Starting Material Oriented Retrosynthetic Analysis in the LHASA Program. 1. General Description

A. Peter Johnson,* Chris Marshall, and Philip N. Judson

The Maxwell Institute for Computer Applications in the Molecular Sciences, University of Leeds, Leeds LS2 9JT, U.K.

Received January 16, 1992

This paper introduces and gives an overview of the starting material oriented analysis package in LHASA.[6-9] The separate modules of the package will be described in detail in following papers.[1-4]

INTRODUCTION

This paper is the first in a series entitled Starting Material Oriented Retrosynthetic Analysis in the LHASA Program.[1-4] The papers describe a strategy incorporated into the LHASA retrosynthetic analysis program which enables the program to select appropriate starting materials (SM) for a target structure and to direct retrosynthetic analysis toward that structure. The later papers will describe the modules of the package in more detail and the algorithms used.

When devising a route for the synthesis of a complex organic molecule, a chemist may be guided to select a particular SM by recognition of structural similarities between the SM and the target. This has been called an ASSOCIATIVE STRATEGY,[5] or simply a STARTING MATERIAL STRATEGY—the name currently used in the LHASA system.

In addition to considering structural similarities between a potential SM and a target, a chemist also uses knowledge about chemical reactions to decide if the SM is a good one. Account is taken of the nature of functional groups present in the SM and their positions relative to carbon–carbon bond forming sites, and the position and chirality of any asymmetric centers. In general, if the problem is not trivial, the synthetic path from the SM to the target will involve reactions at several sites, and if there are two or more structures in the SM they will be joined together at one or more specific positions in each. For an elegant synthesis, these steps must be ordered to accomplish the overall conversion cheaply and efficiently, avoiding the excessive use of protecting groups or functional group interconversion reactions.

LHASA

LHASA (logic and heuristics applied to synthetic analysis)[6-10] is a suite of programs designed to help a chemist planning the synthesis of a complex organic molecule. It is an expert system in which an inference engine proposes solutions to synthesis problems on the basis of information contained in a knowledge base.

All parts of the LHASA program are designed to maximize the benefits of interaction with the user. As the analysis of the problem proceeds, proposed retrosynthetic steps are displayed to the user. At any stage the user can redirect the analysis (for example to concentrate on one route rather than another) or specify a change of strategy for the target structure or a precursor that is of particular interest. The growing tree of synthetic routes to the target can be accessed to monitor progress and pruned if necessary.

THE STARTING MATERIAL STRATEGY

At the start of a retrosynthetic analysis using LHASA, the chemist draws in the target compound and then makes a selection from the LHASA menu shown in Figure 1. The first five menu options correspond to the strategies listed in Table I, which were recently described by Corey and Cheng.[11,12] If the sixth option "LHASA suggestions" is chosen, the inference engine makes its own recommendations on the choice of strategy. Options that LHASA considers unsuitable for the target compound are displayed in parentheses. The implementation of the fifth option in the LHASA menu, the STARTING MATERIAL STRATEGY, is the subject of this series of papers.

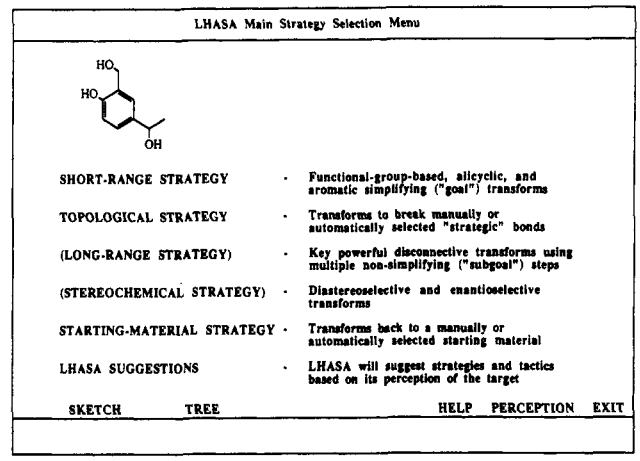


Figure 1. LHASA strategy selection menu.

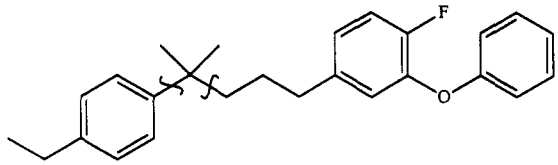
When choosing a starting material, a chemist uses knowledge of chemical reactions and looks for skeletal similarity between the target and starting material. The locations of functional groups relative to sites at which carbon–carbon bonds are formed or broken, the substitution patterns on aromatic systems, and the positions and chirality of asymmetric centers are all taken into account. There may be more than one starting material to be joined together, and some orderings of the reaction steps in the synthetic sequence will be better than others.

Some work on the recognition of starting materials in computer-based retrosynthetic analysis has been reported in the past. Gelernter et al.[14] described the comparison of structures generated by the SYNCHEM program with a catalog of starting materials to detect completion of a retrosynthetic sequence, but the retrosynthetic route was not guided by information about the SM. Wipke et al.[15] described how a numerical estimate of the similarity between a target and SM could be used to select a SM but not how it could be used to direct the retrosynthesis. Hendrickson et al.[16] described the division of the target into several SMs by breaking bonds consistent with a convergent synthesis, but this method cannot

0095-2338/92/1632-0411$03.00/0

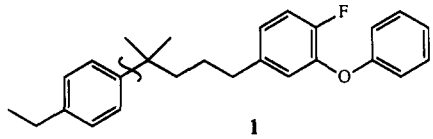
Table I. Five Major Synthesis Strategies

strategy	name in LHASA	description
FUNCTIONAL GROUP BASED	SHORT-RANGE	takes opportunistic advantage of the presence of functional groups in the target structure to favor particular retrosynthetic simplifications
TOPOLOGICAL	TOPOLOGICAL	recognizes the breaking of certain bonds in a target structure as critical to the retrosynthetic creation of chemically simpler precursors and seeks methods to break those bonds
TRANSFORM BASED	LONG-RANGE	seeks to apply a particularly powerful transform (e.g., the Robinson annelation) by using multiple retrosynthetic steps to introduce the necessary chemical features for that transform into the target structure
STEREOCHEMICAL	STEREOCHEMICAL	finds ways to create precursors that are stereochemically simpler than the target structure
STRUCTURE GOAL	STARTING MATERIAL	recognizes the suitability of a particular starting material (or intermediate) for the synthesis of a target structure and seeks a good retrosynthetic route between them



1

Figure 2. Likely first choice of strategic bonds for target **1** using a topological analysis.



1

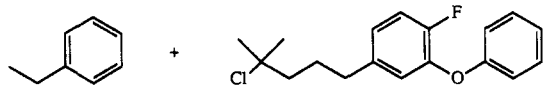
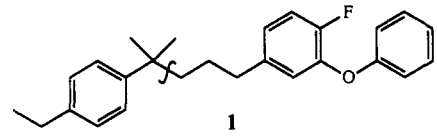


Figure 3. Possible first disconnection of target **1** involving one of the strategic bonds.

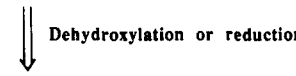
take account of rearrangements and degradative syntheses, for example. The STARTING MATERIAL STRATEGY in LHASA is unique in being directed by information about the relationship between the target and a selected SM.

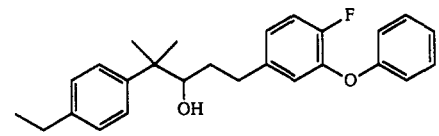
The synthesis of the target structure **1** in Figure 2 illustrates an application of the strategy. The target lacks any obviously useful functionality for a functional group based strategy. Using a topological strategy, a chemist might start by looking for ways to break the bonds marked in Figure 2, leading to first retrosynthetic steps such as those in Figures 3 and 4. However, the patentees[13] of one synthetic route had access to the aldehyde **2** shown in Figure 5. The availability of this aldehyde makes alternative synthetic routes attractive, such as the published sequence shown in Figure 6. The retrosynthetic breaking of the appropriate bond is worthwhile, even though this first necessitates a functional group addition to set up the retron for the Wittig reaction. (A retron is a substructure in a product that is characteristic of a reaction.)

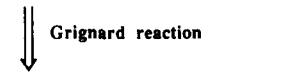
LHASA proposes this route automatically if aldehyde **2** is chosen as SM. The LHASA topological strategy rules for acyclic bonds also recognize this bond as strategic, along with several others, and the program would eventually look for ways to break it in an acyclic topological analysis unless otherwise directed by the user, but only an analysis directed toward a specific SM can produce the result so succinctly.



1







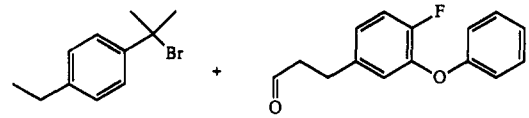
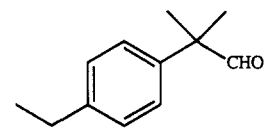


Figure 4. Alternative first disconnection of target **1** involving the other strategic bond.



2

Figure 5. Aldehyde **2** available as a starting material for the synthesis of target **1**.

The LHASA SM-oriented strategy package interfaces with the user in a similar way to other LHASA features. The target structure is drawn in in the usual way, and the SM-oriented strategy is selected from the strategy menu shown in Figure 1. A starting material is specified by the user or selected from a SM pool. Modules map the SM onto the target, assess the feasibility of the chemical modifications implied by the mapping, order the steps into the sequence most likely to be favored from a chemical point of view, and then attempt to perform the sequence using the LHASA knowledge base, as follows.

1. Selection of Appropriate Starting Materials. The menu shown in Figure 7 allows the user to choose to draw in a specific SM or to instruct the program to refer to sets of potential SM (SM pools) suiting different strategic considerations—a chiral pool for enantioselective synthesis of targets that are stereochemically complex, a pool for synthesis of isotopically-labeled compounds, an aromatic pool for targets

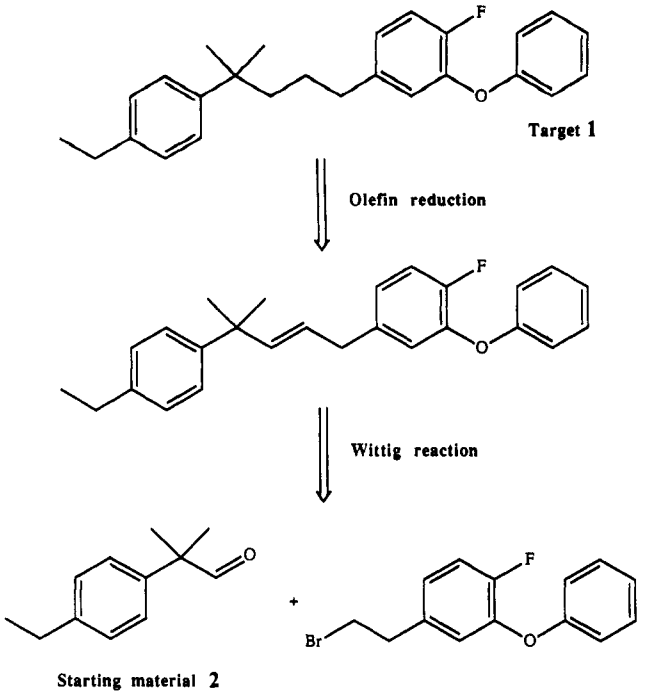


Figure 6. Starting material directed retrosynthetic sequence from target **1** to starting material **2**.

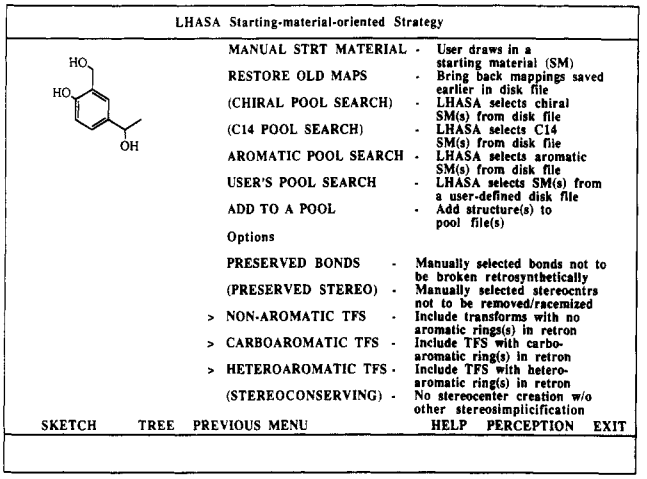
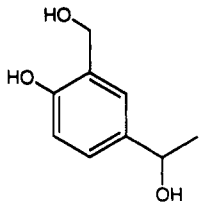


Figure 7. LHASA starting-material oriented strategy menu.



3

Figure 8. Target 3.

which contain benzenoid aromatic rings, or any pool created by the user. For example, for the synthesis of the bis-(hydroxyalkyl)phenol **3** shown in Figure 8, if the aromatic pool is selected, LHASA lists the compounds shown in Figure 9 which it has chosen from the currently available pool of aromatic starting materials. The user can then select structures one at a time from the list for processing.

Research is currently being carried out into the improvement of performance when large SM pools are used (i.e., pools containing thousands of compounds).

2. Recognition of Where the Starting Material Best Maps onto the Target. A chosen SM may map onto the target in more than one way. The program attaches a rating—a measure of the goodness of mapping—to each map and displays the maps to the user in order of priority.

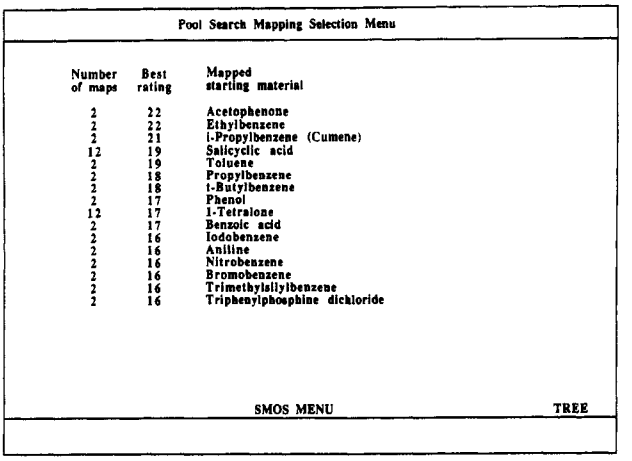


Figure 9. LHASA menu for selection from a typical starting material pool.

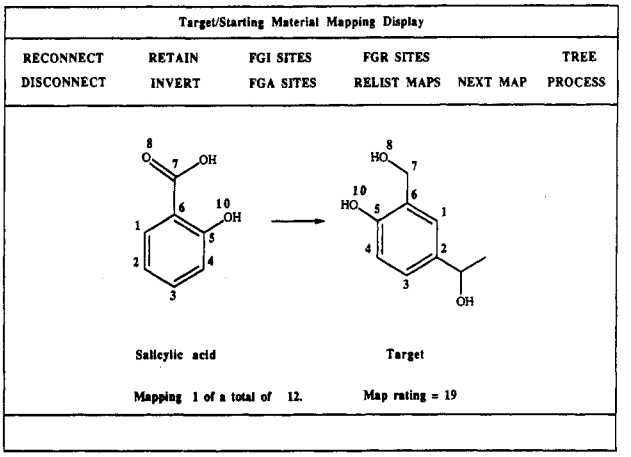


Figure 10. Highest-rated mapping of salicylic acid onto target **3**.

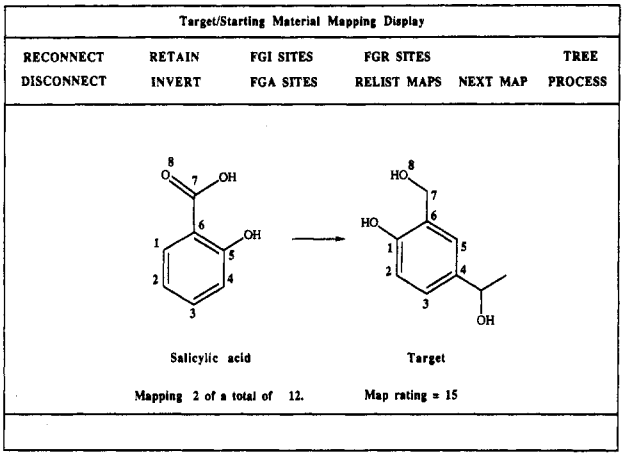


Figure 11. Second mapping of salicylic acid onto target **3**.

For example, the best mapping found for salicylic acid as SM for target **3** is shown in Figure 10. However, there are 11 less attractive possibilities, two of which are shown in Figures 11 and 12. The only mismatches implied by the mapping in Figure 10 are the absence of a carbon–carbon bond in the SM where there is one attached to atom 2 in the target, and the presence of a carboxylic acid on atom 7 in the SM in place of an alcohol in the target. In Figure 11, in addition to these mismatches, there is a hydroxyl group on atom 5 in the SM where there is none in the target and a hydroxyl group on atom 1 in the target where there is none in the SM. In Figure

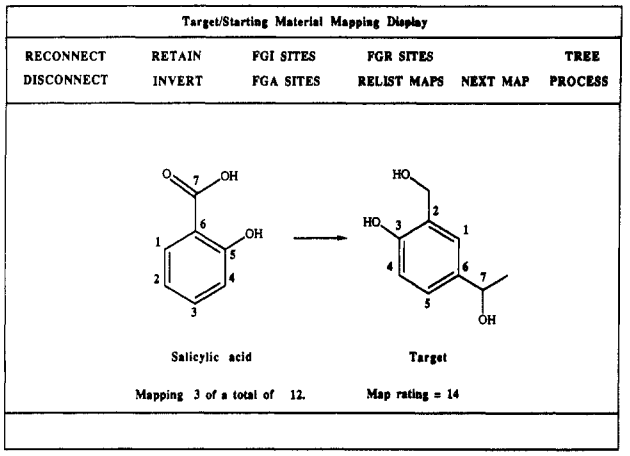


Figure 12. Third mapping of salicylic acid onto target 3.

12, the side chains have been mapped differently. As a result, retrosynthetically, carbon–carbon bonds to atoms 2 and 7 must be broken, an alcohol must be removed from atom 3, an alcohol must be introduced on atom 5, and the functional group must be changed on atom 7.

3. Perception of the Retrosynthetic Goals—the Alterations Required To Convert the Target into the Starting Material. The initial mapping establishes atom-to-atom correspondences between parts of the target and SM. The next step taken by the program is to recognize the changes to functional groups and bonding implied by the chosen mapping. For example, the mapping in Figure 10 implies that retrosynthetic conversion of the target into the SM requires the breaking of the bond between the side chain and the ring and conversion of the alcohol group to a carboxyl group.

4. Selection of the Goal To Be Attempted. Except in trivial cases, the retrosynthetic conversion of a target to a SM requires structural modifications at several sites. The modifications can rarely be planned completely independently, either because functionalities at one or more of the sites are sensitive to the conditions needed to modify other sites or because of direct chemical interaction between sites. For example, in the complete sequence shown in Figure 13, the side chain must be removed, in a retrosynthetic sense, before the alcohol group is converted to a carboxylic acid, because the presence of the deactivating carboxyl group would hinder the introduction (in the synthetic direction) of the side chain by an electrophilic substitution reaction.

Strategic rules are therefore used in LHASA to determine which required change to the structure should be carried out first. These rules describe general principles of chemistry rather than specific reactions. They may be relevant to all kinds of target compounds or just to particular classes of compounds. An example of a strategy rule that is relevant to all targets is the principle of convergent synthesis, the objective of which is to divide the compound into similarly sized fragments. On the other hand, the LHASA aromatic electrophilic substitution rules apply specifically to aromatic target structures. (These rules are nevertheless general to electrophilic substitution transforms—they are not specific to particular electrophiles.)

A priority is determined for each goal, based on the sums of the weightings for each strategy rule applied to it. The goal with the highest priority is chosen.

5. Selection of Transforms from the LHASA Knowledge Base To Perform the Goal Retrosynthetic Conversion. Once a goal has been choosen, LHASA refers to its knowledge base for ways to carry out the required structural change. Usually, it is

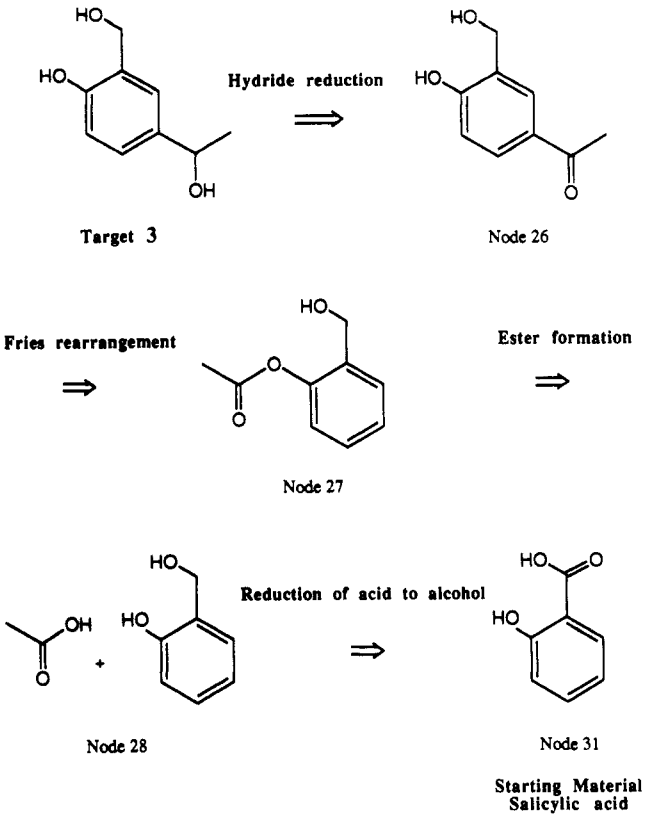


Figure 13. Retrosynthesis from target 3 to salicylic acid.

impossible to achieve a structural change in one step by applying a single transform. LHASA automatically applies other *subgoal* transforms as necessary, calling modules that are used throughout the LHASA program. For example, in the sequence shown in Figure 13, the Fries rearrangement is chosen to achieve the goal of removing the hydroxyethyl side chain, but first, a reduction transform has to be applied to set up the retron needed for the Fries rearrangement.

There is usually more than one way to achieve the goal structural modification. All alternatives are considered and the routes that are generated are added to the retrosynthetic tree. Once this task has been completed, the most promising node on the whole of the retrosynthetic tree is selected for further processing, taking account of the number of steps to the node from the target, the LHASA transform ratings for these steps, and the estimated remaining "chemical distance" from the node to the SM.[2,3] This node becomes the new target for analysis and processing recommences from step three above.

MANAGING THE GROWING RETROSYNTHETIC TREE

The primary objective of any strategy in retrosynthetic analysis is to limit the search space to a manageable size. The selection of a specific SM is not sufficient in itself to meet this objective. Three key techniques mentioned above prevent the retrosynthetic tree generated by the LHASA SM strategy from becoming too large. Firstly, goal structural changes are identified and attempted in the order most likely to lead to a complete, successful sequence, on the basis of chemical considerations. Secondly, transforms are not applied arbitrarily—only those that can achieve identified goal structural changes are attempted. (Throughout LHASA, subgoal transforms are called to set up required retrons only if no goal transforms can be applied directly, or at the user's request.) Thirdly, after each goal structural change has been achieved, all the nodes on the growing tree are reassessed and processing

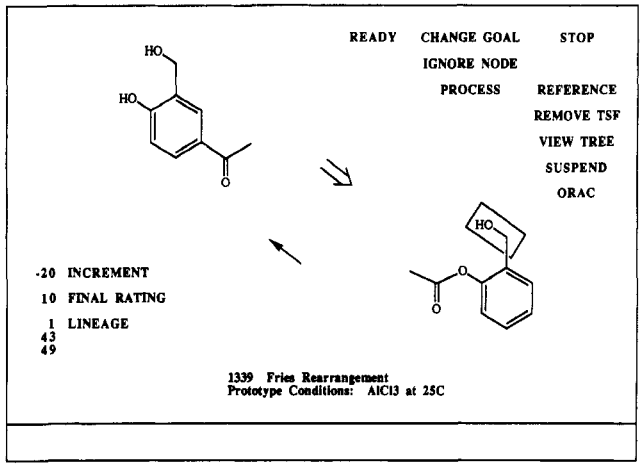


Figure 14. Display of a LHASA transform.

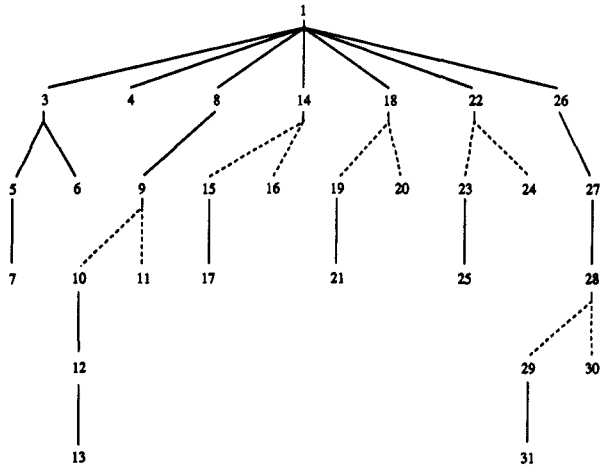


Figure 15. Part of the tree from the retrosynthetic analysis of target 3.

continues from the node judged likely to lead most quickly to an efficient complete sequence.

As LHASA selects transforms, they are displayed to the user, who can redirect the program in order to restrict the search further, or to expand it, as appropriate. For example, Figure 14 shows the display of the Fries rearrangement. The box around the benzylic alcohol group indicates to the user that protection may be necessary. A separate program, PROTECT, can be called to analyze a complete sequence and to advise on suitable choice of protecting groups.

FURTHER DISCUSSION OF EXAMPLES

Figure 15 shows part of a retrosynthetic tree for the bis-(hydroxyalkyl)phenol 3 mentioned above. (In the LHASA tree, if a retrosynthetic step leads to a node containing more than one structural fragment, the fragments are also shown as separate nodes connected to the main node by dotted lines. For example, node 28 is 2-(hydroxymethyl)phenol and acetic acid, which are also shown as nodes 29 and 30, respectively.)

Nodes 3–7 were generated when acetophenone was selected as the SM. The route from the target, node 1, to acetophenone, node 7, is shown in Figure 16. LHASA perceived retrosynthetic breaking of the bond between the hydroxymethyl group and the ring as the first priority and found a transform that could achieve this goal in a single step, leading to node 3; 4-(1-hydroxyethyl)phenol, as shown in the first step of the sequence in Figure 16.

Another transform was found that could achieve the goal of breaking the carbon–carbon bond—the Grignard type

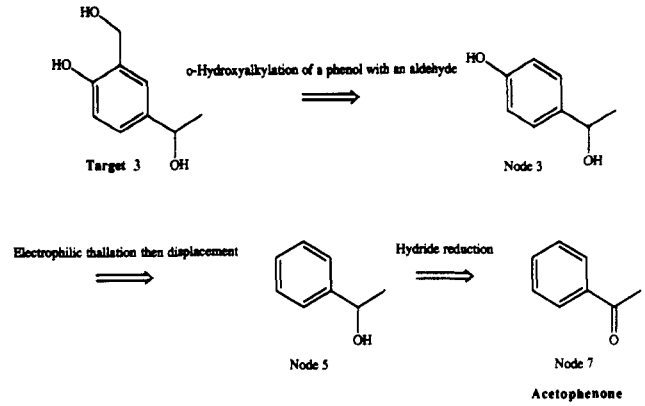


Figure 16. Retrosynthesis from target 3 to acetophenone.

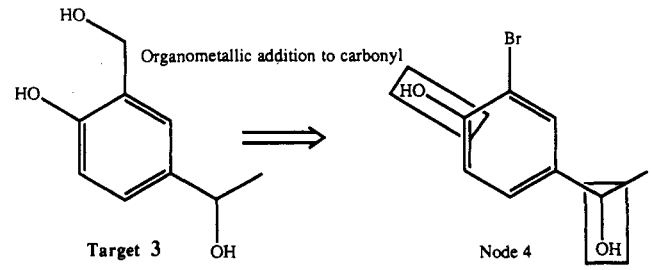


Figure 17. Retrosynthetic removal of the hydroxymethyl group from target 3 by a route involving an organometallic addition transform.

reaction shown in Figure 17—adding node 4 to the tree. However, a second step (the bromination transform) would be needed in order to complete the removal of functionality from this ring position. (The two-step sequence is illustrated in the analogous removal of the hydroxyethyl group in Figure 19.) LHASA recognized that the route it had already found achieved the greater simplification in a single step, and went on to process node 3. This time the goal was to remove the phenolic hydroxyl group. Again, one way of achieving the goal in a single step was found, leading to node 5, 1-phenylethanol. LHASA applied a functional group interconversion transform to node 5, completing the sequence in Figure 16. Processing was stopped manually at this point.

The remaining nodes on the tree were created when salicylic acid was chosen as the SM. The first goal was the removal of the hydroxyethyl group. This time LHASA found no transforms that could operate directly on the target to achieve the goal in a single step, but it successfully applied five transforms by first carrying out subgoal steps to create the retrons required by the goal transforms. These goal transforms created nodes 13, 17, 21, 25, and 29 which are all 2-(hydroxymethyl)phenol. The route to node 13, based on the goal of applying the nitration transform, is shown in Figure 18. The routes to nodes 17 and 25 are minor variations of the route to node 21, shown in Figure 19, in that they use chlorine and iodine in place of bromine for the goal halogenation transform. The route to node 29, directed toward the Fries rearrangement, is the one shown in Figure 13.

LHASA went on to select node 29 as the most promising for further processing, and this gave the SM, node 31. The route to node 13 was less favored because it is longer than the route to node 29, and it involves less promising chemistry. The routes to nodes 17, 21, and 25 are shorter, but the individual steps are not rated very highly—protection is needed in the organometallic addition step and LHASA considers that it would be difficult to ensure monohalogenation at the required site. If the user had requested continuation of the analysis, LHASA would have applied the transform used for node 29 to the

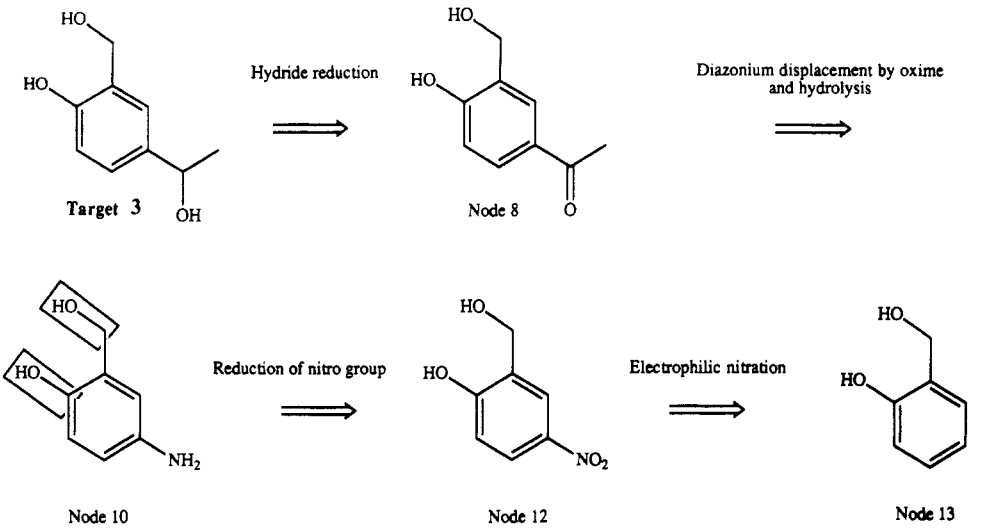


Figure 18. Retrosynthetic removal of the hydroxyethyl group from target **3** by a route involving the nitration transform.

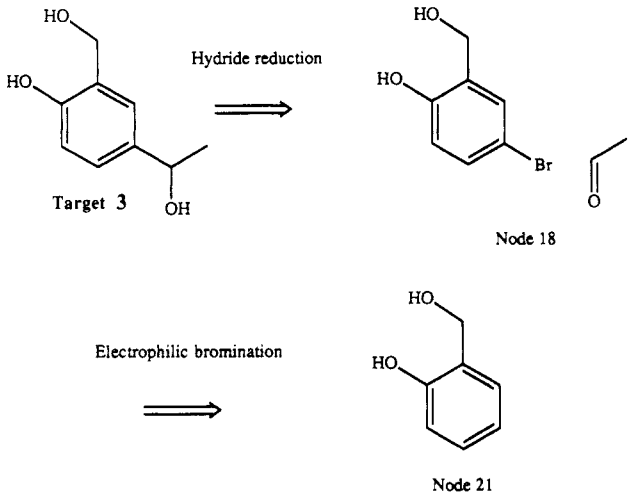


Figure 19. Retrosynthetic removal of the hydroxyethyl group from target **3** by a route involving the halogenation transform.

other, equivalent nodes, as well as searching for new, less direct routes from the target and other intermediates.

CONCLUSION

This paper describes the overall strategy of directing retrosynthetic analysis toward suitable SMs which has been implemented in the LHASA program. The starting material strategy modules can select potential SMs for the synthesis of a target structure drawn by the user and find which portion of the target is best fitted by each SM. The modules find the most similar parts of the carbon skeletons of the structures, and then the comparison is extended to the complete structures. One module provides a chemically-reasonable estimation of the similarity of the structures.

The SM can be chosen by the user or selected automatically from a pool of structures. Work is currently in hand to permit the use of large pools such as one containing all of the structures in the *Fine Chemicals Directory*.[17]

A list is produced of the chemical steps (goals) required to convert the target structure retrosynthetically to the SM. Meta-level knowledge of chemical strategy is used to select the most important goal, and the standard LHASA knowledge base is used to achieve the goal. An A*-like algorithm is used to evaluate the growing retrosynthetic tree and to decide which is the most promising node for further processing, and the most important goal is again selected. This process is repeated until the structure of the SM is generated.

The starting material strategy module is sufficiently fast to allow convenient interactive use on a machine such as a VAX 6310 computer. Finding the best match between the target and the SM typically takes 1–10 s of elapsed time. The time taken for a complete analysis depends strongly on the choice of target and SM and on the time taken by the user to monitor or guide the analysis. As an example, the analysis shown in Figure 15 was completed in approximately 30 min of elapsed time.

ACKNOWLEDGMENT

We extend thanks to all those who have contributed to the work described in this paper, particularly to E. J. Corey, A. K. Long, and S. Rubenstein for useful discussions. This work has been supported by the Science and Engineering Research Council, the Department of Education and Science, Imperial Chemical Industries, the Wolfson Foundation, and the Maxwell Foundation.

REFERENCES AND NOTES

(1) Johnson, A. P.; Marshall, C. Starting Material Oriented Retrosynthetic Analysis in the LHASA Program. 2. Mapping the SM and Target Structures. *J. Chem. Inf. Comput. Sci.*, second of three papers in this issue.

(2) Johnson, A. P.; Marshall, C. Starting Material Oriented Retrosynthetic Analysis in the LHASA Program. 3. Heuristic Estimation of Synthetic Proximity. *J. Chem. Inf. Comput. Sci.*, third of three papers in this issue.

(3) Johnson, A. P.; Marshall, C.; Hopkinson, G. A.; Judson, P. N. Starting Material Oriented Retrosynthetic Analysis in the LHASA Program. 4. Selecting Good Retrosynthetic Routes. *J. Chem. Inf. Comput. Sci.*, in press.

(4) Johnson, A. P.; Marshall, C.; Su, L.; Judson, P. N. Starting Material Oriented Retrosynthetic Analysis in the LHASA Program. 5. Searching a Pool of Starting Materials. *J. Chem. Inf. Comput. Sci.*, in press.

(5) Corey, E. J.; Wipke, W. T. Computer-Assisted Design of Complex Organic Syntheses. *Science* **1969**, *166*, 178–92.

(6) Corey, E. J.; Wipke, W. T.; Cramer, R. D. III; Howe, W. J. Computer-Assisted Synthetic Analysis. Facile Man-Machine Communication of Chemical Structure by Interactive Computer Graphics. *J. Am. Chem. Soc.* **1972**, *94*, 421–30.

(7) Corey, E. J.; Wipke, W. T.; Cramer, R. D. III; Howe, W. J. Techniques of Perception by a Computer of Synthetically Significant Structural Features in Complex Molecules. *J. Am. Chem. Soc.* **1972**, *94*, 431–9.

(8) Corey, E. J.; Wipke, W. T.; Cramer, R. D. III; Howe, W. J. Computer-Assisted Synthetic Analysis for Complex Molecules. Methods and Procedures for Machine Generation of Synthetic Intermediates. *J. Am. Chem. Soc.* **1972**, *94*, 440–59.

(9) Corey, E. J.; Petersson, G. An Algorithm for Machine Perception of Synthetically Significant Rings in Complex Cyclic Organic Molecules. *J. Am. Chem. Soc.* **1972**, *94*, 460–5.

(10) The LHASA system is available from A. K. Long, Harvard Chemistry Department, 12 Oxford Road, Cambridge, MA, and, in Europe, from

J. Langowski, LHASA UK, School of Chemistry, University of Leeds, Leeds LS2 9JT, England.
(11) Corey, E. J. The Logic of Chemical Synthesis: Multistep Synthesis of Complex Carbogenic Molecules. *Angew. Chem.* **1991**, *30*, 455–65.
(12) Corey, E. J.; Cheng, Xue-Min. In *The Logic of Chemical Synthesis*; John Wiley and Sons: New York, 1989.
(13) Mitsui Toatsu Chemicals Inc. Insecticidal and Acaricidal Arylalkanes. Jpn. Kokai Tokkyo Koho JP 60, 42,339 [85 42,339] (Cl C07C15/12) 6 Mar 1985. Appl. 83/149,646, 18 Aug 1983, 45 pp.
(14) Gelernter, H.; Sanders, A. F.; Larson, D. L.; Agarwal, K. K.; Boivie, R. H.; Spritzner, G. A.; Searleman, J. E. Empirical Explorations of SYNCHEM. *Science* **1977**, *197*, 1041–9.
(15) 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*, 71–80.
(16) Hendrickson, J. B.; Bernstein, Z.; Miller, T. M.; Parks, C.; Toczko, A. G. New Directions in the SYNGEN Program for Synthesis Design. In *Expert System Applications in Chemistry*; Hohne, B. A., Pierce, T. H., Eds.; ACS Symposium Series 408; American Chemical Society: Washington, DC, 1989; pp 62–81.
(17) The *Fine Chemicals Directory* is supplied by Molecular Design Ltd, 2132 Farallon Drive, San Leandro, CA 94577.