Description of Organic Reactions Based on Imaginary Transition Structures. 5. Recombination of Reaction Strings in a Synthesis Space and Its Application to the Description of Synthetic Pathways

SHINSAKU FUJITA

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

Received May 27, 1986

Several new concepts have been introduced for the description of synthetic pathways: (1) three modes of recombination of two reaction strings, i.e., degeneration, accumulation, and compensation; (2) multiplication of complex bond numbers; (3) synthesis spaces; and (4) synthesis classes.

Imaginary transition structures (ITS's), which we have proposed in the preceding papers, are unitary representations of structural information on organic reactions.[1] Main advantages of the ITS approach are as follows: (1) ITS's can be stored and manipulated in computer systems in the form of ITS connection tables as well as in the form of figures (graphs). (2) Starting and product stages are easily reproduced from ITS's by operations defined as projection to the starting stage (PS) and to the product stage (PP). (3) Perceptions of ring-opening reactions, ring closures, and rearrangements are translated to that of ring structures of ITS's. (4) Reaction types are presented by reaction graphs (or graphs of reaction centers of various levels), which are the subgraphs of the corresponding ITS's. (5) The concept of reaction strings is easily derived for ITS's and affords a versatile tool to describe details of organic reactions. These advantages stem from the fact that the ITS's can be regarded as structural formulas having three kinds of colored bonds, i.e., in-bonds (—o—), out-bonds (—//—), and par-bonds (—).

In addition to these advantages, another significant feature of the ITS approach should be described here. Reference between the corresponding nodes of a starting stage and of a product stage is important to describe organic reactions. But this is difficult for conventional methods, unless the nodes of both the stages are numbered correspondingly.[2] On the other hand, the node reference is achieved a priori in the present ITS approach.

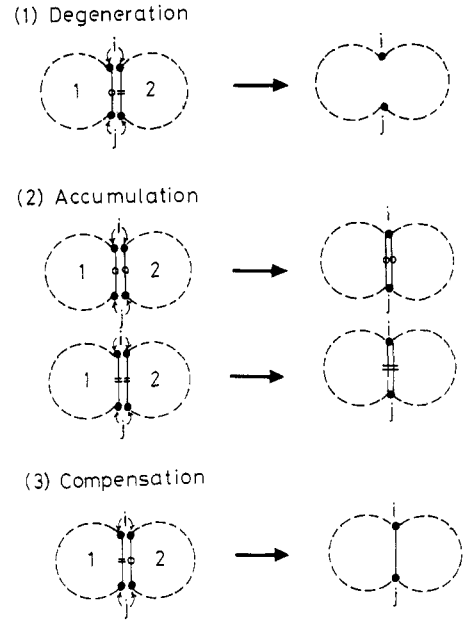
The node reference becomes more complicated in the description of synthetic pathways, which contain several unit reactions. I will propose here the concept of a synthesis space, which makes the node reference easier and hence it becomes a versatile medium to describe the synthetic pathways. I will also discuss behavior of reaction strings in such a synthesis space.

RECOMBINATION OF REACTION STRINGS FOR DESCRIBING SYNTHETIC PATHWAYS

A synthetic pathway involves two or more unit reactions. Conventional methods regard the synthetic pathway as a sequential diagram of several intermediates combined with arrows. This representation of organic synthesis deals only with organic compounds combined rather than with organic reactions themselves.

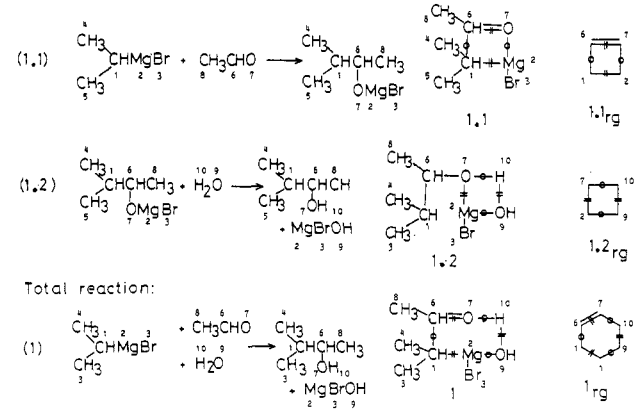
On the other hand, the ITS approach perceives a synthetic pathway as an overall ITS. This ITS is generated by recombination of ITS's, each of which corresponds to the respective unit reaction. When two ITS's are concerned with two successive unit reactions, the reaction strings of the two ITS's are recombined with each other. Thus, we can describe a synthetic pathway by means of recombination of reaction strings. The purpose of this paper is to clarify the relationship between the overall ITS and the ITS's of unit reactions.

Scheme I



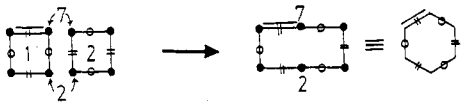
First of all, let us consider the simplest case of synthetic pathway, i.e., two successive unit reactions, each of which has one reaction string. Suppose that the reaction string of the first step (loop 1) and that of the second step (loop 2) are fused at a common edge linking adjacent nodes i and j (Scheme I). There are three modes of recombination of two reaction strings as shown in Scheme I. In the following sections, I exemplify the three modes using several representation pathways.

Degeneration. The reaction of entry 1[3] consists of a Grignard addition (entry 1.1) and the subsequent hydrolysis (entry



0095-2338/86/1626-0238$01.50/0

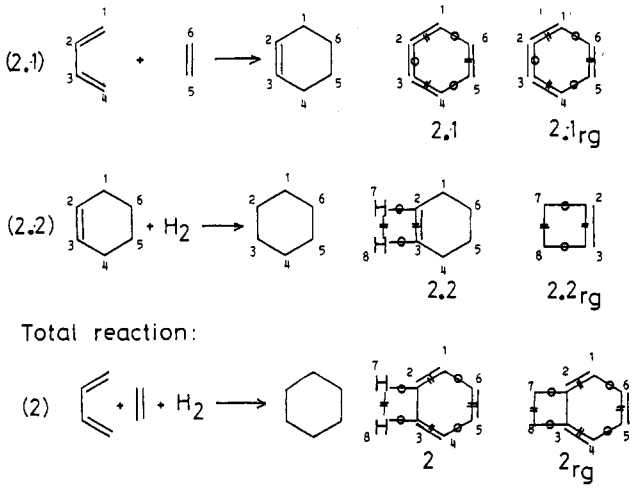
Scheme II



1.2). It is noted that a single bond between nodes 2 and 7 (O–Mg) is formed in the first step and cleaved in the second step. The overall effect provides no bond between these nodes. Thus, the overall reaction (entry 1) is represented by ITS **1** of one-string, whereas the unit reactions are represented by the corresponding ITS's **1.1** and **1.2**, respectively, each of one-string.

For the purpose of computer manipulation, I regard this process diagrammatically. The overall ITS **1** is derived from the ITS's **1.1** and **1.2** if an in-bond multiplied by an out-bond (at nodes 2 and 7) is presumed to give no bond. We call this derivation multiplication of ITS's. In an abstract fashion, multiplication of reaction strings $\mathbf{1.1}_{rg}$ and $\mathbf{1.2}_{rg}$ is represented by Scheme II, in which an in-bond and the next out-bond between the same two nodes are multiplied to result in no bond. Such a recombination of two reaction strings is defined as *degeneration of reaction strings*.

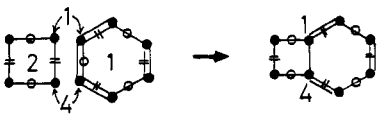
Another example of degeneration is a series of Diels–Alder reactions (**2.1**) and the subsequent hydrogenation (**2.2**). The



two reaction strings in ITS's **2.1** and **2.2** are recombined to result in degeneration to afford a reaction string of ITS **2**. This multiplication of the reaction strings is illustrated in Scheme III. When one considers an ITS bond between nodes 2 and 3 in this case, multiplication of the in-bond (exactly a (1+1) bond) of **2.1** and the out-bond (exactly a (2−1) bond) of **2.2** results in no change of bond multiplicity, as shown in ITS **2**.

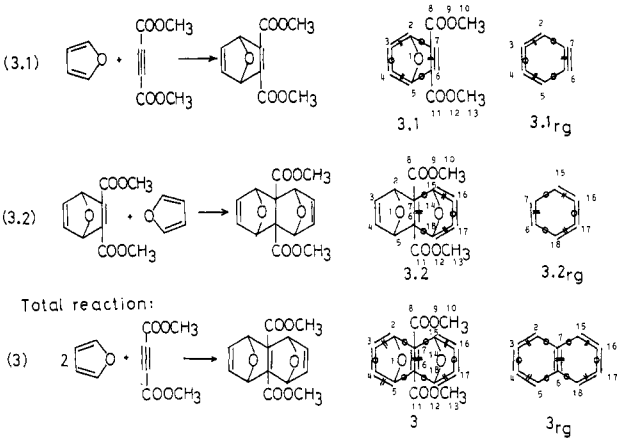
The degeneration of two reaction strings is represented schematically in Scheme I1. The two nodes of inquiry may be modified by par-bond(s) as exemplified in Schemes II and III. Two looped reaction strings are degenerated to form one

Scheme III



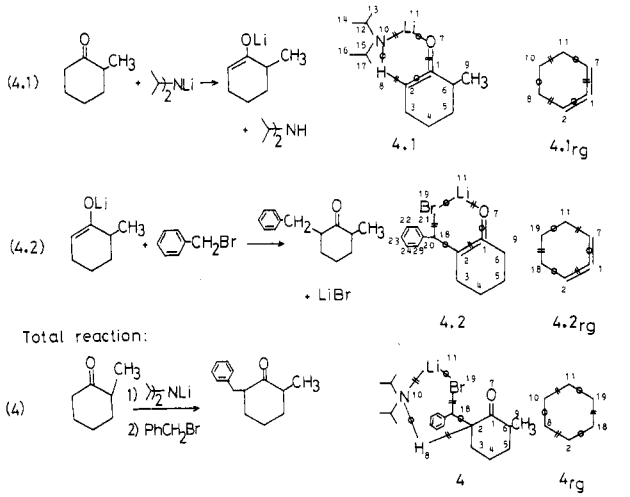
looped reaction string in most cases. A looped reaction string and a linear reaction string form one reaction string on degeneration. And degeneration of two linear reaction strings affords two new reaction strings.

Accumulation. An example of the accumulation of two reaction strings appears in a double Diels–Alder reaction,[4] in which two reaction strings corresponding to **3.1** and to **3.2** are



accumulated at two common nodes. The accumulation is illustrated in Scheme I2 schematically. Two or more reaction strings are accumulated to form a reaction graph of multistring, the details of which are discussed elsewhere.

Compensation. The compensation of two reaction strings is shown schematically in Scheme I3. An out-bond of the first reaction string is multiplied by an in-bond of the second reaction string to leave a par-bond. The enolization of a cyclohexanone (entry 4.1) and the following alkylation (**4.2**)[5]



are represented by ITS's **4.1** and **4.2**, respectively. When one combines ITS's **4.1** and **4.2**, the ITS bond between nodes 1 and 7 can be characterized by the term *compensation of reaction strings*. The chemical meaning of the compensation is obvious. Thus, the double bond between nodes 1 and 7 is converted into a single bond at the first step and then finally to a double bond. The net change is none of the bond multiplicity. The example cited involves other recombinations of strings. Thus, the ITS bonds between 1 and 2 and between 7 and 11 are characterized by degeneration of reaction strings.

It is to be noted that the difference between degeneration and compensation stems from the order of multiplication of an in- and an out-bond. This situation is obvious, if one considers chemical meanings of these operations, and will be formulated in the next section.

MULTIPLICATION OF COMPLEX BOND NUMBERS

The next subject is to transform the diagrammatical expressions described above into machine-readable ones. This can be done in terms of complex bond numbers (*a b*), which are defined to denote imaginary bonds appearing in ITS's.[1,6] Here, the mode of recombination of reaction strings is transformed to the corresponding algebraic representation, i.e., multiplication of complex bond numbers.

Table I. Multiplication of ITS Bonds in the Reaction of Entry 1

node	ITS 1.1		ITS 1.2		ITS 1		
i–j	ITS bond	$(a_{ij}^{(1)}\ b_{ij}^{(1)})$	ITS bond	$(a_{ij}^{(2)}\ b_{ij}^{(2)})$	ITS bond	$(a_{ij}\ b_{ij})$	remarks
1–2	⊣⊢	(1–1)	none	(0+0)	⊣⊢	(1–1)	
1–4	—	(1+0)	—	(1+0)	—	(1+0)	invariant
1–5	—	(1+0)	—	(1+0)	—	(1+0)	invariant
1–6	⊸	(0+1)	—	(1+0)	⊸	(0+1)	
2–3	—	(1+0)	—	(1+0)	—	(1+0)	invariant
2–7	⊸	(0+1)	⊣⊢	(1–1)	none	(0+0)	degenerate
2–9	none	(0+0)	⊸	(0+1)	⊸	(0+1)	
6–7	⇌	(2–1)	—	(1+0)	⇌	(2–1)	
6–8	—	(1+0)	—	(1+0)	—	(1+0)	invariant
7–10	none	(0+0)	⊸	(0+1)	⊸	(0+1)	
9–10	—	(1+0)	⊣⊢	(1–1)	⊣⊢	(1–1)	

Table II. Connection Table of ITS R_3 Defined in the Synthesis Space **R**

node	atom or group	coordinate		neighbor 1		neighbor 2		neighbor 3		neighbor 4		neighbor 5	
		x	y	node	$(a\ b)$	node	$(a\ b)$	node	$(a\ b)$	node	$(a\ b)$	node	$(a\ b)$
1	C	0	200	2	(1+0)	7	(1+0)	8	(1+1)	15	(1–1)		
2	CH	173	100	1	(1+0)	3	(1+0)	15	(1+0)				
3	CH_2	173	−100	2	(1+)	4	(1+0)						
4	CH_2	100	−286	3	(1+0)	5	(1+5)						
5	CH_2	−100	−286	4	(1+0)	6	(1+0)						
6	CH_2	−173	−100	5	(1+0)	7	(1+0)						
7	CH_2	−173	100	1	(1+0)	6	(1+0)						
8	O	0	400	1	(1+1)	9	(1–1)						
9	Si	0	600	8	(1–1)	11	(1+0)	12	(1+0)	13	(1+0)	20	(0+1)
10	Cl	0	800										
11	CH_3	100	773	9	(1+0)								
12	CH_3	173	700	9	(1+0)								
13	CH_3	200	600	9	(1+0)								
14	H	314	−41										
15	CH_2	200	200	1	(1–1)	2	(1+0)	19	(0+1)				
16	I	373	300	18	(1+0)								
17	I	373	100	18	(1+0)								
18	Zn	546	200	16	(1+0)	17	(1+0)						
19	H	200	400	15	(0+1)								
20	OH	−200	600	9	(0+1)								

Suppose that n ITS's (R_1, R_2, ..., and R_n) represent a serial set of n reactions. The complex bond number $(a_{ij}^{(k)}\ b_{ij}^{(k)})$ represents an ITS bond between nodes i and j of ITS R_k. The total reaction is represented by the equation $R = R_nR_{n-1} \ldots R_2R_1$.[7] If the complex bond number $(a_{ij}\ b_{ij})$ is for the ITS bond between nodes i and j of the total ITS R, the following equation can be obtained[8]

$$(a_{ij}^{(1)}b_{ij}^{(1)}) + (a_{ij}^{(2)}b_{ij}^{(2)}) + \ldots + (a_{ij}^{(k)}b_{ij}^{(k)}) + \ldots + (a_{ij}^{(n)}b_{ij}^{(n)}) = (a_{ij}b_{ij}) \quad (1)$$

wherein

$$a_{ij} = a_{ij}^{(1)} \quad (2)$$

$$b_{ij} = b_{ij}^{(1)} + b_{ij}^{(2)} + \ldots + b_{ij}^{(n)} \quad (3)$$

for all i and all j.

The following conditions for consecutive reactions are obtained easily for all i and j of each R_k and R_{k+1}:

$$a_{ij}^{(k)} + b_{ij}^{(k)} = a_{ij}^{(k+1)} \geq 0 \ (1 \leq k \leq n-1) \quad (4)$$

In order to exemplify eq 1–4, I discuss the reactions of entries 1.1 and 1.2 (total: entry 1) in detail. Table I shows the change of all ITS bonds in the series of reactions (two steps). For example, the ITS bond between nodes 2 and 7 is characterized by degeneration, which is represented algebraically as (0+1) + (1–1) = (0+0). All ITS bonds collected in Table I satisfy eq 1–4.

The three modes of recombination of reaction strings are represented algebraically when one uses complex bond numbers. Suppose that nodes i and j are combined by ITS bond $(a_{ij}^{(1)}b_{ij}^{(1)})$ in ITS R_1 by ITS bond $(a_{ij}^{(2)}b_{ij}^{(2)})$ in ITS R_2, and as a result, by $(a_{ij}b_{ij})$ in the whole ITS (R_2R_1). Then the following equations are obtained for the three cases:

degeneration

$$b_{ij}^{(1)} + b_{ij}^{(2)} = b_{ij} = 0 \qquad b_{ij}^{(1)} > 0 \quad (5)$$

accumulation

$$b_{ij}^{(1)}b_{ij}^{(2)} > 0 \quad (6)$$

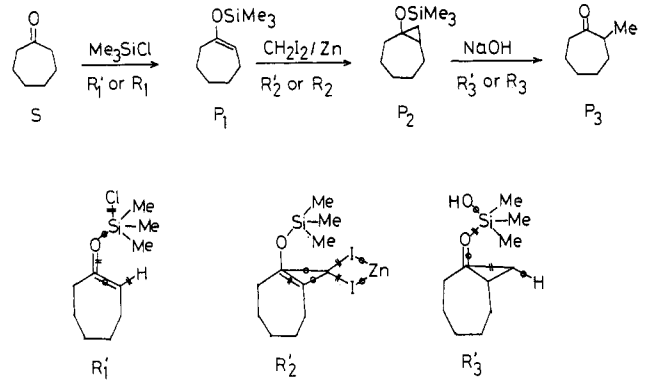
compensation

$$b_{ij}^{(1)} + b_{ij}^{(2)} = b_{ij} = 0 \qquad b_{ij}^{(1)} < 0 \quad (7)$$

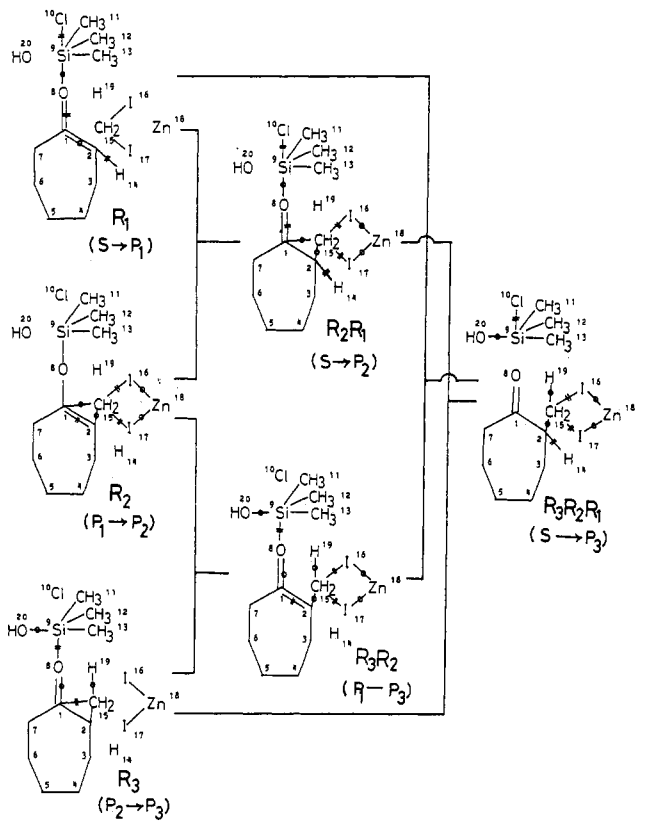
SYNTHESIS SPACES

Let us examine the synthetic pathway[9] shown in Scheme IV. The ITS (or the abbreviated ITS) for each step can be written easily. If one tries to treat the three steps as a synthetic pathway on the whole, one would meet some difficulties, since the ITS's R_1', R_2', and R_3' do not always contain all nodes in common. Hence, the numbering of nodes would be complicated without some new concepts.

Scheme IV



Scheme V



In order to avoid these difficulties, I propose a synthesis space, which contains all nodes appearing in the synthetic pathway and said nodes are numbered in common. For example, the above-described pathway has a synthetic space containing 20 nodes, i.e., 1C, $2CH_2$, $3CH_2$, $4CH_2$, $5CH_2$, $6CH_2$, $7CH_2$, 8O, 9Si, 10Cl, $11CH_3$, $12CH_3$, $13CH_3$, 14H, $15CH_2$, 16I, 17I, 18Zn, 19H, and 20OH. The ITS of each step is defined in this synthesis space, i.e., R_1, R_2, and R_3 (Scheme V). It should be noted that a synthesis space spanned by R_1', R_2', or R_3' is the subspace of the synthesis space spanned by R_1, R_2, and R_3. Thus, the relationship of R_1' and R_1 is illustrated in Scheme VI.

ITS's can be stored and manipulated in terms of an ITS connection table (ITS-CT) as described above. For example, ITS R_3, which is defined in the synthesis space **R**, is represented by the ITS-CT in Table II. In accordance with the fact that ITS R_3 is divided into four parts, an appropriate operation on ITS-CT (e.g., similarity transformation of the corresponding matrix representation) divides all nodes into four sets of nodes (i.e., nodes 10, 14, 16, 17, and 18, and other nodes). Then the reaction string 20+9-8+1-15+19 is extracted from the last set of nodes.

Scheme VI

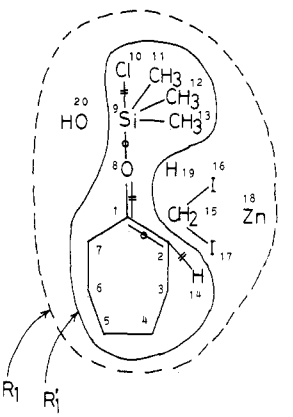


Table III. Synthesis Class Describing the Synthetic Pathway

$$S \xrightarrow[R_1]{} P_1 \xrightarrow[R_2]{} P_2 \xrightarrow[R_3]{} \cdots \xrightarrow[R_n]{} P_n$$

steps	ITS's contained				
n	$R_nR_{n-1}\cdots R_2R_1$	$R_{n-1}\cdots R_2R_1$	$R_{n-2}\cdots R_2R_1$	$\cdots$ R_2R_1	R_1
$n-1$	$R_nR_{n-1}\cdots R_2$	$R_{n-1}\cdots R_2$	$R_{n-2}\cdots R_2$	$\cdots$ R_2	
$n-2$	$R_nR_{n-1}\cdots R_3$	$R_{n-1}\cdots R_3$	$\cdots$	$\cdots$	
$\cdots$					
$\cdots$					
$\cdots$					
2	R_nR_{n-1}	R_{n-1}			
1	R_n				

SYNTHESIS CLASSES OF ITS'S

Graphical recombination of two or three ITS's selected from R_1, R_2, and R_3 (or, equivalently, the corresponding multiplication of complex bond numbers based on such an ITS-CT as Table II) affords a class of ITS's (Scheme V) that are related closely to each other. This class containing $R_3R_2R_1$, R_2R_1, R_3R_2, R_1, R_2, and R_3 is called a synthesis class based on R_1, R_2, and R_3. Suppose that a set of ITS's R_1, R_2, ..., R_n, which satisfy eq 1-4 and span the synthesis space **R**, represents the synthetic pathway below in general.

$$S \xrightarrow[R_1]{} P_1 \xrightarrow[R_2]{} P_2 \xrightarrow[R_3]{} P_3 \rightarrow \cdots \xrightarrow[R_n]{} P_n$$

They can be combined to generate ITS's collected in Table III. For example, R_2R_1 represents combined reaction S to P_2, and $R_3R_2R_1$ represents combined reaction P_1 to P_3. I define a synthesis class as a class of all ITS's (Table III) that are generated from the ITS's of said synthetic pathway. This synthesis class is based on synthesis space **R**. The concept of synthetic class is useful to describe synthetic pathways.

When one tries to retrieve some reaction and hits, for example, a reaction represented by ITS $R_{n-1}...R_2$, one can obtain the synthesis class of Table III. Thus, the reaction $R_{n-1}...R_2$ can be divided into unit reactions represented by R_2, R_3, ..., and R_{n-1} in the light of Table III. Furthermore, all reactions related to the reaction $R_{n-1}...R_2$ appear inside the triangle constructed by $R_{n-1}...R_2$, R_2, and R_{n-1} of Table III as vertices.

It should be emphasized that even $R_nR_{n-1}...R_1$ is an ITS, and thus even a synthetic pathway can be represented by an appropriate ITS. This feature of the present ITS approach makes the retrieval of synthetic pathways easy. Compare Scheme III and entry 4. Scheme III can be regarded totally as α-alkylation of a ketone. This corresponds to the fact that the ITS ($R_3R_2R_1$ of Scheme V) involves H—//—C(CO)—O—C as a substructure. The reaction of entry 4 is another type of α-alkylation of a ketone. The corresponding ITS (**4**) contains the same substructure, i.e., H—//—C(CO)—O—C. These facts

reveal that overall α-alkylation of a ketone can be retrieved by means of the above substructure as a clue.

CONCLUSION

Two reaction strings combine with each other in three ways, i.e., degeneration, accumulation, and compensation, which are useful concepts to describe synthetic pathways. Multiplication of complex bond numbers, synthesis spaces, and synthesis classes are now introduced in order to apply the ITS concept to the description of synthetic pathways.

REFERENCES AND NOTES

(1) Fujita, S. *J. Chem. Inf. Comput. Sci.* preceding papers of this series in this issue.
(2) Only Ugi's system solves this problem in terms of "isomeric ensemble of molecules". See: Jochum, C.; Gasteiger, J.; Ugi, I. *Angew. Chem., Int. Ed. Engl.* **1980**, *19*, 495 and references cited therein.
(3) Drake, N. L.; Cooke, C. B. *Org. Synth.* **1943**, *Coll. Vol. 2*, 406.
(4) Diels, O.; Alder, K. *Justus Liebigs Ann. Chem.* **1931**, *490*, 243.
(5) Gall, M.; House, H. O. *Org. Synth.* **1972**, *52*, 39.
(6) An ITS bond can be denoted by a set of three integers (o, i, p), wherein integer o represents a number of out-bonds, i is that of in-bonds, and p represents that of par-bonds. The condition implied by this expression is $oi = 0$. The following two equations are obtained easily: $a = p + o$ and $b = i - o$.
(7) The ITS (R_2R_1) denotes the multiplication of the first ITS (R_1) by the second ITS (R_2). In general, $R = R_nR_{n-1}...R_1$ means the multiplication of successive ITS's, R_1, ..., and R_n.
(8) It is to be noted that all nodes are numbered commonly throughout the reaction pathway.
(9) Conia, J. M.; Girard, G. *Tetrahedron Lett.* **1973**, 2767.

Coding of Relational Descriptions of Molecular Structures

KARL WIRTH
Fachbereich Mathematik, KME, CH-8001 Zürich, Switzerland

Received November 27, 1985

This is an approach to a naming procedure for molecular structures, which are understood to be stereochemical models of molecules. The procedure is both general and uniform; i.e., it provides any aspect of any molecular structure and codes by using a fully unified criterion. Structural elements of such aspects as, for instance, atoms, bonds, connection angles, dihedral angles, and orientations, are represented by tuples to arrive at relational descriptions of molecular structures. The coding of those relational descriptions is expounded, including the presentation of a canonization algorithm in particular, which is based on minimalization. The result of the coding, which is a systematic name for a described molecular structure, comprises the absolute configuration and enables the symmetry group to be determined.

INTRODUCTION

Why develop a new naming procedure for chemical entities if several already exist?[1] I agree that many of these procedures work very well, which, however, cannot conceal the fact that they are, in a way, patchwork. This is quite understandable as they gradually developed from daily requirements of applied chemistry. As a result existing procedures have been affected in two ways. They are bound to special aspects and classes of molecules and are therefore not general enough. On the other hand, new requirements led to modifications that made these procedures less and less uniform.

Owing to its deductive approach, this paper offers a solution to these two disadvantages. To achieve this, stereochemical models are resorted to. In these models molecules are understood to be (mobile or rigid) arrangements of atoms in space: The atoms are represented by points, and the relationships between them by internal coordinates, e.g., bond lengths, bond angles, and dihedral angles. To avoid any possible misunderstanding it must be pointed out that in this article the term *molecular structure* refers to these limited idealized models of molecules.

We consider a naming procedure of molecular structures to consist of two steps: a first one, which we call *describing*, followed by a second one, which we call *coding*. Describing means representing a molecular structure by a linear sequence of symbols; coding then brings the resulting description into canonical form. The advance of the present procedure over others can now be specified as follows: It is *general*, which means that, in principle, the suggested describing aims to take into account any conceivable aspect of any conceivable molecular structure; and it is *uniform*, which means that the expounded coding functions for any resulting description regardless of whatever structural aspects it represents.[2]

In order to achieve generality and uniformity describing and coding are completely separated, which the existing naming procedures hardly ever try for. In those procedures describing and coding are mixed, coding, moreover, frequently being based on a different criterion for each structural aspect (e.g., IUPAC Nomenclature[3]). Such mixing may be very efficient with specific classes of fairly simple molecular structures. However, not only does it prevent generality and uniformity, with more complex molecular structures it may also be a source of errors with regard to uniqueness.

Generality is achieved in the present naming procedure by using relational descriptions as the result of describing. The term relational description corresponds to that of a finite relational system in mathematics.[4] Such a system is based on tuples, which we use to describe relationships between atoms[5] of molecular structures. A tuple can consist of any number of atoms in any order, and even repeated atoms are admitted. This enables any element of structural aspects to be expressed: constitutional elements such as atom types, bonds, chains, rings, etc. and stereochemical ones such as bond angles, dihedral angles, orientations, topicities,[6] etc.

Uniformity is achieved by consistent use of minimalization for all structural aspects as the basis of coding. This criterion was applied by other naming procedures that, however, usually involve only the constitution of molecular structures and operate with adjacency matrices.[7] A mathematical examination shows that, besides minimalization, any other criterion may be possible.[4] It ought to be an interesting question whether certain criteria have chemical relevance according to the represented molecular structures.

This article is primarily about coding and not describing, but presupposes relational descriptions as the result of describing. It is, of course, impossible to solve all the problems

0095-2338/86/1626-0242$01.50/0