(40) Speziale, A. J.; Freeman, R. C. *Organic Syntheses*; Wiley: New York, 1973; Coll. Vol. 5, p 387.
(41) Werner, N. W.; Casnova, J., Jr. *Organic Syntheses*; Wiley: New York, 1973; Coll. Vol. 5, p 273.
(42) Shulenberg, J. W.; Archer, S. *Org. React.* **1965**, *14*, 1.
(43) Sundberg, R. J.; Pearce, B. C. *J. Org. Chem.* **1985**, *50*, 425.
(44) Hasek, R. H.; Clark, R. D.; Mayberry, G. L. *Organic Syntheses*; Wiley: New York, 1973; Coll. Vol. 5, p 456.
(45) Kita, M.; Yoshida, H.; Sakurai, H. *J. Am. Chem. Soc.* **1985**, *107*, 7767.
(46) Salaün, J. *Tetrahedron Lett.* **1984**, *25*, 1269, 1273.
(47) Krow, G. R.; Reilley, J. *J. Am. Chem. Soc.* **1975**, *97*, 3837.
(48) Shriner, R. L.; Ford, S. G.; Roll, L. J. *Organic Syntheses*; Wiley: New York, 1943; Coll. Vol. 2, p 140.
(49) Crockett, G. C.; Koch, T. H. *J. Org. Chem.* **1977**, *42*, 272.
(50) Brasen, W. R.; Hauser, C. R. *Organic Syntheses*; Wiley: New York, 1963; Coll. Vol. 4, p 585.
(51) Hayashi, M. *J. Chem. Soc.* **1927**, 2516.
(52) Carlson, R. G. *J. Chem. Soc., Chem. Commun.* **1973**, 223.
(53) Allen, C. F. H.; Gates, J. W., Jr. *Organic Syntheses*; Wiley: New York, 1955; Coll. Vol. 3, p 418.
(54) Howard, W. L.; Lorrette, N. B. *Organic Syntheses*; Wiley: New York, 1961; Coll. Vol. 5, p 25.
(55) Funk, R. L.; Munger, J. D., Jr. *J. Org. Chem.* **1985**, *50*, 707.
(56) Rhods, S. J.; Raulins, N. R. *Org. React.* **1975**, *22*, 1.
(57) Evans, D. A.; Golob, A. M. *J. Am. Chem. Soc.* **1975**, *97*, 4765.
(58) DeVolfe, J. R.; Young, W. G. *Chem. Rev.* **1956**, *56*, 763.
(59) Diels, O.; Alder, K. *Ber.* **1929**, *62*, 2337.
(60) Arnold, R. T.; Smolinsky, G. *J. Am. Chem. Soc.* **1960**, *82*, 4918.
(61) Oppolzer, W. *Helv. Chim. Acta* **1973**, *56*, 1812.
(62) Carroll, M. F. *J. Chem. Soc.* **1940**, 704.

"Structure–Reaction Type" Paradigm in the Conventional Methods of Describing Organic Reactions and the Concept of Imaginary Transition Structures Overcoming This Paradigm

SHINSAKU FUJITA

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

Received October 6, 1986

The description of organic reactions is discussed from the viewpoint of a structure–reaction type (SRT) paradigm, in which structural information and reaction type are stored and manipulated more or less independently. This paradigm must be overcome in order to construct an integrated system that will support both retrieval of organic reactions and synthetic design. The concept of imaginary transition structures (ITS) is introduced as a unitary representation free from the SRT paradigm.

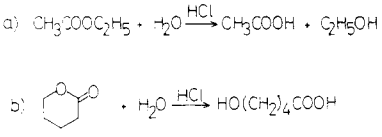
Sellow pointed out in his critical review[1] that "an important amount of success was achieved in the field of codification of structure but not in that of codification of reactions". He ascribed this to two strongly correlated reasons: "the lack of a real and efficient methodology for analyzing the theoretical aspects of reactions and the necessity of translating them from usual language to code". Although his criticism is to the point, the difficulty in describing organic reactions should be discussed more thoroughly from another point of view.

There are two types of computer systems manipulating organic reactions: systems for the retrieval of organic reactions and systems for the design of synthetic pathways. The former systems deal with information on *individual reactions*. The latter concern *reaction types* rather than individual reactions. These two types have grown separately, although both are based on information on organic reactions. In this paper, we discuss why these two types of systems have never been integrated. We reveal a structure–reaction type (SRT) paradigm that restricts the conventional methods of description of organic reactions and, as a result, hinders the integration of the systems.[2]

DUALITY IN THE DESCRIPTION OF ORGANIC REACTIONS: STRUCTURE–REACTION TYPE (SRT) PARADIGM

Let us work out a simple reaction shown in Scheme Ia. How do we describe this reaction? First, we recognize the structure of the substrate as ethyl acetate (or the corresponding coded name). Next, we discern between the substrate and the product and recognize this reaction as hydrolysis (or the corresponding code) from our own knowledge. And then we describe this as a combination of ethyl acetate and hydrolysis. In the process of recognition, we do not describe the individual reaction in itself, but we replace the description of the reaction

Scheme I



by (1) that of the structure of the substrate and (2) that of the reaction type.[3]

Figure 1 summarizes the duality in the conventional description of organic reactions. In general, structural data of a given reaction are divided into information on the structure and that on the reaction type, which are stored separately, more or less independently, in a structure file and in a reaction-type file, respectively. We call this type of duality a structure–reaction type (SRT) paradigm. This paradigm has never been formulated, so that the previous efforts to develop new representations of organic reactions resulted unintentionally in intimate combination of (1) and (2) at the utmost. There have been no attempts to overcome this.

In the SRT paradigm, retrieval of organic reactions requires dual reference or access to the structure file and to the reaction-type file as shown in the right-hand side of Figure 1. In this case, the correspondence between the two files is crucial to construct an effective system of retrieval. However, this is difficult as discussed below so long as we use conventional methods of description.

On the other hand, synthetic design is simpler than retrieval of organic reactions from the viewpoint of the SRT paradigm.[4] Once an expert extracts reaction types from the data of individual reactions, synthetic design requires only single access to the reaction-type file. In other words, structural information is *given as a target molecule* in the case of synthetic design. If our efforts are limited to synthetic designe field, the SRT

0095-2338/87/1627-0120$01.50/0

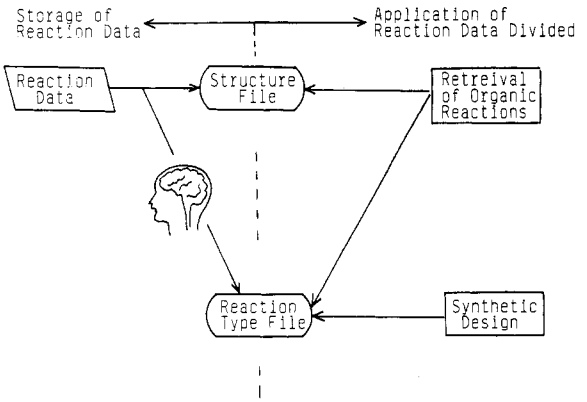


Figure 1. Structure-reaction type paradigm in the conventional methods of description of organic reactions.

```
46. REACTION

Abstraction reaction              ..Reduction, electrochemical
Addition reaction                 ...Electrolysis
.Acylation                        ..Reduction, photochemical
..Carbonylation                   Reforming
..Carboxylation                   Ring cleavage
..Hydroformylation                Ring closure and formation
.Alkenylation                     .Chelation
..Vinylation                      .Cycloaddition reaction
.Alkylation                           .
..Cyanoethylation                     .
      .                               .
      .                           Solvolysis
      .                           .(Acidolysis)
..(Oxidative substitution)        ..Acetolysis
...Ammoxidation                   ..Formolysis
..Oxygenation                     ,Alcoholysis
..Ozonization                     ..Ethanolysis
..Peroxidation                    ..Methanolysis
.Reduction                        .Aminolysis
..Birch reduction                 .Hydrazinolysis
..Deoxidation                     .Hydrolysis
..Hydrogenation                   Substitution reaction
..Hydrogenolysis                      .
..Methanation                         .
..Nitrogen fixation, syntheic         .
```

Figure 2. Reaction hierarchy used by Chemical Abstracts Service.

paradiam is sufficiently effective by itself.

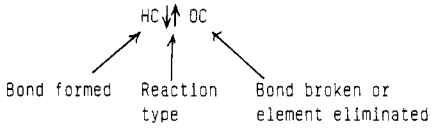
In the next sections, we discuss how conventional methods of description are caught in the SRT paradigm. If we remain in the SRT paradigm, the structure file and the reaction-type file should be combined as intimately as possible. Alternatively, we should overcome this paradigm in order to integrate retrieval field and synthetic design.

NATURAL LANGUAGE TERMS AND REACTION CODES

Among the most conventional methods of description of organic reactions, natural language terms have been and will be used widely and conveniently in the field of retrieval. The linkage between this type of language term and the structural information of the reaction is incomplete and indirect. Even if the structural information is given in the form of a connection table, it is difficult for a computer to recognize what part of the molecule changes during the reaction.

Another disadvantage is also noted. We can describe the example of Scheme Ib as a combination of 5-pentanolide (or the corresponding coded name) and hydrolysis (or the corresponding code). This recognition fails to notice another category of reaction features, e.g., ring cleavage, unless we reexamine the structures of the participating molecules. When the reaction type is called "lactone hydrolysis", ring cleavage is implied in this usage. But the terms "(ester) hydrolysis" and "lactone hydrolysis" must be correlated by some algorithm, since it is desirable to retrieve reactions a and b of Scheme I at the same time.

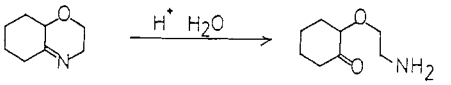
Chemical Abstracts Service[5] uses a reaction hierarchy to search organic reactions effectively (Figure 2). In the CAS system, two descriptors may be selected, i.e., hydrolysis and ring cleavage for reaction b of Scheme I. Although these



Addition: A + B ⟶ C, no loss of material
Rearrangement (Isomerization): A ⟶ B, no loss of material
Exchange: A + B ⟶ C (+D), e.g. condensations, substitutions
Elimination: A ⟶ B (+D), e.g. dehydration of alcohols

H, O, N, Hal (halogen), S, Rem (other elements) and C, in this order.

Figure 3. Constitution of the Theilheimer code.



"C6C6/N1O4"E(10)[-]>>RO

>>"C6"O(1)OR(2)NH_2(T2B)[-]

NC: no change in molecular skeleton, functional group interconversion only.
C+: carbon atoms added, forming a new C-C bond.
C-: loss of carbon atoms, cleaving a C-C bond.
RF: ring formation.
RO: ring opening.
RE: ring expansion.
RC: ring contraction.
RR: rearrangement.

Figure 4. Cohen's coding system of organic reactions.

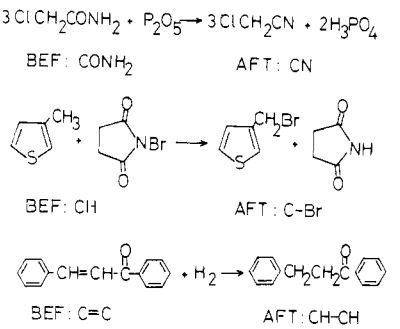


Figure 5. Before and after codes of RIRS.

natural language terms are effective for the purpose of retrieval, these are caught with the SRT paradigm.

Figure 3 shows the reaction symbol notation used by Theilheimer in *Synthetic Methods of Organic Chemistry*. The Theilheimer codes are a combination of the symbols of bonds cleaved and formed and of four symbols of reaction features.

This coding system is promoted by *Journal of Synthetic Methods* and has developed into an online service called CRDS (The Chemical Reaction Documentation Service).[6] From the viewpoint of the SRT paradigm, the same situation as above is also the case in Theilheimer's coding system.

Cohens's coding system[7] is based on a combination of the linear notation of participating molecules and of the code of a reaction type (Figure 4). This coding cannot overcome the SRT paradigm, although it may appear to be a unitary representation of organic reactions. However, it is a simple recombination of a structure and a reaction type divided previously. Such recombination would provide no further aspects on retrieval of organic reactions. For example, other descriptors such as "imide hydrolysis" (or the corresponding code) cannot be abstracted without expert knowledge or some tedious algorithm. Morever, the reaction types selected are limited

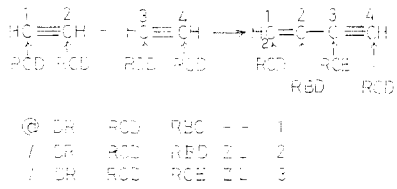
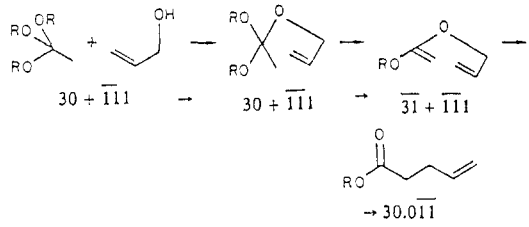


Figure 6. GREMAS code of acetylene dimerization.



Claisen (ZZ) (ΠZ·ΠH) (ZΠ·RΠ·RΠ·ΠΠ·ΠZ)

Figure 7. Hendrickson's codes for describing the Claisen rearrangement.

to descriptors concerning changes of carbon skeletons.

In RIRS (Roche Integrated Reaction System),[8] reaction centers are coded as BEF (before) and AFT (after) as shown in Figure 5. This coding system is specifically designed for retrieval and is not sufficient to support the integrated systems of manipulating organic reactions.

GREMAS, HENDRICKSON'S, AND LITTLER'S CODE

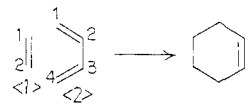
In the case of GREMAS code,[9] each reaction center is coded as a combination of three alphabets (Figure 6). The first letter, R, corresponds to a carbon atom without substitution of heteroatoms. The second letter represents an acetylenic carbon (B) or an olefinic carbon (C). The third letter is a descriptor of branching at the carbon. This system divides a reaction into changes of respective atoms but does not combine them to form the total aspect of the reaction. And so the GREMAS code is effective for the retrieval of reactions but lacks feasibility in the field of synthetic design.

Hendrickson's method[10] is the same as the GREMAS code from the viewpoint of methodology. In his coding, substituents on each carbon atom are classified into four categories, i.e., H (hydrogen-like atom), R (carbon atom), Π (unsaturation), and Z (heteroatom). Then the change of substitution on each carbon is represented by a pair of such letters, ZΠ (Π changes to Z), and the total reaction is recognized as a set of such unit changes (Figure 7). Apparently, molecularity of the reaction is unclear in his coding. In the series of the Claisen rearrangement (Figure 7), the first step contains two molecules, but the code (ZZ) corresponds only to one molecule that contains a carbon atom participating the reaction.

Morever, the third step (of the Claisen rearrangement) involves six reaction centers, only five of which are coded by Hendrickson's method. On the other hand, the related Cope rearrangement contains six reaction centers and has a set of six pairs of letters (RΠ.ΠΠ.ΠR.ΠR.ΠΠ.RΠ). Although his coding method is useful for a wide range of organic reactions, this lack of generality cannot meet requirements in the retrieval field.

Littler's code[11] is based on *primitive changes* as exemplified in Figure 8. For example, the term "as" represents formation of a bond, "pde" corresponds to breakage of a π-bond, etc. The methodology of this system is that a reaction is reduced to a set of bond changes. Contrary to this, the above two coding systems cited in this section regard a reaction as a set of substitution changes.

In summary, the three coding systems discussed in this section give incomplete correspondence between structural



C1<1>cs C1<2>, C2<2>pde C1<2>
C2<2>pc C3<2>, C4<2>pde C3<2>
C4<2>as C2<1>, C1<1>pde C2<1>

Figure 8. Littler's codes for the Diels–Alder addition.

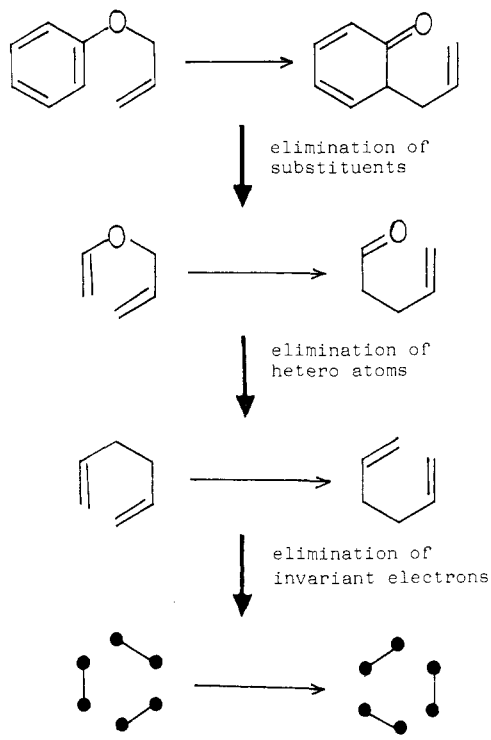


Figure 9. Roberts's coding system for the Claisen rearrangement of allyloxybenzene.

information and reaction-type information. They are limited within the SRT paradigm.

METHODS BASED ON GRAPHS, CONNECTION TABLES, AND MATRICES[12]

Roberts[13] considered the reaction as a series of concerted processes (CP). In his coding, the Claisen rearrangement is reduced to a CP skeleton as shown in Figure 9. This coding discards all residual information other than that of reaction centers. The recognition of this reaction as rearrangement and distinction between this and the Diels–Alder reaction need further comparison of the molecules participating in the reactions. This fact reveals that the Roberts coding is not free from the SRT paradigm.

In the Dubois DARC system,[14] the difference between the structure of the substrate and that of the product is detected automatically, wherein the structural information is given by connection tables or equivalents (Figure 10). The success of this methodology depends upon the algorithm that gives chemically correct results. For instance, although the Favorskii rearrangment is well-known to have a cyclopropane intermediate, the Dubois system is unsuccessful in detecting this reaction feature. Moreover, recognition of rearrangement or ring contraction is based on reexamination of the molecules participating in this reaction. Thus, this is also caught within the SRT paradigm.

Ugi[15] regards a reaction as an isomeric change between the starting stage and the product stage (Figure 11). When the two stages are given in the form of connection matrices (B and E, respectively), the reaction is represented by a reaction (R)

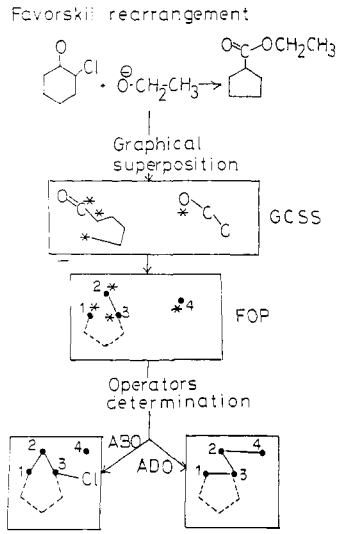


Figure 10. Detection of a reaction center in the DARC.

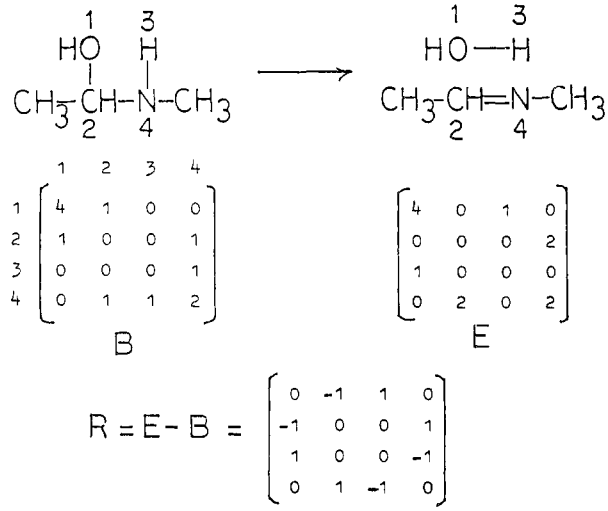


Figure 11. *R* matrix of a dehydration reaction.

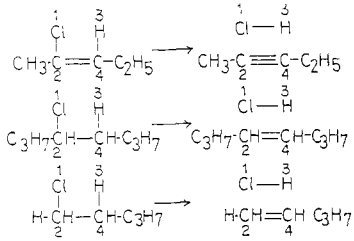


Figure 12. Reactions affording the same *R* matrix as described in Figure 11.

matrix ($R = E - B$). Since the *R* matrix represents the net change of a reaction, the same results are obtained in all examples shown in Figure 12. Thus, the *R* matrix is unable to discriminate double-bond formation from triple-bond formation unless one refers to the starting molecule. This fact stems from the SRT paradigm and is a disadvantage in applying this coding system to the retrieval field. Arens's system (Figure 13)[16] can be regarded as a canonical form of the *R* matrix as pointed out by Brandt.[17] And thus, this system also cannot overcome the SRT paradigm.

IMAGINARY TRANSITION STRUCTURES (ITS) AS UNITARY REPRESENTATION OF ORGANIC REACTIONS

We have proposed the concept of *imaginary transition structures* (ITS), which overcomes the SRT paradigm.[18,19] By

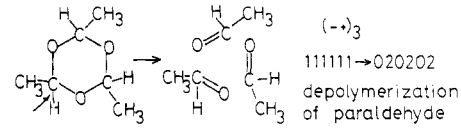


Figure 13. Arens's coding system for describing depolymerization of paraldehyde.

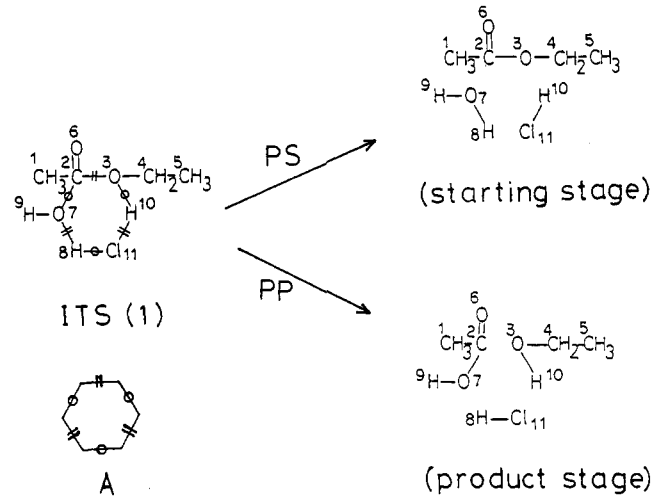


Figure 14. ITS of hydrolysis of ethyl acetate and projection to starting and product stages.

Table I. Types of ITS Bonds

b = -3	-2	-1	0	+1	+2	+3
		(1-1)	(1+0)	(0+1)		
	(2-2)	(2-1)	(2+0)	(1+1)	(0+2)	
(3-3)	(3-2)	(3-1)	(3+0)	(2+1)	(1+2)	(0+3)

this concept, an individual organic reaction is described in itself without discarding any part of the structural information.[20]

The ITS of a given reaction is a structural formula in which molecules of the starting stage are superposed topologically upon those of the product stage, and the bonds are distinguished and classified into three categories (colors), i.e., *out-bonds*, *in-bonds*, and *par-bonds*. The out-bonds (-//-) are bonds appearing only in the starting stage, in-bonds (-O-) appear only in the product stage, and par-bonds (—) are invariant bonds appearing in both stages. For example, the acidic hydrolysis of ethyl acetate is represented by ITS **1**. Hydrolysis of ethyl acetate corresponds to ITS **1** in one-to-one fashion (Figure 14).

The bonds appearing in ITS's are referred to as ITS bonds (Table I). Each ITS bond is denoted by a pair of integers (*a b*) (*a complex bond number*), wherein the integer *a* is the bond multiplicity of the corresponding bond of the starting molecule and the integer *b* is the difference in the bond multplicity between the product and the starting material. The connection table of ITS's is shown in Table II. By extension of the concept of chemical bonds to ITS bonds (Table I), ITS's can be stored and manipulated as usual structural formulas.[21]

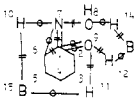
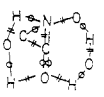
When in-bonds in the ITS are canceled, the starting stage can be regenerated as exemplified in Figure 14. This operation is referred to as the *projection to starting stage* (PS). The PS operation is the displacement of each ITS bond of (*a b*) by the corresponding usual bond of multiplicity *a*.

The *projection to product stage* (PP) is defined as the deletion of out-bonds or displacement of each ITS bond of (*a b*) by a usual bond of multiplicity $a + b$. This operation provides the product stage.

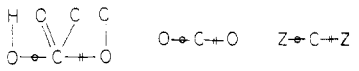
Table II. Connection Table of ITS **1**

node	atom or group	coordinate		neighbor 1		neighbor 2		neighbor 3		neighbor 4	
		x	y	node	$(a\ b)$	node	$(a\ b)$	node	$(a\ b)$	node	$(a\ b)$
1	CH_3	0	0	2	(1+0)						
2	C	200	0	1	(1+0)	3	(1−1)	6	(2+0)	7	(0+1)
3	O	400	0	2	(1−1)	4	(1+0)	10	(0+1)		
4	CH_2	600	0	3	(1+0)	5	(1+0)				
5	CH_3	800	0	4	(1+0)						
6	O	200	200	2	(2+0)						
7	O	200	−200	2	(0+1)	8	(1−1)	9	(1+0)		
8	H	200	−400	7	(1−1)	11	(0+1)				
9	H	0	−200	7	(1+0)						
10	H	400	−200	3	(0+1)	11	(1−1)				
11	H	400	−400	8	(0+1)	10	(1−1)				

Table III. ITS's and Their Subgraphs of Representative Reactions

entry	reaction name	ITS	RC graph	reaction graph
2	Diels–Alder reaction			
3	retro Diels–Alder reaction			
4	Claisen rearrangement			
5	Beckmann rearrangement B = OSO_2OH			

The various subgraphs of ITS's are useful to characterize organic reactions. Three-nodal subgraphs of ITS's are good descriptors of reaction types. For example, *three-nodal subgraph* **2a** is extracted from ITS **1**, wherein its neighbors are



2a 2b 2c

considered. This subgraph (**2a**) represents ester hydrolysis. When the neighbors are omitted, resulting subgraph **2b** is a more general descriptor with respect to substitution of an oxygen group by another oxygen group. When we recognize the two oxygens as heteroatoms (Z) in a more abstract fashion, subgraph **2c** corresponds to a general substitution reaction. Thus, a reaction hierarchy is obtained in terms of subgraphs of ITS's.

A graph of the combined reaction centers is defined as a *graph of reaction centers* (an RC graph). A *reaction graph* is defined as an RC graph in which every node is regarded as balls in an abstract fashion. Table III shows several representative examples.

An ITS, an RC graph, or a reaction graph consists of strings possessing alternate out- and in-bonds. For example, the Diels–Alder reaction (entry 2) gives a hexagon, 1–2+3–4+5–6+1, as such a string. We define this type of string as a *reaction string*. The Beckmann rearrangement (entry 5) has two reaction strings, i.e., 1+9–11+15–10+7–1 and 1+9–11+13–14+8–7+6–1. In general, organic reactions can be classified into one-string, two-string, three-string, ..., and multistring reactions in accord with the numbers of reaction strings.

Reaction graphs shown in Table III (entries 2–4) are regarded as isomers of basic graph **A** modified by the number (m) of double par-bonds and the number (n) of single par-bonds ($m = 0$, $n = 4$). Enumeration of reactions is now replaced by counting isomers, which can be done by the straightforward use of Polya's theorem.[22] Thus, the coefficients of $x^m y^n$ in $G(x,y)$ are the numbers of reaction graphs.

$$G(x,y) = 1 + 2x + 2y + 4x^2 + 6xy + 4y^2 + 6x^3 + 12x^2y + 12xy^2 + 6y^+ + 4x^4 + 12x^3y + 18x^{2,2} + 12xy^+ + 4y^4 + \ldots$$

The ITS's are extended structural formulas that consist of several extended concepts: bonds → ITS bonds, bond multiplicities → complex bond numbers, connection tables → connection tables of ITS's, and so on. In light of these extensions, the manipulation techniques developed for usual structural formulas are applicable to the present ITS's. Table IV examines ring structures appearing in the Beckman rearrangement of entry 5. The counting of ITS bonds of type A ($a + b \neq 0$ and $a \neq 0$), B ($a + b = 0$), and C ($a = 0$) provides a clue for perception of ring opening, ring closure, and rearrangement. Thus, ring 1 is perceived as a bridge of ring opening of order 1 (BO_1), which corresponds to ring opening of a six-membered ring. The presence of ring 2 (bridge of ring closure of order 1 (BC_1)) indicates a ring closure to form a seven-membered ring. The feature of rearrangement is characterized by ring 3 (bridge of rearrangement (BR)).

Wilcox and Levinson[23] have reported another type of unitary representation independently. Their idea is based on a bond-centered labeled graph as illustrated in the case of the Diels–Alder addition (Figure 15, left). The same reaction is represented by the ITS approach for comparison (Figure 15, right).

As can be seen easily, our ITS is more straightforward and may be familiar to organic chemists. Among various advantages of our ITS approach over Wilcox's approach, the following items should be mentioned.

(1) The ITS's embrace usual structural formulas (for compounds) as a subset. Wilcox's bond-centered labeled graphs can be applied to describe organic compounds, but the representations differ from usual structural formulas.

Table IV. Rings Appearing in the ITS of the Beckman Rearrangement of Entry 5

ring	nodes	ring size	A $a + b \neq 0$ and $a \neq 0$	B $a + b = 0$	C $a = 0$	ring type
1	1–2–3–4–5–6–1	6	5	1	0	BO_1
2	1–2–3–4–5–6–7–1	7	6	0	1	BC_1
3	1–6–7–1	3	1	1	1	BR
4	1+9–11+15–10+7–1	6	1	2	3	RS
5	1+9–12+13–14+8–7+6–1	8	1	4	3	RS

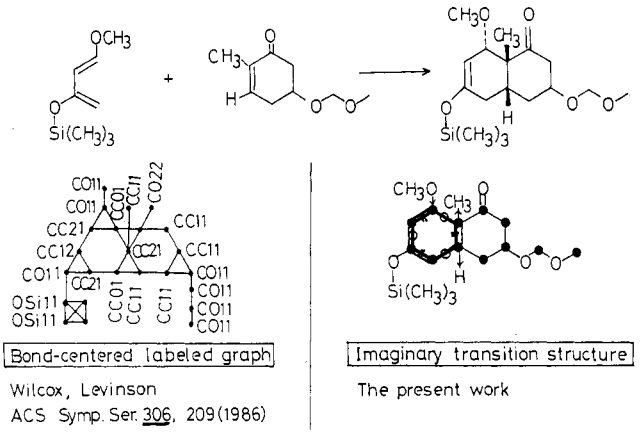


Figure 15. Comparison between Wilcox's bond-centered labeled graph and the present imaginary transition structure.

(2) The counterparts corresponding to PS and PP of our ITS approach may be complex operations in Wilcox's approach, and such operations would afford bond-centered labeled graphs that are different from usual structural formulas.

(3) The enumeration of reactions is replaced by the operation of counting isomers of a basic graph in terms of $G(x,y)$ in the ITS approach. On the other hand, a bond-centered labeled graph requires more complex manipulation to obtain the same results.

(4) The ring structures in a bond-centered labeled graph are not always rings in the corresponding structural formulas. Thus, perception of ring opening, of ring closure, and of rearrangement needs a more complex algorithm than in the ITS approach.

PERSPECTIVES

In *compound* manipulation, a structural formula corresponds to an object (an individual organic compound) in one-to-one fashion. Naming or coding is based on the structural formula, which is regarded frequently as equivalent to the object. The presence of the structural formula as a one-to-one representation provides simplicity and economy of thinking in this field. On the other hand, the lack of such a one-to-one representation has produced much confusion in the field of *reaction* manipulation. Now that ITS is proposed as a unitary representation, naming or coding of organic reactions is based on ITS as an extended structural formula.

The perspectives of the ITS approach are as follows:

(1) The ITS concept forms the common basis of retrieval and design of organic reactions, which are formulated as manipulation of ITS's as extended structural formulas with three colored bonds (Figure 16).

(2) The retrieval of organic reactions is replaced by subgraph or substructure search of ITS's.

(3) Since reaction types are recognized as subgraphs of ITS's, the reaction-type file is constructed automatically as a subfile of the ITS file. As a result, synthetic design obtains a common basis with the retrieval field. Compare this (Figure 16) with the systems described in Figure 1.

(4) The examination of ITS's affords information on reaction features of other categories such as ring opening, ring closure, and rearrangement.

(5) The canonical coding of ITS's would give an effective means for exact matches in registration and retrieval of organic reactions.

(6) Extraction of reaction strings from a reaction graph and examination of behavior of reaction strings would be helpful to understand synthetic pathways.

We have now obtained a sound basis for constructing the integrated computer system FORTUNITS (Fuji Organic Reaction Treating UNity based on Imaginary Transition Structures), which will support both retrieval and synthetic design fields.

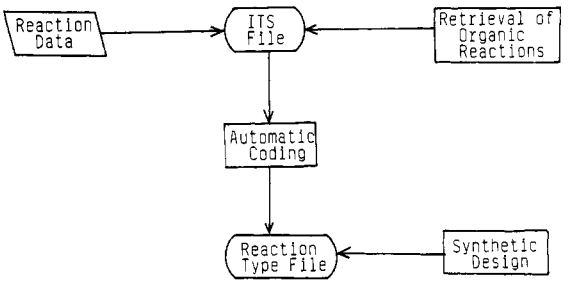


Figure 16. Integrated system of retrieval of organic reactions and synthetic design based on unitary representation of organic reactions.

CONCLUSION

We have pointed out the structure–reaction type (SRT) paradigm that is concealed in the conventional methods of describing organic reactions and hinders the integration of the retrieval and design fields of organic reactions. We have proposed imaginary transition structures (ITS's) to overcome the SRT paradigm.

REFERENCES AND NOTES

(1) Sello, G. "Question of Data Format in Organic Chemistry". *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 249–254.
(2) Fujita, S. "The Description of Organic Reactions". *Yuki Gosei Kagaku Kyokaishi* **1986**, *44*, 354–364.
(3) In an alternative methodology, an individual reaction is described by a combination of (1) the structure of the substrate and (2) that of the product. That is to say, this involves another type of duality, which is essentially equivalent to the SRT paradigm.
(4) The synthetic design has other difficulties than those discussed in this paragraph. For example, retrosynthetic logic would reveal complex and challenging problems. We do not deal with these problems in this paper.
(5) Beach, A. J.; Dabek, H. F., Jr.; Hosansky, N. L. "Chemical Reaction Information Retrieval from Chemical Abstracts Service Publications and Services". *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 149–155.
(6) Finch, A. F. "The Chemical Reactions Document Service. Ten Years On". *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 17–22.
(7) Cohen, B. J. "User-Oriented Approach to a Computerized Organic Reaction Catalog". *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 195–200.
(8) Ziegler, H. J. "Roche Integrated Reaction System (RIRS). A New Documentation System for Organic Reactions". *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 141–149.
(9) Lobeck, M. A. "Use of the IDC System". *Angew Chem., Int. Ed. Engl.* **1970**, *9*, 576–583.
(10) Hendrickson, J. B. "A Systematic Organization of Syntheic Reactions". *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 129–136.
(11) Littler, J. S. "An Approach to the Linear Representation of Reaction Mechanisms". *J. Org. Chem.* **1979**, *44*, 4657–4667.
(12) Various commercially available systems for retrieval of organic reactions (ORAC, SYNLIB, REACCS, etc.) are based probably on two correlated connection tables or the equivalent. See: (a) Johnson, A. P. *Chem. Br.*, **1985**, *21*, 59. (b) Chadosh, D.; Mendelson, W. L. *Dru. Inf. J.* **1983**, *17*, 231. (c) French, S. E. CHEMTECH **1987**, 106.
(13) Roberts, D. C. "A Systematic Approach to the Classification and Nomenclature of Reaction Mechanisms". *J. Org. Chem.* **1978**, *43*, 1473–1480.
(14) Dubois, J.-E. "Computer Assisted Modelling of Reactions and Reactivity". *Pure Appl. Chem.* **1981**, *53*, 1313–1327.
(15) Ugi, I.; Bauer, J.; Brandt, J.; Freidrich, J.; Gasteiger, J.; Jochum, C.; Schubert, W. "New Application of Computer in Chemistry". *Angew. Chem., Int. Ed. Engl.* **1979**, *18*, 111–123.
(16) Arens, J. F. "A Formalism for the Classification and Design of Organic Reactions. I. The Class of $(-+)_3$ Reactions". *Recl. Trav. Chim. Pays-Bas* **1979**, *98*, 155–161.
(17) Brandt, J.; von Scholley, A. "An Efficient Algorithm for Computation of the Canonical Numbering of Reaction Matrices". *Comput. Chem.* **1983**, *7*, 51–59.
(18) (a) Fujita, S. "The Description of Organic Reactions Based on Imaginary Transition Structures". Presented at the 52nd Annual Meeting of the Chemical Society of Japan (Kyoto), April 1–4, 1986. (b) Fujita, S. "Imaginary Transition Structures. Unitary Representations of Organic Reactions". Presented at the 192nt ACS National Meeting, Anaheim, CA, Sept 7–12, 1986. (c) Fujita, S. "Description of Organic Reactions Based on Imaginary Transition Structures. 1. Introduction of New Concepts". *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 205. (d) Fujita, S., "2. Classification of One-String Reactions Having an Even-Membered Cyclic Reaction Graph". *J. Chem. Inf. Comput. Sci.*

1986, 26, 212. (e) Fujita, S. "3. Classification of One-String Reactions Having an Odd-Membered Cyclic Reaction Graph". J. Chem. Inf. Comput. Sci. 1986, 26, 224. (f) Fujita, S. "4. Three-Nodal and Four-Nodal Subgraphs for a Systematic Characterization of Reactions". J. Chem. Inf. Comput. Sci. 1986, 26, 231. (g) Fujita, S. "5. Recombination of Reaction Strings in a Synthesis Space and Its Application to the Description of Synthetic Pathways". J. Chem. Inf. Comput. Sci. 1986, 26, 238. (h) Fujita, S. "6. Classification and Enumeration of Two-String Reactions with One Common Node". J. Chem. Inf. Comput. Sci., first of five papers in this issue. (i) Fujita, S. "7. Classification and Enumeration of Two-String Reactions with Two or More Common Nodes". J. Chem. Inf. Comput. Sci., second of five papers in this issue. (j) Fujita, S. "8. Synthesis Space Attached by a Charge Space and Three-Dimensional Imaginary Transition Structures with Charges". J. Chem. Inf. Comput. Sci., third of five papers in this issue. (k) Fujita, S. "9. Single-Access Perception of Rearrangement Reactions". J. Chem. Inf. Comput. Sci., fourth of five papers in this issue. See also: Chem. Eng. News 1986, 64(39), 75.

(19) The term "imaginary" transition structure stems from the analogy of imaginary numbers which are counterparts of real numbers. The ITS may be a real transition structure when we consider a concerted reaction, and so, the term "complex" transition structures would be more suitable if we consider that complex numbers contain real and imaginary numbers. However, the term "complex" has been used widely in the chemical field. Therefore, we have adopted "imaginary" transition structures in our case.

(20) After submission and acceptance of this paper, the author heard Vladutz's report: Vladutz, G. In "Modern Approaches to Chemical Reaction Searching". Willet, P., Ed., Gower: Aldershot, U.K., 1986; p 202. The author thanks Dr. Vladutz for sending his article. See also: Chem. Eng. News 1987, 65(6), 2. The main differences between Vladutz's *superposed reaction graph* (SRG) and our ITS are as follows. (a) The SRG contains no catalysts even if they participate the reactions. Thus, the SRG counterpart of the reaction of Figure 14 does not contain hydrochloric acid. (b) The ionic character of a bond is represented by the concept of "charge space" in our ITS approach.[18] On the other hand, an ionic SRG and the corresponding nonionic SRG are not integrated.

(21) Information on stereochemistry can be treated by three-dimensional ITS's, which will be discussed elsewhere.

(22) Balaban, A. T., Ed. *Chemical Application of Graph Theory*; Academic: London, 1976.

(23) Wilcox, C. S.; Levinson, R. A. In *Artificial Intelligence. Applications in Chemistry*; Pierce, T. H., Hohne, B. A., Eds.; ACS Symposium Series 306; American Chemical Society: Washington, DC, 1986; pp 209-230.

Computer Storage and Retrieval of Generic Chemical Structures in Patents. 8. Reduced Chemical Graphs and Their Applications in Generic Chemical Structure Retrieval†

VALERIE J. GILLET, GEOFFREY M. DOWNS, AI LING, MICHAEL F. LYNCH,* PALLAPA VENKATARAM, and JENNIFER V. WOOD

Department of Information Studies, University of Sheffield, Sheffield S10 2TN, U.K.

WINFRIED DETHLEFSEN

BASF, Ludwigshafen am Rhein, Federal Republic of Germany

Received February 18, 1987

Reduced chemical graphs for specific chemical substances comprise summary descriptions of the gross structural features of these substances; an example is summarization in terms of only the ring and nonring components, giving a tree structure in which each node is either a cyclic or an acyclic component. Other bases for graph reduction are possible, including those in which the components are formed from the separate aggregates of carbon atoms and of heteroatoms. The reduced graph of a generic chemical structure is usually a multigraph, in which the identities of the variables are summarized in similar terms. The varieties and distributions of several types of reduced graphs created from almost 50 000 specific chemical substances in the Fine Chemicals Directory are characterized. The data provide qualitative guidance on the power of reduced graphs as retrieval keys when a database of generic structures described in this way is searched for queries that are complete specific or generic structures. The performance of several types of reduced graph, taken both singly and in combination, as retrieval keys for searches of generic chemical structures is reported. The test database is a small set of generic structures in which the variables are specific partial structures; the queries comprise both specific structures from patents and the generic structures of the test database itself. The results confirm the potential of reduced chemical graphs for high performance.

INTRODUCTION

Substantial progress has already been reported by us in the development of methods of representing generic chemical structures in machine-readable form for retrieval purposes.[1-10] It was evident from the beginning of our work that a powerful tool kit combining a variety of search representations, search algorithms, and high-performance hardware would be necessary in order to provide workable and economic solutions to searching in the most general sense. Our approaches to search representations have, thus far, been "bottom-up" in design; i.e., we have sought to describe the generic structures in terms of atoms and bonds and their aggregates, in much the same way as the screens employed for substructure searching of databases of specific substances are derived and used. These and similar search screens have played and will continue to play an important role in our work since they provide the most generally applicable description of generic structures for the range of types of searches required.

Two important classes of searching in files of generic chemical structures are those that involve complete structures; i.e., the queries are either specific structures or generic structures. When the query is a specific structure, the purpose is to discover which file structures include the query. When

* Author to whom correspondence should be addressed.

† Paper presented at the Division of Chemical Information Symposium on Generic Chemical Structure Searching, 192nd National Meeting of the American Chemical Society, Anaheim, CA, Sept 9, 1986.

0095-2338/87/1627-0126$01.50/0