A Novel Approach to Retrosynthetic Analysis Using Knowledge Bases Derived from Reaction Databases

Koji Satoh and Kimito Funatsu*,†

Chemical Technology Laboratories, Daiichi Pharmaceutical Co., Ltd., Edogawa-ku, Tokyo 134-8630, Japan, and Department of Knowledge-Based Information Engineering, Toyohashi University of Technology, Tempaku, Toyohashi, Aichi 441-8580, Japan

Received September 10, 1998

A novel approach to empirical retrosynthetic analysis of computer-assisted synthesis design is proposed using the KOSP (Knowledge base-Oriented system for Synthesis Planning) system. KOSP has four functions: (1) strategic site pattern perception, (2) precursor skeleton generation, (3) retrosynthetic scheme evaluation, and (4) retrosynthetic analysis termination. This system is based on knowledge bases in which reactions are abstracted by structural and/or stereo characteristics of reaction sites and their environments. This article describes the framework of KOSP and the results of its execution.

1. INTRODUCTION

Today, in designing synthetic routes for a desired organic compound, organic chemists usually use the concept of retrosynthetic analysis advocated by Corey.[1,2] This is a phased problem-solving method for transforming the structure of a desired organic molecule to simpler structures, such as previously reported or commercially available compounds. This concept is embodied in empirical knowledge-oriented synthesis design systems such as LHASA[3] and SECS[4] in the chemical information field; the LHASA system especially is evaluated as a prospective system.[5] These systems require large transform knowledge bases with sufficient quality and quantity to solve important synthetic problems and yield proper results. Although some ideas have been suggested to solve this burdensome problem, it is difficult to construct knowledge bases on a practical scale. Furthermore, even if monumental knowledge bases that involve sufficient fundamental reactions are constructed at a certain point of time, it is necessary to continuously update knowledge bases for registering novel and effective reactions being developed in the organic synthesis field. Hence, these systems are very useful but have appeared to reach certain conceptual limits, as noted in a recent review article.[6]

On the other hand, reaction databases potentially include much information about chemical reactions (e.g., reaction conditions, yields, catalysts, reagents, and references) and are continuously updated from year to year. Reaction databases have been widely accepted and used in many chemical research laboratories,[6] and an approach toward the synthesis of target molecules using reaction databases such as IRDAS[7] was reported recently.[8] Users must, however, already have contrived the general outline of synthesis in database-oriented synthesis planning before they search whether a planned reaction scheme has an apparent connection to a literature precedent in databases. Empirical knowledge base-oriented synthesis planning systems are therefore attractive for chemists because they can propose retrosynthetic paths based on their knowledge bases without the user's framework of the synthesis of the desired target molecule.

Further, a knowledge base, which could be automatically converted from a Beilstein Reaction Database,[9,10] was reported by Lawson and Kallies.[11] Reactions in this knowledge base are represented by the RABBIT coding[11] of extended reaction centers. This concept is the first incorporation of a large knowledge base, which is different from transform-knowledge bases in the representation of reactions, with the synthesis planning technique. In this regard, this concept is unique. However, this approach has an obvious drawback because reaction conditions and stereochemistry cannot be considered in synthesis planning.

Considering these issues, we began to develop a novel empirical synthesis design system by use of knowledge bases that are free from the disadvantages of transform-knowledge base. This system is called KOSP (Knowledge base-Oriented system for Synthesis Planning). We have concretized a similar concept in the reaction prediction system SOPHIA.[12,13] However, this approach is a first attempt, to our knowledge, for retrosynthetic analysis. Here, the knowledge bases of Artificial Intelligence for Planning and Handling Organic Synthesis (AIPHOS),[14] in which reactions are abstracted by structural and/or stereo characteristics around reaction sites, are applied to this approach because these knowledge bases can be derived automatically from reaction databases. AIPHOS is designed to propose suitable retrosynthetic routes from the standpoint of both novelty and practicality based on knowledge bases. Because KOSP proposes empirically more practical retrosynthetic schemes than those of AIPHOS, the concept of KOSP is different from that of AIPHOS.

2. KNOWLEDGE BASES OF AIPHOS

The current AIPHOS system proceeds stepwise, interactively, determining at each step the synthetic precursors from molecules of the former step. Better proposals are suggested by using the user's synthetic strategies.

† Toyohashi University of Technology.

10.1021/ci980147y CCC: $18.00
Published on Web 02/10/1999

The user communicates with the program via a graphical interface, which is used for drawing the desired target molecule, for selecting proposed strategic site patterns and precursors, for adding suitable leaving groups (SLG) with the leaving-group knowledge base (LGKB) automatically, for evaluating retrosynthetic schemes with the reaction knowledge base (RKB), and for displaying proposed retrosynthetic schemes, which are evaluated that those occur, with related reaction schemes in the reaction database (RDB). In acquiring strategic site patterns, the topological strategy and the functional group-based strategy, which are described by Corey,[1] are applied to the target molecule. If necessary, the user can specify strategic site patterns and add appropriate leaving groups according to his/her synthetic strategies manually.

Thus, the current AIPHOS uses two knowledge bases: the RKB and the LGKB. Although the details of these knowledge bases are described in previous articles,[12,13] a brief outline of the RKB and the LGKB is needed to profitably follow the development of KOSP.

2.1. Reaction Knowledge Base of AIPHOS. Reaction schemes stored in the RKB are represented as information characterized by the location of key substructures and chiral centers around reaction sites of the reactant (reactant reaction site, RR site) and product (product reaction site, PR site) molecules and by type of reaction condition. Each set of the characterized reaction sites of each structure on the reactant or product side has an ID number (ID number of the Characterized Molecule, IDCM). For example, IDCMs are m_00001, m_00002, m_00003, and so on. The numerals in IDCMs are keys of the RKB and have no specific meanings. The classification of reaction conditions summarized by Greene[15] is used, with the reaction conditions divided into 16 categories as the primary classification, and each category further subdivided into several subclasses (Table 1). Thus, every individual reaction is represented as a pair of the transformation of the structural and stereo characteristics of the reaction site during the reaction and the abstracted reaction condition in the RKB, as shown in Figure 1. Information on structural and stereo characteristics of reaction sites and their environments (Structural and Stereo Characteristics of a reaction site and its Environments, SSCE) is represented as an aggregate of bit sequences. The SSCE are computed on the basis of structural characteristics keys (Table 2) and stereo characteristics keys (Table 3). By using these characteristics keys, AIPHOS can deal with stereoselective retrosynthetic analysis. An example of PR-site representation of Figure 1 in the RKB is shown in Figure 2. For the product structure given in Figure 2, the fourth bit is set because the PR site is a single bond, and the atoms forming this PR site are carbons; hence, the sixth bit is set (Figure 2, [1]). The structural characteristic of the atoms belonging to the PR site is —C(=O)—R, so the bit corresponding to key no. 34 is set (Figure 2, [2]). The structural characteristics on atoms at the α, β, and γ positions are obtained in the same manner (Figures 2, [3], [4], and [5]). Further, the second bit is set because the product structure given in Figure 2 has a chiral center, and this reaction, as shown in Figure 1, proceeds with retention of configuration of the chiral center. The chiral center in this reaction exists on an atom at the β position of the PR site, and an oxygen atom is attached to the node; therefore, the bits corresponding to key nos. 7 and 48 are set (Figure 2, [6]). Finally, these six-bit sequences are concatenated as the characterized information on this PR site (Figure 2, [7]).

Table 1. Part of Reaction Condition Keys in the RKB

ID no.	reaction conditions
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	thermal reactions
PRO_M	oxidizing reagents
PRO_N	carbenoids
PRO_O	miscellaneous
PRO_P	electrophiles
PRO_Q	others
Part of Second Classification PRO_B: Nonaqueous Bases	
PRO_B01	NaH
PRO_B02	Ph_3CNa
PRO_B03	$C_{10}H_8Na$
PRO_B04	CH_3SOCH_2Na
PRO_B05	*t*-BuOK
PRO_B06	LDA
PRO_B07	pyridine; Et_3N
PRO_B08	$NaNH_2$; NaNHR
PRO_B09	others

This RKB is composed of four direct access files which hold data links as a network structure (Figure 3). They are a reaction site file (Figure 3A), a reaction scheme file (Figure 3B), an auxiliary accelerator file (Figure 3C), and a reaction data file (Figure 3D). The IDCMs and their SSCEs are represented in file A. The characterized reaction schemes, their ID numbers, and their reaction condition key numbers are described in file B. File C describes how many bit sequences each of the IDCM has and on which side in the characterized reaction schemes in file B this IDCM appears: on the reactant side and/or on the product side. Each of the characterized reaction scheme ID numbers with ID numbers of the individual reaction schemes in the RDB deriving this characterized reaction scheme is represented in file D. For this network structure as shown in Figure 3, all related pieces of information can be selected from a single piece of input information. The RKB is derived from the RDB, which was incorporated from SYNLIB[16] and constructed with the atom-mapping resulting from the set of manipulation instructions of SYNLIB. The KOSP system uses this RKB in strategic site pattern perception and retrosynthetic scheme evaluation.

2.2. LGKB of AIPHOS. In the AIPHOS system, leaving groups are defined as *atoms and/or atomic groups present in reactants molecules and absent in product molecules*. This definition allows AIPHOS to automatically recognize leaving groups and to construct the leaving-group database of AIPHOS (LGDB) from the RDB. The structural information (e.g., atomic species, the connectivity, and the coordinates) of LG, numbers of LG, and ID numbers of the individual reactions from which the LG were extracted are stored in the LGDB as a direct access file. For example, —H and —

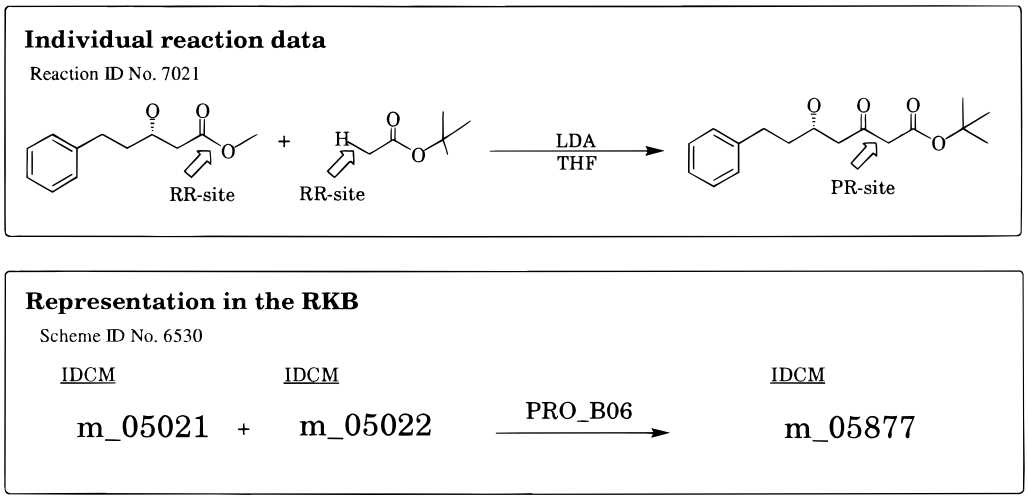

Figure 1. Representation of a reaction scheme in the RKB. LDA, lithium diisopropylamide; THF, tetrahydrofuran.

Table 2. Part of the Structural Characteristics Keys

key no.	structural characteristics	key no.	structural characteristics	key no.	structural characteristics
1	$-CH_3$	48	$-N(H)-C(=O)-NH-$	100	$-N=C<$
2	$-CH_2R$	49	$-C(=O)-S-$	101	$-C=N-$
3	$-CHR_2$	50	$-SO_2OH$	102	$-C\equiv C-$
4	$-CR_3$	⋮	⋮	103	$N\equiv C-$
5	$-CH(CH_3)_2$			⋮	⋮
⋮	⋮	77	△		
34	$-C(=O)R$	⋮	⋮	108	$-C(=O)H$
35	$-C(=O)H$	78	□	⋮	⋮
36	$-CO_2H$	⋮	⋮	111	$-NO_2$
37	$-CO_2R$	80	⬡		
⋮	⋮	⋮	⋮		

Table 3. Part of the Stereo Characteristics Keys

key no.	stereo characteristics	
1	Generation of new chiral center(s)	
2	Progress of a reaction via retention	
3	Progress of a reaction via inversion	
4	Existence of a chiral center on a reaction site	
5	Existence of chiral centers on a reaction site	
6	Existence of chiral center(s) at the α-position on a reaction site	
7	Existence of chiral center(s) at the β-position on a reaction site	
8	Existence of chiral center(s) at the γ-position on a reaction site	
⋮	⋮	
29	External induction of chiral center(s) via reagents and/or catalysts	
⋮	⋮	
40	Connection with oxygen atom(s)	Chiral center(s) at the α-position
41	Connection with sulfur atom(s)	
42	Connection with nitrogen atom(s)	
43	Connection with phosphorus atom(s)	
⋮	⋮	
48	Connection with oxygen atom(s)	Chiral center(s) at the β-position
49	Connection with sulfur atom(s)	
50	Connection with nitrogen atom(s)	
51	Connection with phosphorus atom(s)	
⋮	⋮	
72	Connection with oxygen atom(s)	Chiral center(s) at the ϵ-position
73	Connection with sulfur atom(s)	
74	Connection with nitrogen atom(s)	
75	Connection with phosphorus atom(s)	

OMe are recognized as LG of the individual reaction, which is the same in Figure 1, shown at the top of Figure 4. (The LG are illustrated in circles in the reactant structure; bold lines in the reactant and product structures represent RR and PR sites, respectively.) The lower left portion of Figure 4 shows stored information: structural LG, ID number of the individual reaction (no. 7021), and ID numbers of the LG (nos. 1 and 39).

The LGKB is also derived from the RDB automatically. This LGKB is organized into two files for product and reactant. The product file accumulates an ID number of an individual reaction and information of structural characteristics of the PR site and its environments (Structural Characteristics of a reaction site and its Environments, SCE) (SCE of PR site) concerning LG elimination. The reactant file stores an ID number of an individual reaction, a pairing number, LG ID numbers, and SCE of the RR site to which LG are attached. The pairing number is the ID number of pairing LG; here, pairing LG is a group of LG related to each other in electron shift when they are removed from a reactant skeleton. Other types of pairing LG are also considered, such as intramolecular multiple LG between which the distance is more than vicinal, and intermolecular multiple LG. Because the reactant file is used when SLG are determined and added to free bonds, which are bonds with dummy atoms, of a precursor skeleton, the method of RR-site characterization is different from that of PR-site characterization in the product file; SCE of the RR site in which LG are replaced with dummy atoms are stored in the reactant file. If multiple LG are perceived in a reactant structure, structural characteristics are computed for each reactant skeleton (reactant structure with dummy atoms) from which each of the LG has been removed. The lower right portion of Figure 4 shows knowledge derived from the same reaction as used for explanation of LGDB derivation. In this example, the LG are —H and —OMe, which are pairing LG. The ID number of the individual reaction (no. 7021) and the SCE of the PR site (a C—C bond of a product structure) are stored in the product file. The reactant file stores the pairing no. 3100, the LG ID nos. 1 and 39, and the SCE of

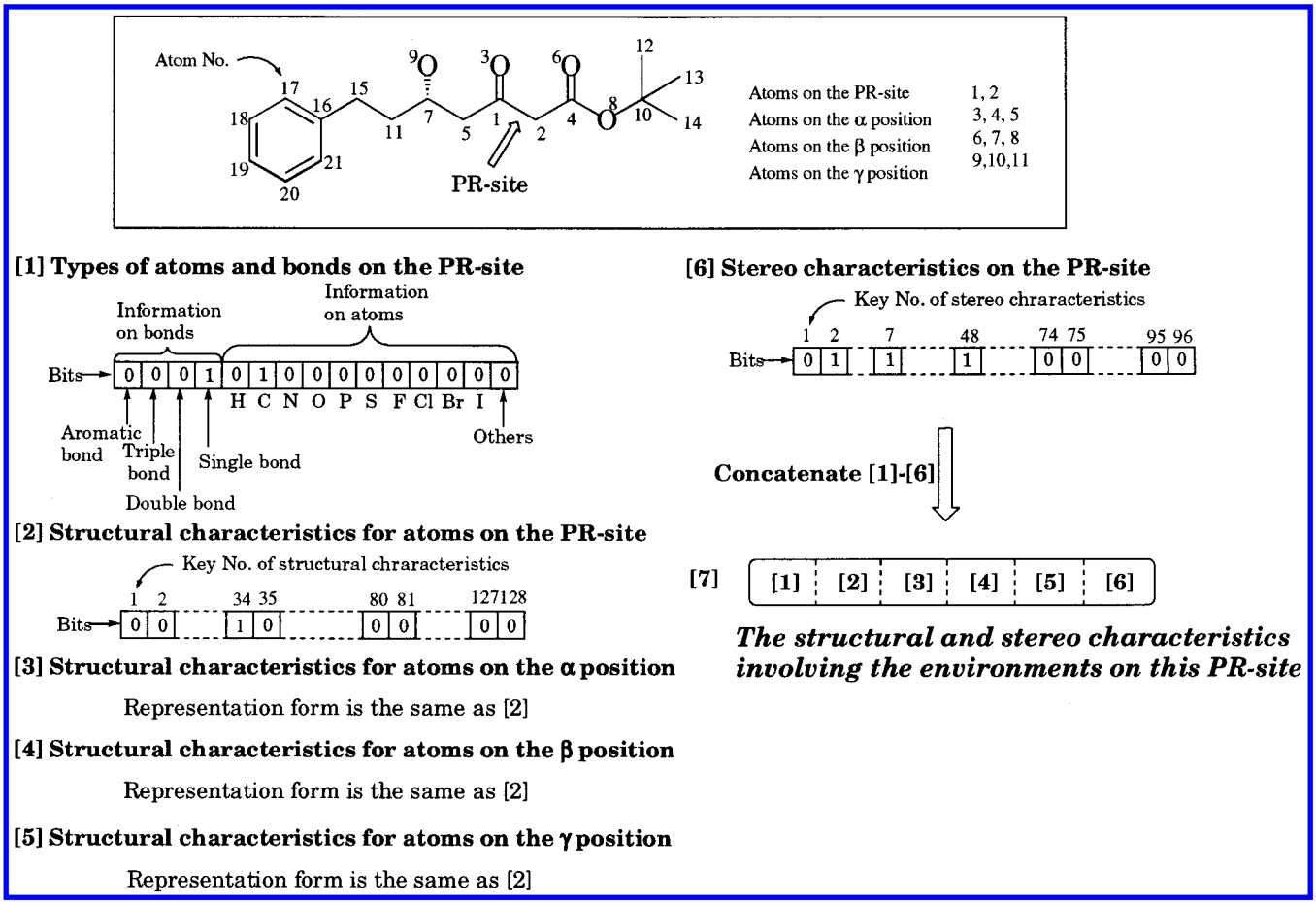

Figure 2. Representation of a PR site in the RKB.

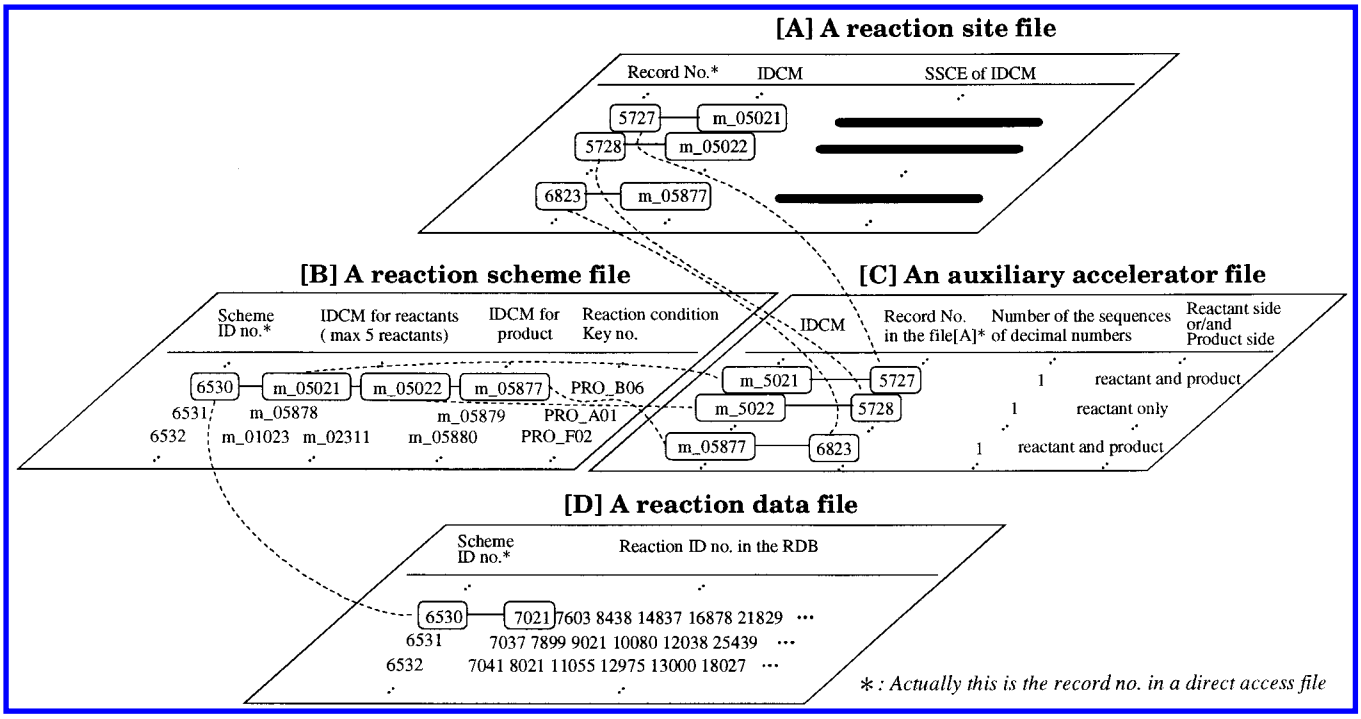

Figure 3. The architecture of the RKB.

the RR site to which dummy atoms are attached (bold lines of the chemical structures at the center of Figure 4), and the ID number of the individual reaction (no. 7021). All files organizing the LGDB and LGKB are direct access files, and the data in these files are linked with each other.

3. RETROSYNTHETIC ANALYSIS PROCEDURES IN KOSP

The KOSP system is roughly divided into four stages, as shown in Figure 5. First, KOSP automatically acquires all possible strategic site patterns for an input target molecule

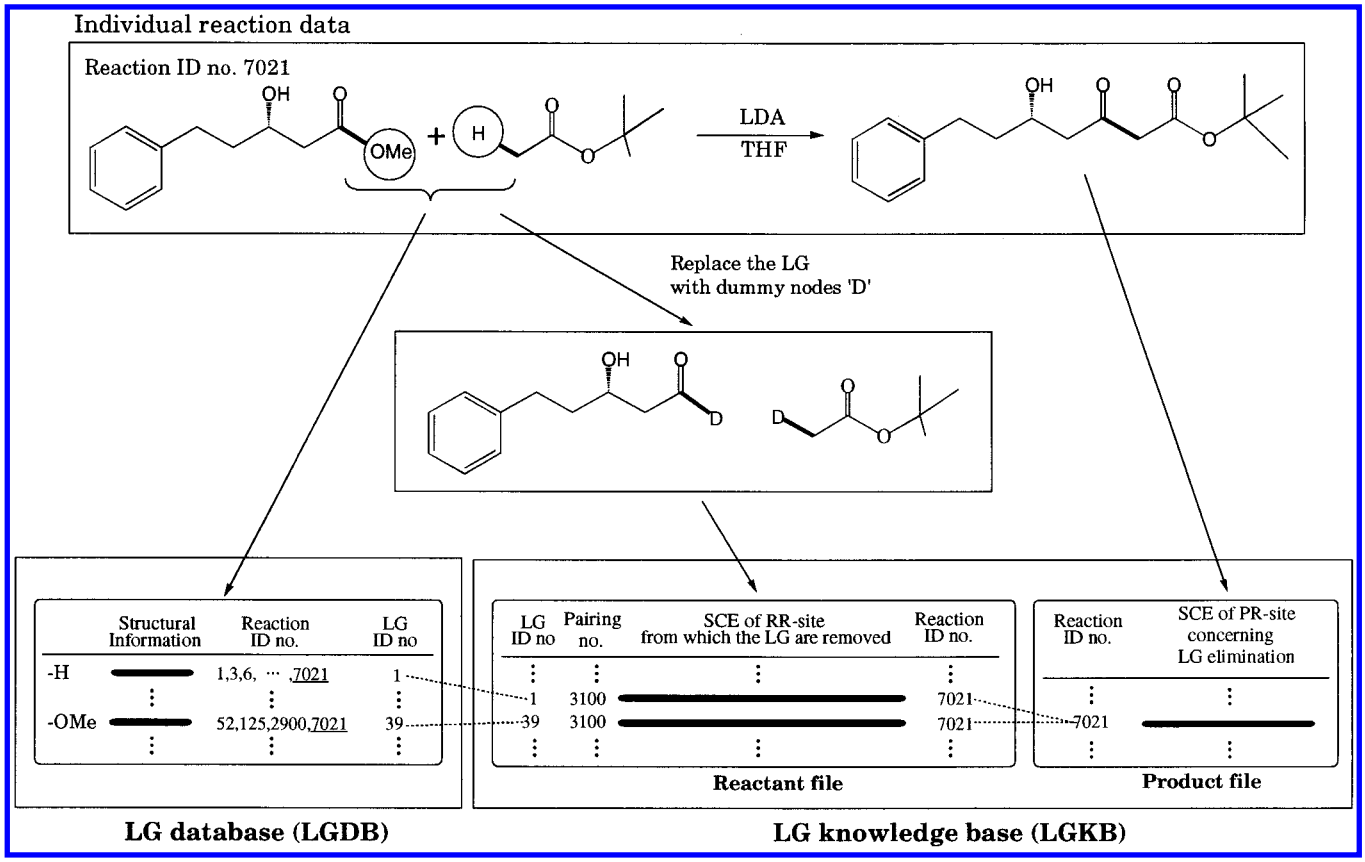

Figure 4. The structure of LGDB and LGKB.

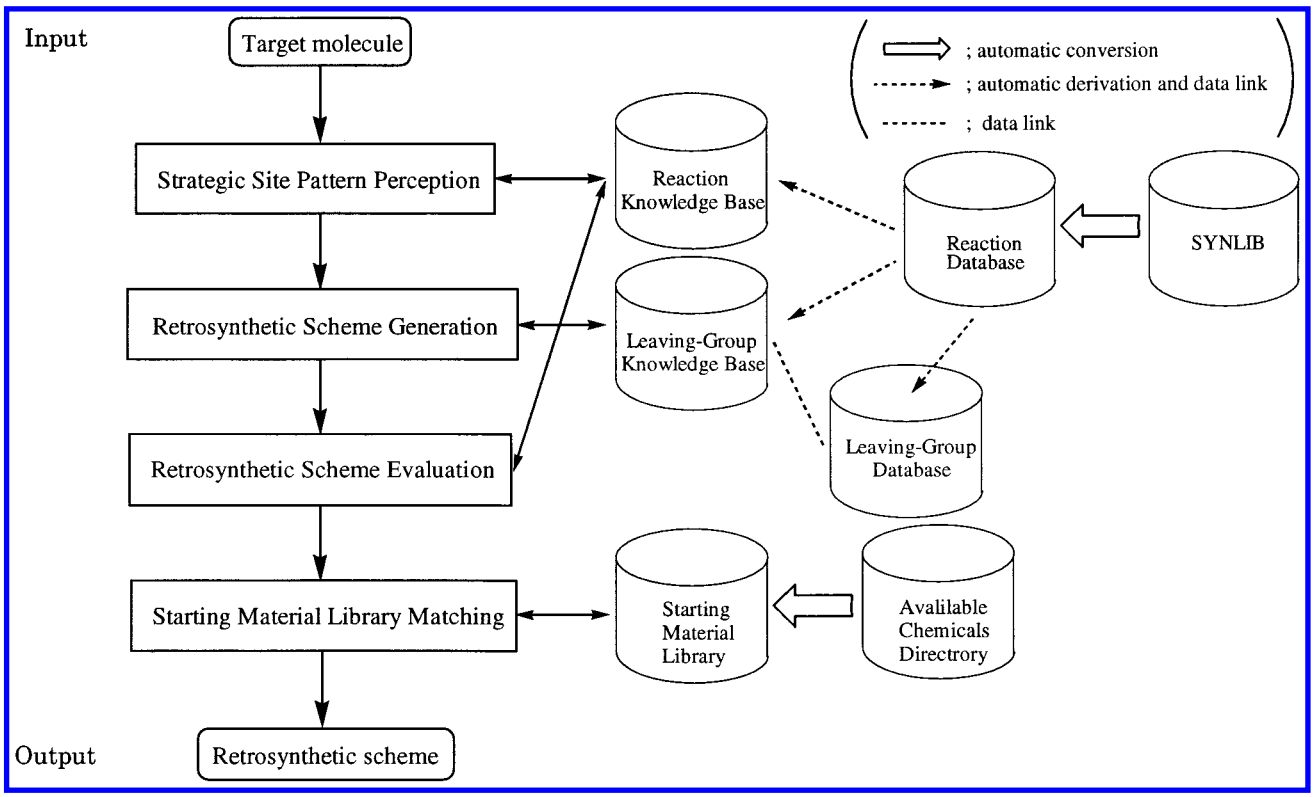

Figure 5. The outline of KOSP.

by using the RKB. In the next step, bonds for constructing each strategic site pattern are cut, and all possible reconnections of cut patterns and/or addition of appropriate atoms or atomic groups to them are performed to generate all possible precursors. Then, each generated retrosynthetic scheme is evaluated to determine whether it occurs by the RKB. Finally, the system judges whether retrosynthetic analyses can be terminated by comparing precursor structures with the starting material library of AIPHOS.[17] The details are described below.

3.1. Automatic Perception of Strategic Site Pattern. Acquisition of strategic site patterns is important for retrosynthetic analysis because it determines whether proposed retrosynthetic schemes are prospective ones. To make detection of attractive and plausible strategic site patterns from the standpoint of chemists possible using the RKB, in the KOSP system a strategic site pattern is defined as a set of bonds that is similar to characterized reaction schemes in the RKB. With this definition, KOSP can acquire prospective strategic site patterns automatically. The actual procedure for strategic site pattern perception is as follows: Input of a desired target structure with a graphical editor is performed, the user specifies whether stereoselective retrosynthetic analyses are considered, and basic structural information (e.g., types of atoms and bonds, connectivity) is computed. The smallest set of smallest ring (SSSR)[18,19] and the aromaticity based on Hückel's rule are determined. With this information, the structural and stereo characteristics (Tables 2 and 3) are recognized, and the bit sequences for every bond in the target molecule are obtained. Next, every combination of the bit sequences from one bond to six bonds of the target molecule is generated in turn, and the bit sequences of the generated combination are compared with the bit sequences in the reaction site file (Figure 3A) by one-to-one correspondence. Five levels can be considered in this method of comparison. The system judges a match according to one of five levels: (1) minimum matching level 1 is a match of type of bond and type of atoms belonging to this bond; (2) minimum matching level 2 is a match of SCE of the atoms of level 1; (3) minimum matching level 3 is a match of SCE of the atoms of level 1 including the information around their α position; (4) minimum matching level 4 is a match of SCE of the atoms of level 1 including the information around their β position; (5) minimum matching level 5 is a match of SCE of the atoms of level 1 including the information around their γ position. Level 2 is default, although the user can select another level if necessary. If the user specifies consideration of stereoselective retrosynthetic analyses, the system combines these five levels and stereo characteristics into a criterion for this matching procedure.

This comparison procedure continues until all possible combinations up to six bonds of the target molecule are generated. By this procedure, sets of all bonds judged to be disconnective are automatically perceived as strategic site patterns. Here, if necessary, acquired strategic site patterns can be ordered with the depth of matching level and/or number of original data in the RDB corresponding to the characterized reaction schemes of the RKB that were the basis for acquiring each strategic site pattern.

3.2. Retrosynthetic Scheme Generation. This procedure is required for KOSP to obtain precursor structures as concrete forms in using the RKB in which reactions are described as abstract forms by SSCE. All possible structure of precursors are generated from the perceived strategic site pattern on the basis of the cyclic permutation also used in the reaction generator of AIPHOS. This generating procedure is continued for all acquired strategic site patterns.

3.2.1. Precursor Skeleton Generation. KOSP cuts all bonds involving the strategic site pattern and generates all possible precursor skeletons by reconnecting free bonds and by adding dummy atoms to the free bonds. If the user specifies consideration of stereoselective retrosynthetic analyses at the strategic site pattern perception, the system generates all possible precursor skeletons corresponding to reactions that proceed with retention and inversion. If an aromatic bond is cut, the system considers this bond both as a single bond and double bond. Furthermore, if a multiple bond is cut, recombination can also be considered.

3.2.2. Determination of SLG. KOSP determines SLG for the proposed precursor skeletons by using the LGKB. SLG determination is composed of three steps: (1) matching original data in the RDB corresponding to characterized reactions of the RKB that were the basis for proposal of the strategic site pattern, (2) matching the precursor skeletons, and (3) determining the SLG. Details of these steps are described below.

Matching Original Data in the RDB Corresponding to Characterized Reactions of the RKB That Were the Basis for Proposal of the Strategic Site Pattern. The system judges whether original reaction data in the RDB corresponding to characterized reactions of the RKB, which were the basis for acquiring the strategic site pattern, have the leaving groups by comparison with the LGDB. If the data are involved in the LGDB, the precursor skeletons are sent to the next step. If not, the SLG determination procedure is discontinued, and only precursor structures that have no dummy atoms can be sent to the retrosynthetic scheme evaluation step (section 3.3).

Matching the Precursor Skeletons. The system compares SCE of RR sites concerning the LG elimination of the free bonds that were involved in the generated skeletons with those in the LGKB. Here, only SCE of RR sites related to LG elimination in the LGKB, corresponding to the ID numbers of original data, which were the basis for proposing the strategic site patterns, are considered. This matching procedure is performed for both cases of the pairing LG being considered and those not being considered. This procedure also considers the five levels of matching used for the acquiring strategic site patterns (section 3.1).

Determination of SLG. SLG for the free bonds of each precursor skeleton are determined by three levels of priority: Priority 1, LG of a matched characterized reaction in the LGKB are pairing, and they were extracted from a common individual reaction; Priority 2, LG of a matched characterized reaction in the LGKB are pairing, and the depth of matching level is high and the number of matchings is maximum as computed in matching of the precursor skeletons; Priority 3, the depth of matching level is high, and the number of matchings is maximum as computed in the matching of the precursor skeletons. Pairing LG is not considered.

3.2.3. Addition of SLG to Precursor Skeleton. The system replaces the corresponding dummy atoms of the precursor skeletons with the substructures of the determined SLG obtained from the LGDB. Thus, the system generates the precursor structures.

3.3. Retrosynthetic Scheme Evaluation. Whether each of the retrosynthetic schemes corresponding to all of the generated precursor structures occurs is determined. The evaluation procedure is as follows: First, basic structural information for a set of precursors in a retrosynthetic path is obtained, and SSSR and aromaticity are recognized in the same manner as for strategic site pattern perception (section 3.1). Second, with use of the basic structural information or

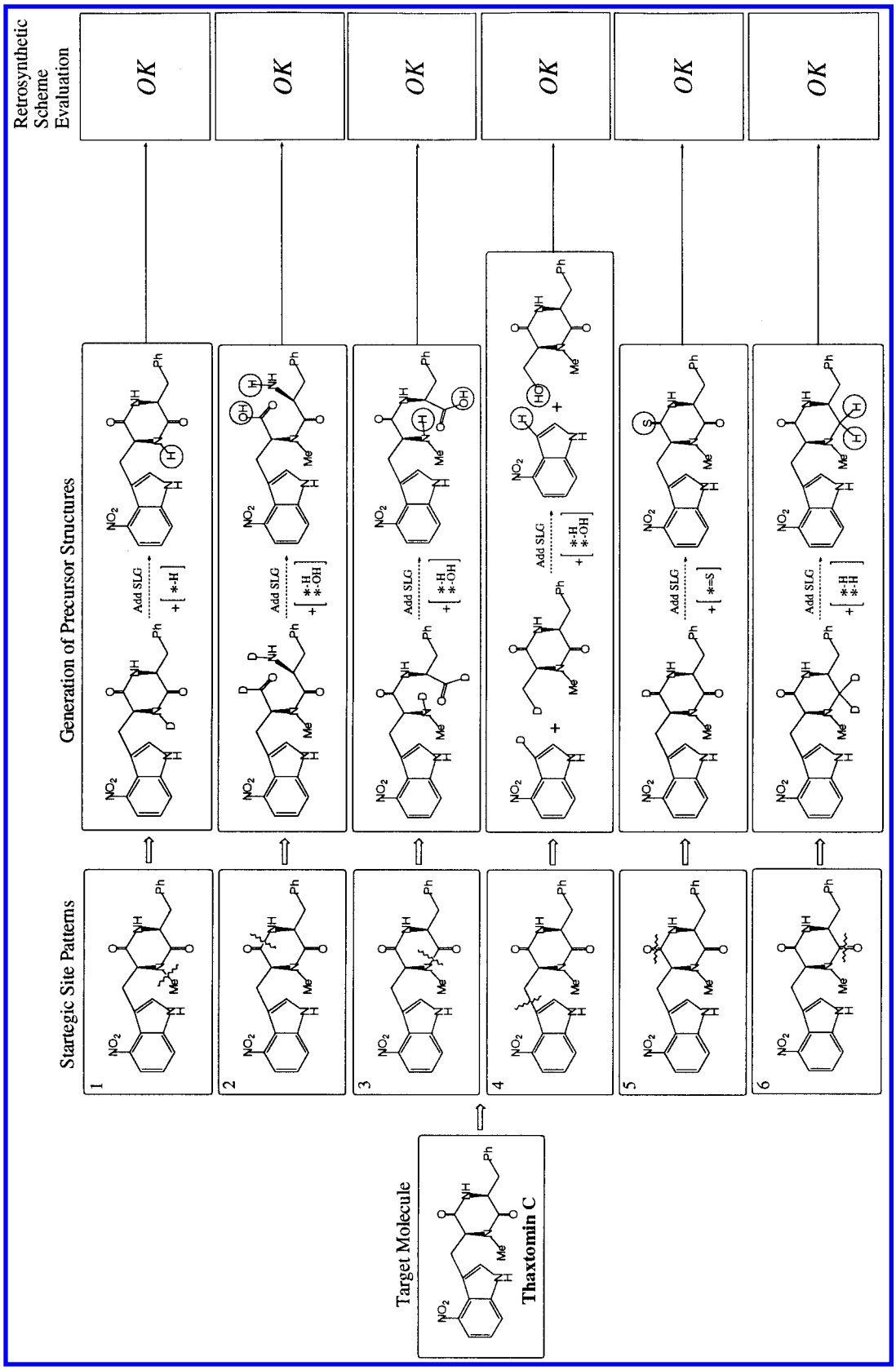

Figure 6. KOSP execution, example 1.

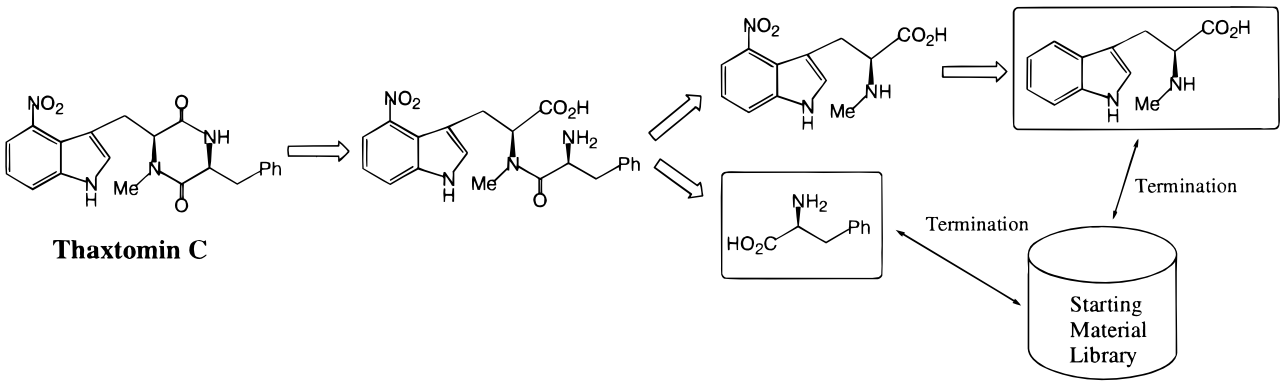

Figure 7. One of the proposed retrosynthetic routes of Thaxtomin C from KOSP.

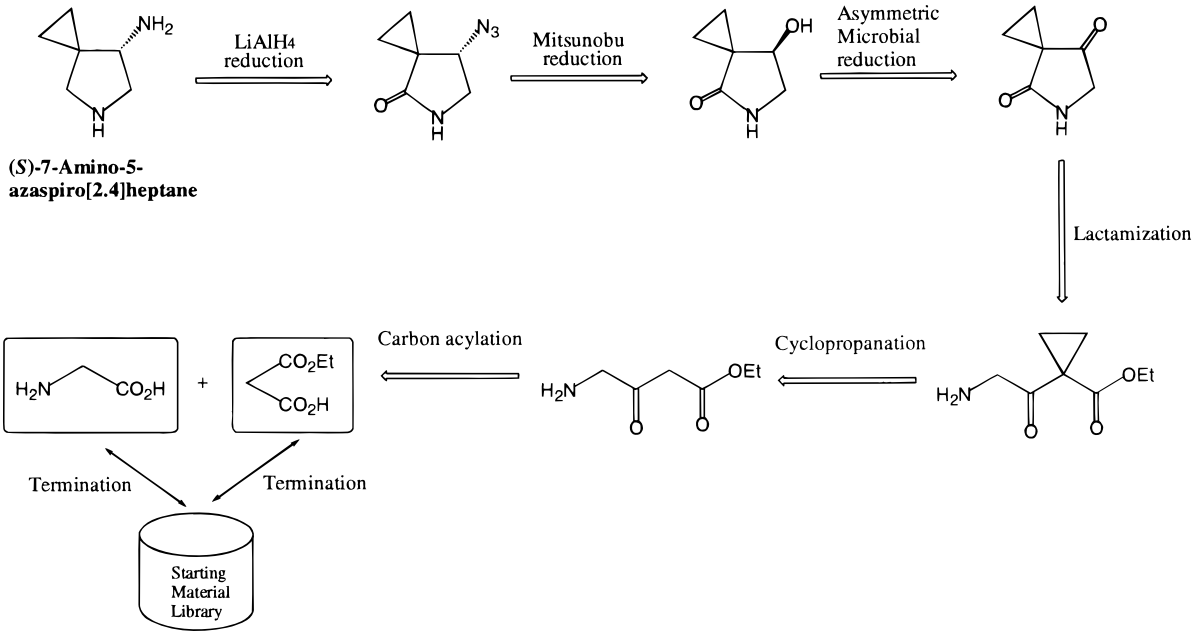

Figure 8. One of the proposed retrosynthetic routes of (*S*)-7-amino-5-azaspiro[2.4]heptane from KOSP.

the RR sites, the bit sequences are also computed in the same manner as in section 2.1. Finally, it is judged that when the bit sequences for the RR sites corresponding to a set of precursor structures match with those in the RKB by one-to-one correspondence, this retrosynthetic scheme occurs. In this step, if the user specifies consideration of stereoselective retrosynthetic analyses of strategic site pattern perception, stereo characteristics are also considered. As mentioned above, generated precursor skeletons that have dummy atoms are ignored in this evaluation procedure, because computation of SSCE is incomplete for a bond to which a dummy atom is attached.

3.4. Matching Precursor with Starting Material Library. The system judges whether the proposed retrosynthetic analysis can be terminated by comparison with the starting material library[17] of AIPHOS, which is retrieved from the chemical inventory database, Available Chemicals Directory (ACD).[20] The library comprises a hierarchical structure. Four levels of abstract structures are used: (1) The S. M. Data level consists of completely specified starting materials. (2) The low level contains information about positions and kinds of functional groups. (3) The high level contains only information about positions of functional groups. (4) The top level comprises basic skeletons of starting materials. Data for about 10 thousand are registered in the current library. The matching operation is performed as described below. In the first stage, the precursor structure involved in the evaluated retrosynthetic scheme is changed to abstract graphs corresponding to those of the library by recognizing functional groups with the indicated structural characteristics (Table 2). Next, the system determines whether the abstract graphs of the evaluated precursor structure exist by set reduction.[21,22] If the evaluated precursor exists in the S. M. Data level of the library, the retrosynthetic analysis of this retrosynthetic scheme is terminated.

4. EXECUTION EXAMPLE

Two examples of KOSP executions are given below.

These executions use the RKB and LGKB prepared from 28 000 reactions of the RDB, which was constructed from SYNLIB containing 81 000 reactions. The current KOSP system can be executed without consulting with the user after the input to the system is the desired target compound. This system has been written in FORTRAN77 and is currently running on an SGI-INDIGO2 system (R4400, 250 MHz).

4.1. Example 1. The input target molecule is a phytotoxin, thaxtomin C.[23] Although 11 strategic site patterns were

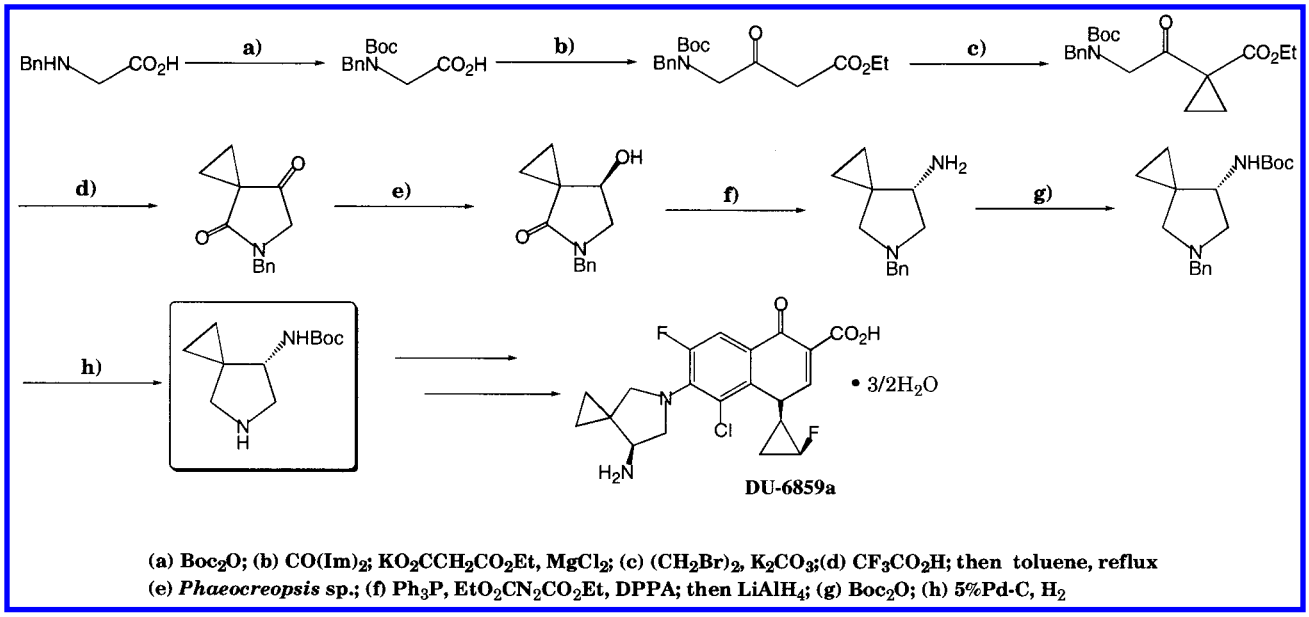

Figure 9. A reported synthetic method of the C-7 substituent of DU-6859a based on the retrosynthetic route (Figure 8).

perceived at the strategic site perception, five strategic site patterns among them could not generate precursor structures, because the SLG for proposed precursor skeletons was not determined (section 3.2.2). The residual six strategic site patterns were shown within the second rectangle from the left side of Figure 6, where strategic site patterns are represented by wavy lines. Precursor skeletons and structures are illustrated within the third rectangle from the left side of Figure 6. — H was perceived as SLG for the free bond obtained by cutting the bond of the first strategic site pattern (one N—C bond); —H and —OH were perceived as SLG from the second to fourth strategic site patterns (one N—C bond, one N—C bond, and one C—C bond, respectively); ═S was perceived as SLG for the fifth strategic site pattern (two C—O bonds); and two —H were perceived as SLG for the sixth strategic site pattern (two C—O bonds). All retrosynthetic schemes were evaluated to occur, but not to terminate retrosynthetic analyses, because none of the precursor structures were in the starting material library. The existing synthetic method[23] is nearly the same as the second retrosynthetic scheme, except that protecting groups are not considered. Almost the same retrosynthetic paths were proposed as a result of continuous executions of KOSP, as shown in Figure 7. It took about 15 min to propose this retrosynthetic path.

4.2. Example 2. The input target molecule is (*S*)-7-amino-5-azaspiro[2.4]heptane, which is the C-7 substituent, not considering the protecting group, of a new-generation antibacterial quinolone carboxylic acid, DU-6859a.[24] One of the proposed routes with KOSP is shown in Figure 8. It was about 30 min to find this retrosynthetic route. This route is attractive, because it involves the synthesis of a single enantiomer by means of asymmetric microbial reduction. The retrosynthetic paths were also suggested by AIPHOS, and one of the authors has developed an efficient, stereoselective route for the C-7 substituent of DU-6859a based on this proposal. This reported procedure is illustrated in Figure 9.[25]

5. CONCLUSION

The aim of the KOSP system is to adjust the reaction knowledge base, which is represented by information about the SSCE and the LGKB, which is represented by information about SCE, to an empirical retrosynthetic approach. To achieve this purpose, four functions are required: strategic site pattern perception, retrosynthetic scheme generation, retrosynthetic scheme evaluation, and retrosynthetic analysis termination. By using these functions, we have established a novel approach to empirical retrosynthetic analysis. One of the advantages of KOSP is that knowledge bases can be immediately derived from reaction databases in cases in which reaction data increase; KOSP can thereby consider novel and effective reactions which are being developed in the organic synthesis field. KOSP can thus use reaction data efficiently.

One plan to enhance KOSP is noninteractive generation of retrosynthetic sequences. For noninteractive generation of retrosynthetic sequences, a method of pruning of a synthesis tree is required to prevent an increase in retrosynthetic proposals. As noted in section 3.1, KOSP orders strategic site patterns with the depth of matching level and/or the number of original data in the RDB suggesting each strategic site pattern. However, this function is not sufficient for ordering retrosynthetic paths from the chemist's point of view. Hence, we are now working on the development of a novel ordering technique of proposals by KOSP.

REFERENCES AND NOTES

(1) Corey, E. J.; Cheng, X.-M. *The Logic of Chemical Synthesis*; John Wiley & Sons: New York, 1989.

(2) Corey, E. J. The Logic of Chemical Synthesis: Multistep Synthesis of Complex Cabogenic Molecules (Nobel Lecture). *Angew. Chem., Int. Ed. Engl.* **1991**, *30*, 455−465.

(3) Pensak, D. A.; Corey, E. J. LHASA—Logic and Heuristics Applied to Synthetic Analysis. In *Computer-Assisted Organic Synthesis*; ACS Symposium Series 61; American Chemical Society: Washington, DC, 1978; pp 1−32.

(4) Wipke, W. T.; Ouchi, G. I.; Krishnan, S. Simulation and Evaluation of Chemical Synthesis-SECS. *Artif. Intell.* **1978**, *11*, 173−193.

(5) Lee, T. V. Expert Systems in Synthesis Planning: A User's View of the LHASA Program. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 259−272.
(6) Ihlenfeldt, W. D.; Gasteiger, J. Computer-Assisted Planning Organic Syntheses: The Second Generation of Programs. *Angew. Chem., Int. Ed. Engl*. **1995**, *34*, 2613−2633.
(7) *IRDAS*: *ISIS Reaction Database Access System*; MDL Information Systems, Inc.: San Leandro, CA.
(8) Tsueng, S.-S. Computer-Assisted Reaction Searching Directed toward the Synthesis of Target Molecules. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 1138−1145.
(9) Jochum, C.; Lawson, S. *Modern Approaches to Chemical Reaction Searching*; Willett, P., Ed.; Gower Press: Aldershot, 1986.
(10) Hicks, M. G. Reactions in the Beilstein Information System: Nonaporic Organic Synthesis. *J. Chem. Inf. Comput. Sci*. **1990**, *30*, 352−359.
(11) Lawson, A. J.; Kallies, H. Multistep Reactions: The RABBIT Approach. *J. Chem. Inf. Comput. Sci*. **1990**, *30*, 426−430.
(12) 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.
(13) Satoh, H.; Funatsu, K. 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. *J. Chem. Inf. Comput. Sci*. **1996**, *36*, 173−184.
(14) Funatsu, K.; Sasaki, S. Computer-Assisted Synthesis Design and Reaction Prediction System AIPHOS. *Tetrahedron Comput. Methodol*. **1988**, *1*, 27−38.
(15) Greene, T. W. *Protective Groups in Organic Synthesis*; John Wiley & Sons: New York, 1981.
(16) Distributed Chemical Graphics, Inc. permitted us to use SYNLIB for research work on AIPHOS.
(17) Satoh, K.; Azuma, S.; Satoh, H.; Funatsu, K. Development of the Program for Construction of a Starting Material Library for AIPHOS. *J. Chem. Software* **1997**, *4*, 101−112.
(18) Wipke, W. T.; Dyott, T. M. Use of Ring Assemblies in a Ring Perception Algorithm. *J. Chem. Inf. Comput. Sci*. **1975**, *15*, 140−147.
(19) Schmidt, B.; Felischauser, J. Fortran IV Program for Finding the Smallest Set of Smallest Rings of a Graph. *J. Chem. Inf. Comput. Sci*. **1978**, *18*, 204−206.
(20) *ACD*: *Available Chemicals Directory*; MDL Information Systems, Inc.: San Leandro, CA.
(21) Susseguth, E. H. Graph-Theoretic Algorithm for Matching Chemical Structures. *J. Chem. Doc*. **1965**, *5*, 36−43.
(22) 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.
(23) King, R. R. Synthesis of thaxtomin C. *Can. J. Chem*. **1997**, *75*, 1172.
(24) Kimura, Y.; Atarashi, S.; Kawakami, K.; Sato, K.; Hayakawa, I. (Fluoro-cyclopropyl)quinolones. 2. Synthesis and Stereochemical Structure-Activity Relationships of Chiral 7-(7-Amino-5-azaspiro[2.4]-heptane-5-yl)-1-(2-fluorocyclo-propyl)quinolone Antibacterial Agents. *J. Med. Chem*. **1994**, *37*, 3344−3352.
(25) Satoh, K.; Imura, A.; Kanai, K.; Miyadera, A.; Yukimoto, Y. An Efficient Synthesis of a Key Intermediate of DU-6859a via Asymmetric Microbial Reduction. *Chem. Pharm. Bull*. **1998**, *46*, 587−591.

CI980147Y