (a) and (b).

REDUCED IMAGINARY TRANSITION STRUCTURES

The CANITS of an ITS is convenient for the purpose of registration, since exact matching is necessary to check dual storing. On the other hand, a more general search would be desirable in the case of retrieval of organic reactions. For example, in searching pinacol rearrangements, we may expect to retrieve reactions that belong to the same category as ITS **2** but have other catalysts than the hydrochloric acid. For this purpose, we introduce reduced imaginary transition structures that are subgraphs of ITS's. The procedure to select the reduced ITS is as follows.

Step 1. Select a reaction kernel. The reaction kernel is a set of reaction centers that is one of the sets listed in descending priority as follows: carbon reaction centers, N–N, N–O, N–S, N–P, O–O, O–S, O–P, S–S, S–P, P–P, N reaction centers, O reaction centers, S reaction centers, and P reaction centers.

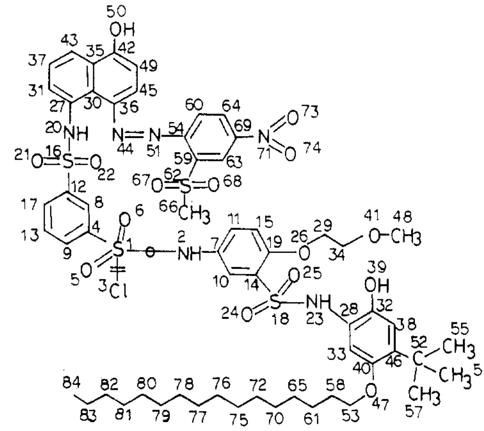
Step 2. Select a non-hydrogen node if each imaginary bond in the path through said node to the reaction kernel (adopted in step 1) contains at least one par-bond.

Step 3. Select a non-hydrogen reaction center that is adjacent to the nodes adopted in steps 1 and 2.

Step 4. If the reaction kernel consists of carbon reaction centers, select an intrastring hydrogen atom that is attached to the carbon reaction center.

The reduced ITS selected from an ITS is numbered canonically in the same method as described above. The reduced ITS **2r** of the pinacol rearrangement is selected from ITS **2** as shown in Figure 17. The canonical numbering and the corresponding linear code are computed. This example has a reaction kernel that has carbon reaction centers.

ITS **6** of the cyan dye formation provides a reduced ITS (**6r**) that contains an N–S reaction center as a reaction kernel. The results are collected in Figure 18. The reduced ITS **6r**



6r

```
CANONICAL NAME OF THE REDUCED ITS
0994/084/006/001001001001001002004004007007008009010011012O1
2014014016016016018018018019020023026027027028028029030030O3
1032032033034035035036036038040041042042044046047051052052O5
2053054054058059059060061062062062063065069070071071072075O7
6077078079080081082083/0130170150190370430400460450490640069/
0012214444444414141221221111444414444141441441141211111.11441
14411224111221111111111444444/100000000000000000000000000000
00000000000000000000000000000000000000000000000000000000000/
0100000000000000000000000000000000000000000000000000000000000
0000000000000000000000000000/160717060808060606060606060606
1606160607080807080808060606060606060606060608060806060706
0606060608070606060606060606060616060606060808060606070608806
060606060606060606/00011111111111111111111111111111111111111
11111111111111111111111111111111111111111111/00000000000000
0000000000000000000000000000000000000000000000000000000000000
00000000/0000000000000000000000000000000000000000000000000000
0000000000000000000000000000000000000/
```

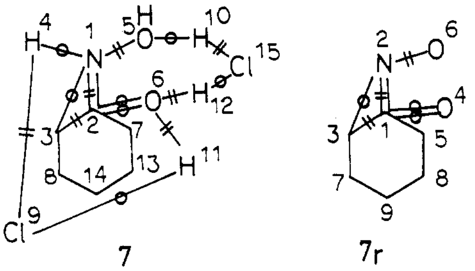
Figure 18. Structure of the reduced ITS **6r** and its canonical code.

has no representation of pyridine but contains the other essential pieces of information. The PS operation on **6r** produces the substrates dye–SO_2Cl and Ar–NH_2 and the PP operation provides the product dye–SO_2NH–Ar. The reduced ITS **6r** also contains information on S–Cl bond cleavage and on S–N bond formation. The canonical code collected in Figure 18 is a one-to-one representation that embraces all items derived from ITS **6r**. It is noted that there are cases in which a reduced ITS is the same as the original ITS. The ITS's **1** and **3** are such cases.

The Beckmann rearrangement represented by ITS **7** provides the corresponding reduced ITS **7r**. The first step of the process selects a reaction kernel that consists of carbon reaction centers (nodes 2 and 3). In the second step, nodes 7, 1, 3, 8, 13, and 14 are selected as members of the reduced ITS. The third step adopts nodes 5 and 6. The fourth step gives no additional nodes. As a result, the reduced ITS **7r** is obtained. The canonical codes of ITS **7** and reduced ITS **7r** are also found in Figure 19.

APPLICATION OF THE PRESENT METHOD TO THE CANONIZATION OF MOLECULAR GRAPHS

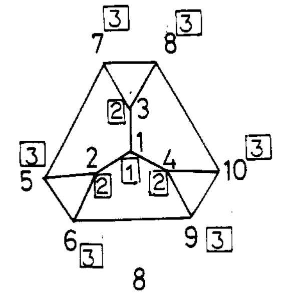
Usual structural formulas of compounds can be regarded as a kind of ITS that consists only of par-bonds. The present method of canonical coding is hence applicable to organic structures (molecular graphs). Among the four kinds of extended connectivities, the two extended connectivities EC2 and EC4 are effective to the partial partitioning process of molecular graphs. Since EC4 is the number of second neighbors, the present method is more discriminating than the original Morgan procedure is. For example, our method divided the vertices of a regular graph (**8**) into three classes, but the Morgan procedure gives one class. As a result, our method requires only 48 comparisons in this case.[18] Since the sorting method of the present approach uses as a restrictive condition the order of the last sorting in each iteration (i.e., the top three-character section (999 – SETNO(i) of EC(i)), the oscillation in the partitioning process is prevented.[19]



7 7r

```
*******************************************************
* CANONICAL NAME FOR IMAGINARY TRANSITION STRUCTURE *
*******************************************************
[TEST 16] BECKMANN REARR: CYCLOHEXANONE OXIME, FULL
NUMBER OF ITERATIONS=           1
NODE REORDERED:  2  7 13 14  8  3  1  4  6  5 12 11 15 10  9
         SETNO:  2 12 14 15 13  4  1  5  3  6  8  7 11 10  9
CANITS:
0217/015/004/001001001001002002003004005006006007008010/0020
0300901101201501301 4/100001100001100001/011020001000000110/1
001000101100011000/070606010808060617010101060617/00000011000
0110/22000000000000/00000000000000/
 *********************************
 * CANONICAL NAME FOR REDUCED ITS *
 *********************************
NUMBER OF ITERATIONS=           1
SELECTED NODE T T T T T T T F T T F F F F F
NODE REORDERED:  1  5  8  9  7  3  2  0  4  6  0  0  0  0  0
         SETNO:  1  6  8  9  7  3  2  0  4  5  0  0  0  0  0
0133/009/002/001001001001002003005007/002003008009/100101110
1/002000010/110010000/06070608060806 06/00010111/2200000
00/000000000/
```

Figure 19. ITS **7** and the corresponding reduced ITS **7r**. The numbers attached to the nodes are the canonical ones. The canonical codes are found below.



8

```
0126/010/006/001001001002002003003004004/005006005007006009O
0700600801000 9010/11111111111111/0606060606060606060606/33333
33333/
```

Figure 20. Canonical numbering and coding of a regular graph (**8**). The present procedure divides the 10 nodes into three classes as assigned by the numbers in the squares.

IMPLEMENTATION AND RESULTS

The steps described above are programed in FORTRAN 77 and implemented on VAX 11/750 (Digital Equipment Co.). An example of the output form is shown in Figure 5. The other results have been already shown in Figures 5, 7, 9, and 12–15. The canonical name for a compound (e.g., Figure 20) omits the lists g, h, j, and l.

CONCLUSION

The canonical numbering and coding of an imaginary transition structure (ITS) are described for the purpose of the unambiguous description of individual organic reactions. The concept of reduced ITS's is proposed.

REFERENCES AND NOTES

(1) (a) Fujita, S. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 205. See also *Chem. Eng. News.* **1986** (Sept 29), 75. (b) Fujita, S. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 212. (c) Fujita, S. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 224. (d) Fujita, S. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 231. (e) Fujita, S. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 238. (f) Fujita, S. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 99. (g) Fujita, S. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 104. (h) Fujita, S. *J. Chem. Inf. Comput. Sci.*

1987, *27*, 111. (i) Fujita, S. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 115. (j) Fujita, S. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 120. (k) Fujita, S. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 1. (l) Fujita, S. *J. Chem. Soc., Perkin Trans. 2* **1988**, 597.

(2) (a) Gluck, D. J. *J. Chem. Doc.* **1965**, *5*, 43. (b) Shelley, C. A.; Munk, M. E. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 247. (c) Herndon, W. C.; Leonard, J. E. *Inorg. Chem.* **1983**, *22*, 544. (d) Schubert, W.; Ugi, I. *J. Am. Chem. Soc.* **1978**, *100*, 37. (e) Jochum, C.; Gasteiger, J. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 49. (f) Uchino, M. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 201. (g) Balaban, A. T.; Mekenyan, O.; Bonchev, D. *J. Comput. Chem.* **1985**, *6*, 538.

(3) (a) Morgan, H. L. *J. Chem. Doc.* **1965**, *5*, 107. (b) Petrarca, A. E.; Lynch, M. F.; Rush, J. E. *J. Chem. Doc.* **1967**, *7*, 154. (c) Wipke, W. T.; Dyott, T. M. *J. Am. Chem. Soc.* **1974**, *96*, 4834.

(4) For the terms "intrastring" and "extrastring", see ref 1d. For a glossary of the ITS approach, see ref 1k.

(5) It should be noted that all of the newly defined extended connectivities (EC1–EC4) are invariant on an operation in which in-bonds and out-bonds are interchanged with each other. For further discussions on this point, see the subsequent paper.

(6) The number 999 is selected as the maximum number of nodes treated in the present method.

(7) The sorting in terms of EC(i) corresponds to a multiple sorting that is accomplished successively by using EC1(i), EC2(i), EC3(i), and EC4(i) in this order of priority.

(8) There is no case that NTEC is smaller than NEC. This fact stems from the restrictive condition that each iteration always refers to the SETNO of the last iteration. This is a different point of methodology from the original Morgan procedure.[3a]

(9) It should be noted that node 1 and node 2 are not equivalent but pseudoequivalent.

(10) For the codification of stereochemistry of organic compounds, see ref 3b and 3c.

(11) For the full representation, a charge space is attached to the synthesis space of the ITS. See ref 1h. See also Figures 15 and 16.

(12) A radical space is attached to the synthesis space of an ITS for description of radical character of a reaction. This is analogous to a charge space defined previously. See ref 1h.

(13) (a) Fujita, S. *J. Org. Chem.* **1983**, *48*, 177. (b) Fujita, S. *Yuki Gosei Kagaku Kyokaishi* **1982**, *40*, 307.

(14) For the terms "unique" and "unambiguous", see: Davis, C. H.; Rush, J. E. *Information Retrieval and Documentation in Chemistry*; Greenwood: Westport, CT, 1974; pp 145–152.

(15) Fujita, S. *Yuki Gosei Kagaku Kyokaishi* **1986**, *44*, 354.

(16) See also ref 1b.

(17) Vladutz, G. In *Modern Approaches to Chemical Reaction Searching*; Willet, P., Ed.; Gower, Aldershot, U.K., 1986; p 202.

(18) The number of iterations increases dramatically in the case of a highly symmetrical structure. The same phenomena were reported for the Morgan procedure.[20] Fortunately, most ITS's have low symmetry because of the presence of out- and in-bonds. Hence, the CANITS procedure of this work would be effective for most ITS's.

(19) For the oscillation phenomena, see ref 2g.

(20) O'Korn, L. J. *ACS Symp. Ser.* **1977**, *44*, 122.

Canonical Numbering and Coding of Reaction Center Graphs and Reduced Reaction Center Graphs Abstracted from Imaginary Transition Structures. A Novel Approach to the Linear Coding of Reaction Types

SHINSAKU FUJITA

Research Laboratories, Ashigara, Fuji Photo Film Co., Ltd., Minami-Ashigara, Kanagawa, Japan 250-01

Received October 9, 1987

A reaction center (RC) graph is the subgraph of an imaginary transition structure. Procedures for abstracting the RC graph and for canonical coding are developed in order to give an unambiguous description of the reaction type. The concept of a reduced RC graph is also introduced and the reduced graph canonized in the same manner.

The systematic characterization of reaction types is important in the construction of an effective computer system for manipulating organic reactions. This subject can be divided into two aspects from the practical point of view: (1) the abstraction of information on the reaction types and (2) the (canonical) coding of any pieces of the information. The first aspect has been formulated as abstraction of various subgraphs from an imaginary transition structure (ITS).[1-12] We have introduced reaction center (RC) graphs of various levels that have all reaction centers of an imaginary transition structure (ITS) and information on various levels of neighboring atoms.[2] These RC graphs correspond to reaction types.

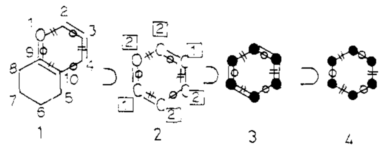
However, the second aspect is open for further discussion. A number of coding systems of reaction types have been reported for this purpose. Hendrickson's method is based on the description of the change of substitution at a carbon reaction center, where an oxidative substitution, for example, is represented by a code ZH.[13] Brandt et al. reported a coding method based on Ugi's reaction matrices.[14] Although Vladutz proposed superimposed reaction skeleton graphs for the representation of reaction types, the linear coding of these graphs has not been reported.[15] Roberts reported the coding system of organic reactions based on concerted process (CP) skeletons.[16a] Several groups proposed their own coding systems based on reaction diagrams.[17,18] All of these systems have paid little attention to multistring reactions,[19] though there are many name reactions classified as multistring.[6,7]

We discussed the coding of the RC graph of level 1 in a previous paper.[2] However, this method of coding is applicable only to cyclic or linear RC graphs having one reaction string. Hence, a novel method is necessary to be able to give a canonical name even to the RC graph that contains two or more reaction strings.[6,7]

In the preceding paper,[20] we have discussed the canonical names of ITS's that afford the unambiguous description of *individual organic reactions*. As a continuation of the work, this paper describes a novel method giving the unambiguous description of *reaction types*. This is based on the canonical numbering and coding of the RC graphs of level 1. This paper also deals with the abstraction and the canonical coding of a reduced RC graph.

ABSTRACTION AND CANONICAL CODING OF AN RC GRAPH

An RC graph of level 1 can be abstracted from an ITS by collecting the nodes to which out- and/or in-bonds are incident.[2] For example, ITS **1**, which represents the Claisen



1 2 3 4

0095-2338/88/1628-0137$01.50/0 © 1988 American Chemical Society