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.

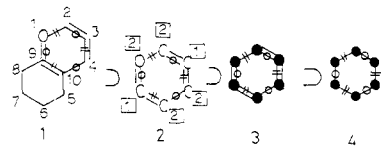
However, the second aspect is open for further discussion. A number of coding systems of reaction types have been reported for this purpose. Hendrickson's method is based on the description of the change of substitution at a carbon reaction center, where an oxidative substitution, for example, is represented 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.

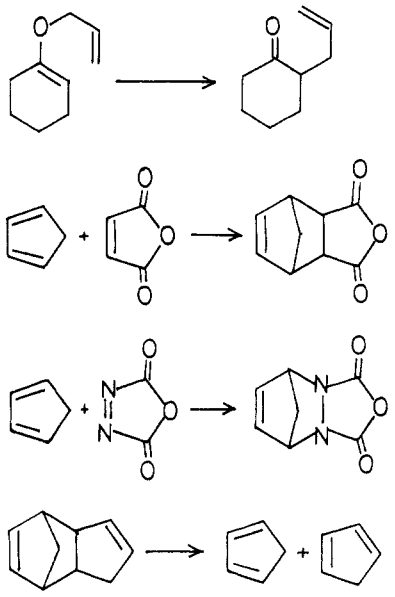
## ABSTRACTION AND CANONICAL CODING OF AN RC GRAPH

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




0095-2338/88/1628-0137$01.50/0 

**Scheme I**



rearrangement of 1-(allyloxy)cyclohexene forming 2-allylcyclohexanone (Scheme I), provides an RC graph of level 1 (**2**). The subgraph **2** corresponds to the generic name of the corresponding reaction type, i.e., the Claisen rearrangement. If the node values of **2** are omitted, the resulting reaction graph (RC graph of level 0) **3** affords a more generic name of the reaction type, i.e., a 3,3-sigmatropic rearrangement. The corresponding basic reaction graph **4** is obtained by omitting par-bonds to indicate a net pattern of electron transfer.

These graphic representations are thus useful to indicate the corresponding reaction types. The next problem is the development of their computer-readable linear codes. RC graph **2** can be coded C(2-1)C(1+1)C(1-1)O(1+1)C(2-1)-C(0+1)C by the method that we reported previously.[2] However, this method is limited to the coding of one-string reactions.[19] Here, we present a novel method of coding that is applicable to two-string or multistring reactions as well as to one-string reactions.

The canonical numbering and coding of the RC graph is based on the method described in the preceding paper.[20]

(1) Partial partitioning of the nodes of an RC graph to (pseudo)equivalent classes is achieved by the iterative calculation of four kinds of extended connectivities: EC1($i$), the number of adjacent reaction center atoms linked with an in- or out-bond to the current node ($i$); EC2($i$), the number of adjacent reaction center atoms linked to the current node ($i$) with any kind of bond; EC3($i$), the number of adjacent reaction center atoms linked with a par-bond to the current node ($i$); and EC4($i$), the number of the second-neighbor reaction center atoms of the current node ($i$).

In each iteration, the nodes $i$ (for all $i$) are divided into (pseudo)equivalent classes by comparing EC($i$) with each other, wherein EC($i$) = |A|EC1|EC2|EC3|EC4| and $A$ = 999 - SETNO($i$). The divided classes are numbered sequentially, and the sequential number assigned to node $i$ is stored in an array SETNO($i$).

(2) Canonical numbering is based on a spanning tree rooted to the node of the highest class.

(3) If two or more spanning trees are possible due to the (pseudo)equivalent nodes, nominated names (codes) based on the respective trees are compared with each other, and then the best (lexicographically smallest) name is selected from them.

The canonical name of an RC graph contains the following lists:[21] (a) the length of the canonical name; (b) the number of nodes considered; (c) the number of rings contained in the RC graph; (d) FROM list; (e) RING-CLOSURE list; (f) PAR-BOND list; (g) IN-BOND list; (h) OUT-BOND list; (i) ATOM list. The detailed description of these processes appeared in the preceding paper.[20] It should be noted that a spanning tree of a graph is an acyclic subgraph of the original graph that contains all of the nodes but not necessarily all of the bonds and that ring-closure bonds are the ones which are not contained in the spanning tree.

**Table I.** Initial Values of Extended Connectivities (EC1-EC4) of Each Node of the RC Graph **2**[a]

| node | EC1 | EC2 | EC3 | EC4 |
|---|---|---|---|---|
| 1 | 2 | 2 | 1 | 2 |
| 2 | 2 | 2 | 1 | 2 |
| 3 | 2 | 2 | 2 | 2 |
| 4 | 2 | 2 | 1 | 2 |
| 9 | 2 | 2 | 2 | 2 |
| 10 | 2 | 2 | 1 | 2 |

[a] Iteration = 0; number of classes = 2.

**Table II.** First Trial Values of Extended Connectivities of Each Node of the RC Graph **2**[a]

| node | EC1 | EC2 | EC3 | EC4 |
|---|---|---|---|---|
| 1 | 4 | 4 | 3 | 4 |
| 2 | 4 | 4 | 3 | 4 |
| 3 | 4 | 4 | 2 | 4 |
| 4 | 4 | 4 | 3 | 4 |
| 9 | 4 | 4 | 2 | 4 |
| 10 | 4 | 4 | 3 | 4 |

[a] The number of divided classes is equal to that of the initial one.

**Table III.** Final Values of EC Based on the First Iteration and the Resulting Partitioning into Two Classes of the Nodes[a]

| | | EC | | | | |
|---|---|---|---|---|---|---|
| node | class | 999-SETNO | EC1 | EC2 | EC3 | EC4 |
| 1 | 2 | 997 | 4 | 4 | 3 | 4 |
| 2 | 2 | 997 | 4 | 4 | 3 | 4 |
| 3 | 1 | 998 | 4 | 4 | 2 | 4 |
| 4 | 2 | 997 | 4 | 4 | 3 | 4 |
| 9 | 1 | 998 | 4 | 4 | 2 | 4 |
| 10 | 2 | 997 | 4 | 4 | 3 | 4 |

[a] Number of classes = 2.

In this paper, the procedure is exemplified by using ITS **1**, the nodes of which are numbered initially as above. The RC graph abstracted contains nodes 1, 2, 3, 4, 9, and 10. Tables I and II show the initial values of EC1-EC4 and the trial values of the first iteration, respectively. The classes divided are the same during this iteration. Table III collects EC($i$) values for each node $i$ and SETNO($i$) indicating the class assigned to the node $i$. Each number surrounded by a square indicates the class (i.e., SETNO($i$)) assigned to the respective node $i$ of the RC graph **2**. The number of classes is two.

There are two nodes (3 and 9) that belong to the highest class (SETNO(3) = SETNO(9) = 1). Since two spanning trees are constructed for each of the highest nodes, a total of four trees (A-D) is obtained as shown in Figure 1.

To begin, trees A and B, both of which are rooted to node 3, are compared. The nominated codes based on trees A and B are obtained by the algorithm described above:

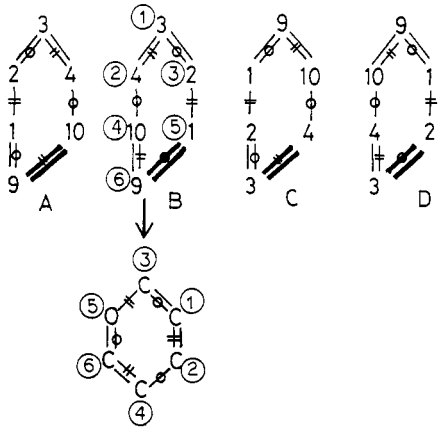
0070/006/001/001001002003004/

110011/100110/011001/060606080606/ for tree A

0070/006/001/001001002003004/

110011/011001/100110/060606060806/ for tree B

Tree B has priority over tree A, since the IN-BOND list of tree A is **100110** and that of tree B is **011001**. This decision is made by adopting the lexicographically smaller code.




0070/006/001/001001002003004/005006/110011/011001/100110/060
606060806/

**Figure 1.** Spanning trees of RC graph **2**, the canonical numbering based on tree B, and the corresponding canonical code. Each bold-faced bond represents a ring-closure bond, which is considered after the formation of the tree. The circled numbers represent the canonical numbering.

Similarly, the trees C and D rooted to node 9 afford the nominated codes as follows:

0070/006/001/001001002003004/

110011/100110/011001/060806060606/ for tree C

0070/006/001/001001002003004/

110011/011001/100110/060608060606/ for tree D

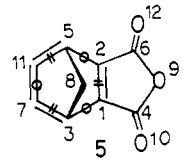
The code for tree D is better (i.e., lexicographically smaller) than that for tree C. The preference of tree D to tree C is decided also at the IN-BOND list.

Finally, comparison between the nominated codes derived from trees B and D selects the name based on tree B as the best name. The selection is made at the last ATOM list, since the ATOM list of tree B (060606060806) is smaller than that of tree D (060608060606).

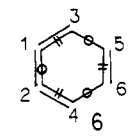
Figure 1 contains the canonical numbering (circled) and the canonical code of the RC graph **2**. The top three sections of the canonical name (Figure 1) are header sections, 0070 being the length of the name, 006 being the number of nodes, and 001 being the number of rings. The subsequent sections of the code are concerned with the connectivities of the RC graph, in which the specification of nodes is based on the canonical numbering obtained now. Thus, the fourth section (FROM list) indicates the parent nodes of node ② to ⑥. For example, the first three-digit value, 001, shows that the node ② is linked to the parent node ①. A pair of three-digit values in the fifth section (RING-CLOSURE list) indicates a ring-closure bond. Thus, the number 005006 corresponds to the presence of a ring-closure bond between nodes ⑤ and ⑥. The sixth to eighth sections designate the modes of imaginary bonds, respectively, in the order of the connectivities described in the FROM and RING-CLOSURE lists. The bond between nodes ① and ② (C±C) is, for example, represented by 1 (the first value of PAR-BOND list), 0 (that of IN-BOND list), and 1 (that of OUT-BOND list). The last section is the ATOM list, which contains atomic numbers of the nodes in the order of the canonical numbering.

## CANONICAL CODES OF HEXAGONAL REACTION GRAPHS

The Diels–Alder reaction of cyclopentadiene with maleic anhydride (Scheme I) is represented by ITS **5**, the canonical numbering and the canonical code of which are collected in Figure 2. From this ITS, the corresponding RC graph **6** is

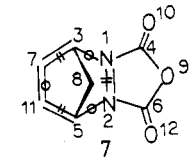


CANITS:
0175/012/003/001001001002002003003004004005006/005008006009
07011/101011112121111/010100000000001/100001001010000/060606060
606060608080608/000101011101/111020100020/122020100010/

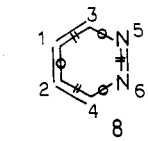


0070/006/001/001001002003004/005006/111001/100110/011001/060
606060606/

**Figure 2.** Canonical numbering and coding of ITS **5** and the corresponding RC graph **6**.

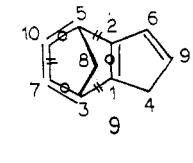


CANITS:
0175/012/003/001001001002002003003004004005006/005008006009
07011/101011112121111/010100000000001/100001001010000/070706060
606060608080608/000101011101/001020100020/001010100010/

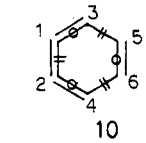


0070/006/001/001001002003004/005006/111001/100110/011001/060
606060707/

**Figure 3.** Canonical numbering and coding of ITS **7** and the corresponding RC graph **8**.



CANITS:
0153/010/003/001001001002002003003004005/005008006009007010/
101011111121/100001001000/010100000001/060606060606060606060/
000101011O/122021101l/111021101 2/



0070/006/001/001001002003004/005006/111001/011001/100110/060
606060606/

**Figure 4.** Canonical numbering and coding of ITS **9** and the corresponding RC graph **10**.

abstracted and given a canonical name as shown also in Figure 2. This canonical code is a computer representation that corresponds to the Diels–Alder addition in a general fashion. Thus, an ITS and its code represent an *individual reaction* that specifies all of the structural changes of substrates, products, catalysts, and their changes. On the other hand, the RC graph and its code express a *reaction type* that indicates generic pieces of information on the reaction.

The results of another Diels–Alder reaction (Scheme I and ITS **7**) are shown in Figure 3. The canonical code of the corresponding RC graph **8** is the same as that of **6** with respect to the lists a–h. This result stems from the fact that both reactions belong to the same category. The difference between them is perceived by subsequent comparison of the ATOM lists. It should be emphasized that the lists a–h represent a

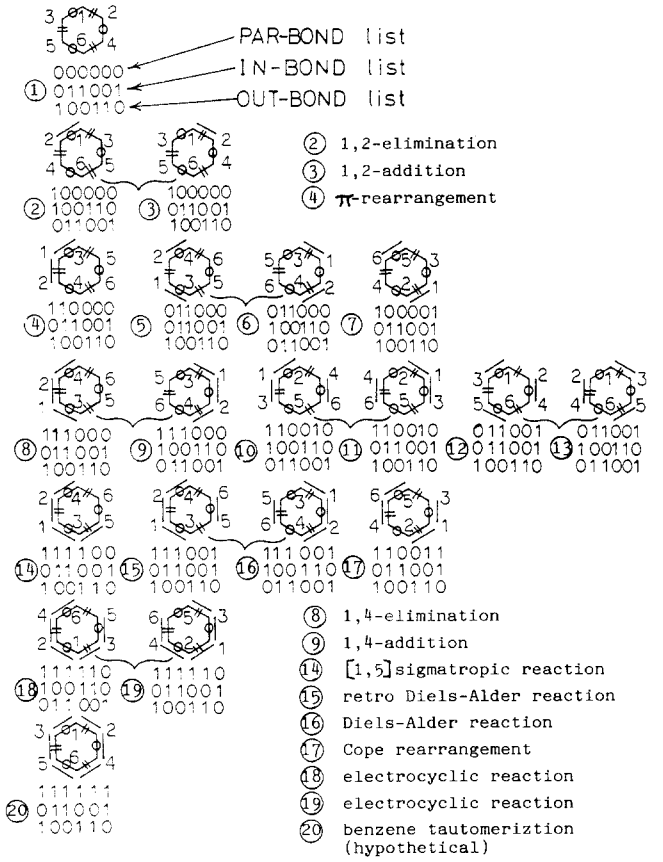



**Figure 5.** Canonical numbering and coding of several reaction graphs of hexagonal class. The PAR-BOND, IN-BOND, and OUT-BOND lists are collected. A couple linked with a brace is a reaction pair. The names of several reactions are also given.

reaction graph (RC graph of level 0) and full lists a–i express an RC graph of level 1.

The retro-Diels–Alder reaction (Scheme I and ITS **9**) affords the corresponding RC graph **10**. The canonical code corresponding to **10** has common lists (a–f) with that of **6** (Figure 4). The subsequent IN-BOND and OUT-BOND lists of **10** are aligned in an interchanged order as compared with those of **6**. This is in accord with the fact that they are reverse reactions to one another.

Figure 5 collects the canonical numbering and codes of some hexagonal reaction graphs as well as some of the corresponding natural language terms.[22] This figure summarizes only the PAR-BOND, IN-BOND, and OUT-BOND lists, since the other lists are common in all the graphs. For example, the full name of the first reaction graph is 0070/006/001/001001002003004/005006/*000000/011001/100110*/060606060606/, in which the italicized string indicates the PAR-BOND, IN-BOND, and OUT-BOND lists collected in Figure 5.[23] A pair linked with a brace is called a reaction pair,[2] the PAR-BOND lists of which are the same. The OUT-BOND list of one of the pair is the same as the IN-BOND list of the other and vice versa. This is the case in general for reaction–reverse reaction pairs.

The reaction pair is characterized by a transformation to a reverse reaction (TRR). This is defined as an operation in which all in-bonds and out-bonds of a reaction graph or of an RC graph are interchanged with each other.[2] The TRR operation affords an interconversion in a forward–backward reaction pair. It should be noted that a par-bond skeleton is invariant on the TRR operation.[2]

The extended connectivities (EC1–EC4) are all invariant on the TRR operation. As a result, the classes divided in the partial partitioning process are invariant on the TRR operation. This means that the selected roots of a reaction and of the corresponding reverse reaction have the same topological environments. Hence, the lists a–f of a reaction are the same as those of the corresponding reverse reaction, since these lists are concerned only with topological and par-bond characters of the graphs. The interchange of the IN-BOND list (g) and the OUT-BOND list (h) of a given RC graph affords the corresponding code of the reverse RC graph.

**Scheme II**

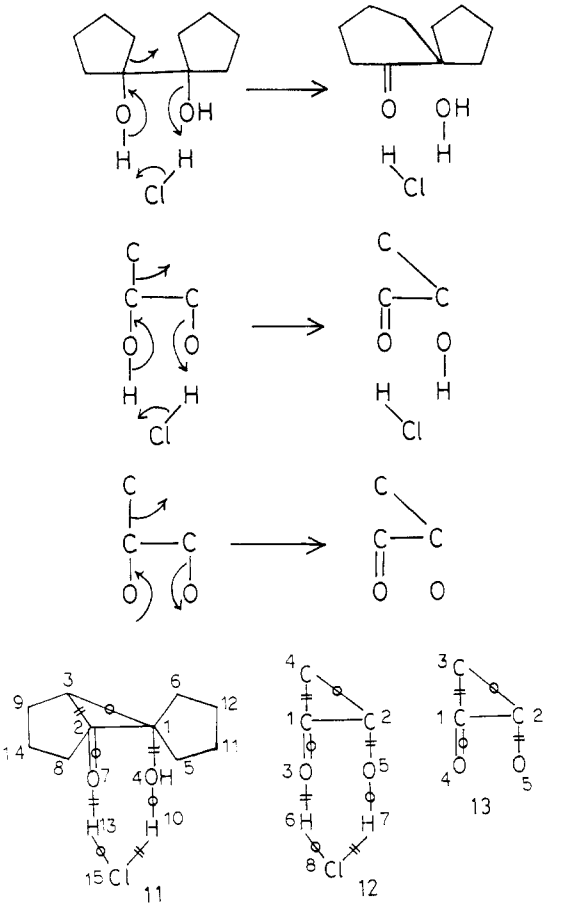



**Figure 6.** Canonical numbering and coding of RC graph **12** and reduced RC graph **13** derived from ITS **11**.

## REDUCED RC GRAPHS AND THEIR CANONICAL CODES

An RC graph of level 1 contains information on a catalyst-participating reaction. In the case of the pinacol rearrangement (Scheme II) represented by ITS **11**, for example, the RC graph **12** holds hydrochloric acid as a catalyst. The catalyst may be other protonic acids such as sulfuric acid, phosphoric acid, and so on. If one attempts to search a pinacol rearrangement without respect to information on the catalysts, the above RC graph is too specific to accomplish the broader searching. As we have considered reduced ITS's abstracted from ITS's,[20] we propose reduced RC graphs that collect

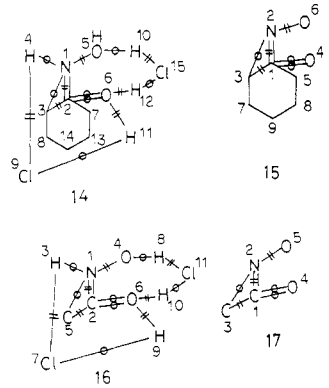



**Figure 7.** ITS **14**, reduced ITS **15**, RC graph **16**, and reduced RC graph **17** for the Beckmann rearrangement.

essential parts from the corresponding RC graphs (Figure 6).

The process to abstract reduced RC graphs is as follow.

(1) Select a reaction kernel. The kernel is a set of reaction centers selected from the combinations (in descending priority) 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.

(2) Select a non-hydrogen reaction center ($i$) that is adjacent to the nodes adopted in process 1. If the bond between them contains one or more par-bonds, set EAFLAG($i$) = 1.

(3) From the remaining N, O, S, or P nodes, select a node if the node is attached with an in- or out-bond to the node of EAFLAG($i$) = 1.

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

The reduced RC graphs abstracted in this procedure are given the corresponding canonical codes in the same method as above. For example, the reduced RC graph **13** represents essential features of the pinacol rearrangement, the canonical code of which is shown in Figure 6.[25]

It should be emphasized that the graphs **12** and **13** are the subgraphs of the ITS **11**. The corresponding relationship may be abstracted from such conventional representations as Scheme II,[16] since the in-bonds and out-bonds of the ITS approach relate to "electron pushing arrows" in Scheme II. The conventional process of perception, however, depends mainly on chemist's intuition and would require complicated algorithms.[26]

The Beckmann rearrangement of cyclohexanone oxime is represented by ITS **14**, from which the reduced ITS **20**,[20] the RC graph **16**, and the reduced RC graph **17** are abstracted and coded canonically (Figure 7). This reaction is a two-string reaction,[27] the canonical coding of which has never been reported. The oxygen atom attached to a nitrogen atom can be selected as a member of the reduced RC graph through process 3 described above. The canonical numbering and coding of these graphs are collected in Figure 8.

## IMPLEMENTATION AND RESULTS

The steps described above are programmed in FORTRAN 77 and implemented on VAX 11/750 (Digital Equipment Co.). The connectivity of an ITS is input in the form of the corresponding ITS connection table that contains the node values, the $x$, $y$, and $z$ coordinates of each node, complex bond numbers, etc.[28] The initial numbering of the nodes is not canonized but given arbitrarily due to the input procedure. The program analyzes the ITS connection table, abstracts the reduced ITS,[20] the RC graph and the reduced RC graph, and then provides the canonical codes to them. An example of an output form is found in Figure 8. The "number of iterations" shows the number of trees examined. The line "node

```
**************************************************
* 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:
0217/015/004/001001001001002002003004005006006007006010/0020
0300901101201501301471000011000011000011/0110200010000001110/1
0010001011001100/0706060106080606170101010606170000001100
0110/220000000000000/000000000000000/
 ***********************************
 * CANONICAL NAME FOR REDUCED ITS *
 ***********************************
NUMBER OF ITERATIONS=          1
SELECTED NODE    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
1/002000010/110010000/060706080608060606/000010111/2200000
00/000000000/
 ******************************************
 * CANONICAL NAME FOR REACTION CENTER GRAPH *
 ******************************************
NUMBER OF ITERATIONS=          1
SELECTED NODE    T  F  F  F  F  T  T  T  T  T  T  T  T  T
NODE REORDERED:  2  0  0  0  0  5  1  3  6  4 10  9 11  8  7
      SETNO:  2  0  0  0  0  6  1  4  3  5  8  7 11 10  9
0128/011/003/001001001001002003004006006008/002005007009010
11/100000000000/010120100011/101001011110 0/0706010806081701
1010117/
 ************************************
 * CANONICAL NAME FOR REDUCED RC GRAPH *
 ************************************
NUMBER OF ITERATIONS=          2
SELECTED NODE    T  F  F  F  F  T  T  F  T  T  F  F  F  F  F
NODE REORDERED:  1  0  0  0  0  3  2  0  4  5  0  0  0  0  0
      SETNO:  1  0  0  0  0  2  1  0  3  3  0  0  0  0  0
0062/005/001/001001001002/002003/10000/00201/11010/060706080
8/
```

**Figure 8.** Canonical numbering and coding of graphs **14–17**.

reordered" contains the canonical numbers of the nodes that are aligned in the order of the initial numbering. A zero value indicates that the corresponding node is not a member of the subgraph considered. The alignment is "setno" indicates the class assigned to the nodes. The line "selected node" contains $T$ or $F$ in the order of the initial numbering, where $T$ is a member of the subgraph considered and $F$ is not. The last part of each section affords the respective canonical name.

## CONCLUSION

A procedure to abstract a reaction center graph from an imaginary transition structure is described, and a method to provide the canonical coding of the RC graph is proposed. The concept of a reduced RC graph is introduced, and the canonical coding is established to indicate essential features of the reaction.

## REFERENCES AND NOTES

(1) Fujita, S. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 205.
(2) Fujita, S. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 212.
(3) Fujita, S. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 224.
(4) Fujita, S. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 231.
(5) Fujita, S. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 238.
(6) Fujita, S. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 99.
(7) Fujita, S. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 104.
(8) Fujita, S. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 111.
(9) Fujita, S. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 115.
(10) Fujita, S. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 120.
(11) Fujita, S. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 1.
(12) Fujita, S. *J. Chem. Soc., Perkin Trans. 2* **1988**, 597. See also *Chem. Eng. News* **1986** (Sept 29), 75.
(13) Hendrickson, J. B. *J. Am. Chem. Soc.* **1971**, *93*, 6847.
(14) Brandt, J.; Bauer, J.; Frank, R. M.; von Scholey, A. *Chem. Scr.* **1981**, *18*, 53.
(15) Vladutz, G. In *Modern Approaches to Chemical Reaction Searching*; Willet, P., Ed.; Gower, Aldershot, U.K., 1986; p 202.
(16) (a) Roberts, D. C. *J. Org. Chem.* **1978**, *43*, 1473. (b) Zefirov, N. S.; Tratch, S. S. *Chem. Scr.* **1980**, *15*, 4.
(17) Littler, J. S. *J. Org. Chem.* **1979**, *44*, 4657.
(18) Arens, J. F. *Recl. Trav. Chim. Pays-Bas* **1979**, *98*, 155.
(19) For the terms, "one-string", "two-string", and "multistring", see ref 1. See also Appendix (Glossary) of ref 11 for the terminology of the ITS approach.
(20) Fujita, S. *J. Chem. Inf. Comput. Sci.* (preceding paper in this issue).
(21) For a detailed explanation of the lists, see the preceding paper.[20] The INTACT NODE list is omitted since it is obvious. The STARTING and PRODUCT STEREO lists are omitted since an RC graph of level

1 has no information on stereochemistry.
(22) The full list of hexagonal reaction graphs was summarized in ref 2.
(23) The node values are presumed to be 06 (carbon atoms).
(24) The enumeration of hexagonal reaction graphs was reported in ref 2.
(25) For the initial numbering of ITS **11** and the canonical coding, see the preceding paper.[20]
(26) Fujita, S. *Yuki Gosei Kagaku Kyokaishi* **1986**, *44*, 354.
(27) ITS **14** has two reaction strings, i.e., 1-5+10-15-6+2-3+1 and 1-2+6-11+9-4+1. See ref 6.
(28) For the connection table of ITS, see ref 1.

# Ring ID Numbers

MILAN RANDIĆ

Department of Mathematics and Computer Science, Drake University, Des Moines, Iowa 50311, and Ames Laboratory—DOE,† Iowa State University, Ames, Iowa 50011

Received October 29, 1987

There are advantages to a single-number representation of a molecule or molecular fragments, even though from the onset it is known that such characterizations necessarily are accompanied by certain loss of information. We consider construction of a single-number characterization of rings in polycyclic structures to serve as ring descriptors. The number, called a ring ID number, is based on the count of suitably weighted paths for atoms forming the ring. The approach is illustrated on various rings in trimethyltricycloheptanes. The proposed ring descriptor shows a high discrimination power. Moreover, ring ID numbers reflect some inherent structural features of rings. We find that rings which are apparently similar are represented by ID values that are numerically close.

## INTRODUCTION

Present expansion of chemical data, including registration of new compounds, is associated with continuous needs for data retrieval. Searching modest data files, disregarding problems of fragment search, or searching for maximal common substructure can already be time consuming unless highly efficient searching is possible. Already, just finding compounds of interest, regardless of other possible tasks, may be quite difficult unless the correct name or code for the compound is known—something that frequently will not be the case. One strategy in efficient searching through a large data file is to introduce a reduced structure basis that will result in presorted samples. A natural choice is to select the number of atoms and number of rings as preliminary screening parameters; however, in practice such choices are not sufficiently discriminatory. For example, there are thousands of structures having a five-membered ring, yet we may be interested in a selected few having a particular structural feature, such as being fused to another ring or having specified substituents. It is therefore important to narrow down the search by including additional structural characteristics in a code, yet keeping the representation as simple as possible. There are no difficulties in developing a coding system that is comprehensive and *lengthy*, where, for example, a structure is represented uniquely and richly by a long list of specifications. The problem is to design a scheme in which a considerable amount of structural information is contained in a simple, *short* code. In addition, the approach should not be restricted to a special class of compounds. The shorter the code, the faster will be the search based on such codes. Hence, representations of a structure by a single number are highly desirable because they will considerably speed a searching process when compared to schemes using sequential data (e.g., connectivity tables). The issue is, Can a useful single-number condensation of structure, having sufficient discrimination, be developed?

It is useful to distinguish between the two conceptually distinct avenues for representing structures: (i) codes versus (ii) descriptors. Codes, as the name implies, presume prescribed *rules*, which when followed produce a name for the structure. Descriptors, as the name implies, use a selected property as the attribute on which characterization is based. Codes are not invariants; i.e., they require a particular (canonical) labeling of atoms. Descriptors are invariants, i.e., independent of how atoms are labeled. There are other important differences between codes and descriptors as will be outlined shortly.

Read[1] summarized desirable qualities for chemical codes (or names), which include line representation (based on common symbols), uniqueness, possibility for reconstruction (of a structure from a code), brevity, etc. In the case of trees (acyclic graphs) the $n$-tuple code of Knop and collaborators,[2] where for an $n$-atom tree a string of $n$ numbers suffices, is an illustration of codes that satisfy the requirements. The $n$-tuple code can be extended to cyclic structures by adding ring-closure information, producing thus very compact codes.[3] The derived codes, $n$-tuple,[2] compact codes,[3] official IUPAC names,[4] or any of numerous alternative canonical labeling of atoms[5] all require the user to know *rules* that govern code construction, rules which differ in complexity, preferences, and arbitrariness. Typically, the simpler the coding rules, the more difficult it is to derive the code. But this is preferred to rules designed to anticipate and resolve all ambiguities and that thus become lengthy, complex, and cumbersome. While computational difficulties, if necessary, can be delegated to a computer, it is hard to foresee all structural features of unknown novel compounds. Hence, the frequent revisions of "official" nomenclature rules, such as made by IUPAC. An illustration of structural rules is canonical labelings of atoms, such as those based on the smallest binary code for the structures when the entries of the adjacency matrix are read from left to right and from top to bottom.[6]

Alternative to coding based on rules are approaches based on structural invariants, indices, as the basis for a design of molecular descriptors. Invariants are independent of assumed labeling and therefore allow a direct comparison of files in different laboratories. A disadvantage of invariants is that there is no guarantee that the representation for a compound is unique. Moreover, it is generally believed that no finite list of structural invariants suffices to specify a compound uniquely. Hence, there is some loss of structural information when structures are represented by invariants, and consequently such descriptors do not allow reconstruction. On the other hand, since invariants represent "mathematical" properties of structures, they have advantages in structure–property studies.

†Operated for U.S. Department of Energy by Iowa State University under Contract No. W-7405-ENG-82. This work was supported in part by the Office of the Director.

0095-2338/88/1628-0142$01.50/0 