Further Development of a Reaction Generator in the SOPHIA System for Organic Reaction Prediction. Knowledge-Guided Addition of Suitable Atoms and/or Atomic Groups to Product Skeleton

Hiroko Satoh and Kimito Funatsu*

Department of Knowledge-Based Information Engineering, Toyohashi University of Technology,
Tempaku Toyohashi 441, Japan

Received June 19, 1995[⊗]

A reaction generator in SOPHIA (System for organic reaction prediction by heuristic approach) generates all possible product structures by reconnections of free bonds obtained by cutting all bonds of a reaction site which is automatically perceived in the input reactantt structure and additions of atoms and/or atomic groups (AAG) to the free bonds. The reaction generator has been extended to automatically recognize and add suitable AAG for suitable free bonds by utilizing knowledge base derived from a reaction database. Contents of the knowledge base are structural information of AAG and structural characteristics of sites and their environments to which the AAG are added during reactions in a reaction database. The reaction generator considers reaction conditions to recognize suitable AAG for suitable free bonds by utilizing reaction condition groups obtained by classification based on word combinations of reaction condition descriptions in a reaction database. SOPHIA also has been extended to employ the reaction condition groups for interpretation of reaction conditions entered by the user and to consider the reaction conditions in reaction prediction procedures. This paper describes the contents and structure of the knowledge base, the reaction condition classification, and implementation of the reaction prediction using these results.

1. INTRODUCTION

A reaction prediction system SOPHIA (System for organic reaction prediction by heuristic approach) has been being developed to predict possible products and the product ratio from arbitrary reactants and under arbitrary reaction conditions. The SOPHIA system is free from reaction categories (e.g., electrophilic reactions under acidic catalyst and nucleophilic reactions under basic catalyst) and reaction types (e.g., Diels−Alder reaction and Michael reaction) in its reaction prediction procedure. SOPHIA does not establish the reaction categories or reaction types and does not require the user to specify them for reaction prediction. This is because chemical reactions may not occur only inside specified categories or types, and SOPHIA has been designed to consider this nature of chemical reactions. This point is essentially different from the other reaction prediction systems like EROS6.0[1] and CAMEO.[2] The SOPHIA's philosophy, overview, and first results were described in a previous paper.[3] The previous SOPHIA had been realized to automatically predict all possible reaction paths from arbitrary input reactants without user's designation of a specific reaction type or category, although quantitative prediction had not been realized yet.

The reaction prediction procedure of the previous SOPHIA is shown in Figure 1. The first step is automatic perception of a reaction site of the input reactants under input reaction conditions (Figure 1-[1]). Specification of reaction conditions is not always indispensable. The perception employs a reaction knowledge base derived from a reaction database. The reaction knowledge base stores reaction schemes characterized by the location of key substructures around reaction sites of the reactant and product structures and by type of reaction conditions. Next procedure is reaction generation (Figure 1−[2]). A reaction generator cuts all bonds of the perceived reaction site to give free bonds and generates all possible product structures by reconnecting the free bonds and adding atoms and/or atomic groups to the free bonds. The system evaluates whether each of these generated product structures can actually form (Figure 1-[3]). The reaction evaluation also employs the reaction knowledge base.

A role of the reaction generator is giving a *concrete* form to the reaction knowledge in which reactions are *abstracted* by structural characteristics of and around the reaction site.[3] This manner of reaction generation is quite different from that found in the CAMEO and the EROS6.0 systems, in which product structures are generated by transformation of substructures in reactant structures according to chemical reaction knowledge and rules.

Most of the operations in the reaction generator in SOPHIA which was presented in the previous paper have been automatically carried out. However, specification of atoms and/or atomic groups (AAG) for addition to the free bonds were left to the user. If the user had no idea of suitable AAG (SAAG) to be added, the reaction generator just added dummy atoms to the free bonds. However, this specification of the AAG by the user brings some inconveniences: First, a product structure having dummy atoms cannot be treated in the reaction evaluation procedure. This is because computation of structural characteristics is incomplete for a bond to which a dummy atom is attached, and this *incomplete* information about structural characteristics of a reaction site and its environment [Structural Characteristics of *a reaction site* and its Environment: SCE of a reaction site] cannot be correctly matched with the reaction knowledge base, which has *complete* information about SCE of a reaction site.

[⊗] Abstract published in *Advance ACS Abstracts,* January 15, 1996.

0095-2338/96/1636-0173$12.00/0 © 1996 American Chemical Society

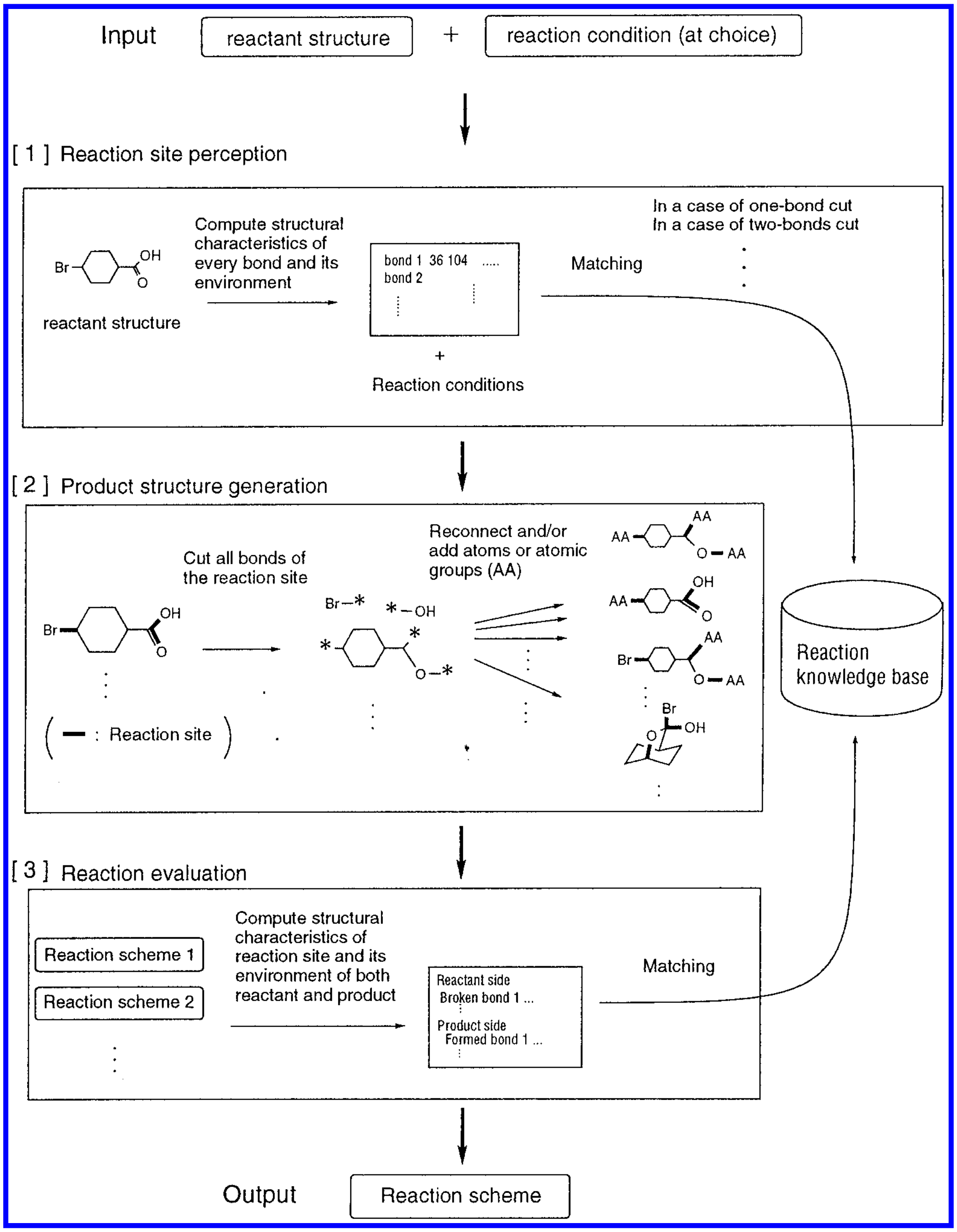

Input
reactant structure
+
reaction condition (at choice)
[1] Reaction site perception
reactant structure
Compute structural characteristics of every bond and its environment
bond 1 36 104 .....
bond 2
+
Reaction conditions
Matching
In a case of one-bond cut
In a case of two-bonds cut
[2] Product structure generation
Cut all bonds of the reaction site
Reconnect and/or add atoms or atomic groups (AA)
Reaction site
Reaction knowledge base
[3] Reaction evaluation
Reaction scheme 1
Reaction scheme 2
Compute structural characteristics of reaction site and its environment of both reactant and product
Reactant side
Broken bond 1 ...
Product side
Formed bond 1 ...
Matching
Output
Reaction scheme

Figure 1. Block diagram of SOPHIA system.

Secondly, the specification means that an idea of the user is reflected in the output from the system. An idea of the user is useful for a synthesis design system but not for a reaction prediction system. This is due to the difference of roles required of these systems. A role of a synthesis design system is to propose a part of possible synthesis routes giving desired target molecules by application of the system's own synthesis strategies and thus better output will be expected by employing the user's synthesis strategies. Actually, the LHASA system,[4] which is designed to maximize the benefits of interaction with the user, makes it possible to propose solutions to various synthesis problems according to suitable strategies for the target to be analyzed. On the other hand, a reaction prediction system must predict all possible reaction products and the product ratio from given reactants under specific reaction conditions. Therefore, every reaction prediction system is absolutely required to give the same answer. Accordingly, it is necessary for a reaction prediction system to run with only entries of reactant structures and reaction conditions by the user, although during development of system the user may cover weak points of the system. The user's idea should not be reflected in output from a reaction prediction system. Thus, a function for automatic determination and additions of SAAG was required for more accurate reaction prediction in SOPHIA. We have developed such function and describe it in what follows.

The function utilizes knowledge derived from a reaction database. Contents of the knowledge are structural information of AAG and SCE of sites to which the AAG are added during reactions in a reaction database. One storing the former information is called an AAG-database (AAG-DB), and another one storing the latter information is called an AAG-knowledge base (AAG-KB).

Furthermore, we have examined how to consider reaction conditions in reaction prediction. Reaction conditions (e.g., solvents, reagents, catalysts, temperature) are important factors controlling a route of reaction. The same reactants may produce different products or product ratio under different reaction conditions. Appropriate consideration of reaction conditions is indispensable for accurate reaction prediction. The reaction knowledge base of SOPHIA in the previous paper describes reaction conditions by reaction condition categories summarized by Greene:[5] these are based on functional group protection in organic synthesis.[3] However, such classification, in which a specific factor (in this case, protective groups) is emphasized, is not suitable for the purpose of SOPHIA aiming to predict reactions without specification of conventional reaction categories or types. For accomplishing this, it is ideally desired to classify reaction conditions by considering correlation among changes of electronic features during reaction, transformations of structural characteristics during reaction, and reaction conditions. It is important to do this classification free from conventional reaction categories or types.

As a first step of this ideal classification, we have classified reaction conditions based on word combinations of reaction condition descriptions in a reaction database. Here, we have not yet taken into account the correlation mentioned above.

SOPHIA has been extended to employ the results of this reaction condition classification to interpret data input by the user; SOPHIA considers as well these reaction conditions for prediction procedures such as the reaction site perception, determination and addition of the SAAG in the reaction generation, and the reaction evaluation.

Section 2 describes the contents and structure of an AAG-DB and an AAG-KB, and section 3 describes the reaction condition classification. Section 4 describes their utilization in reaction prediction procedures. Furthermore, a relationship between the reaction condition classification and quantitative prediction of product ratio is discussed in section 6.

2. CONSTRUCTION OF AN AAG-DB AND AN AAG-KB

2.1. Method. SOPHIA requires a theory which could exactly predict chemical reaction paths of an arbitrary reactant under an arbitrary reaction condition. But such a general theory has not been established yet because of complexity of chemical reactions. In order to predict chemical reactions without a backing general theory we have investigated a method for deriving knowledge or rules leading to the general theory from a reaction database, which potentially holds a large quantity of information about chemical reactions. SOPHIA has already taken this knowledge-guided approach in reaction site perception and reaction evaluation.

We take this approach also for automatic determination and addition of SAAG for free bonds of the product skeleton: we utilize a reaction database from which knowledge for the procedures was derived. In knowledge derivation from a reaction database, we define AAG as *atoms and/or atomic groups present in reactant structures and absent in product structures*. This definition is opposite to that of leaving groups in an organic synthesis design system AIPHOS.[6,7] Leaving groups are defined as *atoms and/or atomic groups present in product structures and absent in reactant structures*. The reason why we define AAG and leaving groups as above is that a commercially available reaction database, in general, does not store information about atoms and/or small atomic groups being appended to reactant to give product and small molecules (e.g., H_2O and H_2) leaving from reactant to give product or stores them just in reaction condition descriptions. These definitions allow the system to automatically recognize the AAG and leaving groups which are not described as reactant and product molecules in a reaction within a reaction database.

According to the AAG definition, an AAG-DB and an AAG-KB have been automatically derived from a reaction database.

2.2. Contents and Structure. We show contents and structure of the derived AAG-DB and AAG-KB. Through this paper, bonds broken or formed during a reaction are called together *reaction bonds*. Among them, broken ones are called *RR-bonds* (reactant reaction bonds). The formed ones are called *PR-bonds* (product reaction bonds). A site composed by all RR-bonds is called *RR-site*, and a site composed by all PR-bonds is called *PR-site*. A more detailed explanation of this definition can be found in the previous papers.[3]

2.2.1. AAG-DB. We used the reaction database[8] for the AIPHOS system to derive the AAG-DB. We also used the AIPHOS reaction database to derive the reaction knowledge base used in the reaction site perception and reaction evaluation in SOPHIA. The AIPHOS's reaction database was constructed using reactions of the SYNLIB reaction database.[9]

When a reaction database has no atom-mapping between reactant and product structures, it is difficult to recognize AAG according to the AAG definition. But the AIPHOS's reaction database was constructed with the atom-mapping resulting from the set of Manipulation Instructions of the SYNLIB, and the difficulty is not found.

Contents of the AAG-DB are the structural information (atomic species, the connectivity, and the coordinates) of AAG, ID nos. of AAG, and ID nos. of the individual reactions from which the AAG were extracted. For example, −H and −OH are recognized as AAG of an individual reaction shown on top of Figure 2 (the AAG are shown within rectangles in the product structure, here, bold lines in reactant and product structures stand for RR- and PR-bonds to which the AAG are added). The lower right of Figure 2 shows stored information: structural information of the AAG, ID numbers of the AAG (nos. 10 and 11), and ID number of the individual reaction (no. 55). The ID numbers are only reference numbers and have no specific meanings. Actually, data of the AAG−DB are stored in a direct access file, and the record numbers correspond to ID nos. of AAG.

The AAG-DB is automatically derived from a reaction database.

2.2.2. AAG-KB. The AAG-KB is derived from the same reaction database which is used for deriving the AAG-DB. We used the same AIPHOS reaction database to derive the AAG-KB.

The AAG-KB is organized into two direct access files for reactant and product.

The reactant file stores SCE of RR-bond concerning AAG addition. The SCE of RR-bond are perceived based on synthetically important substructures established in the system. These substructures are called structural characteristic keys, and some of them are shown in Table 1. This method of characterization is the same as for a reaction site in the reaction knowledge base that SOPHIA employs in reaction site perception and reaction evaluation.

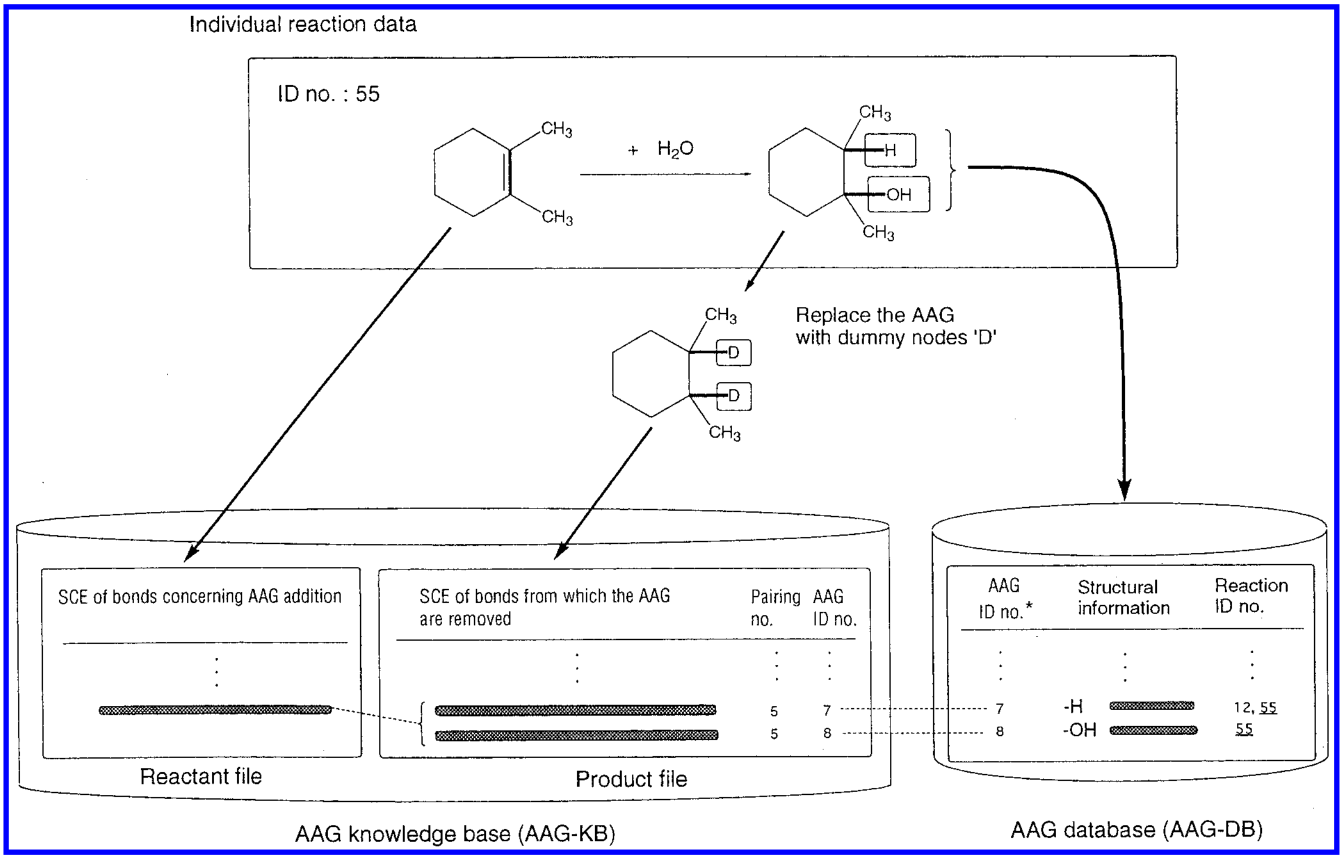
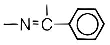
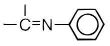
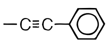
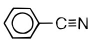
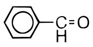
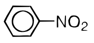


Figure 2. Construction of AAG-KB and AAG-DB.

Table 1. Part of the Structural Characteristics Keys in the Reaction Knowledge Base

Key No.	Structural Characteristics	Key No.	Structural Characteristics
1	$-CH_3$	77	
2	$-CH_2R$		
3	$-CHR_2$	78	
4	$-CR_3$		
5	$-CH(CH_3)_2$	79	
6	$-C(CH_3)_3$		
7	$-F$	80	
•	•	•	•
•	•	•	•
•	•	•	•
33	$-\underset{\underset{O}{\|}}{P}(OR)_2$		
34	$-\underset{\underset{O}{\|}}{C}R$	100	
35	$-\underset{\underset{O}{\|}}{C}H$	101	
•	•		
•	•	102	
•	•		
48	$-\underset{H}{N}-\underset{\underset{O}{\|}}{C}-NH$	103	
49	$-\underset{\underset{O}{\|}}{C}-S-$	•	•
		•	•
50	$-SO_2OH$	108	
•	•	•	•
•	•	•	•
•	•	111	

The product file stores a pairing number, the AAG ID number, and SCE of PR-bond to which AAG are attached. The pairing number is the ID number of *pairing AAG*, and a definition of the pairing AAG is described in the next paragraph. The method for PR-bond characterization is the same as that for RR-bond in the reactant file, differing only in treatment of the dummy atoms attached to the PR-bond.

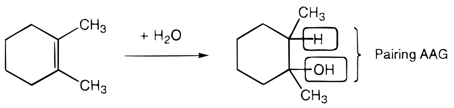


Figure 3. Pairing AAG.

The product file is compared with free bonds, which are bonds with dummy atoms, of a product skeleton to determine and add SAAG for the free bonds. So the product file stores SCE of PR-bonds in which AAG are replaced with dummy atoms. Details of the manner of SAAG determination and addition are described in section 4. When multiple AAG are recognized in a product structure, structural characteristics are computed for each product skeleton (product structure having dummy atoms) from which each of the AAG is removed and all AAG are removed.

Pairing AAG are a group of AAG related to each other in electron shift when they are added to a product skeleton. A simple example of the pairing AAG is shown in Figure 3. Here, a pair of AAG $-H$ and $-OH$ is added to a double bond. Other types of pairing AAG are also considered like intramolecular multiple AAG between which the distance is more than vicinal and intermolecular multiple AAG.

Data in the product and reactant files are linked by their record numbers.

Figure 2 shows an example of the AAG-KB derivation from the same individual reaction as was used for explanation of AAG-DB derivation. The lower left of Figure 2 shows knowledge derived from this reaction. In this example, the AAG are $-H$ and $-OH$, which are pairing AAG. The reactant file stores the SCE of RR-bond concerning the pairing AAG (double bond of reactant structure). The product file stores the pairing no. 5, the AAG ID nos. 7 and 8, and SCE of PR-bond to which dummy atoms are attached (bold line of a chemical structure at the center of Figure 2). Again, the pairing number and the AAG ID number are only reference numbers and have no specific meanings.

The AAG-KB is also automatically derived from a reaction database.

As mentioned above, all files organizing the AAG-DB and -KB are direct access files, and the data in the files are linked with each other by the record numbers. This allows the system rapid access to desired information in SAAG determination and addition.

3. REACTION CONDITION CLASSIFICATION

3.1. Method. We have classified reaction conditions based on combinations of words of reaction condition descriptions of SYNLIB. An overview of the classification is described here, and details of it are available as a supporting information.

Steps of the classification are (1) elimination of multiple-step reactions, (2) editing reaction condition descriptions, (3) word extraction, (4) construction of word lists and word dictionary, and (5) classification of reaction conditions. Contents of these operations are described below.

(1) Elimination of Multiple-Step Reactions. Multiple-step reactions are not useful from which knowledge and rules for reaction prediction are derived. We eliminated multiple-step reactions from SYNLIB.

(2) Editing Reaction Condition Descriptions. Reaction condition descriptions often contain a variety of comments which are not reaction conditions, e.g., “7:1 THIS ISOMER”, “VARIOUS AROMATIC ALDEHYDES ALSO OK”, and often contain descriptions of alternative reaction conditions, e.g., “$SNCL_4$ or $TICL_4$” and “KOH or ETOH”. We edited out the former type of comments and divided the latter descriptions.

(3) Word Extraction. We extracted words from the edited descriptions.

(4) Construction of Word Lists and Word Dictionary. We collected the extracted words to make a word list and classified them on the basis of the kinds of the words. We constructed four kinds of word lists of compounds, temperature, time, and concentration. Since solvents, reagents, catalysts, and so on are not distinguishable from each other in the reaction condition descriptions, they are collected as compounds.

We prepared a dictionary by collecting synonyms of the extracted words. For example, “AcOH”, “HOAc”, “CH_3-COOH”, and “acetic acid” were collected and written into the dictionary.

(5) Classification of Reaction Conditions. We classified the reaction conditions based on the word combinations in the reaction condition descriptions edited in (2). The word lists were used for investigating the word combinations. The investigation used the dictionary of synonyms. Individual reactions having the same word combination in their reaction condition descriptions were regarded to link to the same reaction condition group. The classification was done by all combinations of four word lists (15 combinations) shown in Table 2. This allows rapid reaction prediction processing.

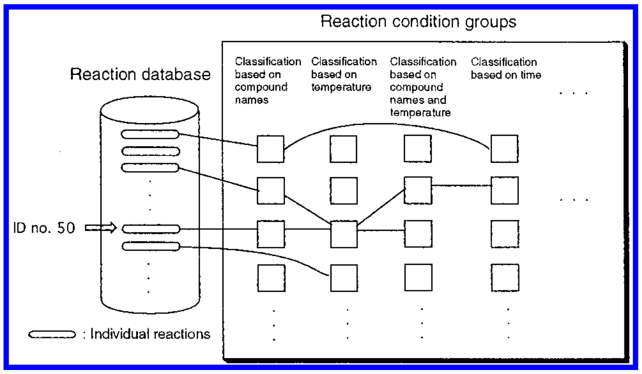

Reaction condition groups
Reaction database
Classification based on compound names
Classification based on temperature
Classification based on compound names and temperature
Classification based on time
. . .
. . .
ID no. 50
: Individual reactions

Figure 4. Overview of the result of reaction condition classification.

Table 2. Reaction Condition Classification Using 15 Ways of Combining Kinds of Word Lists

(a)	classification using word lists for compounds, temperature, time, and concentration
(b)	temperature, time, and concentration
(c)	compounds, temperature, and time
(d)	compounds, temperature, and concentration
(e)	compounds, time, and concentration
(f)	temperature and time
(g)	temperature and concentration
(h)	compounds and temperature
(i)	time and concentration
(j)	compounds and time
(k)	compounds and concentration
(l)	temperature
(m)	time
(n)	concentration
(o)	compounds

3.2. Result. The result of the classification of reaction conditions is a network-like structure in which individual reactions and the reaction groups are linked to each other (Figure 4). For example, a reaction (ID no. 50) having as reaction conditions the compound name and temperature links to three types of reaction condition groups: (1) the reaction condition group obtained using a word list of compound names (Table 2-(o)), (2) the reaction condition group obtained using a word list for temperatures (Table 2-(l)), and (3) the reaction condition group obtained using word lists for the compounds and the temperatures (Table 2-(h)).

4. IMPLEMENTATION OF REACTION PREDICTION

SOPHIA has been improved in the reaction site perception and reaction generation. The reaction site perception has been extended so as to utilize the reaction condition groups for interpretation of input reaction conditions and to consider the reaction conditions in reaction prediction procedures. The reaction generator also has been extended so as to automatically receive suggestions from the AAG-DB, the AAG-KB, and the reaction condition groups and to automatically determine and add SAAG for free bonds of the product skeleton. Other procedures have not been modified since the previous paper.[3]

This section focuses on the extensions and shows reaction prediction in SOPHIA.

4.1. Input Data. Input data are chemical structures of reactants and a set of reaction conditions. Reaction conditions are not indispensable, with means that their input is optional.

When the user wishes to specify reaction conditions, she/he must select one of the following two ways of specification. One is a concrete specification of reaction conditions, and the other is the selection of only some items of reaction conditions.

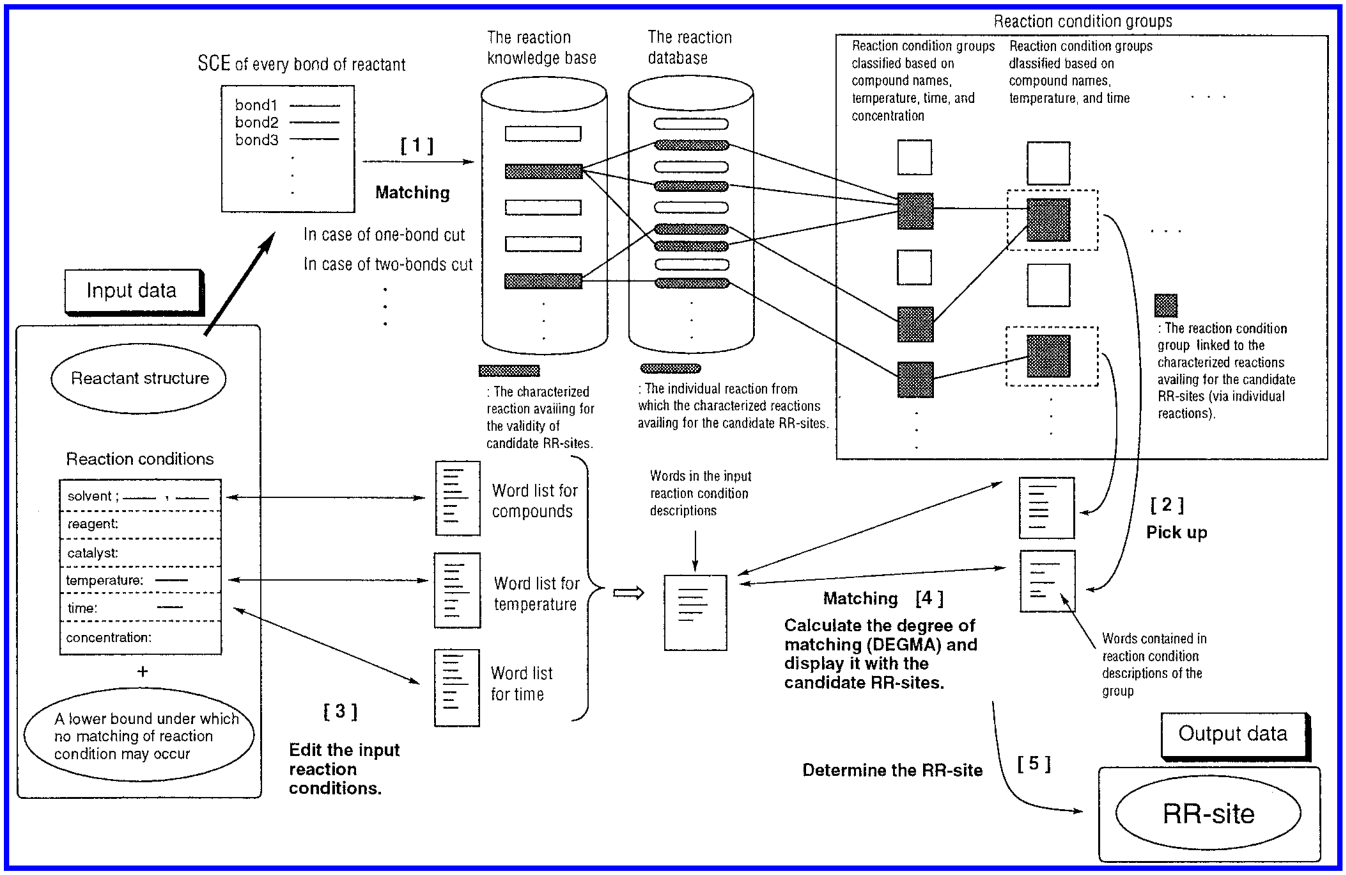

SCE of every bond of reactant
bond1
bond2
bond3
[1]
Matching
In case of one-bond cut
In case of two-bonds cut
Input data
Reactant structure
Reaction conditions
solvent ;
reagent:
catalyst:
temperature:
time:
concentration:
+
A lower bound under which no matching of reaction condition may occur
The reaction knowledge base
The reaction database
: The characterized reaction availing for the validity of candidate RR-sites.
: The individual reaction from which the characterized reactions availing for the candidate RR-sites.
Reaction condition groups
Reaction condition groups classified based on compound names, temperature, time, and concentration
Reaction condition groups classified based on compound names, temperature, and time
. . .
: The reaction condition group linked to the characterized reactions availing for the candidate RR-sites (via individual reactions).
Word list for compounds
Word list for temperature
Word list for time
[3]
Edit the input reaction conditions.
Words in the input reaction condition descriptions
Matching [4]
Calculate the degree of matching (DEGMA) and display it with the candidate RR-sites.
[2]
Pick up
Words contained in reaction condition descriptions of the group
Output data
Determine the RR-site
[5]
RR-site

Figure 5. RR-site perception procedure extended to employ the novel reaction condition classification (when concrete reaction conditions are specified by the user).

In the former the user can freely specify reaction conditions of the six items related to solvent, reagent, catalyst, temperature, time, and concentration by typing them in free format. In this case all of them may not be specified. The user may also specify a lower bound under which no matching of reaction conditions may occur. The system uses the lower bound to automatically judge whether the input reaction conditions are matched with the reaction condition groups or not. When the lower bound is not specified, the judgment is left to the user.

In the latter the user can select any of the six reaction condition items, which she/he wishes to take into account for reaction prediction.

Different ways of reaction condition specification lead to different reaction prediction procedures and results, as will be seen below.

4.2. RR-Site Perception. Reaction prediction processing begins with RR-site perception for the input reactant structure. This procedure takes account of reaction conditions in three cases: (A) concrete reaction conditions are specified by the user, (B) there is no concrete formulation and only some reaction condition items are selected by the user, and (C) neither concrete contents nor items of reaction conditions are specified by the user, as shown below.

(A) The Reaction Conditions Are Specified Concretely (Figure 5). (1) Perception of Candidates of RR-Sites (Figure 5-[1]). Possible combinations of RR-bonds are perceived as candidates of RR-sites based on the SCE of each bond of the input reactant structures. The perception uses the reaction knowledge base. As mentioned in the Introduction (section 1), the reaction knowledge base stores reactions characterized by SCE of the RR- and PR-sites. The knowledge on the RR-site is applied to the RR-site perception. The previous paper presented a method of the characterization of reactions, a perception procedure of the SCE of the bonds being processed, and a way to refer to the reaction knowledge base.[3]

The shaded rectangles within a box entitled “the reaction knowledge base” in Figure 5 stand for the characterized reactions that avail for the validity of the perceived candidate RR-sites.

(2) Picking up Reaction Condition Groups (Figure 5-[2]). The system selects the reaction condition groups related to both the input reaction conditions and the perceived RR-side candidates. First, the system chooses reaction condition groups linked to the characterized reactions availing for the perceived RR-site candidates (they are linked via individual reactions). Here, in Figure 5, the linkage between each of the individual reactions and each of the characterized reactions shows that the characterized reaction was derived from the individual reaction, and then from the chosen reaction groups the system picks up those containing the same types of descriptions as those of the items at the input.

For example, when the user enters reaction conditions containing items such as solvent, temperature, and time, the system picks up reaction condition groups containing descriptions referred to compound, temperature, and time. This is illustrated in Figure 5 as groups within rectangles drawn with dotted lines. This reaction condition classification corresponds to Table 2-(c). The reason why solvents are regarded as compounds is described in section 3.

(3) Editing Input Reaction Conditions (Figure 5-[3]). The system edits the input reaction conditions by extracting words contained in the word lists mentioned in section 3.

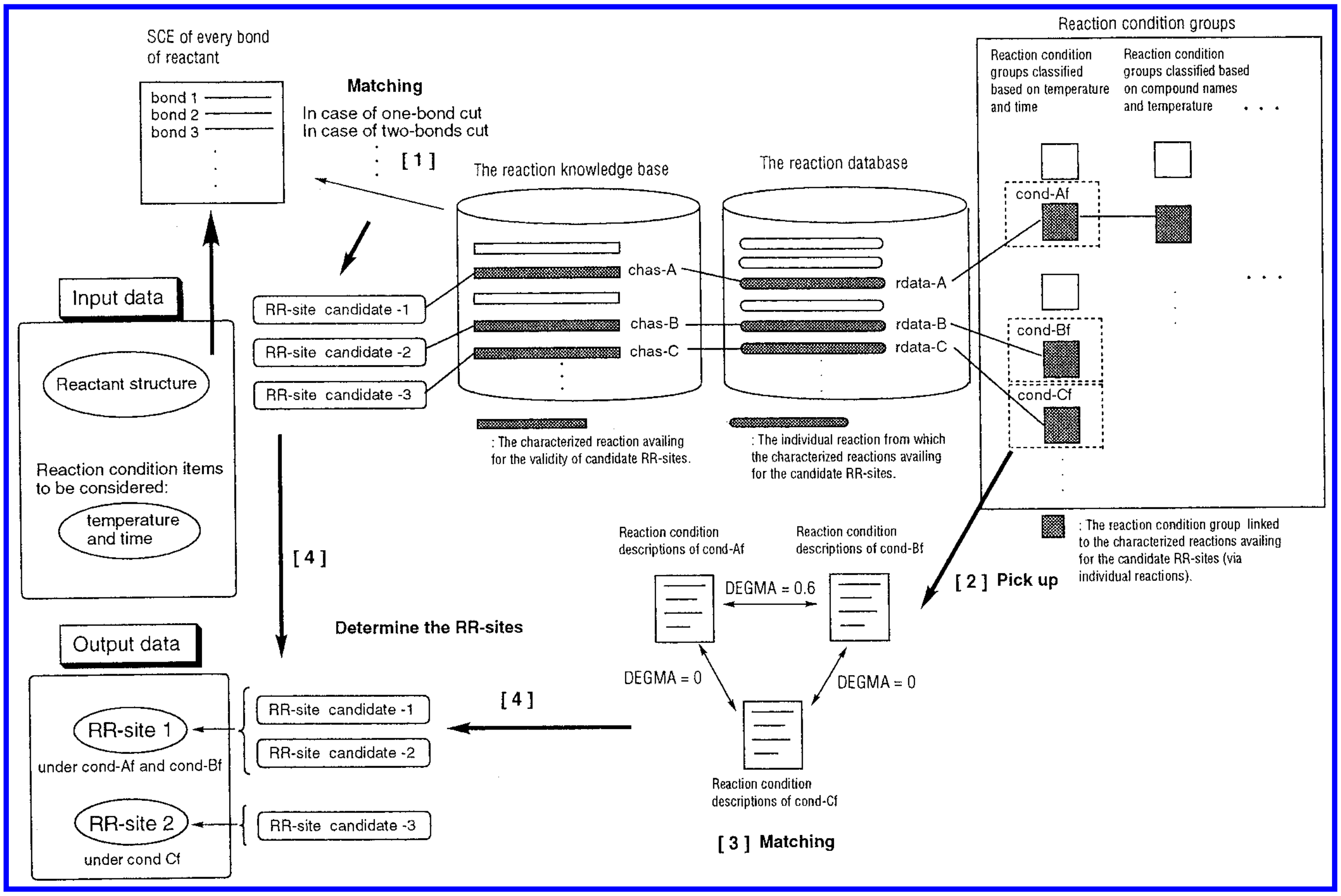

SCE of every bond of reactant
bond 1
bond 2
bond 3
Matching
In case of one-bond cut
In case of two-bonds cut
[1]
The reaction knowledge base
The reaction database
Reaction condition groups
Reaction condition groups classified based on temperature and time
Reaction condition groups classified based on compound names and temperature
cond-Af
cond-Bf
cond-Cf
Input data
Reactant structure
Reaction condition items to be considered:
temperature and time
RR-site candidate -1
RR-site candidate -2
RR-site candidate -3
chas-A
chas-B
chas-C
rdata-A
rdata-B
rdata-C
: The characterized reaction availing for the validity of candidate RR-sites.
: The individual reaction from which the characterized reactions availing for the candidate RR-sites.
: The reaction condition group linked to the characterized reactions availing for the candidate RR-sites (via individual reactions).
[4]
Reaction condition descriptions of cond-Af
Reaction condition descriptions of cond-Bf
DEGMA = 0.6
[2] Pick up
Determine the RR-sites
Output data
DEGMA = 0
DEGMA = 0
[4]
RR-site 1
under cond-Af and cond-Bf
RR-site 2
under cond Cf
RR-site candidate -1
RR-site candidate -2
RR-site candidate -3
Reaction condition descriptions of cond-Cf
[3] Matching

Figure 6. RR-site perception procedure extended to employ the novel reaction condition classification (when only reaction condition items are selected by the user).

The system also standardizes the notations using the dictionary mentioned in section 3. For example, words of “CH3OH” and “MeOH” are standardized to “Methanol”.

(4) Matching between the Reaction Condition Groups and the Input Reaction Conditions (Figure 5-[4]). The system compares descriptions of each of the reaction condition groups picked up in (2) with the input reaction condition descriptions edited in (3). The degree of match (DEGMA) is calculated

$$\mathrm{DEGMA} = 2\mathrm{NM}/(\mathrm{NG} + \mathrm{NU})$$

here, NM stands for the number of matched words, NG stands for the number of words contained in the reaction condition group being compared, and NU stands for the number of words extracted from input reaction conditions in (3). In this comparison, the system judges that two words are “matched” only when they are exactly the same.

The system automatically tells a match between the input reaction conditions and the reaction condition groups, when the DEGMA value is higher than the lower bound (if the lower bound has been input by the user). When the lower bound is not specified, the system displays the descriptions of the input reaction conditions and of each of the reaction condition groups picked up in (2). All of the values of DEGMA are also displayed. In this case the judgment is left to the user.

As a result of this analysis only those groups that match the input reaction conditions survive.

(5) Perception of RR-Site (Figure 5-[5]). When there are the characterized reactions which avail for perceiving the candidate RR-sites and which are linked to the surviving reaction condition groups via the individual reactions, the set of the only candidates perceived by application of the characterized reactions is treated as the RR-site under the input reaction conditions in the following procedure, reaction generation.

(B) When There Is No Concrete Formulation of Reaction Conditions (Only Some Reaction Condition Items Are Selected) (Figure 6). (1) Perception of Candidates of RR-Sites (Figure 6-[1]). Possible combinations of RR-bonds are perceived as candidates of RR-sites based on SCE of each bond of input reactant structures. This step is the same as (A)-(1).

For example, three candidates of RR-sites are perceived: RR-site candidate-1, -2, and -3 shown in Figure 6. Here, shaded rectangles (chas-A, -B, and -C) within a box entitled “the reaction knowledge base” stand for that the characterized reactions avail for the validity the RR-site candidates. These relationships are represented as linkages between the RR-site candidates and the characterized reactions in Figure 6.

(2) Picking up Reaction Condition Groups (Figure 6-[2]). The system picks up the reaction condition groups related to both the input reaction condition items and the perceived RR-site candidates. First, the system chooses reaction condition groups linked to the perceived RR-site candidates via the individual and the characterized reactions. Then, from the chosen reaction groups the system picks up those containing the same types of descriptions as those of the input items.

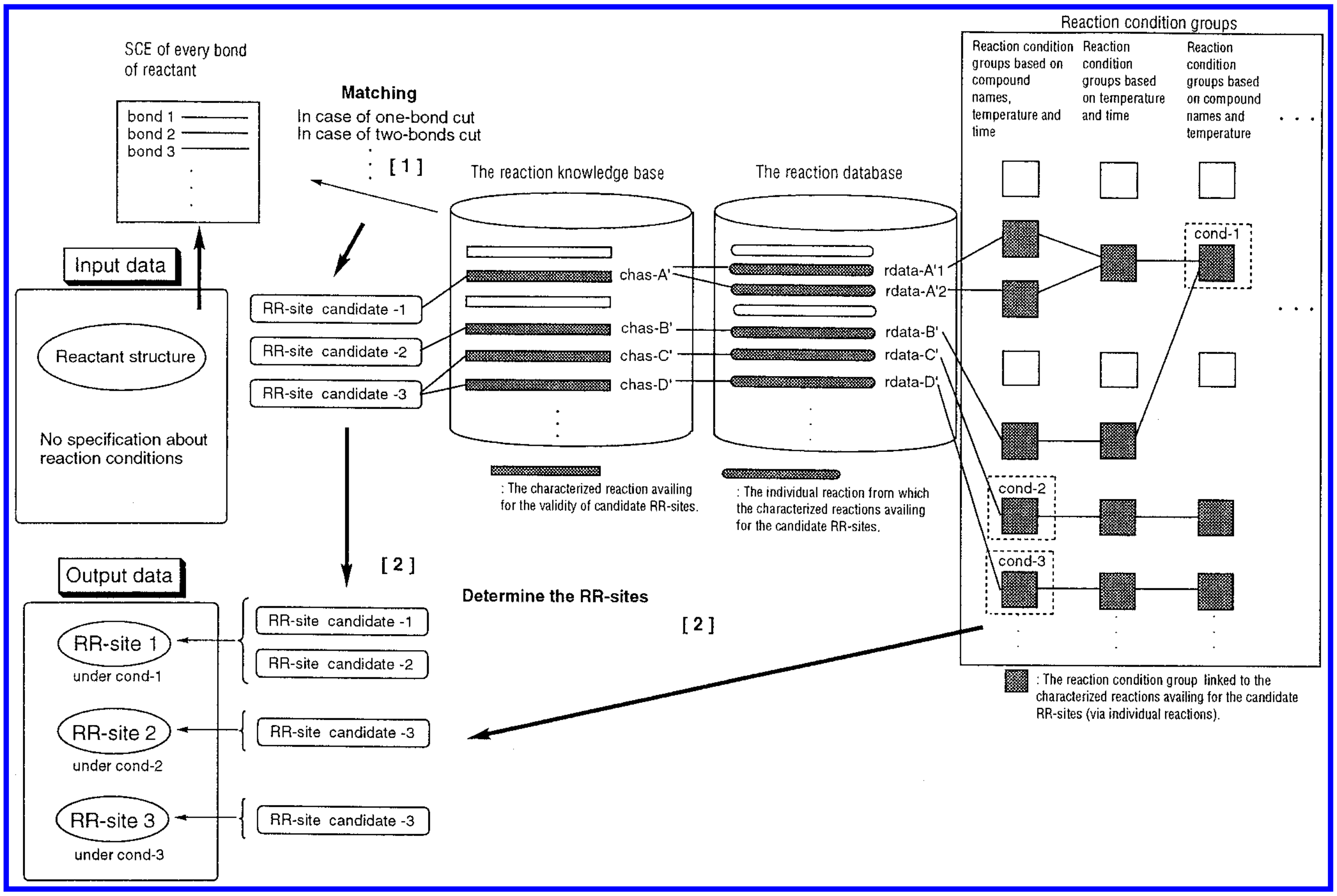

SCE of every bond
of reactant
bond 1
bond 2
bond 3
Matching
In case of one-bond cut
In case of two-bonds cut
[1]
The reaction knowledge base
The reaction database
Reaction condition groups
Reaction condition groups based on compound names, temperature and time
Reaction condition groups based on temperature and time
Reaction condition groups based on compound names and temperature
cond-1
Input data
Reactant structure
No specification about reaction conditions
RR-site candidate -1
RR-site candidate -2
RR-site candidate -3
chas-A'
chas-B'
chas-C'
chas-D'
rdata-A'1
rdata-A'2
rdata-B'
rdata-C'
rdata-D'
: The characterized reaction availing for the validity of candidate RR-sites.
: The individual reaction from which the characterized reactions availing for the candidate RR-sites.
cond-2
cond-3
[2]
Output data
Determine the RR-sites
[2]
RR-site 1
under cond-1
RR-site 2
under cond-2
RR-site 3
under cond-3
RR-site candidate -1
RR-site candidate -2
RR-site candidate -3
RR-site candidate -3
: The reaction condition group linked to the characterized reactions availing for the candidate RR-sites (via individual reactions).

Figure 7. RR-site perception procedure extended to employ the novel reaction condition classification (when any reaction conditions are not specified by the user).

Figure 6 shows the example in which temperature and time are the reaction condition items selected by the user, and three candidates of RR-sites (RR-site candidate-1, -2, and -3) are perceived. The RR-site candidate-1 is linked to the characterized reaction chas-A, namely, the RR-site candidate-1 is perceived by application of the chas-A. The chas-A is linked to an individual reaction rdata-A, namely, the chas-A was derived from the rdata-A. The RR-site candidate-2 is linked to the chas-B linked to rdata-B. The RR-site candidate-3 is linked to chas-C linked to rdata-C. Thus, the system chooses all reaction condition groups linked to the rdata-A, -B, and -C. Then, from the chosen reaction groups the system picks up those containing the same types of descriptions as those the input, temperature, and time (cond-Af, -Bf, and -Cf, which are within rectangles drawn with dotted lines). This reaction condition classification corresponds to Table 2-(f).

(3) Matching Among the Selected Reaction Condition Groups (Figure 6-[3]). The system compares descriptions among every pair of the reaction condition groups selected in (2) and calculates the DEGMA values. The system automatically judges that reaction condition groups are matched for those groups whose DEGMA values are greater than 0.

In the example shown in Figure 6 the comparison is done for three combinations: the cond-Af and -Bf, the cond-Af and -Cf, and the cond-Bf and -Cf. Three DEGMA values are calculated for them (0.6, 0, 0, respectively). Thus, the system judges as matching groups only the cond-Af and -Bf.

(4) Perception of RR-Sites (Figure 6-[4]). The matched reaction condition groups are combined to form a set of reaction condition groups, and each of the others makes a set having only one element. In this way, the system obtains several sets of reaction condition groups.

Then the system assigns each of the sets of the reaction condition groups to all the candidate RR-sites. Candidate RR-sites with different assigned sets of reaction condition groups are perceived as different RR-sites. Consequently, the description of the set of the reaction condition groups is attached to the RR-site and is considered as the condition of the reaction to be predicted.

In the example shown in Figure 6 the system obtains two sets of reaction condition groups: a set formed by the cond-Af and -Bf and a set formed by the cond-Cf. Then the system recognizes two sets of candidate RR-sites: The first is composed by RR-site candidate-1 and -2 under reaction conditions cond-Af and -Bf, which is taken as RR-site 1. The second is composed by RR-site candidate-3 under the cond-Cf, which is taken as RR-site 2.

(C) Neither Concrete Contents Nor Items of Reaction Conditions Are Specified by the User (Figure 7). (1) Perception of Candidates of RR-Sites (Figure 7-[1]). Possible combinations of RR-bonds are perceived as candidates of RR-sites based on SCE of each bond of input reactant structures. This step is the same as (A)-(1).

For example, three candidates of RR-sites are perceived in Figure 7. These are RR-site candidate-1, -2, and -3.

(2) Perception of RR-Sites (Figure 7-[2]). The system collects the RR-site candidates in one set to which it assigns one or more reaction condition groups which are common to all the RR-site candidates, i.e., those reaction condition groups to which all the RR-site candidates are linked via

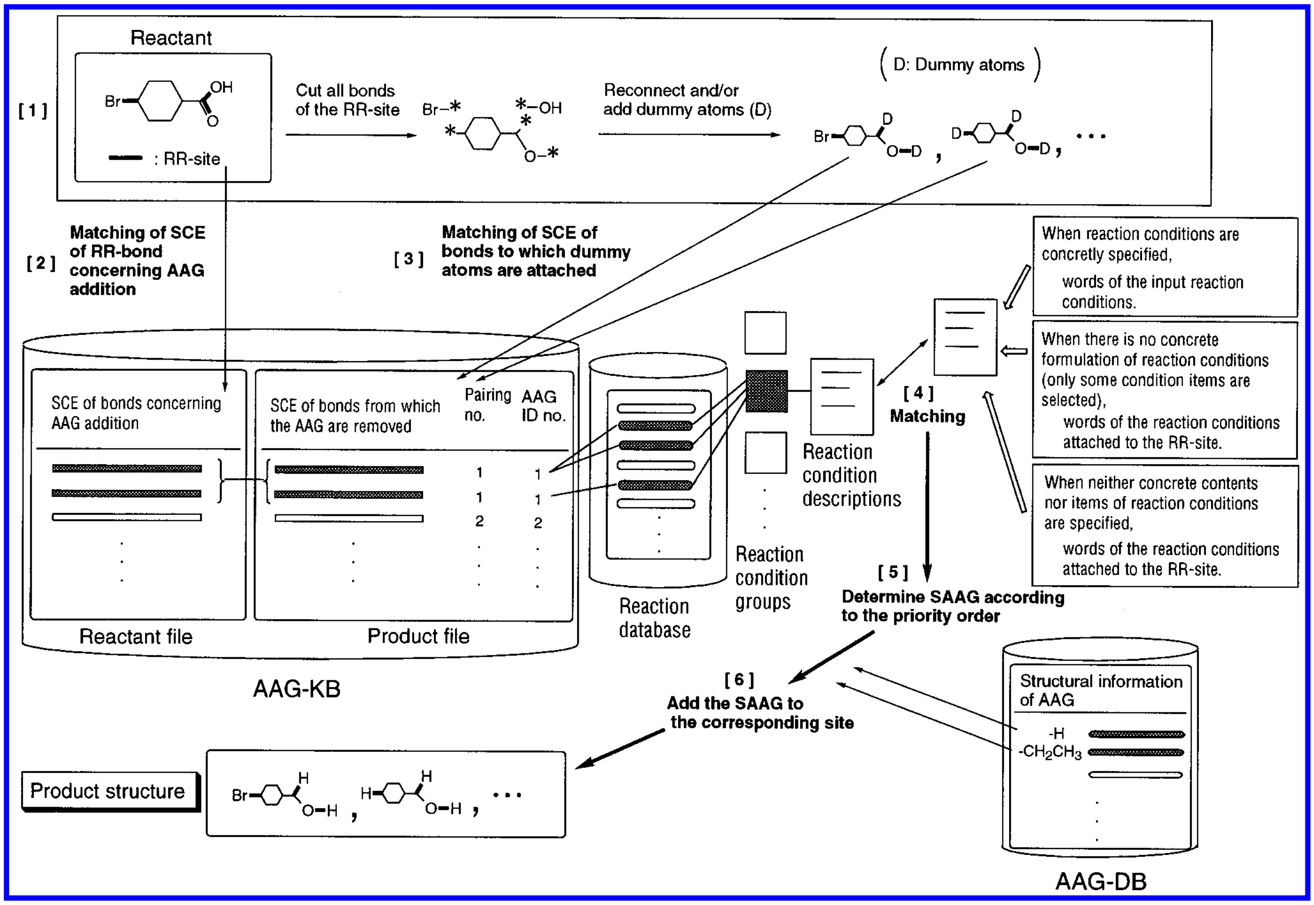

[1]
Reactant
: RR-site
Cut all bonds of the RR-site
Reconnect and/or add dummy atoms (D)
(D: Dummy atoms)
[2] Matching of SCE of RR-bond concerning AAG addition
[3] Matching of SCE of bonds to which dummy atoms are attached
When reaction conditions are concretly specified,
words of the input reaction conditions.
When there is no concrete formulation of reaction conditions (only some condition items are selected),
words of the reaction conditions attached to the RR-site.
When neither concrete contents nor items of reaction conditions are specified,
words of the reaction conditions attached to the RR-site.
SCE of bonds concerning AAG addition
SCE of bonds from which the AAG are removed
Pairing no.
AAG ID no.
Reactant file
Product file
AAG-KB
Reaction database
Reaction condition groups
Reaction condition descriptions
[4] Matching
[5] Determine SAAG according to the priority order
[6] Add the SAAG to the corresponding site
Structural information of AAG
-H
-CH2CH3
AAG-DB
Product structure

Figure 8. Product structure generation procedure in which SAAG are automatically determined and added.

the characterized and individual reactions. Candidate RR-sites with different assigned sets of reaction condition groups are perceived as different RR-sites. When more than one reaction condition group are assigned to a set of RR-site candidates, the system selects the reaction condition group whose description contains the maximum information. Consequently, the description of the selected reaction condition group is attached to the RR-site and is considered as the condition of the reaction to be predicted.

Figure 7 shows an example of three perceived candidates of RR-sites: the RR-site candidate-1, -2, and -3. The RR-site candidate-1 is linked to the characterized reaction chas-A′. The chas-A′ is linked to individual reactions rdata-A′1 and -A′2. The RR-site candidate-2 is linked to chas-B′. The chas-B′ is linked to rdata-B′. The RR-site candidate-3 is linked to chas-C′ and -D′. The chas-C′ is linked to rdata-C′, and the chas-D′ is linked to rdata-D′. Three reactions, the rdata-A′1, -A′2, and B′, are linked to one common reaction condition group cond-1. The rdata-C′ is linked to no common reaction condition group to the other reactions, the rdata-A′1, -A′2, -B′, and -D′. The cond-2 in Figure 7 contains the maximum information among reaction condition groups linked to the rdata-C′. The rdata-D′ is also linked to no common reaction condition group to the other reactions. The cond-3 in Figure 7 contains the maximum information among reaction condition groups linked to the rdata-C′. Thus, the system perceives three RR-sites under different reaction conditions: a set of the RR-site candidate-1 and -2 (RR-site 1 under the cond-1); the RR-site candidate-3 (RR-site 2 under the cond-2); and the RR-site candidate-3 (RR-site 3 under the cond-3).

4.3. Reaction Generation (Figure 8). After the RR-site perception, product structures are generated as described below.

4.3.1. Product Skeleton Generation (Figure 8-[1]). The system generates all possible product skeletons by reconnecting free bonds resulting from all bonds of the RR-site cut and by adding dummy atoms to the free bonds.[3] This generation is done for the every RR-site perceived in the RR-site perception (section 4.2).

4.3.2. Determination of SAAG Using the AAG-KB. The system determines SAAG for the generated product skeletons by receiving suggestions from the AAG-KB, which stores reactions characterized based on the SCE of RR- and PR-bonds concerning the AAG addition. Reaction conditions are also taken into account.

The SAAG determination is organized into four steps: (1) matching the reactant, (2) matching the product, (3) matching the reaction conditions, and (4) determination of SAAG. Details of these steps are described below.

(1) Matching the Reactant (Figure 8-[2]). The system compares SCE of RR-bonds concerning the AAG addition of the input reactants with the SCE of RR-bonds concerning the AAG addition in the AAG-KB.

This matching operation is performed on the RR-bond according to its five levels of its neighboring bonds and atoms, i.e., its environment within the reactant molecule. The system judges a match according to one of the five levels: (1) the minimum match condition level 1 is the match of type of bond and type of atoms attached to this, (2) the minimum match condition level 2 is the match of the SCE of the atoms of level 1, (3) the minimum match condition level 3 is the match of the SCE of the atoms at the α position, (4) the minimum match condition level 3 is the match of the SCE of the atoms at the β position, (5) the minimum match condition level 3 is the match of the SCE of the atoms at the γ position. The default is level 2, though the user can select another level if necessary. These five levels are the same as those considered when matching the input reactant with the reaction knowledge base for recognizing candidate RR-sites.[3]

(2) Matching the Product (Figure 8-[3]). The system compares SCE of PR-bonds concerning the AAG addition of each of the free bonds of the generated skeletons with SCE of PR-bonds concerning the AAG addition in the AAG-KB. Here, SCE of PR-bonds concerning the AAG addition which are linked to the SCE of RR-bonds concerning the AAG addition matched with the input reactant in the matching on reactant side (1). Figure 8 represents the SCE of PR-bonds concerning the AAG addition as shaded rectangles in the product file and the SCE of RR-bonds concerning the AAG addition as shaded rectangles in the reactant file. This matching is done in both cases of the pairing AAG being considered and not being considered. The matching also considers the five levels of matching which are the same as those considered in the matching on reactant side (1).

(3) Matching the Reaction Conditions (Figure 8-[4]). The system compares reaction conditions for the RR-site generating the product skeleton being processed with reaction conditions of each of the characterized reactions in the AAG-KB which are matched on both reactant and product sides in the matching (1) − (2). The latter reaction conditions are those of reaction condition groups linked to the individual reactions deriving these characterized reactions. The former reaction conditions depend on a way of specification of reaction conditions by the user: (A) in the case of reaction conditions concretely specified, (B) in the case of no concrete reaction conditions specified and only some reaction condition items selected by the user, or (C) in the case of neither concrete contents nor items of reaction conditions specified.

Each of matching procedures in these three cases are described below.

(A) The Reaction Conditions Are Concretely Specified. The input reaction conditions are compared. The manner of match is the same as in the RR-site perception (section 4.2.-A)). The DEGMA is also calculated and is used in the determination of the SAAG (section 4.3.2-(4)).

(B) When There Is No Concrete Formulation of Reaction Conditions (Only Some Reaction Condition Items Are Selected). The reaction conditions attached to the RR-site perception (section 4.2.-B)) are compared. The DEGMA is also calculated and is used in the determination of the SAAG (section 4.3.2-(4)).

(C) Neither Concrete Contents Nor Items of Reaction Conditions Are Specified by the User. The reaction conditions attached in the RR-site perception (section 4.2.-C)) are compared. The DEGMA is also calculated and is used in the determination of the SAAG (section 4.3.2-(4)).

(4) Determination of SAAG (Figure 8-[5]). SAAG for the PR-site of each product skeleton are determined according to the three levels of priority:

Priority 1: AAG of a matched characterized reaction in the AAG-KB are pairing; they were extracted from a common individual reaction; the DEGMA is high which was calculated in the matching of reaction conditions.

Priority 2: AAG of a matched characterized reaction in the AAG-KB are pairing; environment matched for SCE of RR- and PR-bonds was more deep; the DEGMA is high.

Priority 3: Environment matched for SCE of RR- and PR-bonds was more deep; the DEGMA is high. Pairing AAG is not considered.

4.3.3. Addition of SAAG to Product Skeletons (Figure 8-[6]). The system obtains molecular substructures of the determined SAAG from the AAG-DB and replaces the corresponding dummy atoms of the product skeletons with the substructures to generate the product structures.

4.4. Reaction Evaluation. It is evaluated whether each of the generated products can actually formed. Details were described in the previous paper.[3]

5. RESULTS

This section shows two results of execution of SOPHIA. The same reactant structure (cyclohexanone) is used as the input, differing in reaction condition specification. One is a case in which concrete reaction conditions were specified by the user (result 1), and another is a case in which only reaction condition items were specified (result 2). These results are for showing how SOPHIA runs which has been extended to use the AAG-DB, AAG-KB, and the reaction condition groups. The input reactant of cyclohexanone is used because it is simple to explain.

The executions employ the reaction condition groups prepared using 28 000 reactions of the AIPHOS-DB which was constructed from SYNLIB consisting of 81 000 reactions. However, the reaction knowledge base, the AAG-DB, and AAG-KB used here were derived from 3000 reactions of the AIPHOS-DB which was constructed from 6000 reactions in SYNLIB because of hardware restrictions.

5.1. Result 1 (Figure 9). Input reactant structure and input reaction condition are shown within a rectangle on upper left of Figure 9. The specified reaction condition is $NaBH_4$, and the lower bound is zero, which means that it is automatically judged that a reaction condition group containing the description of $NaBH_4$ is matched with the input reaction condition.

One C−O bond of the carbonyl group was perceived as RR-site represented by a bold line of a molecular structure within a rectangle on the lower left of Figure 9. A reaction condition group is matched with the input reaction condition, among reaction condition groups linked—via individual reactions—to the characterized reactions supporting the RR-site. Descriptions of the reaction condition group are "$NaBH_4$, MeOH". The DEGMA is 0.7.

The system generated one product structure by cutting the C−O bond giving two free bonds (represented as bonds to which dummy nodes "D" are attached in Figure 8), by determining the SAAG for the free bonds (−H for the both free bonds), and by adding the SAAG to the free bonds. The generated product skeleton and the product structure are shown within a rectangle on the lower center of Figure 8.

The structural characteristics of the two formed bonds of the product, which are represented by bold lines of a molecular structure shown within a rectangle on the lower right of Figure 8, were matched with the reaction knowledge base. Thus it was judged that this product can be actually

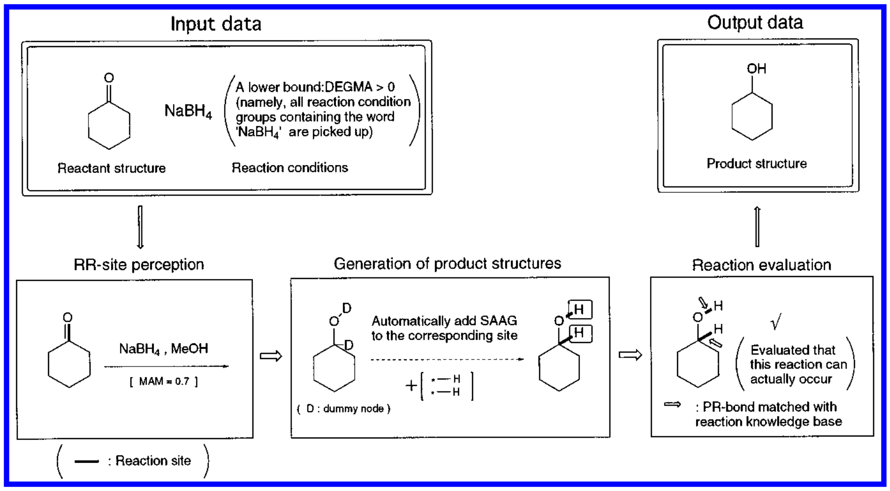

Input data
NaBH4
A lower bound:DEGMA > 0 (namely, all reaction condition groups containing the word 'NaBH4' are picked up)
Reactant structure
Reaction conditions
Output data
OH
Product structure
RR-site perception
NaBH4 , MeOH
[MAM = 0.7]
Generation of product structures
D
Automatically add SAAG to the corresponding site
(D : dummy node)
H
Reaction evaluation
√
Evaluated that this reaction can actually occur
: PR-bond matched with reaction knowledge base
: Reaction site

Figure 9. Result 1—a case in which concrete reaction conditions are specified by the user.

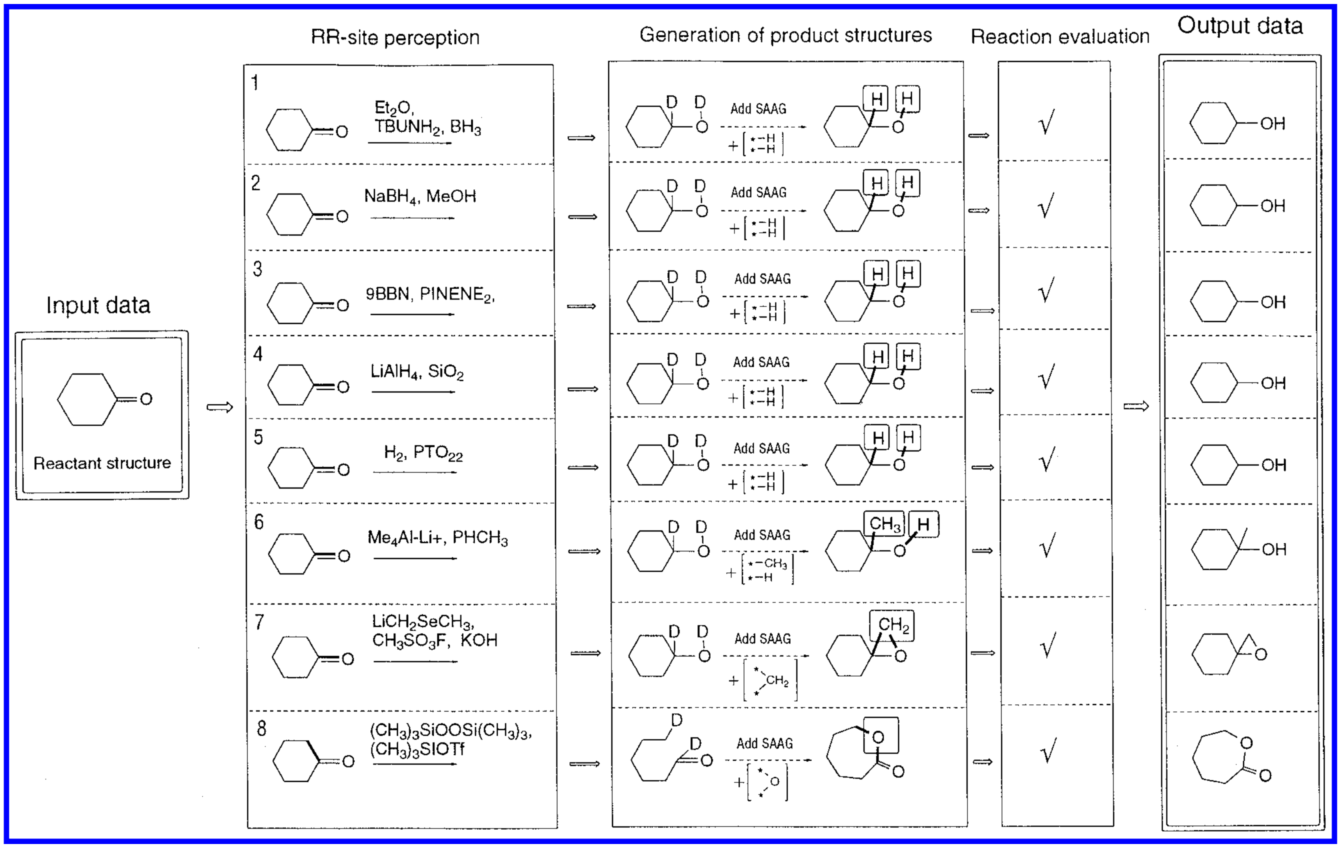


Figure 10. Result 2—a case in which only reaction condition items are selected by the user.

formed. Finally, SOPHIA predicted one product structure shown within a rectangle on the upper right of Figure 8.

5.2. Result 2 (Figure 10). Input reactant structure is shown within a rectangle on the left side of Figure 10. Selected reaction condition items were solvent, catalysts, and reagents.

Eight sets of RR-site were perceived as shown within the second rectangle from the left side of Figure 10, where RR-sites are represented by bold lines. The first to seventh RR-sites are the same, which are one C−O bond of the carbonyl groups, differing in reaction conditions. The eighth RR-site is one C−C bond neighboring the carbonyl group.

Product structures generated from each of these eight RR-sites are shown within a rectangle on the center of Figure 10. The same product structures were generated from the first to fifth RR-sites. In these cases, SAAG were two −H. −CH3 and −H were perceived as SAAG for free bonds obtained by cutting a bond of the sixth RR-site (one C−O bond), and $-CH_2-$ was perceived as SAAG for the seventh RR-site (one C−C bond). The last case is a Baeyer−Villiger reaction. Consequently, four kinds of product structures were generated.

All products were evaluated that they can actually be formed. Finally, SOPHIA predicted eight reactions with the reaction conditions shown within the right rectangle of Figure 10.

6. DISCUSSION

Further development of the reaction generator makes it possible for SOPHIA to automatically predict products from only reactants and reaction conditions (at choice), and utilization of results of reaction condition classification makes it possible for SOPHIA to automatically predict products by considering conditions as appropriately as possible at present.

The extended reaction generator automatically determines SAAG for product skeletons and adds the SAAG to suitable sites of the product skeletons. *Addable* kinds of SAAG have also been extended to those which have more than two free bonds (e.g., $-CH_2-$, $-O-$) and contain more than two atoms except hydrogen atoms (e.g., $-CH_2CH_3$, $-Si(CH_3)_3$), although the previous SOPHIA[3] allowed to specify SAAG which contain only one atom except hydrogen atoms (e.g., $-HC_3$, $-OH$) or a hydrogen atom $-H$. Further the extended generator considers reaction conditions to determine SAAG by utilizing the results of reaction condition classification. For instance, as shown in result 2 in section 5, the reaction generator recognizes different SAAG for the same product skeleton generated from the same RR-site differing only in the reaction conditions.

We have investigated reaction condition descriptions in a reaction database and obtained reaction condition groups (it is important that this investigation is free from any preconception). Application of the result allows the system to utilize reaction condition descriptions of currently available reaction databases so as to interpret reaction condition descriptions freely entered by the user and to take account of the conditions in prediction as appropriately as possible at present.

The next development of SOPHIA will concentrate on classifying chemical reactions and utilizing the result of reaction classification to quantitatively predict product ratio. For this purpose we will investigate relationships among the reaction condition groups, changes of structural characteristics around the reaction site during the reaction, and changes of electronic features around the reaction site during the reaction. Again, it is important that the classification of reactions is done without introduction of preconceived ideas about chemical reactions. The results will be reported in a future paper.

ACKNOWLEDGMENT

We wish to thank Dr. Carlos A. Del Carpio for his useful discussion in preparation of this paper. This work has been supported by the Hori Information Science Promotion Foundation.

Supporting Information Available: Reaction condition classification details (9 pages). This material is contained in many libraries on microfiche, immediately follows this article in the microfilm version of the journal, can be ordered from the ACS, and can be downloaded from the Internet; see any current masthead page for ordering information and Internet access instructions.

REFERENCES AND NOTES

(1) Röse, P.; Gasteiger, J. Automated derivation of reaction rules for the EROS 6.0 system for reaction prediction. *Anal. Chim. Acta* **1990**, *235*, 163−168.

(2) Salatin, T. D.; Jorgensen, W. L. Computer-Assisted Mechanistic Evaluation of Organic Reactions. 1. Overview. *J. Org. Chem.* **1980**, *45*, 2043−2057.

(3) Satoh, H.; Funatsu, K. SOPHIA, a Knowledge Base-Guided Reaction Prediction System—Utilization of a Knowledge Base Derived from a Reaction Database. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 34−44.

(4) Corey, E. J.; Wipke, W. T.; Cramer III, R. D.; Howe, W. J. Computer-Assisted Synthetic Analysis. Facile Man-Machine Communication of Chemical Structure by Interactive Computer Graphics. *J. Am. Chem. Soc.* **1972**, *26*, 421−430. Corey, E. J.; Cramer III, R. D.; Howe, W. J. Computer-Assisted Synthetic Analysis for Complex Molecules. Methods and Procedures for Machine Generation of Synthetic Intermediates. *J. Am. Chem. Soc.* **1972**, *26*, 440−459.

(5) Greene, T. W. Protective Groups in Organic Synthesis.; John Wiley & Sons: New York, 1981.

(6) Funatsu, K.; Sasaki, S. Computer-Assisted Synthesis Design and Reaction Prediction System AIPHOS. *Tetrahedron Comput. Method.* **1988**, *1*, 27−38.

(7) Funatsu, K.; Kitamura, S.; Watanabe, M.; Negishi, Y.; Takahashi, Y.; Horiuchi, K.; Dogane, I.; Sasaki, S. Development of Organic Synthesis Design System AIPHOS (11)—Structure and Functions of Knowledge Base for Proposing Suitable Leaving Group. *In Proceedings of the 16th Symposium on Chemical Information and Computer Sciences/21st Symposium on Structure−Activity Relationships*; Tsukihara, T., Terada, H., Eds.; Tokushima University: Tokushima, Japan, 1993; pp 101−104.

(8) Funatsu, K.; Kitamura, S.; Tiba, M.; Watanabe, M.; Uchida, T.; Tanaka, A.; Oue, T.; Dogane, I.; Sasaki, S. Development of Organic Synthesis Design System AIPHOS (12)—Development of Program for Automatic Introduction of Commercially Available Reaction Database to AIPHOS Database. *In Proceedings of the 16th Symposium on Chemical Information and Computer Sciences/21st Symposium on Structure−Activity Relationships*; Tsukihara, T., Terada, H., Eds.; Tokushima University: Tokushima, Japan, 1993; pp 105−108.

(9) Distributed Chemical Graphics, Inc. permitted us to use SYNLIB for research work of AIPHOS and SOPHIA.

CI950058A