Computer-Aided Synthesis Design at RISC-Linz: Automatic Extraction and Use of Reaction Classes

EDWARD S. BLUROCK
RISC-Linz, Johannes Kepler University, A-4040 Linz, Austria

Received May 27, 1990

The RETROSYN system, a program (written in LISP) for retrosynthetic planning is introduced. Emphasis is placed on its use of reaction classes. The source of information for the reaction classes is a database of reactions (represented simply as a list of reactants and products), and a means by which this information can be extracted, organized, and then used in RETROSYN is elaborated. The organization of the reaction classes involves establishing a hierarchy of classes and subclasses. Further uses and organizations of reaction classes for use in synthetic analysis are given.

INTRODUCTION

Reaction databases are in widespread use in the organic chemistry laboratory.[1,2] The quality of these systems depends not only on the chemical information contained within the databases but also on the retrieval system supplied with the databases.[3-6] The creation of user-friendly systems which allow for high-level queries are of vital importance. To achieve this, AI techniques can be effectively employed. A typical search in a reaction database using standard systems now available is described.

Standard Existing Technique for Reaction Search. Given a target molecule, the search for a set of suitable retrosynthetic[7] reactions serving as suggestions to be actually tried in the laboratory normally consists of several stages of interaction with a database system. The first stage is to ask whether the target is a product in any of the reactions. If not, then the target must be retrosynthetically analyzed for important groups and reactive centers. This analysis, done by hand, consists of determining the substructures around one of the reactive centers within the target which are involved in a key step in the synthesis. These substructures are the chemist's representation of the class of reactions that are needed to synthesize the molecule. The result of the search for these substructures in the database system is the set of specific reactions in the literature corresponding to this class. Finding a suitable substructure that yields a reasonable number of suggestions is a trial and error process. If too many "hits" are found, then the substructure must be expanded to include more of the target so that less reactions will be included. If no hits are found, then groups of atoms must be eliminated from the substructure to make the search more general.

Retrosynthetic Search Using RETROSYN. RETROSYN is a system developed at RISC-Linz for reaction database search [the long term goal is a complete CAOS (computer-aided organic synthesis) systems[8]]. The RETROSYN module is part of an expert chemical system, written in LISP, which is used to extract, calculate, and organize chemical information (with emphasis on organic synthesis) from the raw information contained in chemical databases.[9] The system is now in usable prototype form.

The following is a description of a how a single step of a retrosynthetic search can be made by using the RETROSYN module and the results that can be obtained:

User Input: The input is given through a menu-driven graphic display using a mouse. The procedure is as follows:

1. Input target molecule
2. Mark the bond to be broken or made or where functionality should be added

Output: An ordered list of possible reactions which could accomplish this task. With each reaction one becomes the following information:

1. The reaction as stated in the database
2. The subtargets resulting from the retrosynthetic application of this reaction
3. The reaction class (graphically represented) to give an indication of the applicability of this class on the target

REACTION PATTERNS

The key to the procedure is the set of reaction patterns. A reaction pattern is defined as two sets of substructures (the product and reactant sides of a reaction) representing the essential structural information needed to adequately describe the reactive center (see refs. 10-14 and references therein for descriptions) of the reaction. In our system, a set of reaction patterns is automatically determined in a preprocessing step from a database of specific synthetic reactions from the literature. These reaction patterns are used in the runtime search procedure. Each reaction pattern holds the following information:

1. A set of pointers to specific reactions from the original reaction database which belong to this class.
2. A graphical description of the reaction center describing this class. This description consists of the *essential* atoms and bonds making up the reactive center and the surrounding activating groups. The reaction class description is a compromise between completeness and economy of information.
3. The set of atoms and bonds which change in character as a result of the reaction.

The *essential* atoms and bonds describing the class can be defined in various ways. Common to all definitions is the *reactive center* which is the set of atoms and bonds which have changed in character between the reactants and products. (For example, Figure 2b is the reactive center of the reaction shown in Figure 2a.) Although in the literature the representation of the reactive center (both on paper and in the computer) is quite varied, the information content is relatively the same. The reactive center alone, however, does not describe the driving force of the reaction (e.g., activating groups). The description of the reactive center is thus *expanded* to include a certain set of the surrounding atoms and bonds. Most definitions of this set are slight variations of that by Wilcox and Levinson.[15] The definition used here and how it is derived is described in the next section.

Each reaction pattern represents a set or *class* of reactions. Due to the graphical[16] nature of the internal representation of the reaction classes, the words reaction class and *reaction pattern* are used here interchangeably. The word class is used in context of classification, and pattern in the context of its concrete representation as a molecular graph. In a later section a hierarchy of reaction classes, resulting from the derivation

0095-2338/90/1630-0505$02.50/0

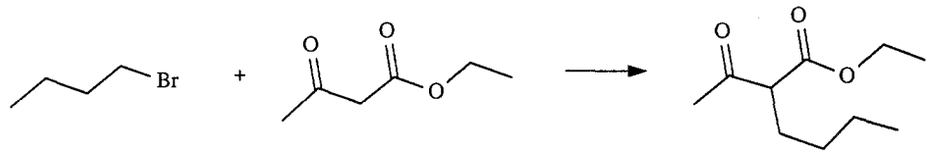

Figure 1. The original reaction from the databank is a set of molecules, each of which being represented, for example, as a connection table.

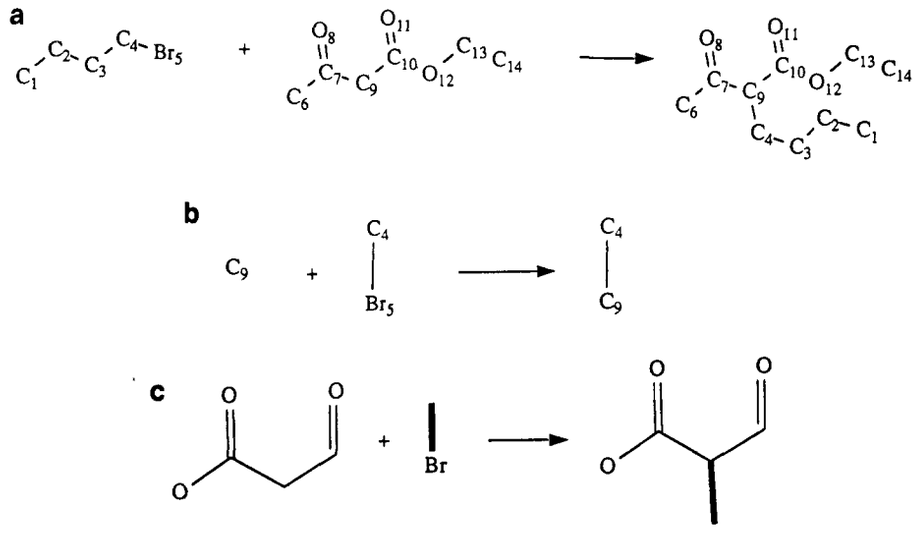

Figure 2. (a) Graph difference algorithm finds the atom–atom correspondences. (b) Using the atom–atom correspondences, remove all atoms and bonds which did not change in character in the course of the reaction. What remains is the reaction pattern as derived from the graph difference algorithm (compare with Figure 1. (c) The reaction pattern is expanded to include the activating groups local to the reactive center (compare with Figure 1 and parts a and b of this figure).

of the reaction pattern from a database of reactions, will be introduced.

Derivation of the Reaction Patterns. The main thrust of the work at RISC-Linz is to develop methods of extracting all the essential information needed for synthetic planning automatically (with little user intervention) from the raw databases of chemical information (in this case reaction databases). The following is a short description of how the reaction patterns are automatically extracted:

Input: The only required input is a set of reactions represented as a set of connection tables (for example, MOL files[3])—see Figure 1.

Abstraction Algorithm: The following is a short outline of how the reaction classes are extracted from a database of reaction information in a preprocessing step and then organized for use in a runtime synthetic planning system.

1. Derivation of a reaction pattern from a single reaction in database
   a. Use a modified maximal common subgraph technique,[13,14,17,18] the *graph difference* algorithm, to find atom–atom correspondences between the reactions and products (Figure 2a).
   b. Select out that atom correspondence leading to the simplest change.
   c. Determine the set of bond changes, valence changes, and leaving groups. This is done by eliminating all atoms and bonds which do not change in character (see Figure 2b).
   d. *Expand* the reaction pattern to include the activating groups[15] which are *local* to the reactive center found (see Figure 3). The term local is used in a general sense in that entire aromatic and resonant systems bordering on the reactive center are also included. Those atoms and bonds which are included is that set along the path from the reaction center to (but not including) the first singly bonded carbon.

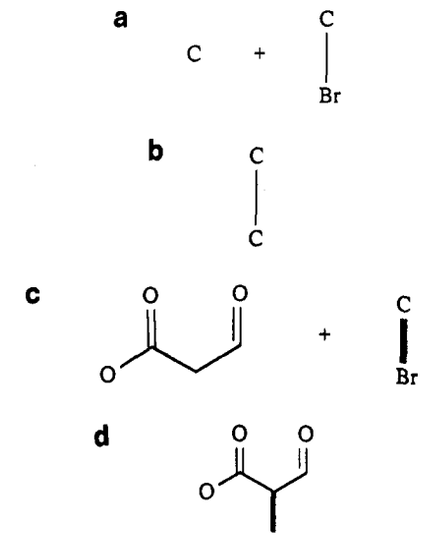

Figure 3. (a) Reactant reaction center class (reactant–RCC) of the reaction in Figure 2. A significant feature which this particular example shows is that the reaction involves bromine as a leaving group (i.e., this information could be used to decide early whether problems could result when using this class). (b) Product reaction center class (product–RCC) for the reaction in Figure 2. The significant feature of this example is that the class represents those reactions which only break a carbon–carbon bond and nothing else. Note that since the bromine (as a leaving group) is not included in this class description the reactions involving the other halogens are also included in this class and would then differentiate with the reactant–RCC description. (c) Reactant expanded reaction center class (reactant–ERCC) of the reaction in Figure 2. Here the surrounding structures are included. (d) Product expanded reaction center class (product–ERCC) of the reaction in Figure 2. This is the most significant class for retrosynthetic analysis. To be able to use this class, one has find this entire structure in the target.

2. Organization of derived reaction patterns (retrosynthetic analysis)

a. Group classes according to their functionality (e.g., by which bond is broken in the reaction).
b. By use of a partial order definition, determine equivalent and similar reaction classes (for possible definitions of similar, see section below).
c. Group those patterns which have the same product subgraph.
d. Organize the set of unique product subgraphs in order to facilitate efficient subgraph search within a target (rooted subgraph search using graph invariants[19]).

HIERARCHY OF REACTION CLASSES

Through the methods of extraction and the use of the reaction information from a database of reactions, a natural hierarchy of classes and subclasses of reaction characterizations is established. This is outlined below. The specific objects in each class are characterized by a molecular graph description. Many of the classes are represented by sets of molecular graphs (such as a set of connected graphs for the reactant and product sides of the reaction, respectively). In most descriptions of classes (see references in Zeferov[10]), the class is represented as a single object where the changes are represented by labeled bond (i.e., whether the bond is to be broken or made). In the hierarchy presented here, the product side and the reactant side are separated, and each class is represented as a static graph. In essence, the information content of both is the same. This separation arose mainly from the use of the reaction patterns in the retrosynthetic planning system and the distance meanings that arose in various stages of analysis.

The concept of classes, subclasses, and superclasses is used in the descriptions. One uses this organization to structure the set of individual reactions. A more structured database yields a more efficient use of the objects within the database. A class represents a set of objects (i.e., reactions) which have the same characteristics (as defined by the class description). A subclass is a subset of the set of objects. The description of the subclass relative to the class is more specific, i.e., it contains more detail. A superclass is to a class as a class is to a subclass, i.e., it has lass specific information than the class and thus describes a larger set of objects.

List of Reaction Classes. The following is a description of the classes that follow from the derivation procedure of the previous section.

Class: Reaction Type. This class represents the operation which the reaction can perform. An example of such a class (as used in RETROSYN) is retrosynthetic bond breaking. A single reaction can perform several operations and thus belong to several *reaction-type* classes. For example, the reaction in Figure 2 belongs to both reaction-type classes carbon–carbon bond making (retrosynthetic breaking) and carbon–bromine bond breaking (retrosynthetic making).

Class: Reaction Center Class (Product–RCC and Reactant–RCC). This is the class as defined by the reaction center of the reaction. The reaction center is defined as the set of bonds and atoms that have changed in character within the course of the reaction.

This class is divided into two subclasses as defined by the molecular subgraph of the products and of the reactants. Since the use of the reaction classes is primarily retrosynthetic, the class defined by the products is a superclass of that defined by the reactants. The product–RCC gives the result of application of the reaction, and the reactant–RCC gives the method by which this result is to be achieved. For example, the reactive center of a reaction could just be a carbon–carbon bond. The reactant–RCC could split this class up into the methods of achieving this bond such as through the use of various leaving groups (see Figure 3b).

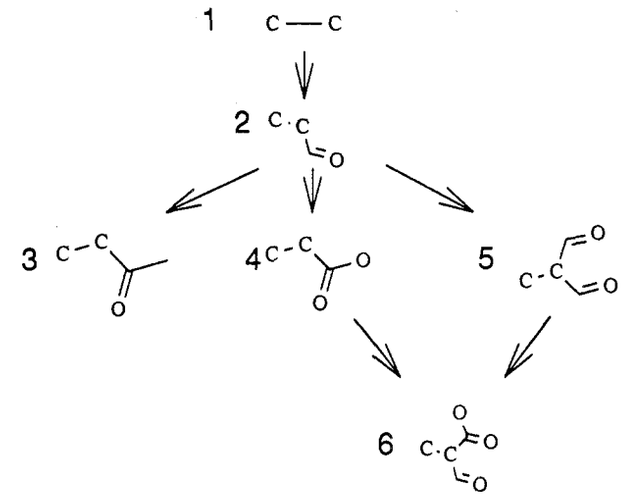

Figure 4. Partial ordering of the product side of the expanded reaction pattern class. Using the definition of similarity, reactions **3**, **4**, and **5** are similar because they have reaction **2** as a common subgraph. Hence using the functional group interchange they could be interconverted.

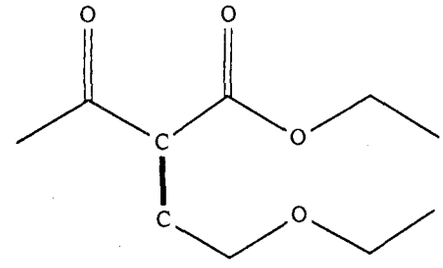

Figure 5. Input menu provides all the usual functions of target input, including an optimizing function which cleans up the by-hand inputted molecule. The molecule shown is the target used in this example, and the marked bond is that which should be broken.

Class: Expanded Reaction Center Class ERCC. This is the class where the reactive center description is *expanded* to include neighboring activating groups. In the use of the reaction classes, it was found that the reaction center description alone did not adequately describe the sense of the specific reaction. What is lost are the activating groups which do not change in character within the course of the reaction (and hence do not appear in the reactive center description), but nevertheless contribute the driving force of the reaction. Therefore, in the description of this class just enough of the surrounding activating groups are included to characterize the sense of the reaction.

The method used to specify which atoms and bonds to include around the reactive center can be defined in many ways. The most common method (which is used here) is that of Wilcox and Levinson.[15] This is a fixed set of rules which essentially specify that all atoms and bonds around the reactive center should be included that are not singly bonded carbons. Work is underway by the author in using *learning* algorithms to specify which groups should be included. Such methods are not based on fixed chemical rules. A group of reactions within the same reactive center class are analyzed, and common substructures are searched for.

As before, this class is divided into two subclasses, one for the products and one for the reactants, and as before, for retrosynthetic purposes, the products are a superclass of the reactants. The separation at this stage is more than academic in that in the RETROSYN analysis the product class description is used to determine whether the class of reactions can be used or not, i.e., if the molecular subgraph of the class

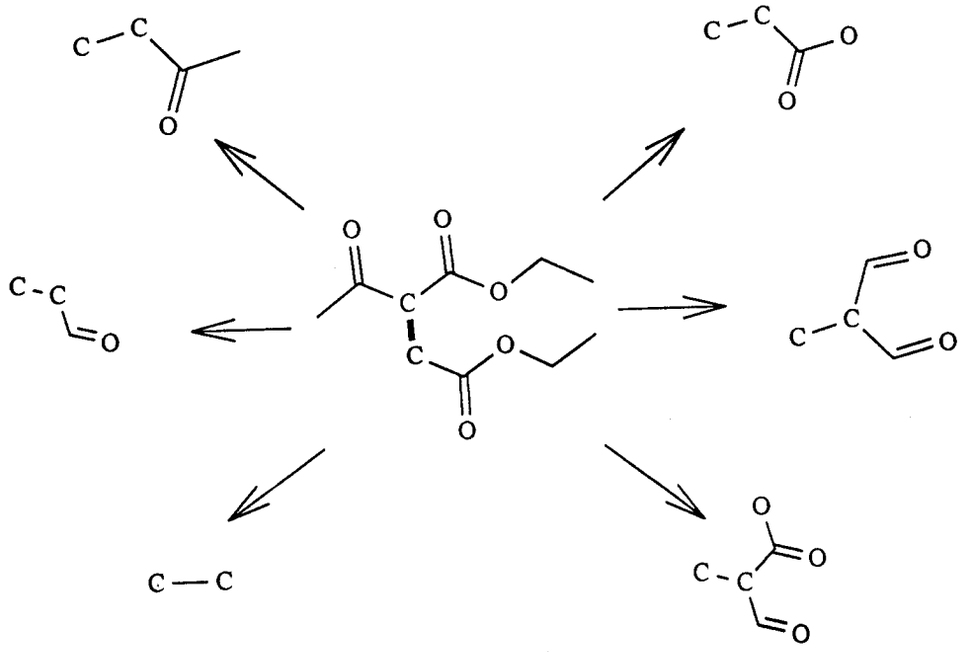

Figure 6. From the marked bonds in the target, the product expanded reaction center classes were searched. A larger pattern implies that the corresponding class of reactions are more specific to the target.

is found within the target molecule, then this class of reactions can, in principle, be used to perform a retrosynthetic step (see RETROSYN section for more details).

Class: The Specific Reaction. At the end of this hierarchal tree is the specific reaction represented as two sets of molecular graphs. In a certain sense it is still a class representing several reactions in that side conditions (e.g., temperature, pressure, reflux, acidity, etc.) are not included in the description.

ADDITIONAL USES OF REACTION PATTERNS

In addition to the simple reaction class search described above, the reaction patterns can be used in other ways to add an extra level of sophistication to a retrosynthetic search.

Definition of Similar Reactions. A definition of one type of similarity (for other definitions see refs 5, 20, and 21) between reactions can be derived either from the concept of reaction classes or from the partial ordering of the reactions.

In reaction classes, two reactions are similar if they have at least one class in common within the hierarchy of reaction classes. The degree of similarity is a function of what class they have in common. For example, two reactions having only the class of carbon–carbon bond-breaking reaction in common are not as similar as two reactions belonging to the class with the same expanded reaction pattern.

A degree of similarity within a class can be given by a certain definition of partial order. With the class objects represented as reaction patterns (i.e., molecular graphs), then a partial ordering can result by using the definition of subgraph, i.e., one pattern is less than another if it is a subgraph of it. No relation exists if neither is a subgraph of the other (hence the term *partial* ordering). Under this ordering, two reactions are similar if, within one of the classes, the reactions share a subgraph (see Figure 4).

This definition of *similar* is restricted to a hierarchy of reactions each of which have common subgraphs. This does not include reactions that are mechanistically similar but have different atoms. For example, the set of reactions involving iodine substituted for bromine in the reaction shown in Figures 2 and 3 would not be considered similar by this definition. This could be (at least partially) eliminated by introducing a generalization step into the reaction organizational procedure. Those reaction classes with the same bonds but with one (or more) different atom(s) substituted could be grouped together (i.e., generalized) into one class where a *wildcard* atom is substituted on the position(s) where the atoms differ. Such methods are being explored at this time (e.g., using learning algorithms to determine whether such a substitution is justified and the possibility of substituting entire substructures).

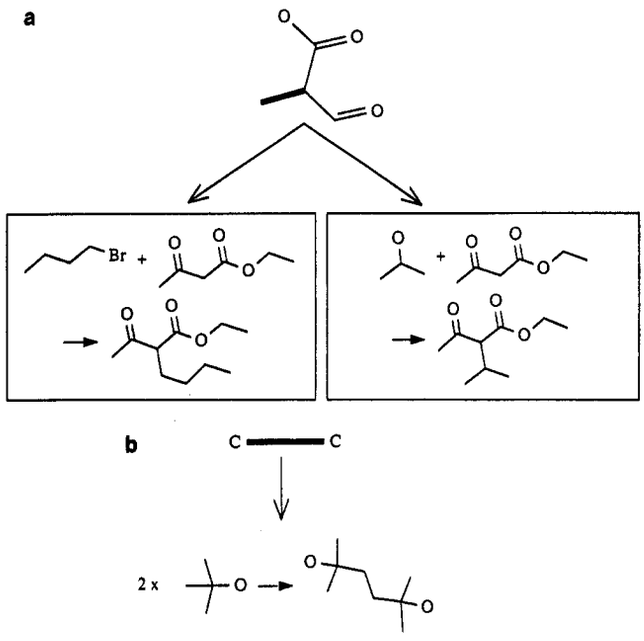

Figure 7. (a) Two examples of the specific reaction examples to break the bond in the target. Note that they differ only in the leaving group. This implies that the reactant side of the reaction pattern can also be used as a further reaction class refinement. (b) Example of a very simple reaction class yielding a general reaction for carbon–carbon bond breaking.

Determination of a Specific Class of Side Reactions. Given a matched reaction pattern (the product–ERCC is a subgraph of the target), one can determine if the subgraphs of the reaction pattern exist elsewhere in the target and retrosynthetic subtargets. Finding matches elsewhere in the target means that a reaction takes place not only on the site intended but also on the other sites, hence side reactions can occur. This

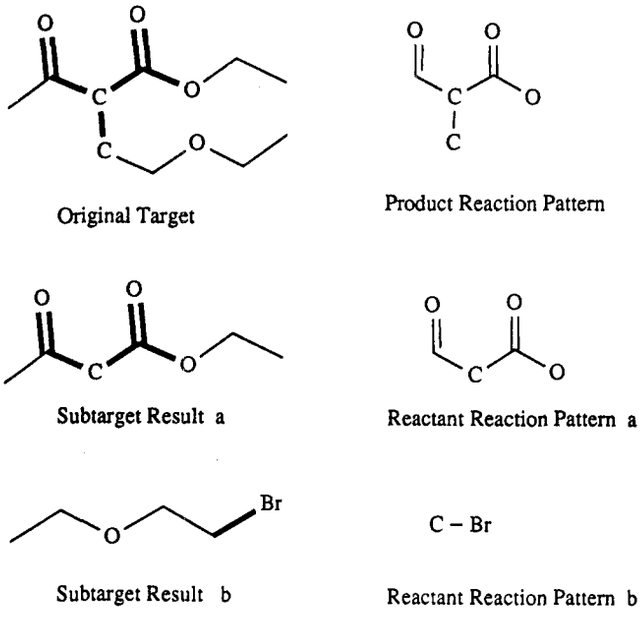

Figure 8. Performance of the reaction through the reaction pattern.

definition gives an example of a class of side reactions that can occur (it is apparent to the author that this does not include all types of side reactions that can occur).

Target Modification over Several Steps. If a reaction class produces too many side effects, using the definition of reaction similarity, a more specific reaction, i.e., one that has more functionality around the reactive center, can be suggested. With these extra functionalities, the reaction can be made specific to the intended site and not applicable to the sites that gave side reactions with the original pattern. The modification of the target molecule to supply the extra functionalities needed can be automatically sent as subgoals to the system. Upon return, the target has the required functionality to perform the more specific reaction class.

Automatic Determination of Leaving Groups. In databases of reactions, there can be sets of atoms (connected together in one group) which appear on the reactant side but not on the product side of the reaction. These groups are the leaving groups. This analysis uses the implicit information within databases where only the *important* molecules are listed on the product side. The leaving groups are defined within the program as those (connected) substructures which have not been matched between the products and the reactants. These leaving groups are stored as part of the definition of the reaction pattern.

A SAMPLE RUN OF RETROSYN

The Database. The current database of RETROSYN is a selection out of the ORGSYN Database of Molecular Design Ltd.[3] (The entire database was analyzed, but only a subset is used.) The ORGSYN Database was chosen because it represents a good, broad base of organic chemistry.

The Disconnection Approach. The current RETROSYN system analyses the target with respect to the disconnection approach as introduced by Warren.[22] The system requires as input a target molecule and a bond to be broken. One inputs the target with a mouse and a menu-driven input window. The usual functions of adding, deleting, and moving atoms and bonds are included. Also included is an "optimize" function which optimizes the input on the screen (see Figure 5).

The Range of Answers. The search for suitable retrosynthetic reactions is performed by means of the expanded product reaction pattern. Figure 6 shows a set of matched reaction patterns that were found within the target. The ordering of the reaction patterns is a function of the size of the matched pattern. The philosophy is that a larger pattern represents a

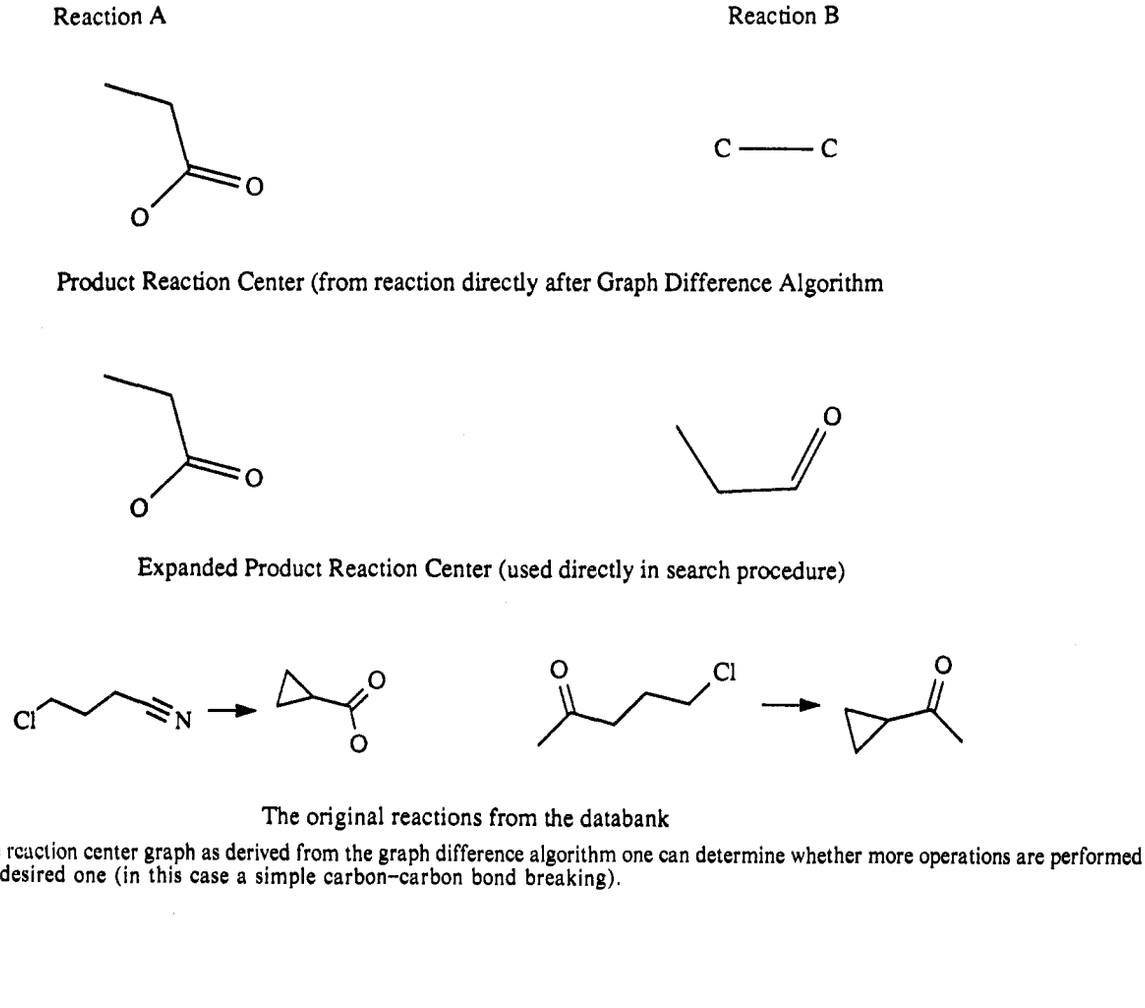

Figure 9. From the reaction center graph as derived from the graph difference algorithm one can determine whether more operations are performed in addition to the desired one (in this case a simple carbon–carbon bond breaking).

reaction that is more specific to the target and thus has a greater chance of being an appropriate choice. The answers range from the relatively specific carbon–carbon bond-breaking reaction requiring specific activated groups (i.e., that with the largest pattern) as in Figure 7a to a very simple and very general reaction with no activating groups as in Figure 7b. It is currently left to the chemist to analyze the conditions under which the specific examples occur and to determine whether they are appropriate for use.

Application of Pattern. As an aid to determining the quality of the answers found, the system *performs* the reaction on the target and shows the resulting subtargets. This is done by the system via the reaction pattern. By using the correspondence between the atoms on the product side of the reaction pattern and the target, those bonds and atoms which change in character in retrosynthetically going from products to reactants in the reaction pattern are noted and modified appropriately in the target. An example is shown in Figure 8.

Refinement of Reaction Classes. In the example, the set of reactions belonging to the class of carbon–carbon breaking reactions were searched. The product side of the expanded reaction pattern forms a subclass (P–ER class) of this class and is used to narrow down and order the set of possible carbon–carbon breaking reactions in determining whether the necessary functionality is present in the target. One sees in Figure 7a that a further and useful refinement the P–ER class is possible, namely that described by the reactant side of the expanded reaction pattern (the R–ER class). This gives a further refinement reflecting the method of the reaction and is used to test whether the functionality of the reactants can have destructive effects.

Several of the intermediate results of the original database analysis can yield further useful refinements of the reaction classes. An example is the reaction center as found by the graph difference routine (i.e., the product reaction center class, P–RC class, and the reactant reaction center class, R–RC class). These classes represent the full functionality of the reaction, e.g., whether, in addition to the desired bond breaking, the reaction breaks or makes other bonds. Figure 9 shows an example where this distinction is made.

CONCLUSION

The use of reaction classes plays an important role in synthetic chemistry. The definition of a hierarchy of reaction classes and the definition of an ordering within a reaction class provide useful tools for the organization of results and give the first hints toward multistep analysis. The extraction of reaction classes through automatic means has the great advantage in that one keeps the connection between the classes and all the original reactions from the literature. In addition, as the knowledge of chemistry increases, a synthetic system based on automatically extracted reaction classes can easily keep pace. Instead of putting the effort in defining reaction classes by hand, one is working at a higher level by developing methods to automatically extract such classes.

REFERENCES AND NOTES

(1) Zefirov, N. S.; Gordeeva, E. V. Computer-Assisted Synthesis. *Russ. Chem. Rev. (Engl. Transl.)* **1987**, *56*, 1002–1014.
(2) Barone, R.; Chanon, M. *Computer Aids in Chemistry*; Ellis Harwood Ltd.: Chichester, 1986.
(3) *REACCS: Reference Manual*, Molecular Design Limited: San Leondro, CA, 1986.
(4) Kos, A. J.; Grethe, G. *Nachr. Chem. Tech. Lab.* **1987**, *35*, 586.
(5) Moock, T. E.; Grier, D. L.; Hounshell, W. D.; Grethe, G.; Cronin, K.; Nourse, J. G.; Theodosiou, J. Similiarity Searching in the organic reaction domain. *Tetrahedron Comput. Methodol.* **1988**, *1*, 117–128.
(6) Johnson, A. P. Computer Aids to Synthesis Planning. *Chem. Br.* **1985**, 59–69.
(7) Corey, E. J. Retrosynthetic Thinking—Essentials and Examples. *Chem. Soc. Rev.* **1988**, *17*, 111–133.
(8) Blurock, E. S. Computer Aided Organic Synthesis: Development and Implementation of a Complete Synthetic Strategy. *Tetrahedron Comput. Methodol.* **1989**, *2*, 207–222.
(9) Blurock, E.; Strelow, T. Automatic Chemical Synthesis Planning. *DECHEMA Monogr.* **1989**, *116*, 531.
(10) Zefirov, N. S. An Approach to Systematization and Design of Organic Reactions. *Acc. Chem. Res.* **1987**, *20*, 237–243.
(11) Fujita, S. The description of organic reactions based on imaginary transition structures. A novel approach to the taxonomy of organic reactions. *Pure Appl. Chem.* **1989**, *61*, 605–608.
(12) Koca, J. A Graph Model of the Synthon. *Collect. Czech. Chem. Commun.* **1988**, *53*, 3108–3130.
(13) Wochner, M.; Brandt, J.; von Scholley, A.; Ugi, I. Chemical Similarity, Chemical Distance, and its Exact Determination. *Chimia* **1988**, *42*, 217–225.
(14) Funatsu, K.; Endo, T.; Kotera, N.; Sasaki, S. Automatic recognition of reaction site in organic chemical reactions. *Tetrahedron Comput. Methodol.* **1988**, *1*, 53–70.
(15) Wilcox, C. S.; Levinson, R. A. A Self-Organized Knowledge Base for Recall, Design and Discovery in Organic Chemistry. *ACS Symp. Ser.* **1986**, *306*, 209–230.
(16) Balaban, A. T. *Chemical Applications of Graph Theory*; Academic Press: London, 1976.
(17) Lynch, M. F.; Willet, P. The Automatic Detection of Chemical Reaction Sites. *J. Chem. Inf. Comput. Sci.* **1981**, *18*, 154–159.
(18) Takahashi, Y. S.; Suzuki, H.; Sasaki, S. Recognition of Largest Common Structural Fragment Among a Variety of Chemical Structures. *Anal. Sci.* **1987**, *3*, 23–28.
(19) Nagy, M. Z.; Kozics, S.; Veszpremi, T.; Bruck, P. Substructure Search on very large files using Tree-Structured Databases. In *Chemical Structures*; Warr, W. A., Ed.; Springer-Verlag: Berlin, 1988; pp 127–130.
(20) Willet, P.; Winterman, V. A Comparison of Some Measures for the Determination of Inter-Molecular Structural Similarity. *Quant. Struct.-Act. Relat.* **1986**, *5*, 18–25.
(21) Wipke, W. T.; Rogers, D. Artificial Intelligence in Organic Synthesis. SST: Starting Material Selection Strategies. An application of Superstructure Search. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 71–81.
(22) Warren, S. *Workbook for Organic Synthesis, The Disconnection Approach*; Wiley: Chichester, 1982.