

## Fragmentations and the "Feeling of Goal" in Computer-Assisted Synthesis

Dmitry E. Lushnikov and Nikolai S. Zefirov\*

N. D. Zelinsky Institute of Organic Chemistry, Leninsky Prospect 47, 117913, Moscow, Russia

Received November 4, 1991

Many successful short syntheses actually include the cleavage of carbon-carbon bonds as well as their formation. To incorporate such fragmentation reactions (as well as functional group interchanges) in synthesis design, one should state the goal of their usage. For this purpose, key reactions in multistage syntheses are divided into a certain strategic kernel (which defines the remote, final goal of their use and reflects all important changes, from a strategic point of view, reached after their application), and its tactical environment (some functional, activating, protecting groups, and multiple bonds which are necessary for carrying out the reaction successfully but may not be necessary for the following steps of the synthesis). The hierarchical goal-driven search proposed in the STORM microcomputer program consists of two steps. At the top level, the program tries to use the kernel of the key transformation, ignoring the discrepancies between the transformation environment and the target structure. It allows possible fragmentations and functional group interchanges to hide to avoid their explicit enumeration. If this new direction seems promising, the program can search for the actual ways of modifying the target structure with the aim of carrying out the key transformation at the selected place. Several chemical examples are shown.

### INTRODUCTION

For the last 20 years, significant attention has been paid to the development of computer-assisted synthesis projects (see refs 1 and 2). The crucial point of every such program is the formalization of knowledge about organic reactions and about the principles concerning the planning of multistep syntheses. Synthetic plans created by chemists frequently consist of a sequence of steps united by an implicit strategic goal. In order to carry out powerful simplifying transformations (retroreactions), such as the Diels-Alder reaction, the Robinson annulation, and halolactonization, the methodology of long-range search was developed in LHASA.<sup>3-5</sup> This methodology permits one to find sequences of preliminary transformations, such as functional group interchanges (FGI) or functional group additions, for subsequent successful application of the key retroreaction.

The general trend in synthetic planning is successive simplification of the target structure. The most consistent development of this principle was described by Hendrickson et al. in the course of the creation of the SYNGEN program.<sup>6-8</sup> This program implements the strategy of an "ideal", fully convergent synthesis which consists, for the major part, of constructive reactions. The problem of finding the optimal forming sequence for the given set of strategic bonds in the convergent retrosynthesis was solved heuristically in the LILITH<sup>9</sup> program. However, it should be mentioned that one can encounter numerous difficulties during the practical realization of even the shortest convergent plan.

That is why various roundabout ways are so common in organic synthesis. One of the most exciting ways is the cleavage of carbon-carbon bonds in the course of the synthesis. In the retrosynthetic direction, such a destruction means the complication of the target structure by the addition of new atoms and/or bonds. However, in the case of such fragmentation transformations, it is very difficult to formalize the goal of their usage. The only known program which includes them is LHASA. It uses the strategy of "reconnective transforms" to connect appendages with rings for subsequent stereospecific synthesis.<sup>10</sup> Applications of fragmentation reactions and of

rearrangements for skeleton assembly were logically analyzed by Hendrickson.<sup>11</sup>

We focused our efforts on the search for the ways of target simplification via fragmentation. For the present moment, in the STORM microcomputer program discussed below, we do not use the procedures for the automatic selection of strategic bonds and for the automatic evaluation of precursors. We have tried to study the possibilities and limitations for the application of fragmentation reactions before constructive transformations.

### LONG-RANGE SEARCH

Initially we tried to include fragmentation transformations in the long-range search technique within the framework of the COMPASS program.<sup>12</sup> The program carries out the search for sequences of retroreactions that begin with functional group interchange (FGI) transformations and may include a fragmentation (destructive transformation) (Figure 1). The last transformation in each sequence must be constructive. The crucial point in rejecting unfruitful lines of analysis is the conservation of connectivity for the whole sequence. It means that the changes in the target structure which are caused by a certain transformation from the sequence should be used by other transformations and finally by the constructive transformation from the same sequence.

The COMPASS program was shown to reproduce the sequence of key steps (involving fragmentation reactions) for the major part of published total syntheses of sesquiterpenes such as quadron and sirenin.<sup>12</sup> The plan proposed by the COMPASS program and reproducing the key steps for the total synthesis of quadron by Burke et al.<sup>13</sup> is shown in Figure 2. The tree branching points are marked by asterisks. Types of transformations are indicated above the arrows.

Nevertheless, this data-driven exhaustive enumeration of transformations has many shortcomings. First, the procedure is time-consuming, especially for the use of fragmentation transformations. The possibilities for preliminary modifications of the target structure are considerably restricted by the

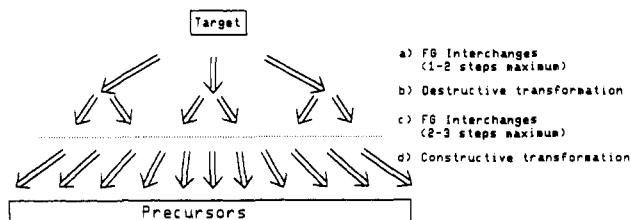


Figure 1. Application of the long-range search in the COMPASS program.

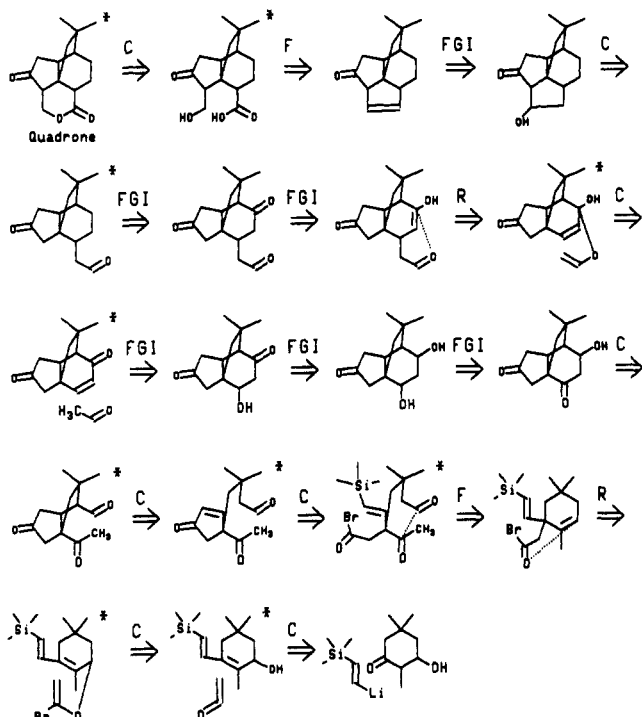


Figure 2. Reproduction of the total synthesis of quadrone published in ref 13. FGI, functional group interchange; F, fragmentation transformation; C, constructive transformation; R, rearrangement.

available set of FGI and fragmentation transformations. Last, some syntheses (even for such a test compound as quadrone) cannot fit this scheme of long-range search. A more rational approach is suggested in the STORM program, although the long-range search is also included as an optional search technique.

#### NEW METHODOLOGY OF FRAGMENTATION USAGE

Key reactions in multistage syntheses can be divided into a certain strategic kernel and its tactical environment. The kernel defines the remote, final goal for the use of these reactions. The environment is necessary for carrying out the reaction successfully, but it can be changed or removed at the following steps of synthesis. The kernel is not retroreaction itself but its essence, which reflects all important (from a strategic point of view) changes reached after its application in the structure. To obtain a kernel, one should remove environment from the retroreaction description, namely, those functional, activating, protecting groups, some multiple bonds that are allowed to be changed at the following steps of synthesis.

The planning of complicated syntheses seems to be hierarchical: first of all, the chemist tries to think up how to use the kernel of the key transformation, ignoring the discrepancies between the transformation environment and the target structure. If this new direction seems promising, one

can search for the actual ways of modifying the target structure with the aim of carrying out the key transformation at the selected place. Before considering this strategy of hierarchical search in more detail, we should describe the representation of transformations in the STORM program.

#### REPRESENTATION OF TRANSFORMATION RULES

In the STORM program, as in an expert system, all transformations are separated from the body of the program and written down in the knowledge base in the form of production rules. Each rule represents a labeled graph of arbitrary complexity (it can be even disconnected). Its vertices and edges correspond to lists of condition checks (the IF part of the production rule) and to lists of necessary changes (the THEN part of the production rule). An edge may have zero multiplicity, which denotes the absence of the corresponding bond in the chemical structure. Thus, one can easily describe both the cleavage and the creation of new bonds. Moreover, this way one can describe not only retrosynthetic transformations but also synthetic reactions, elementary steps of reaction mechanisms, R-matrices,<sup>14</sup> and cyclic redistributions of bonds.<sup>15</sup>

Each check of condition consists of three elements: the name (the identifier) of the value to be checked, the symbol of the relation ( $=$ ,  $<$ ,  $>$ ,  $<=$ ,  $>=$ ,  $<=>$ ), and the number (or the symbol) that is to be compared with this value.

The vertices of the rule may include the following checks (the allowed relations are placed in parentheses):

- the symbol of the chemical element ( $=$ ,  $<$ ,  $>$ )
- the number of hydrogens at the atom in question (all relations)
- the charge of the atom in question ( $=$ ,  $<$ ,  $>$ )
- the multiplicity of activation (only for carbon atoms; denotes the number of bonds with heteroatoms, their multiplicity taken into account) (all relations)
- the hybridization ( $sp^3$ ,  $sp^2$ ,  $sp$ ) ( $=$ ,  $<$ ,  $>$ )
- the degree of substitution (the number of bonds with non-hydrogen atoms) (all relations)
- the number of valence electrons (only s-, p-, d-electrons, all relations)
- the number of lone electron pairs (all relations)
- the number of vacant orbitals (all relations)

The allowed changes which can act on atoms (the vertices of the rule) point out (for a certain parameter) either the increment or the decrement of its current value or its new value. The list of changes can include the following directions:

- a change in the number of hydrogens at the atom in question
- a change in the multiplicity of activation (the addition/removal of the simplest functional groups:  $-OH$ ,  $=O$ )
- a change in the charge of the atom in question
- the addition of the given structural fragment to the atom in question

Checks at the edges of the rule may include:

- the bond multiplicity (all relations)
- whether the bond has been marked by the user ( $=$ ,  $<$ ,  $>$ )

For bonds, only the change of multiplicity is allowed.

#### PARTIAL MATCHING OF RULES AND IMAGINARY STRUCTURES

The search for precursors means the search for all non-equivalent isomorphous embeddings of the subgraph (the

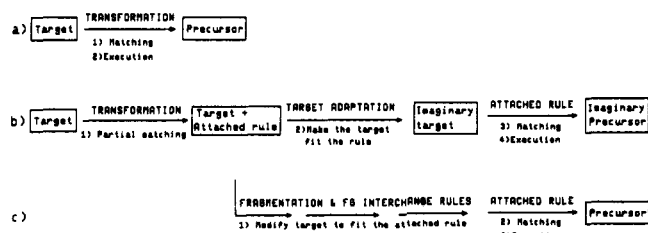


Figure 3. Flow chart of the hierarchical search in the STORM program.

transformation rule) into the graph which corresponds to the target structure. If all the checks at the atom/bond which corresponds to a certain vertex/edge of the rule are true, this vertex/edge becomes matched to the corresponding atom/bond in the target structure. If the matching of the whole rule was successful, the program modifies the target structure according to the lists of changes and forms a precursor (Figure 3a).

Both checks of conditions and edges/vertices themselves can be characterized as either hard (if they are related to the transformation kernel) or soft (if they are related to the environment). In general, the matching of soft checks, edges, and vertices with the target structure is not necessary. If some soft checks are false or if some soft edges or soft vertices remain unmatched, this matching of the rule with the structure will be called partial.

At present, the program includes three modes of partial matching:

**Soft Checks.** Some soft checks on edges and vertices may remain unmatched, but all vertices and edges should be matched.

**Soft Bonds.** Some soft edges of the rule may be unmatched to actual bonds.

**Soft Atoms.** Some soft vertices and all the adjacent bonds may remain unmatched.

Thus, during partial matching, the program carries out the search for all nonequivalent embeddings of only the corresponding hard part of the rule. The soft part may be partially or completely unmatched. The soft bonds and soft atoms modes include the soft checks mode, but the former two are independent.

Unfortunately, a partially matched rule cannot be carried out directly. In order to obtain the result, the program must make an "intuitive leap", i.e., change the target structure so that all the unmatched checks, edges, and vertices get matched, and then form the precursor. Such a new adapted target structure will be referred to as an "imaginary target", and the precursor obtained after the rule is carried out becomes an "imaginary precursor" (Figure 3b).

To build an imaginary target from the target structure and the partially matched rule, the program tries to modify the target either until all the remaining soft checks are satisfied or until all the possibilities of modification have been considered. Such an adaptation may include a change of bond multiplicity, the insertion/removal/modification of simple functional groups, and the addition of necessary bonds and atoms. The format (the representation) of checks in the rule allows one to control the causes of failures (nonfulfillment of the rule) without difficulty and to modify the structure in the desired direction.

Each pair of imaginary structures (target precursor) clearly reconstruct an example of the necessary environment of the transformation kernel. Therefore, the user can evaluate whether the key transformation can be regarded as promising in this or the other place of the target structure. All the

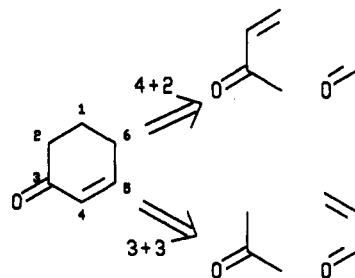


Figure 4. Rules of retro-Robinson annulation.

necessary FGI and fragmentation transformations are imitated by the module of imaginary structure generation, and at the top level they have no actual chemical analogues.

At the bottom level of the hierarchy, promising variants from the list can be investigated in detail. The goal of such "environment search" is to find sequences of actual FGI/fragmentation transformation, in order to construct the necessary environment about the matched kernel. As a result, the partially matched rule should be fully matched and a precursor should be generated (Figure 3c). Let us consider some examples.

## EXAMPLES

**Soft Checks.** The division of the key transformation into the kernel and the environment allows one to carry out the search for nonobvious applications of powerful synthetic methods of skeleton construction. The LHASA long-range search<sup>5</sup> and the MARSEIL/SOS computer program<sup>16</sup> imply the same aim. To compare the results, we have studied the test case from the paper,<sup>5</sup> namely, the use of retro-Robinson annulation for the planning of valeranone synthesis.

For this purpose, two rules corresponding to the 4 + 2 and 3 + 3 annulation (Figure 4) have been defined. The rule of the 4 + 2 annulation is presented in Table I, with numbers of atoms in Figure 4 corresponding with items from this table. As many checks in the rules as possible were specified as soft in order to expand the transformation environment and to make a wider search possible.

The program found 28 ways of applying the kernel of retro-Robinson annulation to the valeranone skeleton. Half of solutions corresponds to the 4 + 2, and half to the 3 + 3 annulation. From this set, only eight imaginary targets have been now selected; for them, the direct Robinson annulation of the imaginary precursor yields only one product (the imaginary target). Two solutions, where one of the precursors was represented by acetaldehyde, were rejected. Other pairs are presented in Figure 5. At the bottom of Figure 5, the three most convergent ways are given. Paths 1 and 2 were suggested earlier by LHASA among the lowest ranked ways.<sup>5</sup>

The imaginary targets and precursors do not contain any information about the actual ways of modifying the imaginary target into the target structure. They show only the results of the use of the transformation kernel. An example of the lower-level analysis of imaginary structures will be discussed below.

**Soft Bonds.** The suggested hierarchy allows one to formalize an original technique proposed by Indian chemists for triquinanes syntheses.<sup>17</sup> Its peculiarity is the use of [2 + 2]-cycloaddition, followed by the cleavage of another pair of bonds in the four-atom ring during [2 + 2]-cyclofragmentation (Figure 6).

For the analysis, the two edges in the retro[2 + 2]-cycloaddition rule, which are to break in the course of cyclof-

Table I. Rule for 4 = 2 Robinson Annulation Transformation<sup>a</sup>

item	IF			THEN
vertex 1	label	= "C"	hard	
	activation	= 0	soft	
vertex 2	label	= "C"	hard	change no. of hydrogen atoms on -1
	activation	= 0	soft	
	no. of hydrogen atoms	> 0	soft	
vertex 3	label	= "C"	hard	
	activation	= 2	soft	
vertex 4	label	= "C"	hard	change no. of hydrogen atoms on 2
	activation	= 0	soft	
vertex 5	label	= "C"	hard	change activation on 2
	activation	= 0	soft	
vertex 6	label	= "C"	hard	change no. of hydrogen atoms on 1
	activation	= 0	soft	
edge 1-2	multiplicity	= 1	soft	change multiplicity on 1
edge 2-3	multiplicity	= 1	soft	
edge 3-4	multiplicity	= 1	soft	change multiplicity on -2
edge 4-5	to be broken	= "yes"	hard	
	multiplicity	= 2	soft	change multiplicity on -1
edge 5-6	multiplicity	= 1	soft	
edge 1-6	to be broken	= "yes"	hard	
	multiplicity	= 1	soft	

<sup>a</sup> Check of condition consists of the identifier of the value to be checked, the symbol of the relation, the number or symbol (in quotation marks) which is to be compared with this value, and belonging to kernel (hard) or environment (soft). "Change on" means increasing (for positive value) or decreasing (for negative value) of the parameter on a specified number of points.

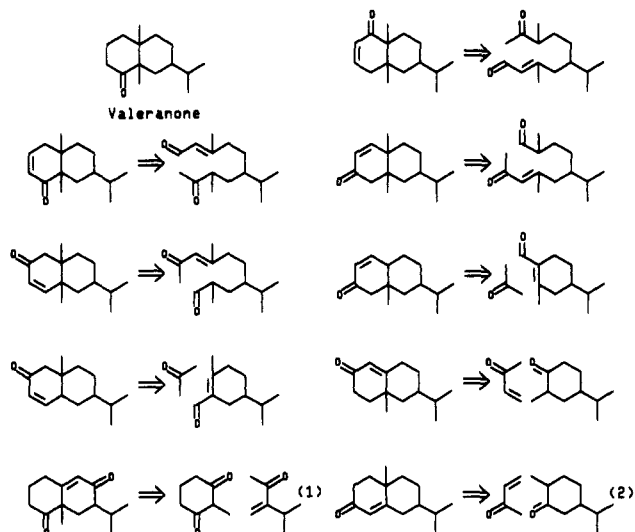


Figure 5. Sample imaginary precursors proposed by the STORM program for the synthesis of valeranone.

ragmentation, have been declared as soft (dashed lines in Figure 6). They are absent in the target triquinane 3 but appear in the imaginary target 4, yielding the imaginary precursor 5. The total number of imaginary solutions, obtained from this rule and the set of marked bonds in 3, is 26, but almost all the imaginary targets are very strained. The actual synthesis looks the best.

**Soft Atoms.** This mode implies that one or several vertices of the rule may remain unmatched within the target structure. In other words, these atoms were removed and disappeared after the application of the corresponding key step. Two closely-related examples of such an approach have been demonstrated in the descriptions of quadron total syntheses.<sup>18,19</sup> Their key steps are outlined in Figure 7. Atom 5

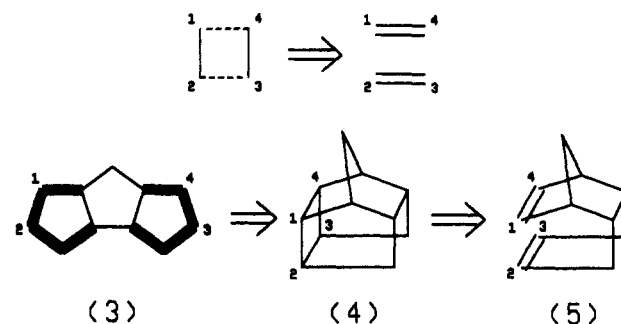


Figure 6. Rule for the retro-[2+2]-cycloaddition and reproduction of the key step of triquinane synthesis published in ref 17.

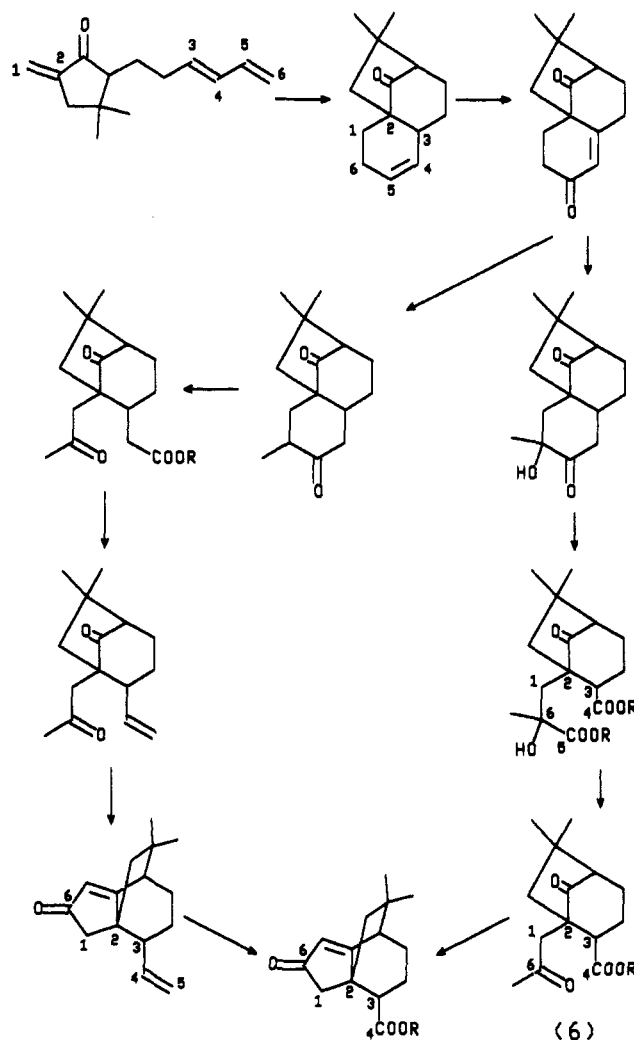


Figure 7. Key steps of the total syntheses of quadron published in refs 18 and 19.

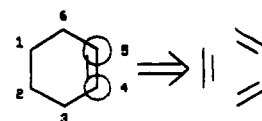
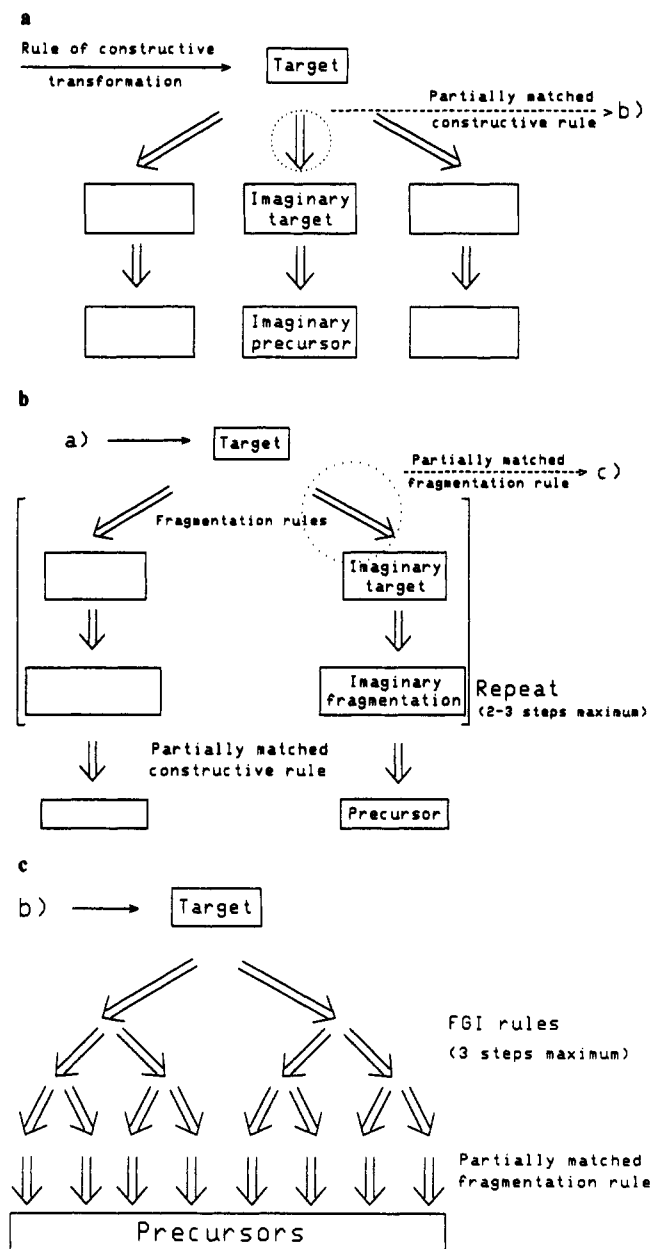


Figure 8. Definition of the soft atoms in the rule of retro-[4+2]-cycloaddition.

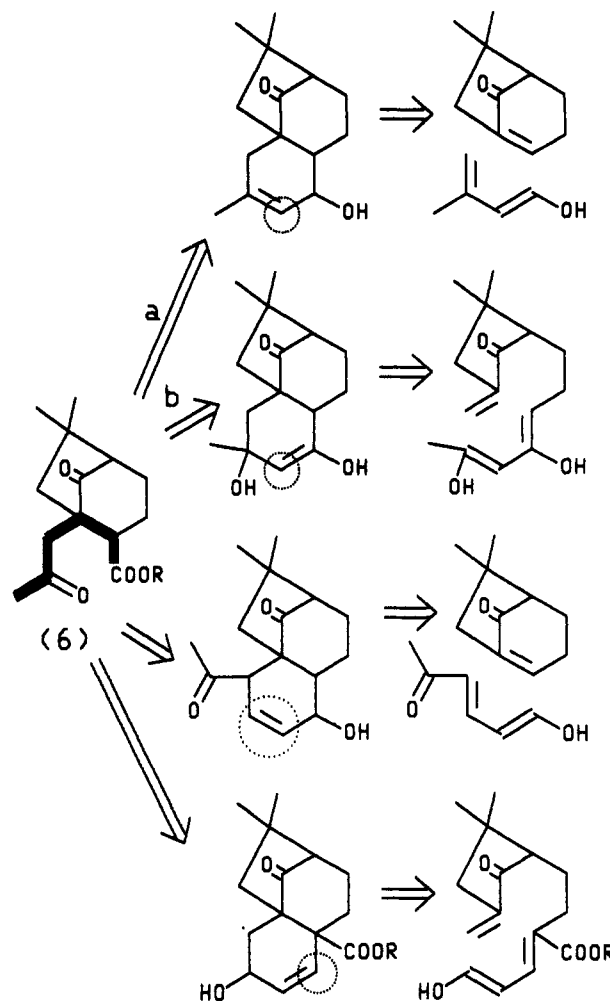
takes part in the [2+4]-cycloaddition and then disappears after fragmentations. Therefore two atoms in the rule (atom 5 and atom 4 which is symmetrical to it) should be declared as soft (Figure 8). It means that they can be omitted during the matching procedure and should be added during the search for an imaginary target. The flow chart of the top-level search is presented in Figure 9a. The application of the rule to the marked set of bonds in target 6 gives eight solutions. Figure



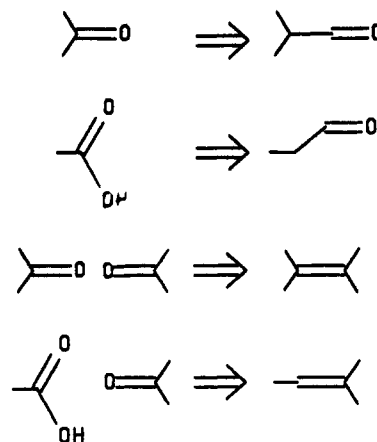
**Figure 9.** (a) Application of constructive transformations at the top level of hierarchical search. (b) Application of the fragmentation transformations at the middle level of hierarchical search. (c) Application of the FGI transformations at the bottom level of hierarchical search.

10 presents some of them where new atoms appear in the imaginary target. The added atoms are marked by dashed circles.

Subsequent analysis of imaginary solutions was carried out according to Figure 9b. Its aim was to find sequences of fragmentation rules resulting in the addition of the unmatched part to the target structure. In this example four simple fragmentation rules were used: two of them are retro-ozonolysis and other two are retro-decarbonylation (Figure 11). The search for sequences of the necessary environment was successful only for the a and b pathways from Figure 10. These sequences are presented in Figure 12. The two pathways leading to b match closely the actual synthesis (see Figure 7). The other two pathways lead to an anti-Bredt structure and should be rejected. In general, the bottom-level environment search is a narrow version of the long-range search algorithm (see Figure 1): at the last step in all sequences of the former technique only one partially matched rule can be applied. The



**Figure 10.** Sample imaginary precursors proposed by the STORM program for the synthesis of quadron precursor (6).



**Figure 11.** Sample retrofragmentation rules used for the Environment Search.

flow chart of the bottom-level search for the environment of each fragmentation rule by the application of FGI transformations is presented in Figure 9c, but it is out of the scope of the present paper.

The above-discussed example allows one to suggest the most feasible ways to use the Soft Atoms mode. If only one terminal vertex of the rule is declared as soft, it can be added in one step according to one of the rules listed in Figure 13. If one or two soft vertices is contained in the body of the rule, then the application of two or even more fragmentation transformations may be necessary.

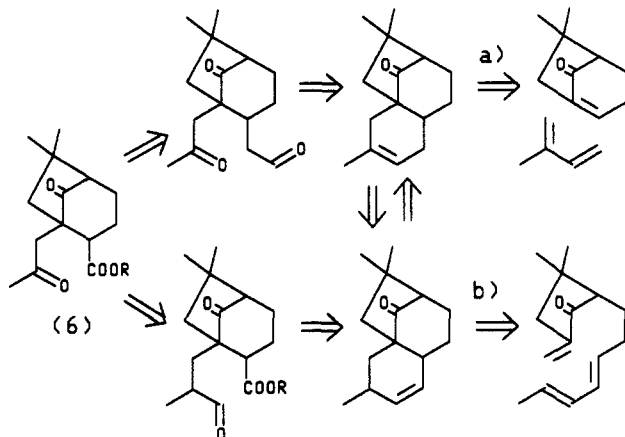


Figure 12. Sample sequences of fragmentation rules which reconstruct the environment for the following retro-[4 + 2]-cycloaddition.

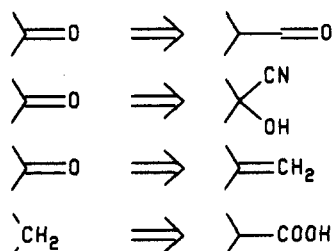


Figure 13. Fragmentation rules presently used in the STORM program.

### COMPUTER IMPLEMENTATION

The program STORM (Soft Transformations for Organic Reactions and Mechanisms) is implemented on an IBM PC-compatible personal computer, running MS DOS 3.30. It is written in PDC Prolog, except for drawing routines written in C language. The program uses 450–500K of main memory. The total number of structures in the tree is restricted only by the capacity of external storage (hard disk) and can reach several thousands. The program has a well-adjusted user interface.

### CONCLUSION

Although one-level search allows us to get the reasonable pathways immediately, it is not goal directed and depends on the completeness of the FGI/fragmentations knowledge base. It requires essential enumeration and therefore is time-consuming. The generation of imaginary structures at the top level of the hierarchy depends only on the assignment of the kernel and the environment for transformation. It is not restricted by low-level transformations and is rather fast. The list of top-level imaginary solutions actually represents all the possible ways of applying the given constructive reaction(s).

As a result, the hierarchical search allows one to focus one's attention at the most interesting and promising branches of the synthesis tree and to see clearly the goal of subsequent detailed search. At present, the decision whether to carry out the detailed search or not depends on the user, after the analysis of imaginary precursors. It should be noted that the top-level search for imaginary targets allows possible fragmentation to hide. Therefore, at this level the most crucial point (that is, the plan of skeleton assembly) can be evaluated according to well-grounded considerations based on target simplification and convergence.

### REFERENCES AND NOTES

- (1) Barone, R.; Chanon, M. *Computer-Aided Organic Synthesis*. In *Computer Aids to Chemistry*; Vernin, G., Chanon, M., Eds.; Ellis Horwood: Chichester, 1986; pp 19–102.
- (2) Zefirov, N. S.; Gordeeva, E. V. *Computer-Assisted Synthesis*. *Usp. Khim. (Russ. Chem. Rev.)* **1987**, *56*, 1002–1014.
- (3) Corey, E. J.; Howe, W. J.; Pensak, D. A. *Computer-Assisted Synthetic Analysis*. Methods for Machine Generation of Synthetic Intermediates Involving Multistep Look-Ahead. *J. Am. Chem. Soc.* **1974**, *96*, 7724–7737.
- (4) Corey, E. J.; Long, A. K. *Computer-Assisted Synthetic Analysis*. Performance of Long-Range Search Strategies for Stereoselective Olefin Synthesis. *J. Org. Chem.* **1978**, *43*, 2208–2216.
- (5) Corey, E. J.; Johnson, A. P.; Long, A. K. *Computer-Assisted Synthetic Analysis*. Techniques for Efficient Long-Range Retrosynthetic Searches Applied to the Robinson Annulation Process. *J. Org. Chem.* **1980**, *45*, 2051–2057.
- (6) Hendrickson, J. B. *Systematic Synthesis Design*. 6. Yield Analysis and Convergency. *J. Am. Chem. Soc.* **1977**, *99*, 5439–5450.
- (7) Hendrickson, J. B. *Approaching the Logic of Synthesis Design*. *Acc. Chem. Res.* **1986**, *19*, 274–281.
- (8) (a) Hendrickson, J. B.; Grier, D. L.; Toczko, A. G. A Logic-Based Program for Synthesis Design. *J. Am. Chem. Soc.* **1985**, *107*, 5228–5238. (b) Hendrickson, J. B.; Toczko, A. G. SYNGEN Program for Synthesis Design: Basic Computing Techniques. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 137–145.
- (9) Baumer, L.; Sala, G.; Sello, G. *Organic Synthesis Planning: An Algorithm for Selection Strategic Bond Forming Sequences*. *Tetrahedron* **1989**, *45*, 2665–2676.
- (10) Corey, E. J.; Jorgensen, W. L. *Computer-Assisted Synthetic Analysis*. Synthetic Strategies Based on Appendages and the Use of Reconnective Transforms. *J. Am. Chem. Soc.* **1976**, *98*, 189–203.
- (11) Hendrickson, J. B. *Fragmentations and Rearrangements in Organic Synthesis*. *J. Am. Chem. Soc.* **1986**, *108*, 6748–6756.
- (12) (a) Lushnikov, D. E. Candidate Dissertation, Moscow, 1991. (b) Gordeeva, E. V.; Lushnikov, D. E.; Zefirov, N. S. COMPASS Program: Combination of Empirical Rules and Combinatorial Methods for Planning of Organic Synthesis. *Tetrahedron Comput. Methodol.* **1990**, *3* (6).
- (13) (a) Burke, S. D.; Murtiashaw, C. W.; Saunders, J. O.; Dike, M. S. A Novel and Efficient Entry to (±)-Quadrone. *J. Am. Chem. Soc.* **1982**, *104*, 872–874. (b) Burke, S. D.; Murtiashaw, C. W.; Oplinger, J. A. *Synthetic Studies Directed at (±)-Quadrone*. A New And Game. *Tetrahedron Lett.* **1983**, *24*, 2949–2952.
- (14) (a) Dugundji, J.; Ugi, I. An Algebraic Model of Constitutional Chemistry as a Basis for Chemical Computer Programs. *Top. Curr. Chem.* **1973**, *39*, 19–67. (b) Blair, J.; Gasteiger, J.; Gillespie, C.; Gillespie, P. D.; Ugi, I. Representation of the Constitutional and Stereochemical Features of Chemical Systems in the Computer-Assisted Design of Synthesis. *Tetrahedron* **1974**, *30*, 1845–1859.
- (15) Zefirov, N. S. An Approach to Systematization and Design of Organic Reactions. *Acc. Chem. Res.* **1987**, *20*, 237–243.
- (16) Mehta, G.; Azario, P.; Barone, R.; Chanon, M. How to Search Original Key Steps in a Synthesis of Complex Structures by Using a Microcomputer as a Chemical Pocket-Like Calculator. *Tetrahedron* **1989**, *45*, 1985–1994.
- (17) Mehta, G.; Murthy, A. N.; Reddy, D. S.; Reddy, A. V. A General Approach to Linearly Fused Triquinane Natural Products. Total Synthesis of (±)-Hirsutene, (±)-Coriolin, and (±)-Caphellene. *J. Am. Chem. Soc.* **1986**, *108*, 3443–3452.
- (18) Schlessinger, R. H.; Wood, J. L.; Poss, A. J.; Nugent, R. A.; Parsons, W. H. An Intramolecular Diels–Alder Route to Racemic Quadrone. *J. Org. Chem.* **1983**, *48*, 1146–1147.
- (19) Dewanckele, J. M.; Zutterman, F.; Vandewalle, M. Sesquiterpene Lactones. A Total Synthesis of (±)-Quadrone. *Tetrahedron* **1983**, *39*, 3235–3244.