Development of a Program for Construction of a Starting Material Library for AIPHOS

Koji SATOH^a, Shukou AZUMA^b, Hiroko SATOH^{b,c},
and Kimito FUNATSU^{b*}

^aChemical Technology Research Laboratories, Daiichi Pharmaceutical Co., Ltd.,
Edogawa-ku, Tokyo, 134-8630 Japan

^bDepartment of Knowledge-Based Information Engineering,
Toyohashi University of Technology
Tempaku, Toyohashi, Aichi, 441-8122 Japan

^cPresent address: Synthetic Organic Chemistry Laboratory,
The Institute of Physical and Chemical Research (RIKEN)
Wako, Saitama, 351-0106 Japan

(Received : July 23, 1997 ; Accepted for publication : October 14, 1997 ; Published on Web : November 8, 1997)

A program to construct a fundamental and original starting material library to embody starting material-oriented retrosynthetic analysis in the AIPHOS system has been developed. The data in the library can be registered automatically from ACD, a commercial chemical structure database, by utilizing this program. This program recognizes synthetic equivalents to categorize functional groups. The resultant library is made up of a hierarchical structure using abstract graphs. The structure, the constructing method of the starting material library for AIPHOS and the registration results of data selected at random from ACD are described. In addition, a brief outline of AIPHOS is included.

Keywords: Computer-assisted synthesis planning system, AIPHOS, Starting material library, Abstract graph, Hierarchical structure, ACD

1 Introduction

Since the seminal publication on computer-assisted organic synthesis by Corey and Wipke[1] was presented, more than fifty computer-assisted synthesis planning systems, as we know, for designing syntheses of organic compounds have been developed during the last twenty five years. There are two categories in computer-assisted organic synthesis planning systems. One is an empirical knowledge-based approach, such as in LHASA[2] or SECS[3]. The other is a logic-centered approach, such as in SYNGEN[4] or EROS[5]. The advantage with regard to the former systems is that their proposals of retrosyntheses can generally be executed in laboratories. The latter ones have the merit that their proposed retrosynthetic paths involve original and novel ones. Both two approaches are attractive for chemists.

AIPHOS (Artificial Intelligence for Planning and Handling Organic Synthesis)[6], which has been developed in our laboratory, combines the merits of the empirical knowledge-based approach and the logic-centered approach. That is, AIPHOS can propose original and new retrosynthetic routes to assure their implementation in laboratories because of its unique reaction knowledge base[7]. For the moment, the AIPHOS system proceeds stepwise, interactively, determining at each step the synthetic precursors from molecules of the former step. The synthesis design procedure of AIPHOS is shown in Figure 1. A series of procedures in the current AIPHOS is as follows: (1) Input of a desired target structure is made. (2) Plausible strategic sites in the target molecule are obtained. In acquiring strategic sites, the topological strategy and the functional group based strategy which are described by Corey[8] are applied to the target molecule. One of the proposed strategic sites is selected by the user. When the user wishes to specify strategic sites according to his/her synthetic strategies, he/she may edit his/her own strategic sites manually. (3) Possible precursors on the basis of the strategic sites selected at step 2 are generated. One set of proposed precursors is selected by the user. (4) Appropriate leaving groups are added to a set of precursors selected at step 3 automatically by utilizing the leaving groups knowledge base. If necessary, the user may add leaving groups manually. (5) The proposed retrosynthetic path is evaluated as to whether it can occur or not with the reaction knowledge base. If the proposed retrosynthetic path can occur, it is displayed to the user with related reaction schemes in the AIPHOS database.

In planning of organic syntheses, chemists look for efficient retrosyntheses analyzing a relationship between a proper unit of the target molecule and the structure of available compounds like starting materials offered in catalogs. Other synthesis planning systems such as LHASA[9], SECS[10] and SYNCHM[11] embody this kind of analysis of chemists by using their original starting material libraries approaches to starting material-based retrosynthetic analysis. In order to introduce a similar approach to AIPHOS, we have developed an efficient program which can incorporate data of a database of chemical structures on the market into the library. This program is also able to recognize synthetic equivalents to be categorized functional groups. And then, the resulting library is made up of a hierarchical structure using abstract graphs. In this paper, we report a structure, a construction method of our original starting material library and the registration results of data selected at random from a commercial chemical structure database. The starting material-oriented retrosynthetic analysis in AIPHOS will be reported in a future paper[12].

2 Abstract Graphs

A hierarchy of abstract search spaces similar to Sacerdoti's hierarchical planning[13] is selected for building the starting material library. The main reason it was chosen is that hierarchy ignores irrelevant connections, the number of compared graphs becomes smaller than that of original graphs for abstraction, and the time required registration of data can be shortened.

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 recognized according to functional group codes. (3) The high level, in which functional groups are replaced with X, contains only information about their positions. (4) The top level comprises basic skeletons of starting materials. If nonaromatic C-C multiple bonds exist in abstract graphs, they are converted into single bonds at this level. In Figure 2,

Figure 1. Block diagram of AIPHOS.

an example of abstract graphs at each level is shown.

3 Recognition of Positions and Kinds of Functional Groups

One of the characteristic points of our starting material library lies in exact recognition of positions and kinds of functional groups. For this recognition, we make use of 111 structural characteristic keys which feature reaction sites and their environments in AIPHOS. Some of them are listed in Table 1. The concept of how to establish them is described in a previous paper[14].

In principle, abstract graphs must contain basic skeletons of compounds in building the starting material library. Hence, only a part of the above keys, which represent functional groups, is used. Further, several new structural characteristic keys (*e.g.*; -NHAc, -NHCO₂R) are added on because many *N*-substituted compounds are commercially available as starting materials for various compounds.

4 Categorization of Functional Groups

When an efficient starting material library is constructed, reducing spaces for abstract spaces is important. Therefore, the concept of synthetic equivalents, which are afforded only for changing functional groups, is applied for abstraction of data. Forty functional groups are categorized into 9 groups shown in Figure 3. If necessary, more functional groups can be additionally registered and the combination of the functional groups in the categories also can be changed by users/chemists.

5 Abstraction of Starting Materials

An example (benzoic acid), for explanation of abstraction, is given in Figure 4, because it is simple to explain. -CO₂H is recognized as a functional group. The structural characteristic key no. is 36, and node number 4 is assigned as a root node of the functional group in recognition of structural characteristic key.

Abstract graphs are derived as follows; The core node comprising the functional group characterized above is replaced with X₇, which is the functional group code corresponding to -CO₂H. The low level abstract graph is prepared by deleting nodes consisting of the functional group. The high level abstract graph is made by changing X₇ in the low level abstract graph to X. The top level abstract graph is given by deleting X in the high level abstract graph (Figure 5). Esters, hetero mixed ethers (*e.g.*; R¹-O-R², R¹-S-R², R¹-NH-R²) and carbamates only consisting of their alkyl groups (methyl, ethyl, *i*-propyl, *t*-butyl) are abstracted, because they are frequently employed as starting materials in organic synthesis. The above functional groups consisting of other alkyl groups are not abstracted, and compounds in which any structural characteristics are not recognized are also registered without abstraction.

Figure 2. Examples of abstract graphs. X_1 is a code arranging functional groups. X is converted from functional groups codes.

- X₁; -OH, -OR (Me, Et, *i*-Pr, *t*-Bu), -F, -Cl, -Br, -I**
- X₂; -NHCl, -NHBr, -NHOH**
- X₃; -SH, -SR (Me, Et, *i*-Pr, *t*-Bu)**
- X₄; -NO₂**
- X₅; -SO₂OH, -SO₂Cl, -SO₂Br**
- X₆; -CN**
- X₇; -CHO, -CO₂H, -CO₂R (Me, Et, *i*-Pr, *t*-Bu), -CONH₂**
- X₈; -NH₂, -NHR (Me, Et, *i*-Pr, *t*-Bu), -NHAc,
-NHCO₂R (Me, Et, *i*-Pr, *t*-Bu)**
- X₉; =O**

Figure 3. Categories of functional groups.

Table 1. Some structural characteristic keys in the reaction knowledge base in AIPHOS.

Key No.	Structural Characteristics	Key No.	Structural Characteristics	Key No.	Structural Characteristics
1	$-\text{CH}_3$	48	$-\text{N}-\text{C}-\text{N}-$ $\text{H} \quad \text{O} \quad \text{H}$	100	$-\text{N}=\text{C}-$
2	$-\text{CH}_2\text{R}$	49	$-\text{C}-\text{S}-$ O	101	$-\text{C}=\text{N}-$
3	$-\text{CHR}_2$	50	$-\text{SO}_2\text{OH}$	102	$-\text{C}\equiv\text{C}-$
4	$-\text{CR}_3$	:	:	103	:
5	$-\text{CH}(\text{CH}_3)_2$	77		:	:
:	:	78		108	$-\text{CH}=\text{O}$ H
33	$-\text{P}(\text{OR})_2$	79		:	:
34	$-\text{CR}$ O	80		111	$-\text{NO}_2$
35	$-\text{CH}$ O	:	:		
36	$-\text{CO}_2\text{H}$				
:	:

Figure 4. Recognition of structural characteristics in the case of benzoic acid.

6 Decision of Identification

A single compound must not be registered with duplication in constructing the starting material library. It is necessary to register only new compounds. At this time, a Set Reduction algorithm[15] for detecting the existence of the same chemical structure is taken. As a Set Reduction algorithm is essentially used for substructure search, a new registering compound and a compound already registered in the library are cross referred by the Set Reduction method for finding structural identification.

Figure 5. An example of abstraction on benzoic acid.

7 Structure of Starting Material Library

The structure of the starting material library is separated into three blocks: i.e. chiral pool, aromatic pool, nonaromatic pool. In particular, a chiral pool is suited to precursors related to an optically active target molecule. Furthermore, each pool is separated according to the number of carbon atoms contained in top level abstract graphs.

8 Construction of the AIPHOS Starting Material Library

The starting material library for our purpose requires thousands of data. As it is inefficient to register such a large amount of data manually, the Available Chemicals Directory (ACD)[16] containing approximately 130,000 entries of fine chemicals and bulks is introduced into the library. The data having aromaticity in ACD are described by nonaromatic bond information, but aromatic parts of precursors exported from AIPHOS have aromatic bond information. If one chemical structure is described in a different manner (*e.g.*; bond information), the structure is recognized as a different structure by the Set Reduction method. For solving this problem,

before registering data of ACD into the library, the bond information of data having aromaticity is converted to aromatic bond information using the aromaticity recognition program[17] based on the Hückel rule. Although our final purpose is to build a large starting material library, five thousand data, which had been selected from ACD at random, were used, to begin with. The numbers of carbon atoms in the materials are from three to ten, since organic chemists use them as starting materials in many cases. The result is shown in Table 2. The percentage of automatically registered data from ACD was 94%. The main factor which did not give 100% registration is that our processing discarded salts such as hydrochloride, potassium and sodium. The percentage of reduction from S.M. data level to top level was 46%. This might be a satisfying result, considering that data were selected at random from ACD.

The flow chart for constructing the starting material library is diagrammed in Figure 6.

Table 2. A result of registration from ACD.

Level	Chiral	Aromatic	Nonaromatic	Total	Reduction(%)
Top	239	1068	869	2176	46
High	298	1541	1396	3235	69
Low	333	1807	1621	3761	80
S.M.Data	352	2446	1891	4689	100

9 Conclusions

We have developed a program for constructing a fundamental starting material library to use starting material-oriented retrosynthetic analysis in AIPHOS. This program can import data of other commercial chemical structure databases as well as data of ACD into the library, if their structural data are described by MDL mole file format or can be converted to its file format. Synthetic equivalents are also recognized by using this program. The resultant library is made up of a hierarchical structure using abstract graphs. This library could be used to decide which step in retrosynthetic routes proposed by AIPHOS should be terminated.

Now, further investigations for registering much more data, such as all data in ACD, in the library automatically are in progress.

10 Experimental

In the starting material library, we can deal with the following atom types (C, H, N, O, P, S, F, Cl, Br, or I) and less than 128 atoms. The program is written in FORTRAN 77. The program is currently running on a SGI-INDY system (R4400, 200MHz).

We thank Dr. Ryutaro Kishimoto and Molecular Design, Ltd., for permission to use data from the Available Chemicals Directory and also acknowledge Mrs. Atomi Yoshida and Mr. Yasuhiko Yotsui of the Information Planning and Administration Department, Daiichi Pharmaceutical Co. Ltd., for assisting in those preparations.

References

- [1] E. J. Corey and W. T. Wipke, *Science*, **166**, 178 (1969).
- [2] D. A. Pensak and E. J. Corey, *Computer-Assisted Organic Synthesis*, American Chemical Society, Washington, DC (1978), pp.1-32.
- [3] W. T. Wipke, G. I. Ouchi and S. Krishnan, *Artif. Intell.*, **11**, 173 (1978).
- [4] J. B. Hendrickson, E. Braun-Keller and G. A. Toczko, *Tetrahedron*, **37**, 359 (1981).
- [5] J. Gasteiger, *Chim. Ind. (Milan)*, **64**, 714 (1982).
- [6] K. Funatsu and S. Sasaki, *Tetrahedron Comput. Method.*, **1**, 27 (1988).
- [7] H. Satoh and K. Funatsu, *J. Chem. Inf. Comput. Sci.*, **35**, 34 (1995).
- [8] E. J. Corey, *Angew. Chem. Int. Ed. Engl.*, **30**, 455 (1991).
- [9] A. P. Johnson, C. Marshall and P. N. Judson, *Rec. Trav. Chim. Pays-Bas.*, **111**, 310 (1992).
- [10] W. T. Wipke and D. Rogers, *J. Chem. Inf. Comput. Sci.*, **24**, 71 (1982).
- [11] H. Gelernter, A. Sanders, D. L. Larson, K. K. Agarwal, R. H. Boivie, G. A. Spritzner, and J. E. Searlman, *Science*, **197**, 1041 (1977).
- [12] K. Funatsu, T. Yoshino and K. Satoh, *in preparation*.
- [13] E. D. Sacerdoti, *Artif. Intell.*, **5**, 115 (1974).
- [14] K. Funatsu, C. A. Del Carpio and S. Sasaki, *Tetrahedron Comput. Method.*, **1**, 39 (1988).
- [15] E. H. Sussenguth, *J. Chem. Doc.*, **5**, 36 (1963).
- [16] The *Available Chemicals Directory* is supplied by Molecular Design Ltd., 2132 Farallon Drive, San Leandro, CA 94577, United States of America.
- [17] B. L. Rose-Kozel and W. L. Jorgensen, *J. Chem. Inf. Comput. Sci.*, **21**, 204 (1981).

合成設計システム AIPHOS 用出発物質 ライブラリー構築プログラムの開発

佐藤 耕司^a, 東 秀行^b, 佐藤 寛子^{b,c}, 船津 公人^{b*}

^a 第一製薬（株）製薬技術研究所 〒134-8630 江戸川区北葛西 1-16-13

^b 豊橋技術科学大学 知識情報工学系 〒441-8122 豊橋市天伯雲雀ヶ丘 1-1

^c 現在の所属: 理化学研究所 〒351-0106 和光市広沢 2-1

合成設計システム AIPHOS において、出発物質を想定した逆合成ルート提案に利用するための基本的かつオリジナルな出発物質ライブラリーを構築するためのプログラムを開発した。本プログラムを用いることにより、市販化学構造データベース ACD より自動的に出発物質ライブラリーを構築することが可能となった。また、本プログラムでは官能基を分類することにより合成等価体を認識することが可能となっている。また、構築されるライブラリーは、抽象化グラフを利用した階層化構造により構成されている。本稿では、合成設計システム AIPHOS 用出発物質ライブラリーの構造、構築方法および ACD から任意に抽出したデータ登録結果を報告する。さらに AIPHOS の簡単な説明も併せて行なう。

キーワード: Computer-assisted synthesis planning system, AIPHOS, Starting material library, Abstract graph, Hierarchical structure, ACD