

Construction of a generic reaction knowledge base by reaction data mining

Ke Wang,* Lisha Wang,* Qiong Yuan,† Shiwei Luo,* Jianhua Yao,*
Shengang Yuan,* Chongzhi Zheng,* and Josef Brandt‡

*Laboratory of Computer Chemistry, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai, China

†Chemical Abstracts Service, Columbus, Ohio, USA

‡Institut für Organische Chemie und Biochemie, Technische Universität München, Garching, Germany

As synthesis by combinatorial chemistry and high throughput screening have become well-established strategies in the drug discovery process, chemists face increased challenges in managing large amounts of data and using these data to design more diverse and focused libraries. As synthesis is an intuitive and empirical process, however, the classical approaches to computer-assisted synthesis planning do not fully satisfy the needs of the synthetic chemist. We describe a novel computational technique for extracting reaction data and building a generic reaction knowledge base (GRKB) to provide chemists with useful and well-organized knowledge. The method consists of three key steps: (1) the automatic recognition of reaction centers, (2) the definition of a hierarchy of reaction patterns, and (3) the organization of the generic reaction knowledge. Significant reaction knowledge has been discovered via mining a subset of the InfoChem Reaction database. A frame system has been constructed to store and retrieve the GRKB. Applications of this GRKB to synthesis planning are illustrated. © 2001 by Elsevier Science Inc.

Keywords: synthesis planning, generic reaction knowledge base, reaction classification, data mining

INTRODUCTION

Combinatorial synthesis and high throughput screening have significantly changed the traditional approach to drug discovery

by making it possible to screen thousands of chemical compounds to identify new lead molecules. With these methods a chemist could synthesize 1,000 compounds per week and large quantities of reaction data have been accumulated.^{1–3} These data need to be managed and analyzed to guide chemists in the design and synthesis of more focused libraries. The past decades have seen the development of computer programs aimed at facilitating organic synthesis. Generally they are classified into two categories: reaction databases, like REACCS,⁴ ChemReact,⁵ CASREACT,⁶ etc; and synthesis planning systems, such as LHASA⁷, SECS⁸, SYNCHEM⁹, SYNGEN¹⁰, EROS¹¹, CAMEO¹². In contrast to the popular acceptance of reaction databases by many chemical companies and research laboratories,^{13,14} none of the synthesis planning systems, have been widely accepted as a routine and practical tool. As these systems are complex and have inadequacies; chemists still tend to regard synthesis as an intuitive and empirical process not amenable to inference from retrosynthetic analysis done by computers.

Although reaction databases are easy to set up and have been generally accepted, the rapidly growing body of reaction data means that chemists are increasingly dissatisfied with traditional retrieval tools, and new methodology is needed to interpret the data. Presently, substructure searching is the only way to resolve a query reaction that has not been included in reaction databases. However, there are issues regarding how a query substructure is defined and how hits are evaluated. Although there are strategies to refine the query substructure and retrieve meaningful results, the problem is how to quickly limit the number of hits to those containing useful and task-oriented knowledge. This situation occurs more often when the target molecule is a novel compound. We have discovered methods to resolve these issues based on the chemist's intuitive reasoning process; our methods allow recognition of the retro-reactionary active part of the target molecule and recall of familiar reactions in which the products are similar to the

Color Plates for this article are on page 469.

Corresponding author: S. Yuan, Laboratory of Computer Chemistry, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, 354 Fenglin Lu, Shanghai 200032, China.

E-mail address: yuansg@pub.sioc.ac.cn (S. Yuan).

target. If the products have the same retro-reactionary active parts as the target molecule, chemists infer an analogous method to synthesize the target according to the known reaction, or heuristic reactions. To implement this process, the key step is to generate a generic reaction knowledge base (GRKB), which stores millions of reactions according to generic schemes. We have developed the programs to build up this generic knowledge from a reaction database automatically. Compared with SOPHIA¹⁵ and KOSP,¹⁶ which have been reported in the literature, the unique and important feature of our technology is that the reaction knowledge is hierarchically represented in four levels and each level represents a different level of abstraction of the knowledge. Such knowledge architecture is flexible in predicting preparative reactions for new molecules. We call this new methodology, used to exploit the knowledge from reaction data, reaction data mining.¹⁷ The experimental database we used is ChemReact from InfoChem (München, Germany), which contains 100,000 reactions.

BASIC TERMS

Reaction Center

The first step in reaction data mining is to recognize the reaction centers for each reaction.^{18,19} We refer to all atoms with bonds being built/broken during the reaction as reaction centers. Most reaction classifications, such as Diels-Alder and Friedel-Crafts alkylation, are based on these atoms centers. In GRKB, reaction centers are considered the core of the reaction knowledge and the basis of the hierarchical model. We have successfully developed a software tool to recognize reaction centers in spite of ChemReact not providing reaction center information for each reaction. To effectively handle large sets of data, we chose not to use conventional atom-atom matching, which is generally time-consuming and impractical.²⁰⁻²² Through testing and comparison, we are convinced that our program is practical and highly efficient. Detailed algorithms and testing results will be presented in a future article.

Reaction Patterns

Reaction patterns are a useful concept as they are an effective method of classifying reactions by computer. A reaction pattern can be understood as a reaction formula, which describes the structure changes from reactants to products of the reaction. However, since the reaction pattern does not involve the mechanism of the reaction, two reactions with totally different reaction mechanisms could have the same reaction pattern. Both reaction graphs show the same structure changes.

Hierarchical Model of Reaction Patterns

The feasibility of a reaction depends on many factors, among which the structures and conditions are most influential. Structures of the reactants, particularly reaction centers and the surrounding active groups, are intrinsic factors that determine whether the reaction can occur, while the conditions are external factors that control the direction and extent of the reaction. We believe that within the reactions contained in the reaction database there is some innate, generic information to be mined. The key is how to mine this generic knowledge and represent it in the correct format. In our system, we use the set of reaction

patterns expressed by a hierarchical model. It consists of four main layers, named reaction type (RT), reaction core (RC), extended reaction core (ERC), and concrete reaction (CR). An ERC consists of four sublayers: smallest ERC(ERC_S), one-step ERC(ERC_O), two-step ERC(ERC_T), and multi-step ERC(ERC_M).²³⁻²⁶ Figure 1 shows the hierarchy of the four layers and the relationship between them. Figure 2 illustrates the process of extracting the reaction patterns from a specific reaction.

Generic Reactions

In GRKB, reactions having the same reaction pattern are classified into a same reaction schema, which is regarded as a generic reaction. We have developed a hierarchical reaction pattern system to represent generic reactions accurately and comprehensively.¹⁹

REACTION KNOWLEDGE DISCOVERY

Reaction Type (RT) Layer

The RT reaction pattern can be generated directly from the reaction database by extracting the reaction centers. If there is no reaction center information stored in the reaction database, our algorithm will identify the reaction centers. For example, Figure 2b shows the RT reaction pattern derived from a reaction instance shown in Figure 2a.

Reaction Core (RC) Layer

Sometimes the RT layer consists of a disconnected pattern (graph), either in the reactant part or product part or in both. As Figure 2a shows, both reactant and product part are disconnected graphs. Modifying this RT to a connected graph on both

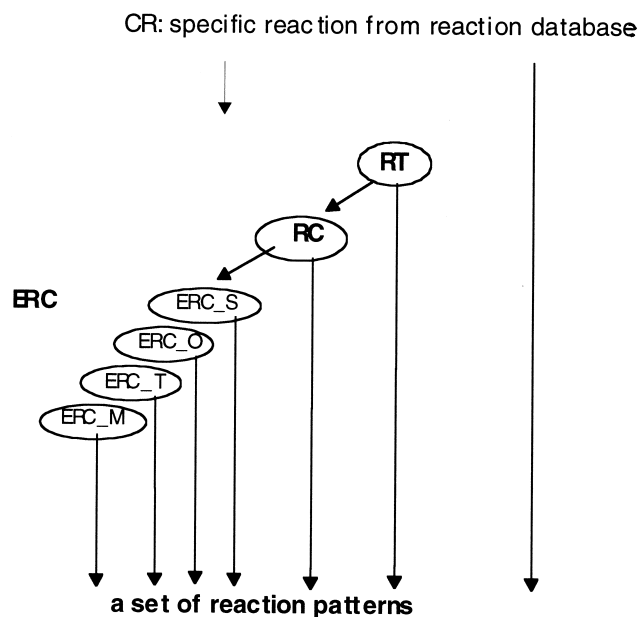
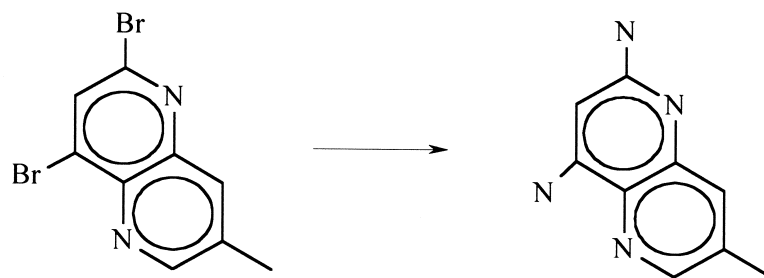
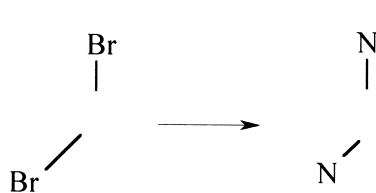


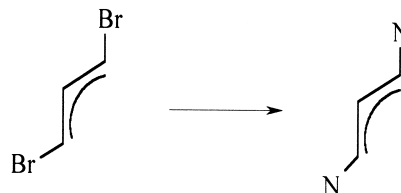
Figure 1. Hierarchy of reaction patterns. The Extended Reaction Core Layer consists of four sublayers. A set of reaction patterns can be acquired and easily organized via this architecture.



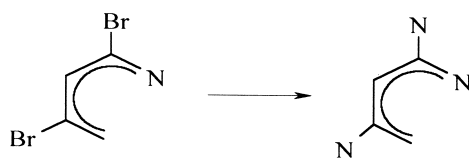
a Example of a reaction (also CR Layer)



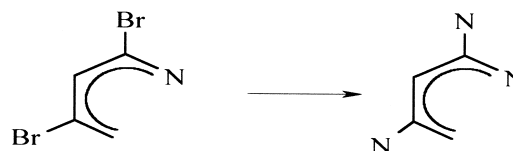
b. RT Layer



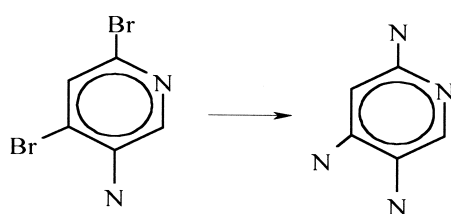
c. RC Layer



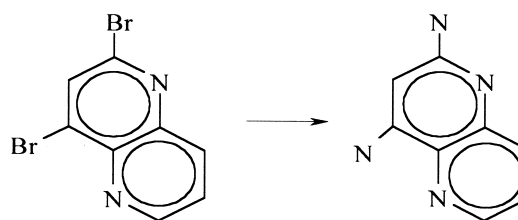
d. ERC_S Layer



e. ERC_O Layer



f. ERC_T Layer



g. ERC_M Layer

Figure 2. An example illustrating the Hierarchy of Reaction Patterns.

sides of the reaction pattern in a corresponding way produces the RC reaction pattern. The algorithm seeking the shortest path of a disconnected graph is used in both sides of RT layer. When one path is added to the disconnected side of RT, the corresponding atoms are also added to the other side. Figure 2c shows the RC pattern developed from RT pattern. However, if RT pattern is a connected graph, the RC pattern will be the same as RT pattern.

Extended Reaction Core (ERC) Layer

The ERC pattern can be divided into four sublayer patterns by using different methods to extend the RC pattern. Two definitions are introduced here before we describe these four sublayers further. A Common Atom is an atom that fulfils the following three criteria: (1) it must be a carbon atom; (2) all its neighbor atoms must be carbon atoms; and (3) every bond

between these atoms must be a single bond. Otherwise, this atom will be called a Non-Common Atom.

A. Smallest ERC (ERC_S): Starting from an RC pattern, extending the active atoms within one topological distance step from the RC pattern generates the ERC_S. The active atoms are the heteroatoms and carbon atoms that connect with the RC pattern by nonsingle bonds. Figure 2d shows the ERC_S pattern of the example reaction.

B. One-step ERC (ERC_O): From the RC pattern, the ERC_O pattern can be developed by extending the non-common atoms within one topological distance step from the RC. Figure 2e shows the ERC_O pattern of the example reaction. In many cases the ERC_O is identical with ERC_S. The example in Figure 2 is one such a case. We define ERC_O to include information about multiple bonds and heteroatoms that are separated by a single bond and a carbon atom with the reaction centers.

C. Two-step ERC (ERC_T): From the RC pattern, the ERC_T pattern can be created by extending the non-common atoms within two topological distance steps continuously from RC. Figure 2f shows the ERC_T pattern of the example reaction.

D. Multi-step ERC (ERC_M): The ERC_M pattern can be developed from the RC pattern by extending the non-common atoms within more than two topological distance steps continuously from RC. Figure 2g shows the ERC_M pattern of the example reaction.

Concrete Reaction Layer (CR)

The last layer, the CR reaction pattern, is the same as the specific reaction, as illustrated in Figure 2a.

Example of Reaction Knowledge Discovery

Our techniques mine reaction data (CRs) to generate generic reaction knowledge. The following example illustrates this mining process using the 4+2 Diels-Alder generic reaction pattern:

1. Search the database and find the reactions have the same kind of reaction centers. In this example, 5,702 reactions were found. Two kinds of intra-molecular Diels-Alder reactions are presented in Figure 3.
2. Derive the reaction knowledge represented in RC and ERC from each reaction.
3. Analyze the RC and ERC from above step, including merging and discarding the duplicates. Finally, the whole generic reaction pattern is obtained.

From Figure 3, the mechanism of our hierarchical model can be easily understood. The further we progress from CR, ERC, RC, to RT, the top level in this hierarchy, the more generic is the reaction knowledge that is being extracted. This process can be regarded as generalization of reaction knowledge and is an effective way to classify the reaction data. Conversely, as layers from RT to CR are spanned, reaction centers and their environments are extended step by step, and more precise and concrete knowledge is described. This creates another mechanism for retrieving the original reaction data starting with the generic reaction knowledge.

IMPLEMENTATION OF GENERIC REACTION KNOWLEDGE BASE

As well as developing techniques to mine reaction data and extract the generic reaction knowledge, we have designed a frame system to: store generic reaction knowledge (GRKB); retrieve the new information from this GRKB; and implement linking to the original reaction data (CRs).

System Structure

The basic data structure of this frame system is adapted from Minsky's frame theory in knowledge engineering.²⁷ Figure 4 shows the organization of a reaction in this system.

Reasoning Mechanism

A progressively abstract mode is adopted in our reasoning process. Going from the CR, ERC, and RC, to RT layers of generic reaction patterns represents an increase in abstraction. The search starts with CR. If no results appear, the search descends to the next layer, and so on. Ultimately, if no result can be found at the RT layer of abstraction, this suggests a new reaction type for the GRKB. Generally, in all cases a result best matching with the known reaction knowledge will be found.

The concrete reasoning process is as follows:

1. Search the target molecule in the CR pattern layer.
2. If there is no hit in the CR level, identify the possible active part in the target by an algorithm, or manually.
3. Search the ERC pattern of the target in the ERC pattern layer, from ERC_M, ERC_T, and ERC_O, to ERC_S.
4. If there is no hit in the ERC level, search the RC pattern of the target in the RC pattern layer.
5. If there is no hit in the RC level, search the the RT pattern of the target in the RT pattern layer.
6. If there is no hit in the RT level, change the possible active part and return to step 3.
7. Go back to step 2 and repeat the steps until a satisfactory precursor is found.

SYNTHESIS PLANNING FROM GRKB

GRKB-S Introduction

As well as helping to analyze reaction databases and manage large amounts of reaction data, an additional advantage of GRKB is that it can be used in synthesis planning. The software we developed for synthesis planning from GRKB is called GRKB-S and has been tested on a GRKB extracted by randomly selecting 100,000 reactions from ChemReact. Color Plate 1 is a screenshot of this system and shows the graphical user interface in the Windows style. We have already used this system to solve some practical synthesis planning problems.²⁸

GRKB-S Operation

1. Input the query structure (see the example shown in Figure 5).
2. Identify the active part, either by computer heuristics or manually. In this example we set the ring as the reacting group (the bonds flagged with #- shown in Figure 5).

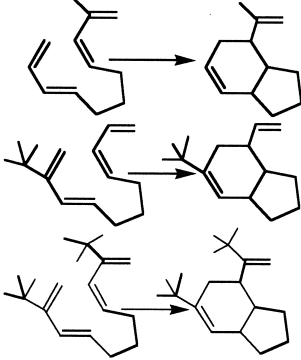
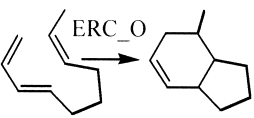
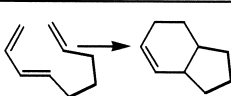
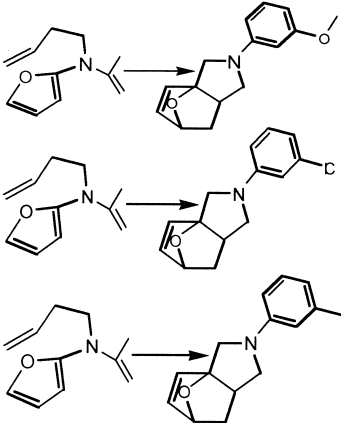
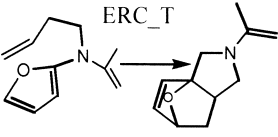
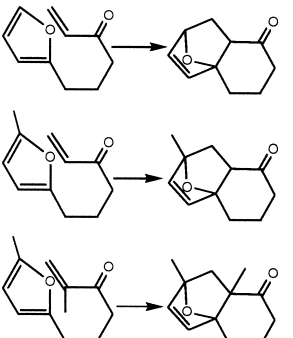
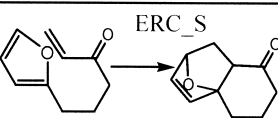
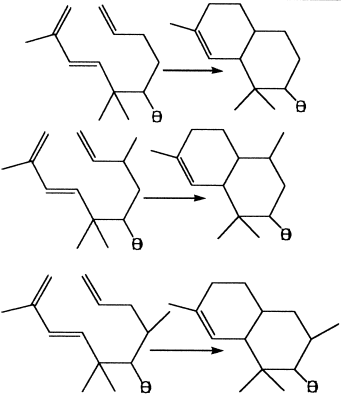
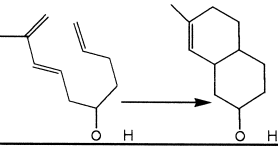
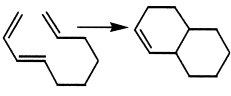
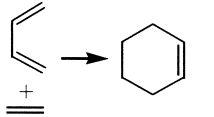
Completed Reaction	Extended Reaction Core	Reaction Core	Reaction Type
			
			
			
			
			

Fig. 3. An example of organizing a generic reaction through a hierarchical model.

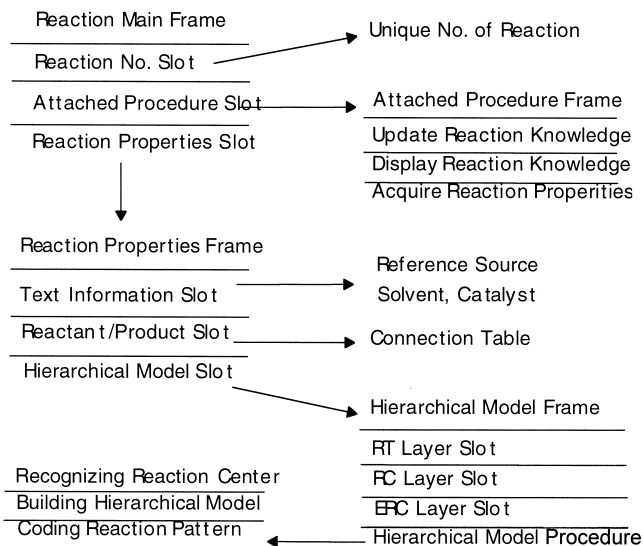


Figure 4. Organization of a reaction in the general frame system. The unique number of the reaction is the access key for the frame.

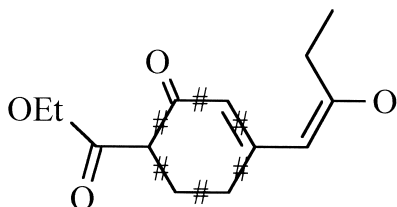


Figure 5. Structure of target molecule which was used as a query to search using GRKB-S. The flagged bonds indicate the proposed reaction centers.

- Click on the corresponding button in the main menu to start searching.
- Hits are retrieved very quickly and can be browsed easily. In this example, due to the GRKB accessing only a small subset of data, only one relevant hit (shown in Color Plate 2) was found in from the ERC_O layer.

From the hit reaction shown in Color Plate 2, we complete the first step of the target molecule synthesis planning, as shown in

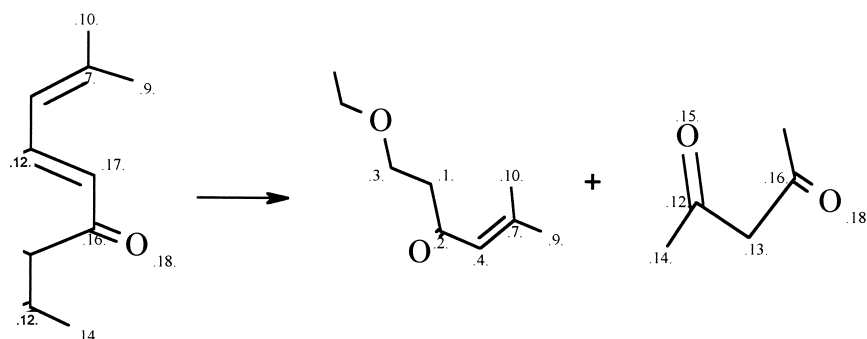


Figure 6. Suggested synthetical reaction for target molecule shown in Figure 5 utilizing the hit reaction shown in Color Plate 2.

Figure 6. Subsequently, these two potential reactants can be used as queries to search the GRKB again and obtain the next step in synthesis planning.

Discussion of GRKB-S

If the target molecule (Query) is a novel compound, it is normally not easy to obtain a satisfactory result because if the compound has never been synthesized it will not be in any reaction database. However, the GRKB-S system, which is based on hierarchical multi-layer reaction patterns searching, can automatically adjust the searching mode, enlarge the matching environment, and find the most relevant reaction for the query. This type of searching greatly expands the usage of knowledge from limited reaction sources. Even though this GRKB is from small subset of the database, we still can use it for synthesis planning and give synthetic chemists useful information. The GRKB-S system has been proved to be a potential valuable, predictive tool for synthesis planning in our laboratory.

CONCLUSION

The terabytes of reaction data that are now being generated and accumulated make classical methods inefficient for knowledge management and discovery. Although there are several commercial reaction databases being updated, the increasing amount of data and old retrieval tools causes them to be underutilized by chemists. Our work shows that data mining, as a newly emerging information technology, can help chemists make better use of reaction data. Principally, our GRKB mining techniques on reaction centers, hierarchy of reaction patterns, and generic reactions capture the essence of reaction knowledge. We have developed a frame system to resolve another critical problem for organization and efficient utilization of the knowledge in GRKB. Furthermore, we have also developed a synthesis planning program to facilitate synthesis planning tasks. During its relatively short period of operation, our GRKB has shown that the general design and methods used in its structuring and building are valid and useful.

ACKNOWLEDGMENT

This work was founded in part by grants from the Minister of Science and Technology of China (Grant No. 96-547-0-02 and

REFERENCES

- Zheng, W., Cho, S.J., Waller, C.L., and Tropsha, A. Rational combinatorial library design. 3. Simulated annealing guided evaluation (SAGE) of molecular diversity: A novel computational tool for universal library design and database mining. *J. Chem. Inf. Comput. Sci.* 1999, **39**, 738–746
- Tratch S.S., and Zefirov, N.S. A hierarchical classification scheme for chemical reactions. *J. Chem. Inf. Comput. Sci.* 1998, **38**, 349–366
- Lewell, X.Q., Judd, D.B., Watson, S.P., and Hann, M.M. RECAP-Retrosynthetic combinatorial analysis procedure: A powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* 1998, **38**, 511–522
- REACCS: Reaction ACCess System: MDL Information Systems, Inc.: San Leandro
- This database was derived from: VINITI/ZIC file: InfoChem GmbH: Munich, Germany
- Blake, J.E., and Dana, R.C. CASREACT: more than a million reactions. *J. Chem. Inf. Comput. Sci.* 1990, **30**, 394–399
- Corey, E.J., Wipke, W.T., Cramer, R.D., and Howe, W.J. Computer-assisted synthetic analysis. *J. Am. Chem. Soc.* 1972, **94**, 421–430
- Wipke, W.T., Ouchi, G.I., and Krishnan, S. Simulation and evaluation of chemical synthesis-SECS. *Artif. Intell.* 1978, **11**, 173–193
- Gelernter, H.L., Sander, A.F., Larsen, D.L., Agarwal, K.K., Boivie, R.H., Spritzer, G.A., and Searleman, J.E. Empirical exploration of SYNCHEM. *Science*. 1977, **197**, 1041–1049
- Hendrickson, J.B. Descriptions of reaction: their logic and applications. *Recl. Trav. Chim. Pays-Bas* 1992, **111**, 323–334
- Gasteiger, J., Hondelmann, U., Rose, P., and Witzendichler, W. Computer-assisted prediction of the degradation of chemicals: hydrolysis of amides and benzoyl-phenylureas. *J. Chem. Soc. Perkin Trans. II.* 1995, **2**, 193–204
- Laird, E.R., and Jorgenson, W.L. Computer-assisted analysis of reactions involving organic free radicals and diradicals. *J. Chem. Inf. Comput. Sci.* 1990, **30**, 458–466
- Ihlenfeldt, W.D., and Gasteiger, J. Computer-assisted planning of organic synthesis: The second generation of programs. *Angew. Chem. (Int. Ed. Engl.)* 1995, **34**, 2613–2633
- Shin-Shyong Tseng. Computer-assisted reaction searching directed toward the synthesis of target molecules. *J. Chem. Inf. Comput. Sci.* 1997, **37**, 1138–1145
- Satoh, H., and 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
- Satoh, K., and Funatsu K. A novel approach to retrosynthetic analysis using knowledge bases derived from reaction databases. *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 316–325
- Wang Ke, M.S. thesis, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai, 1998
- Lynch, M.F., and Willett, P. The automatic detection of chemical reaction sites. *J. Chem. Inf. Comput. Sci.* 1978, **18**, 154–159
- Jauffret, P., Hanser, T., Tonnelier, C., and Kaufmann, G. Machine learning of generic reaction. *Tet. Comput. Methodology*. 1990, **3**, 323
- Andrew, T.B., and Willett, P. Algorithms for the identification of three-dimensional maximal common substructure. *J. Chem. Inf. Comput. Sci.* 1987, **27**, 152–158
- Bron, C., and Kerbosch, J. Algorithm 457: finding all cliques of an undirected graph [H]. *Communications of the ACM*, 1973, **16**, 575–577
- Yuan, S.G., Zeng, F.Y., Zhao, X., and Zheng, C.Z. Identification of maximal common substructure in structure/activity studies. *Analytica Chimica Acta*. 1990, **235**, 239–241
- Blurock, E.S. Computer-aided synthesis design at RISC-linz: automatic extraction and use of reaction classes. *J. Chem. Inf. Comput. Sci.* 1990, **30**, 505–510
- Wilcox, C.S., and Levinson, R.A. A Self-organized knowledge base for recall, design and discovery in organic chemistry, artificial intelligence applications in chemistry, Pierce, T.H., Holme, B.A, Eds., ACS Symposium Series 306, American Chemical Society. Washington, DC, 1986
- Nakayama, T. Computer-assisted knowledge acquisition system for synthesis planning. *J. Chem. Inf. Comput. Sci.* 1991, **31**, 495–503
- Nakayama, T. Building and structuring a large knowledge base for computer-assisted synthesis planning. *J. Chem. Inf. Comput. Sci.* 1995, **35**, 885–893
- Minsky, M. A framework for representing knowledge in the psychology of computer vision, 1975
- Yuan Shen-Gang, Luo Shi-Wei, Chai Ge-Qing, Yao Jian-hua, Chen Hai-feng, and Zheng Chong-Zhi. From pharmacophore to leads: Bioactive compound discovery via computer-aided techniques. *Chinese J. Chem.*, 1999, **17**, 237–243