SOPHIA, a Knowledge Base-Guided Reaction Prediction System—Utilization of a Knowledge Base Derived from a Reaction Database

Hiroko Satoh and Kimito Funatsu*

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

Received March 28, 1994®

A reaction prediction system called SOPHIA (System for organic reaction prediction by heuristic approach) has been developed to predict possible products and the product ratio from arbitrary reactants under arbitrary reaction conditions. As a first step in developing SOPHIA we used the reaction knowledge base of the organic synthesis design system AIPHOS, which was derived from a reaction database, as a general knowledge base for reaction prediction. It became possible to automatically perceive a reaction site and to predict possible reaction paths without the user's designation of a specific reaction type or category. This paper describes the philosophy of SOPHIA and the current level of development together with an overview and first results.

I. INTRODUCTION

In designing a synthesis route for a desired target structure an experimental chemist infers which part of the target structure is synthetically accessible and what starting materials and reaction conditions could make the desired reaction proceed according to the synthesis strategy (backward planning), and the chemist is required to judge possibilities of reactions and side reactions (forward verification). In general, a synthesis design program system has only the former function: it designs backward synthesis routes according to its own synthesis strategies. However, the actual process of designing synthesis routes by an experimental chemist suggests that the latter function of a reaction evaluation is necessary to complete a synthesis design system. A reaction prediction system is a necessary component.

A reaction prediction system needs to answer one question: what product structures will be obtained from an input reactant structure and reaction condition, and what will be the products ratio, in contrast a synthesis design system proposes some of many possible synthesis routes. In general, if the problem is not trivial, a large number of synthesis routes which give the desired target are possible. Therefore, the output from a synthesis design system is dependent on the implemented synthesis strategies of the system and is generally different from the output of other synthesis design systems. However, it is absolutely required that every reaction prediction system should give the same answer (Figure 1).

For exact reaction prediction, it is necessary to use a chemical reaction theory. However, a *general* theory by which chemical reaction paths of an arbitrary reactant under an arbitrary reaction condition can be exactly predicted has not been established yet. At present, quantum mechanical approach to reaction prediction and synthesis design is applicable to only small molecules or molecules of the same series. A chemical reaction occurs as a result of complicated interactions among the attributes of a reactant (e.g., electronic properties, three-dimensional structural features) and the reaction conditions (solvent properties, temperature, concentration, etc.). The degree of contribution of these factors to a reaction varies.

This is the reason why not very many reaction prediction systems have not been developed in spite of their potential importance. The difference of the number of known systems explains this fact: CAMEO[1] and EROS6.0[2] are described as reaction prediction systems, whereas LHASA,[3] SECS,[4] EROS[5] (the versions older than ver.6.0), WODCA,[6] and AIPHOS[7] are described as synthesis design systems.

In order to exactly predict chemical reaction without a backing general theory the EROS6.0 and CAMEO systems have been designed. In the EROS6.0 system chemical reaction knowledge used for reaction prediction is described for each reaction type (e.g., Diels–Alder reaction). Contents of the knowledge for each reaction type are a substructure transformation during the reaction, ranges of allowed values of attributes for each atom and bond (i.e., σ-, π-charge, electronegativity, bond dissociation energy, polarizability, etc.) under which the type of reaction possibly occurs, and chemical reactivity functions which were derived by correlation analysis of experimental data (energy of activation, rate constant) with calculated physicochemical data. The reactivity function is given if rate constant values are obtained from literatures. The knowledge is derived from a reaction database. EROS6.0 requires the user, as indispensable input data, to specify a reactant structure and reaction types to be considered and generates reaction products using the knowledge of substructure transformation if all the attribute values of input reactant structures and reaction conditions are within the ranges of values. Furthermore, if the reactivity function for the reaction type is given, the product ratio is predicted from the relative reaction rate calculated by the reactivity function. Reaction categories established in the CAMEO system are more general and familiar to an experimental chemist (e.g., electrophilic reactions under acidic catalyst and nucleophilic reactions under basic catalyst). In each of the categories, reactions are classified by the intermediate structure, and the classified reactions are divided into the elemental reactions. The elemental reactions are sorted by reactivity and the reaction condition. This data are based

® Abstract published in *Advance ACS Abstracts,* December 15, 1994.

0095-2338/95/1635-0034$09.00/0

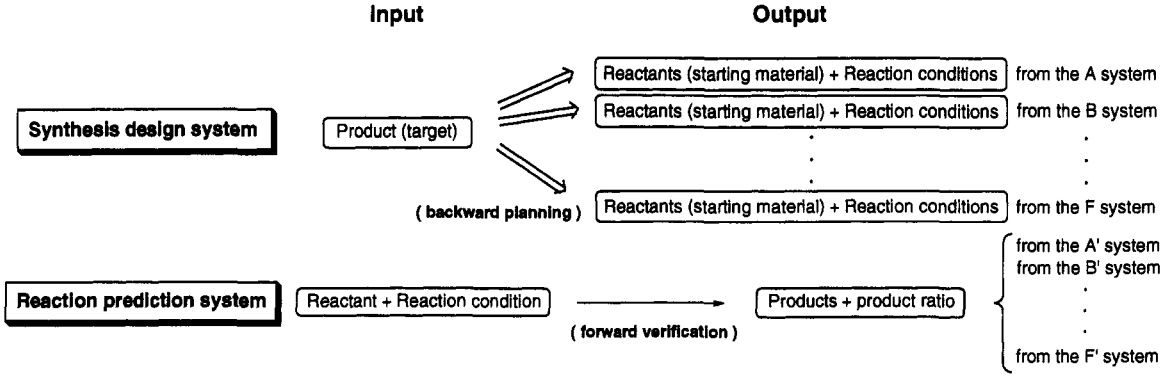

Figure 1. Difference between a synthesis design system and a reaction prediction system.

on literature and textbook chemical knowledge and described by the system developers. CAMEO requires the user, as indispensable input data, to specify a reactant structure, a reaction condition, and one of the reaction categories.

Thus, reaction prediction in the EROS6.0 and CAMEO systems uses the reaction knowledge arranged for each reaction type or category. Therefore, both systems have the common prerequisite that they require the user to specify the reaction type or category as indispensable input data. This prerequisite, which is regarded as a restriction of the system, makes it possible to analyze in detail the reaction belonging to the type or category of interest. However, seen in a different light, it brings no consideration of possibility of any reaction outside of the specified reaction type or category. Furthermore, describing the chemical knowledge dependent on reaction types or categories involves the following problems: First, reactions which does not belong to the established reaction types or categories are not used for deriving reaction rules. Second, it is required to obtain reaction knowledge for all reaction types or categories. Third, some reactions might be difficult to be assigned to an established reaction types or categories.

We began, therefore, to develop a different type of system called SOPHIA (system for organic reaction prediction by heuristic approach), which is able to predict and analyze all possible reaction paths without the user specifying any reaction type or category. A knowledge base derived from a reaction database is applied to this approach.

II. KNOWLEDGE BASE FOR REACTION PREDICTION

A chemical reaction database holds potentially much information about chemical reactions, from which general knowledge, rules, and theories of chemical reactions can be derived. A basic point of the SOPHIA system is that it utilizes a reaction database to derive knowledge for exact reaction prediction. In taking this approach, the efficiency of the system is dependent on the contents of the database, the knowledge derived from it, the knowledge representation, and the style of knowledge base application. The more suitable the overall configuration of these factors the more exact the prediction of reaction paths.

The representation of knowledge on chemical reactions can be classified into two types: *concrete* and *abstract* forms. In the concrete form, a chemical reaction scheme is represented as a transformation of substructures; in the abstract form it is represented as a transformation of structural attributes (e.g., electronic properties).

Reaction prediction using the concrete knowledge is simple: the substructure in a reactant is transformed to give a product according to the knowledge. This manner of application of reaction knowledge is similar to that found in the CAMEO system and the EROS6.0 system.

In contrast, in the case of the abstract knowledge, a computer system requires a more complicated reaction prediction procedure, in which all reaction schemes corresponding to the knowledge should be generated. This means that concretion of knowledge is required. The EROS6.0 quantitatively predicts product ratio with mathematical equations involving multiple physicochemical parameters. The CAMEO system also predicts reactivity. However, in perception of a reaction site and during the generation of reactions they do not use abstract knowledge.

In SOPHIA we use abstract knowledge for reaction site perception and reaction evaluation.

1. Knowledge Base of AIPHOS.[8,9] The knowledge base of AIPHOS (AIPHOS-KB) was originally designed not only for synthesis design but also for reaction prediction. Therefore initially we examined whether the AIPHOS-KB has a suitable structure and content for SOPHIA.

The representation of reactions in the AIPHOS-KB and its structure are described below.

1.1. Representation of Reactions. 1.1.1. Reaction Site. An example of a reaction site representation in AIPHOS-KB is illustrated in Figure 2.

Reaction bonds here are called bonds broken or formed during a reaction. Atoms attached to the reaction bonds are reaction atoms, and a reaction site on the reactant or product side is a set of all sites having reaction atoms and reaction bonds. In the most general case, more than a single molecule is contained on the reactant side and/or a product side of a reaction scheme. For later discussion, in this paper a set of all sites having broken bonds and atoms attached to them is called the reaction site of the reactant (reactant reaction site: RR-site), and a set of all sites having formed bonds and atoms attached to them is called a reaction site of the product (product reaction site: PR-site). Information about a reaction site is represented as information on reaction bonds. The information about a reaction bond is organized into five slots: (1) type of the reaction bond and types of reaction atoms attached to this bond; (2) structural characteristics for these reaction atoms; and (3) structural characteristics for nodes at the α, (4)β, and (5)γ positions of the reaction atoms. The meaning of the structural characteristics is described below.

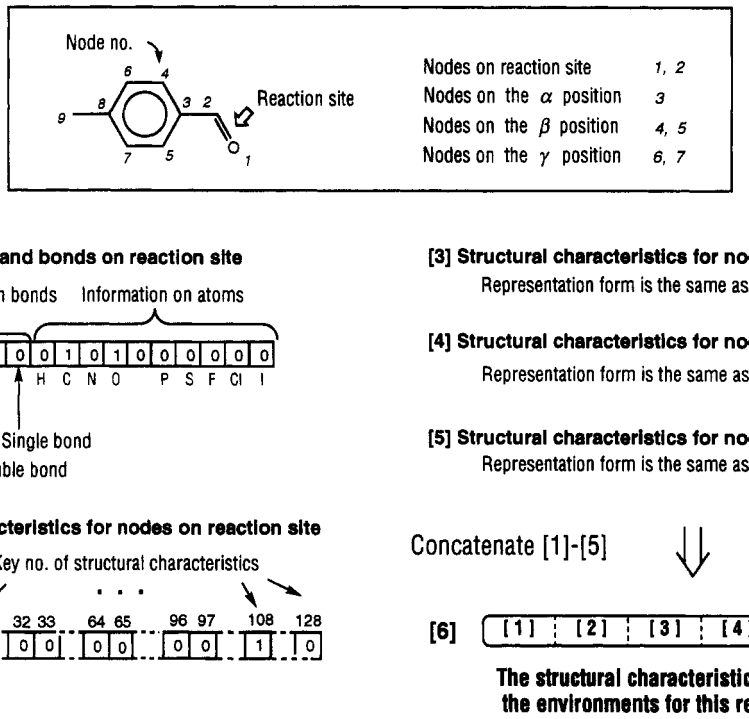

Figure 2. Representation of a reaction site in the knowledge base of AIPHOS.

For a reactant structure given in Figure 2, the type of the reaction bond is a double bond, so the third bit is set (Figure 1-[1]), and the types of the reaction atoms attached to this bond are carbon and oxygen, so the sixth and the eighth bits are set. Information sets (2), (3), (4), and (5) are described by substructures. In these pieces of information, environment features around a reaction bond are considered. The current AIPHOS uses 111 substructures to characterize reaction sites and the environment. These substructures were selected for importance in the field of chemical reactions.[10,11] We call them structural characteristic keys; some of them are listed in Table 1. Bits corresponding to each of these structural characteristic keys are needed. The structural characteristic is computed by integrating basic structural information (e.g., atom and bond types for the nodes, atom, and bond types attached to this nodes). The details of how to recognize the structural characteristics are described in the preceding paper.[10,11] For the reactant given in Figure 2, the structural characteristic of the nodes belonging to the reaction bond is Ar−CHO, so the bit corresponding to this key no. 108 is set (Figure 2-[2]). The structural characteristic information on nodes at the α, β, and γ positions (the information sets (3), (4), and (5)) are obtained in the same manner (Figure 2-[3], [4], and [5]). Finally, these five bit sequences shown in Figure 2-[1]−[5] are concatenated as the characterized information on this reaction bond (Figure 2-[6]).

A reaction bond characterized in this manner is used for characterizing a reaction site. For representation of reaction sites and schemes, each set of the characterized reaction bonds of each molecule in reactant or product side is given an ID number (ID no. of the characterized molecule, IDCM; the characterized molecule, CM). For example, IDCMs are m_00001, m_00002, m_00003, and so on. The numerals in IDCMs are keys of the knowledge base and have no specific meanings. A set of the IDCMs for all molecules of reactants or products represents a reaction site of each side of a reaction scheme. For example, for Friedel−Crafts alkylation shown in Figure 3 the RR-site is a set of the IDCMs

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

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	$-P(OR)_2$ (=O)	•	•
34	$-CR_3$ (=O)		
		100	$-N{=}C-$⬡
35	$-CH$ (=O)	101	$-C{=}N-$⬡
•	•		
•	•	102	$-C{\equiv}C-$⬡
•	•		
48	$-N(H)-C(=O)-NH$	103	⬡$-C{\equiv}N$
49	$-C(=O)-S-$	•	•
		•	•
50	$-SO_2OH$	108	⬡$-C(=O)H$
•	•	•	•
•	•	•	•
•	•	111	⬡$-NO_2$

m_00020 and m_00021), which originate from the characterized reaction bonds of reactant structures, Ar−H bond of phenol and C−Cl bond of *tert*-butyl chloride. The PR-site is the IDCM (m_00022), which is obtained from the characterized reaction bonds of product structure, C−C bond of p-*tert*-butylphenol.

1.1.2. Reaction Scheme. An example of a reaction scheme represented in the AIPHOS-KB is illustrated in Figure 4.

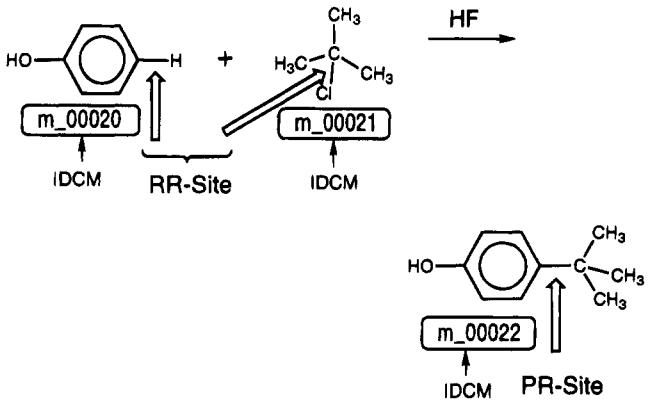

Figure 3. Definition of IDCM, RR-site, and PR-site.

The reaction scheme is represented by the characterized RR-site, PR-site, and a reaction condition class. At present we use the classes of reaction conditions summarized by Greene,[12] where the reaction conditions are classified into 16 categories as the primary classification (Table 2-[1]), and each category is further subdivided into several subclasses (Table 2-[2]). In the AIPHOS-KB these reaction condition classes are used for describing various reaction conditions but not for classifying chemical reactions. Therefore, using these reaction condition classes is not equal to establishing reaction types or categories. For reaction conditions which do not correspond to the classes from PRO_A to PRO_P, PRO_Q (others) is used (Table 2).

The reaction scheme of the reduction reaction shown in Figure 4 is represented by the IDCMs m_00550 for the reactant and m_00525 for the product, with the reaction condition ID no. PRO_H01. Since this product has two reaction bonds, m_00525 consists of two bit sequences characterizing them. In this example, numerals in IDCMs are not successive, because when a new reaction scheme is characterized and stored to the AIPHOS-KB, a set of the characterized bonds for each molecule in the reaction scheme is not stored if the set has already been stored, and the IDCM already stored in another context is used for representing this new reaction scheme.

1.2. Structure. The AIPHOS-KB is organized into six random access files. Those four which are used in SOPHIA (Figure 5) are shown here. They are a reaction site file (Figure 5-[A]), a reaction scheme file (Figure 5-[B]), a reaction data file (Figure 5-[C]), and an auxiliary accelerator file (Figure 5-[D]).

The IDCMs and their environment bit sequences are in file [A]. The characterized reaction schemes and their ID nos. are in file [B]. In file [C] each of the characterized reaction scheme ID no. with ID nos. of the individual reactions deriving this characterized reaction scheme is described. In general, several individual reactions in a reaction database often produce the same characterized reaction scheme. In other words, it is implied that a single characterized reaction scheme extends to several concrete reaction schemes which are similar to each other. Some of these extended reactions will exist in the database used for deriving this characterized reaction scheme. File [D] has a description of how many bit sequences each of the IDCM has and on which side in the characterized reaction schemes in file [C] this IDCM appears: on the reactant side and/or on the product side.

Table 2. Part of the Reaction Condition Keys in the Knowledge Base of AIPHOS

ID no.	reaction condition
(1) First Classification	
PRO_A	aqueous
PRO_B	nonaqueous bases
PRO_C	nonaqueous nucleophiles
PRO_D	organometallic
PRO_E	catalytic reduction
PRO_F	acidic reduction
PRO_G	basic or neutral reduction
PRO_H	hydride reduction
PRO_I	Lewis acids
PRO_J	soft acids
PRO_K	radical addition
PRO_L	oxidizing agents
PRO_M	thermal reactions
PRO_N	carbenoids
PRO_O	miscellaneous
PRO_P	electrophiles
PRO_Q	others
(2) Part of the Second Classification	
PRO_A Aqueous	
PRO_A01	pH < 1, 100 °C: refluxing HBr
PRO_A02	pH < 1: 1 H HCl
PRO_A03	pH 1: 0.1 N HCl
PRO_A04	pH 2−4: 0.01 N HCl; 1−0.01 N HOAc
PRO_A05	pH 4−6: 0.1 N H_3BO_3; phosphate buffer; HOAc−NaOAc
PRO_A06	pH 6−8.5: H_2O
PRO_A07	pH 8.5−10: 0.1 N HCO_3^-; 0.1 N OAc^-; satd $CaCO_3$
PRO_A08	pH 10−12: 0.1 N $(CO_3^-)_2$; 1−0.01 N NH_4OH; 0.01 N NaOH
PRO_A09	pH > 12: satd Ca(OH)
PRO_A10	pH > 12, 150 °C: 1−0.1 N NaOH
PRO_A11	others
PRO_H Hydride Reduction	
PRO_H01	$LiAlH_4$: Li-Selectride
PRO_H02	Li−*sec*-Bu_3BH, −50 °C: disiamylborane
PRO_H03	$Li[(CH_3)_2CHCH(CH_3)]_2BH$
PRO_H04	B_2H_6, 0 °C
PRO_H05	$NaBH_4$: neutral reduction
PRO_H06	$Zn(BH_4)_2$
PRO_H07	$NaBH_3CN$, pH 4−6: Dibal
PRO_H08	$(i\text{-}C_4H_9)_2AlH$, −60 °C
PRO_H09	$Li(O\text{-}t\text{-}C_4H_9)_3AlH$, 0 °C
PRO_H10	others

As shown in Figure 5, these four files contain data links. Following the references all related information to a single piece of input information can be retrieved. This network like structure allows the access all related pieces of information starting from any point.

This knowledge base has been derived automatically from a reaction database.

2. Knowledge Base of SOPHIA. AIPHOS-KB has three important features.

First, no chemical reaction type nor category are taken into account. Again, the chemical reaction classes are used only for describing various reaction conditions, and every individual reaction is represented as a pair of the transformation of the structural characteristics of the reaction site during the reaction and the reaction class.

Second, the characterized reaction sites and schemes potentially cover other individual chemical reactions similar to a reaction used for the characterization. Thus we regard that AIPHOS-KB represents chemical reactions abstractly.

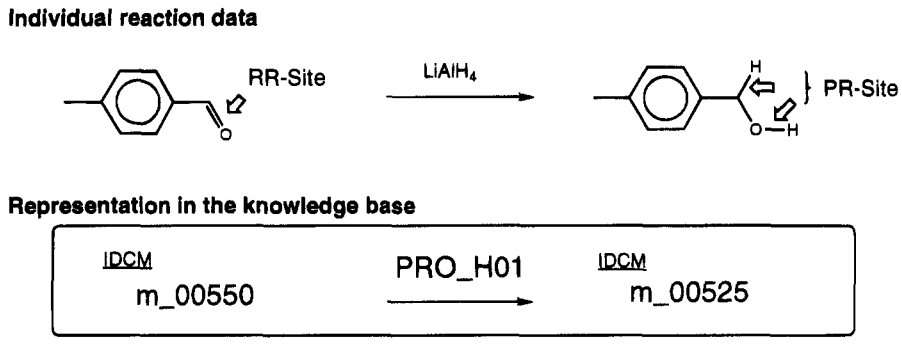

Figure 4. Representation of a reaction scheme in the knowledge base of AIPHOS.

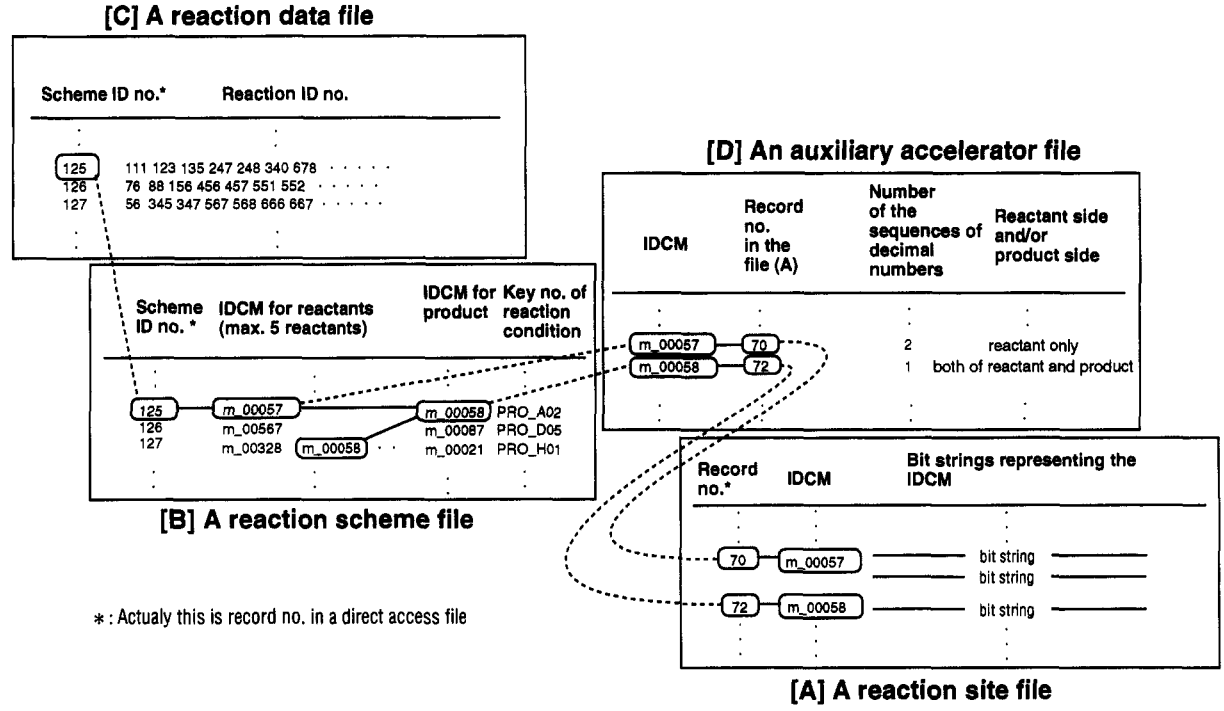

Figure 5. The structure of the knowledge base of AIPHOS.

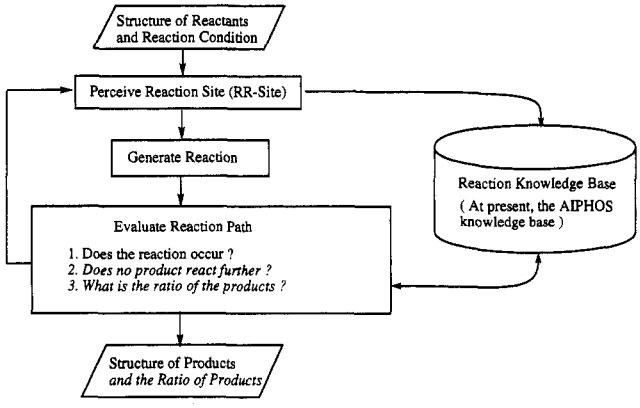

Figure 6. Block diagram of SOPHIA (the functions currently being implemented are written in block letters).

Third, the network structure allows access via multiple paths, for example, via the characterized reaction site, the characterized reaction scheme, a reaction condition, and so on. This is an advantage compared to classical coding schemes such as production rules.

These above features of AIPHOS-KB are useful for SOPHIA. Although this knowledge base is not applicable to the quantitative reaction prediction because information for quantitative evaluation has not been described yet, we decided to use this knowledge because it is handy in the first steps in developing SOPHIA. For quantitative reaction prediction we plan to develope a knowledge base using phisicochemical features discussed in section V.

III. BLOCK DIAGRAM OF SOPHIA

The organization of SOPHIA is diagramed in Figure 6. The functionality already implemented is represented with block letters. The current SOPHIA release automatically perceives all possible RR-sites for an input reactant under consideration of an input reaction condition (a reaction condition is dispensable). AIPHOS-KB is used in this procedure. In the next step, reaction bonds for each perceived RR-site are cut, and all possible reconnection of cut bonds and/or addition of atoms or atomic groups to them are performed to generate all possible reaction reaction products. Finally, each generated reaction from the input reactant to the generated product is evaluated whether it is probable. This evaluation procedure also uses AIPHOS-KB.

SOPHIA has been written FORTRAN77 under MS-DOS on a NEC-PC9801FA 32-bit personal computer.

IV. REACTION PREDICTION PROCEDURES IN SOPHIA

1. Input Data. Input data of SOPHIA are chemical structures of reactants and a reaction condition (at choice).

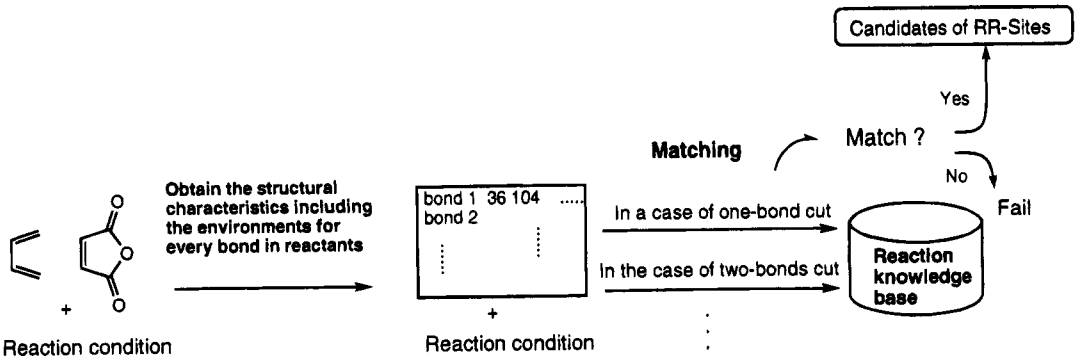

Figure 7. Reaction site (RR-site) perception.

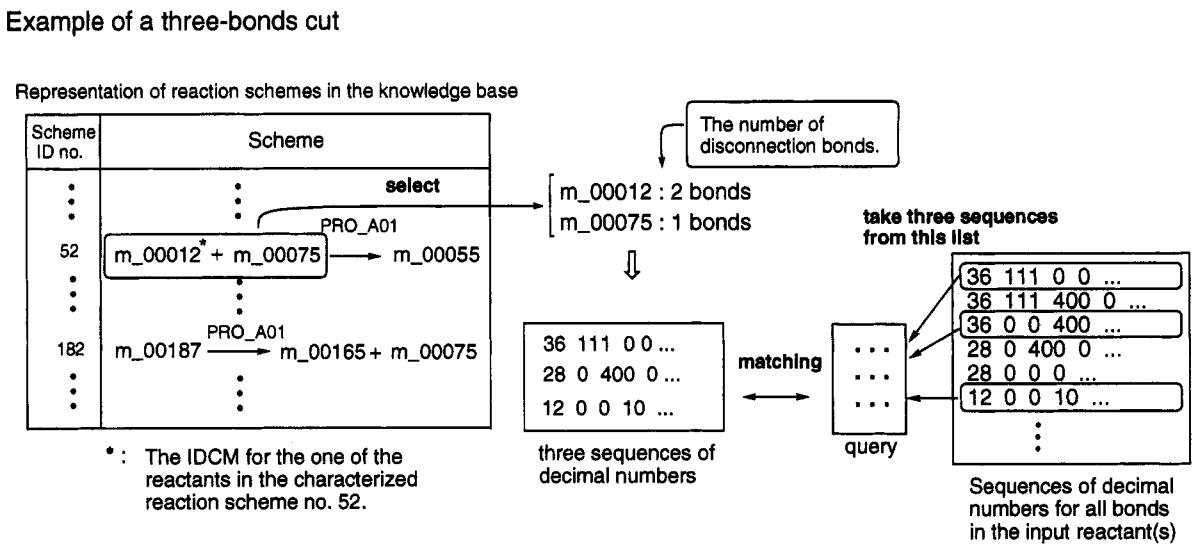

Figure 8. The detail of the matching procedure in the RR-site perception.

A reaction condition may be selected by the user from the list shown in Table 2.

2. RR-Site Perception. All possible RR-sites are automatically recognized. A reaction condition is taken into account if it is specified. An outline of the procedure is shown in Figure 7, and the details are described below.

2.1. Structure Perception. Basic structural characteristics of reactants, for example, types of atoms and bonds, connectivity, etc., are computed, and the SSSR (smallest set of smallest ring)[13] and aromaticity are perceived. The perception of aromaticity is based on Hückel's rule. This information is used in the next step of structural characteristics computation.

2.2. Computation of Structural Characteristics. By using the above basic structural information, the structural characteristics (listed in Table 1) are perceived, and the bit sequences for every bond in the reactants (i.e., characterized bonds) are obtained as in Figure 2-[6].

2.3. Matching. The next step is matching with the knowledge base. Here every combination (e.g., each combination of one-bond, two-bonds, three-bonds, and so on) of the characterized bonds of the input reactant structures is automatically generated in turn, and the bit sequences of the reaction bonds of the combination are compared with the bit sequences in the file [A] in Figure 5. Furthermore, in this matching procedure, five levels of matching are considered. They are

level 1: minimum matching level is type of the bond and type of atoms attached to this bond (Figure 2-[1]),

level 2: minimum matching level is the structural characteristics of these atoms (Figure 2-[2]),

level 3: minimum matching level is the structural characteristics of atoms at the α position (Figure 2-[3]),

level 4: minimum matching level is the structural characteristics of atoms at the β position (Figure 2-[4]), and

level 5: minimum matching level is the structural characteristics of atoms at the γ position (Figure 2-[5]).

Default level is 2, and if necessary the user can select another level.

In order to explain the procedure of matching, an example of a three-bond combination is shown.

In the Case that a Reaction Condition (PRO_A01) is Specified. If the reaction condition, PRO_A01, is specified by the user, the characterized reaction schemes with three reaction bonds on the reactant side and with PRO_A01 as the reaction condition are selected from the knowledge base, i.e., the file [B] in Figure 5. For example, as shown in Figure 8, the characterized reaction scheme ID no. 52 (actually, this is the record no. in the direct access files of the knowledge base) has two reactants represented as their IDCMs, m_00012 and m_00075; m_00012 has two reaction bonds, and m_00075 has one reaction bond, thus the total number of reaction bonds is three. Therefore this reaction scheme is valid.

Next, the matching is performed: a comparison between these three bit sequences characterizing the reactants of the reaction scheme ID no. 52 and three bit sequences taken from those for the all bonds of the input reactants (obtained in

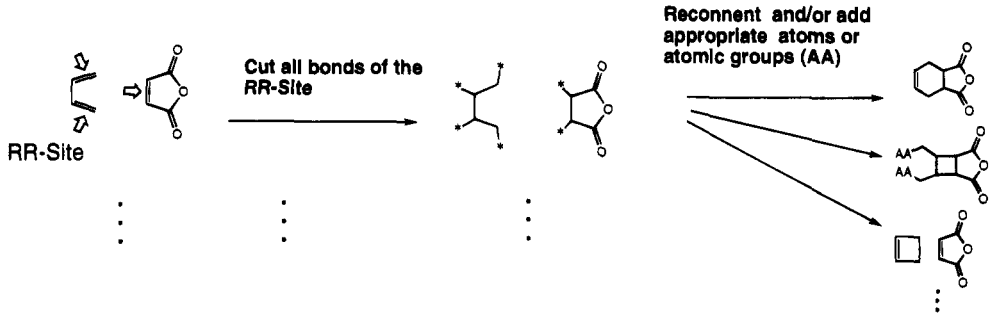

Figure 9. Reaction generation procedure.

the previous step (see IV-2.2)) is made. The latter set of bit sequences is a query pattern for the knowledge base. When one-to-one correspondence is found, it is judged that the processing set of bonds as the query would be reaction bonds, because the one-to-one correspondence means that a set of characterized reaction bonds of a RR-site matched with the query pattern is in the characterized reaction schemes in the knowledge base.

This matching procedure is continued for all query possibilities, three bit sequences selected from those for all bonds of input reactants, until the comparison with all characterized reaction schemes with three reaction bonds in its reactants contained in the knowledge base is finished.

After this matching procedure, a set of all bonds which are judged that they should be reactive is perceived as a RR-site for the input reactants. This procedure makes it possible to generate reactions by a combination of multiple characterized reaction schemes under the same reaction condition in the knowledge base as shown in results 3 and 4 of section V.

In the Case that a Reaction Condition is Not Specified. Even if no reaction condition is specified, the matching procedure is the same as above except for the treatment of reaction conditions: descriptions of reaction conditions are ignored when the characterized reaction schemes are selected from the knowledge base. After the matching procedure, each set of all bonds which are matched with the characterized reaction scheme having the same reaction condition and are judged that they should be reactive is perceived as each RR-site for the input reactant. Each set of bonds is perceived as each RR-site under each reaction condition. Therefore, the number of perceived RR-sites is the same as the number of different classes of reaction conditions described in characterized reaction schemes in the knowledge base which match the query. A set of all bonds which are matched with the characterized scheme having the reaction condition PRO_Q is also perceived as one of the RR-sites; however, this procedure is not proper for PRO_Q because PRO_Q includes all reaction conditions which are not corresponding to any condition description of PRO_A – PRO_P. A countermeasure for solving this problem will be described in section VI.

3. Reaction Generation. The procedure is shown in Figure 9. From the structures of reactants with the perceived RR-sites all possible reaction paths are generated on the basis of cyclic permutation. This procedure concretizes the knowledge.

When no reaction condition is specified and multiple RR-sites are perceived, the following procedure from 3.1–3.4 is performed for each of the RR-sites.

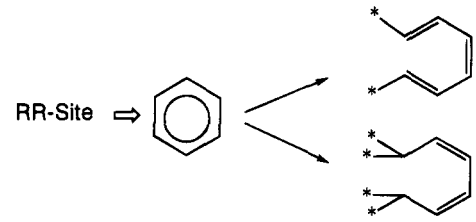

Figure 10. Aromatic bond cuts. * stands for free bond.

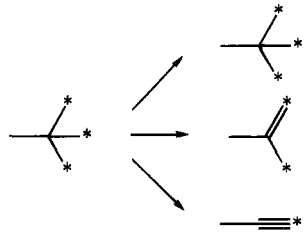

Figure 11. Consideration of multiple bonds during the reconnection phase. * stands for free bond.

3.1. Cutting Bonds. All reaction bonds of the perceived RR-site are cut. If an aromatic bond is cut, it is cut both as a single bond and double bond (Figure 10).

3.2. Reconnection of Cut Bonds and/or Addition of Atoms or Atomic Groups. All possible reconnections of cut bonds and/or addition of atoms or atomic groups are performed. In current SOPHIA, the user must specify the atoms or atomic groups for addition to the cut bonds. If the user has no idea as to how to choose the specific atoms or atomic groups, a dummy node,"*" is used. If multiple cut bonds belong to the same node, recombination is also considered. For example, as shown in Figure 11, if two cut bonds belong to a single node, they may be combined to a double bond; if three cut bonds belong to the same node, they may be transformed to a triple bond or a double bond and a single bond.

4. Reaction Path Evaluation. Each of the generated reaction path is evaluated whether it can actually occur. The procedure is shown in Figure 12. The following procedure from 4.1–4.3 is performed for every generated reaction path.

4.1. Structure Perception. For the products, basic structural information is obtained, and SSSR and aromaticity are perceived in the same manner as for reactants in RR-site perception (section IV-1). This information is used in the next procedure 4.2.

4.2. Computation of Structural Characteristics. With the basic structural information, for PR-site bonds, the bit sequences characterizing these bonds (Figure 1-[6]) are computed; they are obtained in the same manner as in the RR-site perception.

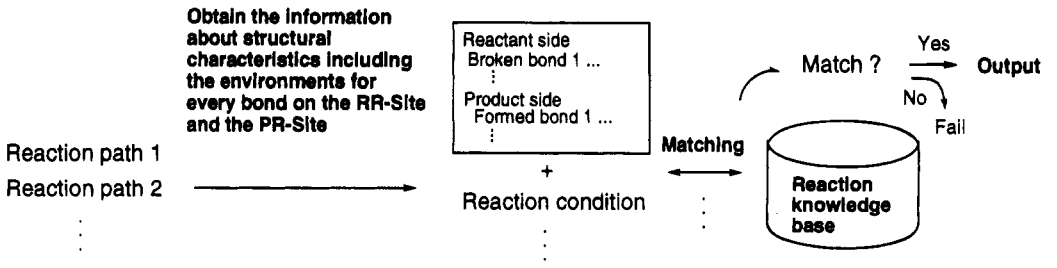

Figure 12. Reaction path evaluation.

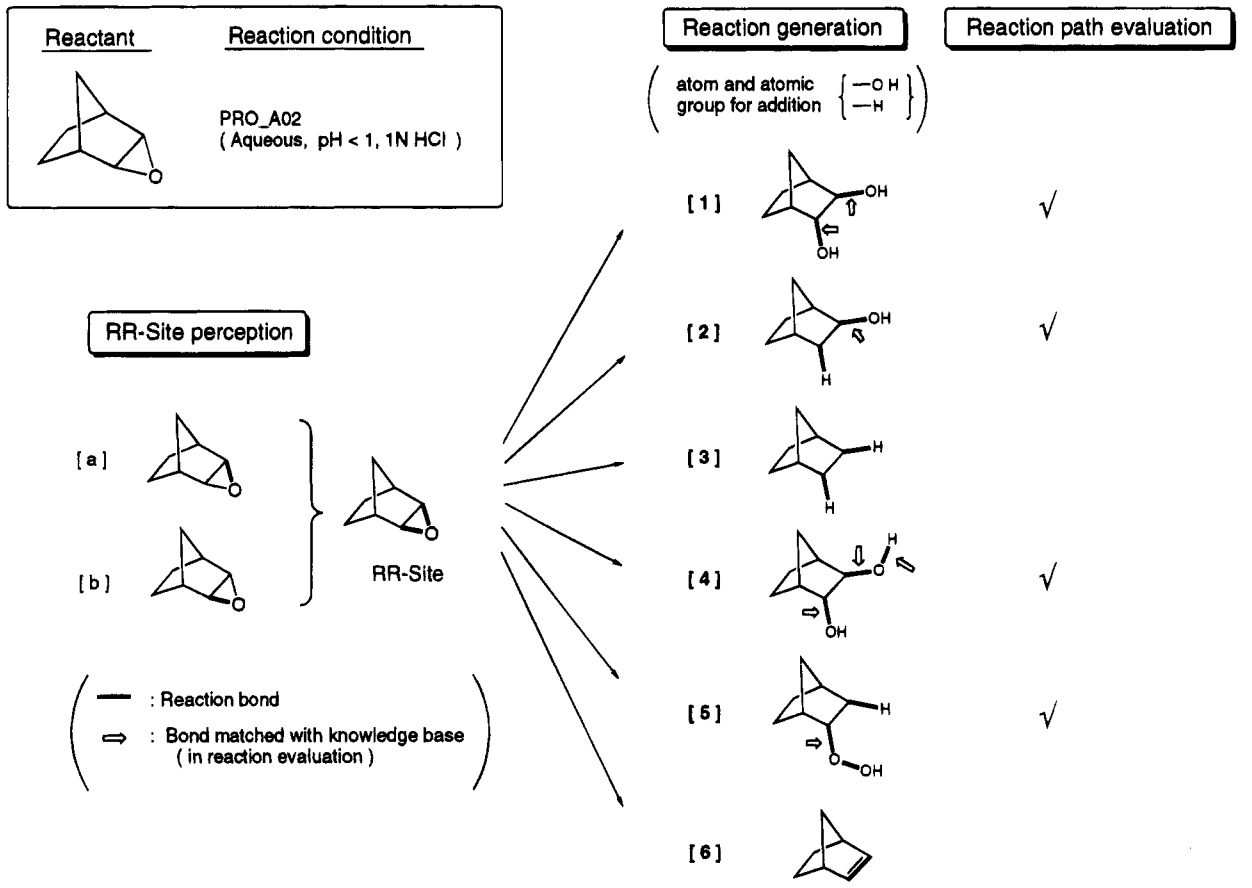

Figure 13. Sample result 1.

4.3. Matching. For any input reactant structure, if it is not trivial, multiple reaction schemes having the same RR-site are obtained in the reaction generation procedure by the combinatorial method (section VI-3). Each of these reaction schemes is evaluated by comparing the PR-site of the reaction scheme to be evaluated with the PR-site of the characterized reaction schemes picked up in the RR-site perception (section IV-2).

It is judged that when at least one of the bit sequences for the reaction bonds of the generated products under evaluation matches with one of the bit sequences in the knowledge base, this reaction path can actually occur. One of the five precision levels can be selected as described in the RR-site perception (section IV-1).

In addition, generated products with dummy nodes cannot be treated in reaction evaluation, because the structural characteristics for a bond with a dummy node cannot be computed. Therefore, it is necessary function to automatically add suitable atoms or atomic groups to suitable sites. More discussion about this point will be made in section V.

5. Output. 5.1. Checking Duplication. A check for duplication of product with the same RR-site is performed by set reduction.[14]

V. RESULTS

Four results of SOPHIA runs are given below.

About 3000 CMs and about 1500 characterized reaction schemes derived from about 2500 individual reactions in SYNLIB[15] and in our original reaction database, which is constructed from textbook reactions, are stored in the knowledge base used here.

All reactants tested here are not included neither in SYNLIB nor in our private reaction database.

1. Result 1 (Figure 13). Input reactant is epoxide, and the reaction condition is PRO_A02 (aqueous, pH < 1, 1 N HCl) shown within a rectangle on upper left of Figure 13.

A set of two C—O bonds of the epoxide ring (the bold line in structures [a] and [b] in Figure 13; the bold line stands for a reaction bond), which were perceived as reaction bonds respectively, was automatically perceived as the RR-site under this reaction condition. —H and —OH were specified

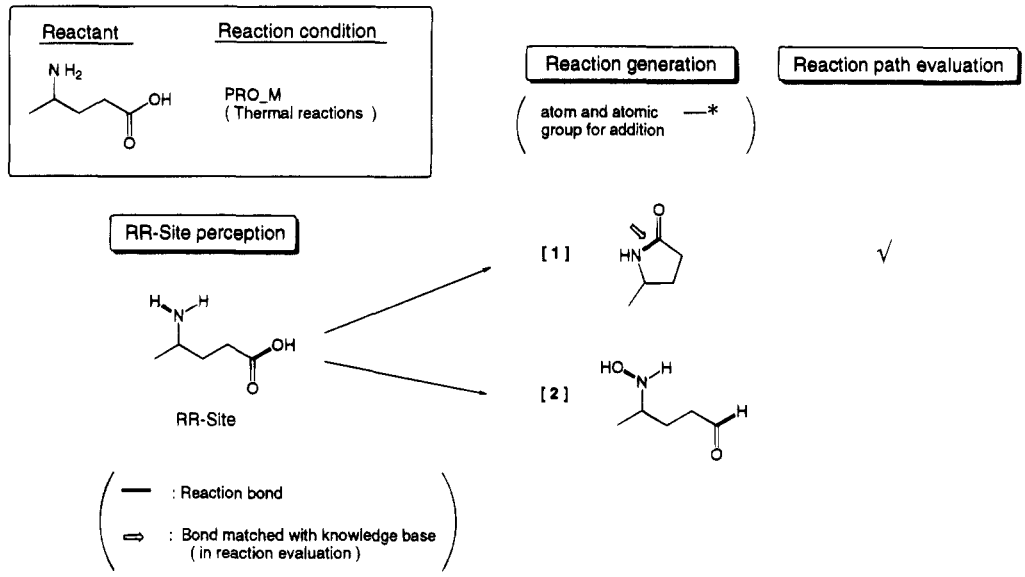

Figure 14. Sample result 2.

as an atom and atomic group for addition by the user considering the reaction condition, and six reaction products [1]−[6] in Figure 13 were generated. In Figure 13 small molecules like H_2O are omitted. Four products [1], [2], [4], and [5] were judged that they can actually be given. Every reaction bond matched with the knowledge base in reaction path evaluation procedure is indicated with an arrow.

2. Result 2 (Figure 14). Input reactant is 4-aminopentanoic acid, and the reaction condition is PRO_M (thermal reaction) shown within a rectangle in Figure 14.

One N−H bond of amine and one C−O bond of the carboxyl group were perceived as the RR-site. A "*" (dummy node) was specified by the user as an atomic group for addition. Two reaction products [1] and [2] in Figure 14 having no "*" were generated; among them it was judged that the reaction path [1] producing lactam can actually occur. In addition, six products having "*" were generated. However, these products were not treated in reaction evaluation because the structural characteristics cannot be computed for a bond with a dummy node.

3. Result 3 (Figure 15). Input reactant is 4-bromo-1-cyclohexanecarboxylic acid, and the reaction condition is PRO_H01 (hydride reduction, $LiAlH_4$) shown within a rectangle in Figure 15.

Several characterized reaction schemes in the knowledge base proposed that one C−Br bond was reactive, and several other characterized reaction schemes proposed that one C−O bond of the carbonyl group in the carboxyl group and one C−O bond attached to the hydroxyl group was reactive. A set of these three bonds was automatically perceived as the RR-site in this reaction condition. Properly, the O−H bond also should be perceived as RR-site; however, no reaction in the database used for deriving the knowledge base used here contains the cleavage of O−H bond under this reaction condition. In general, simple or synthetically trivial reactions are not compiled in a commercially available reaction database. This is a recurring problem in database utilization, and a countermeasure for solving this is required. −H was specified for addition, and 46 reaction products were generated; some of them are shown in Figure 15. It was judged that 39 products can be given in reality. These 39 products contain three proper reactions, which are reduction of only the bromide, reduction of only the carboxyl group, and reduction of both. However, a large number of products passed reaction evaluation procedure. This is because reaction evaluation procedure judges that when at least one of the bit sequences for the reaction bonds of the generated products under evaluation matches with one of the bit sequences in the knowledge base this reaction path can actually occur as mentioned in section IV-4. This evaluation method dose not always give a practical result. Thus, more detail reaction evaluation from another point of view is necessary to scope with this problem.

4. Result 4 (Figure 16). All RR-sites perceived under no specification of reaction conditions for the same reactant used in result 3 are shown in Figure 16. A bold line in a structure stands for a reaction bond.

The knowledge base used here proposed 13 RR-sites. On the left side of the dotted line in Figure 16 every RR-site proposed by one characterized reaction scheme in the knowledge base and the reaction condition are shown; on the right side reaction bonds proposed by multiple characterized reaction scheme, RR-site (a set of these reaction bonds) and the reaction condition are shown. The second RR-site on the right side is the same as that of result 3.

Reaction products are generated from every RR-site shown in Figure 16 and evaluated as shown in results 1, 2, and 3.

VI. DISCUSSION

The automated perception of reaction site without the user having to specify a reaction type or category makes it possible to consider the possibilities of reactions for a given set of reactants exhaustively. The knowledge used for perception of reaction sites and for the evaluation of reaction paths is represented in abstract form. Furthermore the reaction generation procedure, where all possible reconnections and/or additions of atoms or atomic groups are performed for cut bonds, allows generation of all possible concrete products.

These functions can lead ultimately to general reaction prediction. The current SOPHIA contains the basic functions for reaction prediction, perception of reaction sites, generation of reactions, and evaluation of the reaction paths which can be extended in the future.

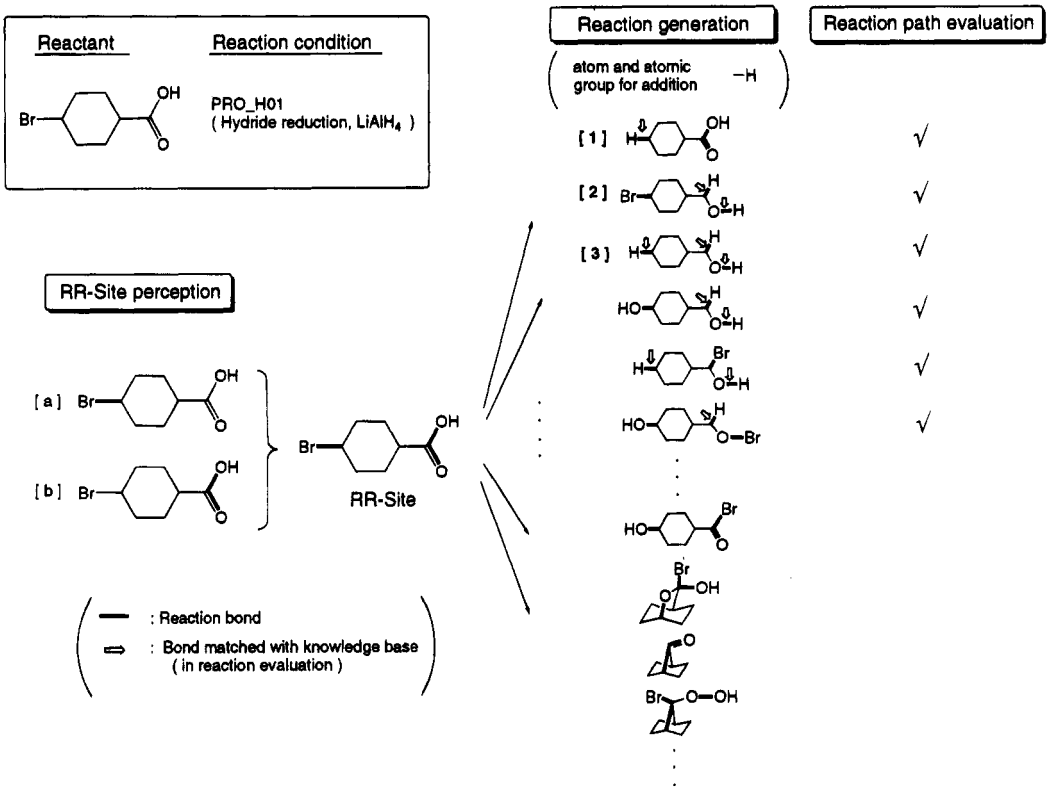

Figure 15. Sample result 3.

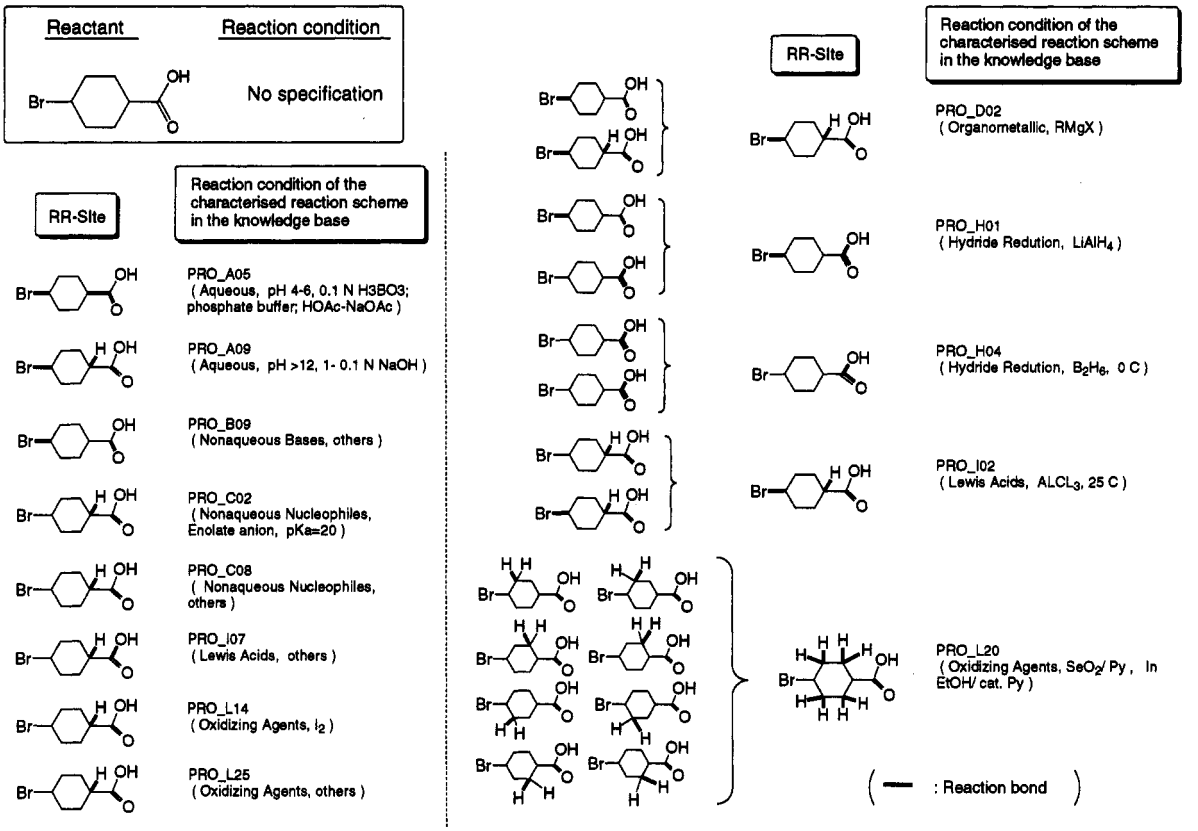

Figure 16. Sample result 4.

Plans to enhance SOPHIA are sufficient consideration of reaction conditions, automatic addition of suitable atoms or atomic groups to reaction products in reaction generation, quantitative reaction evaluation, consideration of steric effects, etc. For consideration of reaction conditions a program which can automatically interpret input reaction conditions (e.g., solvent, reagent, catalyst, temperature, concentration, time) without using reaction condition classes has been developed.[16] In reaction generation, it has been found possible to automatically add suitable atoms or atomic groups to suitable sites of product.[17] These above results will be reported in other papers. We plan to improve the

current knowledge base toward more precise, detailed, and quantitative reaction prediction. This improvement will be realized by the utilization of physicochemical parameters. In AIPHOS some physicochemical parameters (the pK_a values,[18,19] bond dissociation energy,[19] π,σ-charge, residual electronegativity[20,21] and so on) with influence on chemical reactions can already be calculated. It might be necessary to classify and organize chemical reactions from one novel point of view by investigating the correlation among these physicochemical parameters, substructure transformation, and reaction conditions. We also plan to consider steric congestion in the quantitative evaluation of reaction paths.

VII. CONCLUSION

The current SOPHIA made it possible to predict chemical reactions more generally without specification of a conventional type or category of chemical reaction. Although SOPHIA cannot predict the product ratio and not sufficiently consider reaction conditions yet, the framework of the system has been realized, and first results were given.

We expect SOPHIA to complement a synthesis design system like AIPHOS as described in the introduction. Furthermore it has potential to discover unknown reactions by automated perception of reaction sites using a knowledge base derived from a large quantity of individual chemical reaction data and by generation of all possible products.

ACKNOWLEDGMENT

We wish to thank Dr. Wolf-Dietrich Ihlenfeldt for his useful discussion in the preparation of this paper. This work has been supported by the Hori Information Science Promotion Foundation.

REFERENCES AND NOTES

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

(2) 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.

(3) Pensak, D. A.; Corey, E. J. *LHASA—Logic and Heuristics Applied to Synthesis Analysis. In Computer-Assisted Organic Synthesis*; ACS Symposium Series 61; Wipke, W. T., Howe, W. J., Eds.; American Chemical Society: Washington, DC, 1977; pp 1−32.

(4) Wipke, W. T.; Brawn, H.; Smith, G.; Shoplin, F.; Sieber, W. SECS-Simulation and Evaluation of Chemical Synthesis. In *Computer-Assisted Organic Synthesis*; ACS Symposium Series 61; Wipke, W. T., Howe, W. J., Eds.; American Chemical Society: Washington, DC, 1977; pp 97−127.

(5) Gasteiger, J.; Hutchings, M. G.; Christoph, B.; Gann, L.; Hiller, C.; Löw, P.; Marsili, M.; Saller, H.; Yuki, K. A New Treatment of Chemical Reactivity; Development of EROS, an Expert System for Reaction Prediction and Synthesis Design. *Top. Curr. Chem.* **1987**, *137*, 19−73.

(6) Gasteiger, J.; Ihlenfeldt, W. D.; Röse, P. A collection of computer methods for synthesis design and reaction prediction. *Recl. Trav. Chim. Pays-Bas.* **1992**, *111*, 270−290.

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

(8) Funatsu, K.; Shiraishi, Y.; Nishimura, T.; Takahashi, Y.; Nomura, S.; Kitamura, S.; Kimura, S.; Korogi, K.; Chiba, M.; Takabatake, T.; Uchida, T.; Dogane, I.; Sasaki, S. Development of Organic Synthesis Planning System AIPHOS (1)—Development of Knowledge Base and Specific Database System of Organic Reactions. In *Proceedings of the 13th Symposium on Chemical Information and Computer Sciences/18th Symposium on Structure−Activity Relationships*; Sasaki, S., Ed.; Toyohashi Univ. of Tech.: Toyohashi, Japan, 1990; pp 165−168.

(9) Funatsu, K.; Shiraishi, Y.; Nishimura, T.; Takahashi, Y.; Koremoto, T.; Dogane, I.; Sasaki, S. Development of Organic Synthesis Design System AIPHOS (5)—Development of Knowledge Base System. In *Proceedings of the 14th Symposium on Chemical Information and Computer Sciences/19th Symposium on Structure−Activity Relationships*; Hosoya, H., Ed.; Ochanomizu Univ.: Tokyo, Japan, 1991; pp 38−41.

(10) Funatsu, K.; Del Carpio, C. A.; Sasaki, S. Automatic Perception of Structure Characteristics of Organic Compound Enhancing Molecular Reactivity Directed to the Planning of Organic Synthesis. *Tetrahedron Comput. Method.* **1988**, *1*, 39.

(11) Funatsu, K.; Ohno, T.; Isozaki, M.; Horiuchi, K.; Dogane, I.; Sasaki, S. Development of Organic Synthesis Planning System AIPHOS (3)—Development of Aromaticity/tautomer Perception Program. In *Proceedings of the 13th Symposium on Chemical Information and Computer Sciences/18th Symposium on Structure−Activity Relationships*; Sasaki, S., Ed.; Toyohashi Univ. of Tech.: Toyohashi, Japan, 1990; pp 154−157. Funatsu, K.; Ohno, T.; Isozaki, M.; Horiuchi, K.; Dogane, I.; Sasaki, S. Development of Organic Synthesis Design System AIPHOS (7) − Development of Strategy Program. In *Proceedings of the 14th Symposium on Chemical Information and Computer Sciences/19th Symposium on Structure-Activity Relationships*; Hosoya, H., Ed.; Ochanomizu Univ.: Tokyo, Japan, 1991; pp 154−157. Funatsu, K.; Isozaki, M.; Horiuchi, K.; Dogane, I.; Sasaki, S. Development of Organic Synthesis Design System AIPHOS (9)—Development of Program for Acquisition of Strategy site. In *Proceedings of the 16th Symposium on Chemical Information and Computer Sciences/21st Symposium on Structure−Activity Relationships*; Tsukihara, T., Terada, H., Eds.; Tokushima Univ.: Tokushima, Japan, 1993; pp 93−96.

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

(13) Wipke, W. T.; Dyott, T. M. Use of Ring Assemblies in a Ring Perception Algorithm. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 140−147. Schmidt, B.; Felischauer, J. A Fortran IV Program for Finding the Smallest Set of Smallest Rings of a Graph. *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 204−206.

(14) Sussenguth, E. H. A Graph-Theoretic Algorithm for Matching Chemical Structures. *J. Chem. Doc.* **1965**, *5*, 36−43. Funatsu, K.; Miyabayashi, N.; Sasaki, S. Further Development of Structure Generation in the Automated Structure Elucidation System. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 18−28.

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

(16) Satoh, H.; Funatsu, K. Development of Organic Reaction Prediction System SOPHIA (2)—Automatic Interpretation of a Reaction Condition. In *Proceedings of the 17th Symposium on Chemical Information and Computer Sciences/22nd Symposium on Structure−Activity Relationships*; Itai, A., Ed.; Tokyo Univ.: Tokyo, Japan, 1994; pp 66−69.

(17) Satoh, H.; Sano, T.; Funatsu, K. Development of Organic Reactions Prediction System SOPHIA (3)—Knowledge-Guided Automatic Addition of Suitable Atoms or Atomic Groups (SAAG) in Reaction Generation. In *Proceedings of the 17th Symposium on Chemical Information and Computer Sciences/22nd Symposium on Structure−Activity Relationships*; Itai, A., Ed.; Tokyo Univ.: Tokyo, Japan, 1994; pp 90−93.

(18) Gushurst, A. J.; Jorgensen, W. L. Computer-Assisted Mechanistic Evaluation of Organic Reactions. 12. pK_a Predictions for Organic Compounds in Me_2SO. *J. Org. Chem.* **1986**, *51*, 3513−3522.

(19) Funatsu, K.; Ueyama, N.; Watanabe, M.; Negishi, Y.; Oohashi, T.; Sasaki, S. Development of Organic Synthesis Planning System AIPHOS (2)—Estimation of Bond Dissociation Energies. In *Proceedings of the 13th Symposium on Chemical Information and Computer Sciences/18th Symposium on Structure−Activity Relationships*; Sasaki, S., Ed.; Toyohashi Univ. of Tech.: Toyohashi, Japan, 1990; pp 169−172.

(20) Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity − A Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36*, 3219−3288.

(21) Marsili, M.; Gasteiger, J. π Charge Distribution from Molecular Topology and π Orbital Electronegativity. *CCACAA* **1980**, *53*, 601−614.

CI940031B