Similarity Searching in REACCS. A New Tool for the Synthetic Chemist

GUENTER GRETHE* and THOMAS E. MOOCK

Molecular Design Limited, 2132 Farallon Drive, San Leandro, California 94577

Received May 21, 1990

Similarity searching over large reaction databases has been implemented in REACCS very recently. This paper describes several applications in which the new technique has been used advantageously. These include reaction classification by reaction or reagent type and synthesis planning. The complementary nature of the new method adds a new tool to the armament of the synthetic chemist for effectively managing information needs.

INTRODUCTION

Computer-aided reaction retrieval over large databases has become an increasingly important tool for the synthetic chemist. New developments in available software have allowed chemists to ask very precise questions based on well-formulated substructure queries or to search for full structures. But often it is not the exact answers that interest scientists but information on similar reactions or transformations. While the principle of molecule similarity has been applied for some time and has been widely published,[1] reaction similarity has only recently been implemented.[2a,b]

The highly subjective nature of similarity becomes more complicated in the reaction domain because of differences in opinion of what constitutes similarity. Similarity could be based on reaction type (reaction centers), on the nature of the participating molecules, or on reaction conditions (Figure 1).

The similarity program developed for Molecular Design Ltd.'s Reaction Access System (REACCS) has been described by Moock and co-workers in detail.[2] REACCS determines the degree of similarity between two molecules or two reactions by calculating the amount of overlap between structural fragments. These fragments, or keys, for both molecules and reactions consists of atom-pair descriptors and are stored in inverted files. Some of the keys represent multiple occurrences of other fragments. The molecule sector consists of 933 keys; 230 keys are associated with the reaction centers and the immediate environment, up to two bonds away from the reaction bonds. The degree of similarity is defined by the Tanimoto coefficient,[9] expressed in its simplest form by

$$S = \frac{C}{Q + H - C}$$

where C = the sum of squares of weights of keys in common between the query and the reaction/molecule(s); Q = the sum of squares of weights of keys set by the query, and H = the sum of squares of weights of keys set by the reaction/molecule(s).

The weighting scheme is based on the uniqueness of a particular key in a large, representative database. The Theilheimer[21a] Database was used for assigning the reaction center key weights and the Fine Chemicals Directory Database of 64 000 commercially available structures was used for the molecule key weights. In contrast to most other similarity programs, REACCS does not create a sorted hitlist but uses user-supplied threshold similarity parameters for both the reaction and molecular components.

Adjustable parameters have the advantage that users can create an environment tailored to their needs. By changing the parameters, users can study the effects of absolute parameter values as well as the variations in the interrelationship between molecule and reaction center values.

Following are several examples highlighting different aspects of similarity searching and demonstrating the utility of this new tool for the synthetic chemist.

REACTION CLASSIFICATION

Many attempts have been made to solve the important but difficult problem of classifying reactions. The criteria which have been used to group reactions range from product family (e.g., the preparation of pyridines) to reaction types (C–C bond formation, C=O reduction, etc.). A systematic approach to synthetic methodology is important for didactic and indexing purposes. The availability of reaction data in electronic format added a new dimension to the problem:[10] with the rapid growth of reaction databases, it has become increasingly important to provide tools to the user for the extraction of subsets of manageable size. Subsets which address the same type of methodology are definitely valuable to synthetic chemists. Systematization of organic reactions also plays an important part in conjunction with synthesis planning programs, particularly with those based on nonempirical rules. The subject was recently addressed by Herger[11] and Hendrickson.[12]

Heterocycles. Classification of reactions is particularly difficult in the vast area of heterocyclic chemistry. One approach, which we describe here, classifies heterocyclic systems of a given ring size containing one or more heteroatoms of the same or different type at various positions according to their mode of formation. The common denominator for each class or group of reactions is the conceptual similarity in the principle of formation and not the actual reaction type. To demonstrate the validity of this approach, we searched a medium-sized database for those reactions which describe the cyclization of 2- and 3-atom fragments to five-membered heterocycles. We anticipated that classification could be done more effectively with similarity searching than by using the more stringent substructure search methodologies.

For the purpose of our study we selected the 1,3-dipolar cycloaddition reaction shown in Figure 2 as our search query. The heavy lines indicate reaction centers, i.e., bonds which are modified during the reaction. These reaction centers and the mapping of corresponding atoms, not shown in the figure, are automatically assigned by the program.[13] The database used was the Current Literature File[21c] (CLF) with selected reactions covering all areas of synthetic methodology.

Searches were run at different threshold combinations to evaluate the effects of molecule and reaction parameters. The results are shown in Figure 3. The numbers in the matrix represent the hits for each search.

Inspection of the reactions showed three distinctive types: the desired cyclization by (3+2) atom fragments (type A), cyclization to a five-membered heterocycle by other methodologies (type B), and other reactions which either produced

0095-2338/90/1630-0511$02.50/0

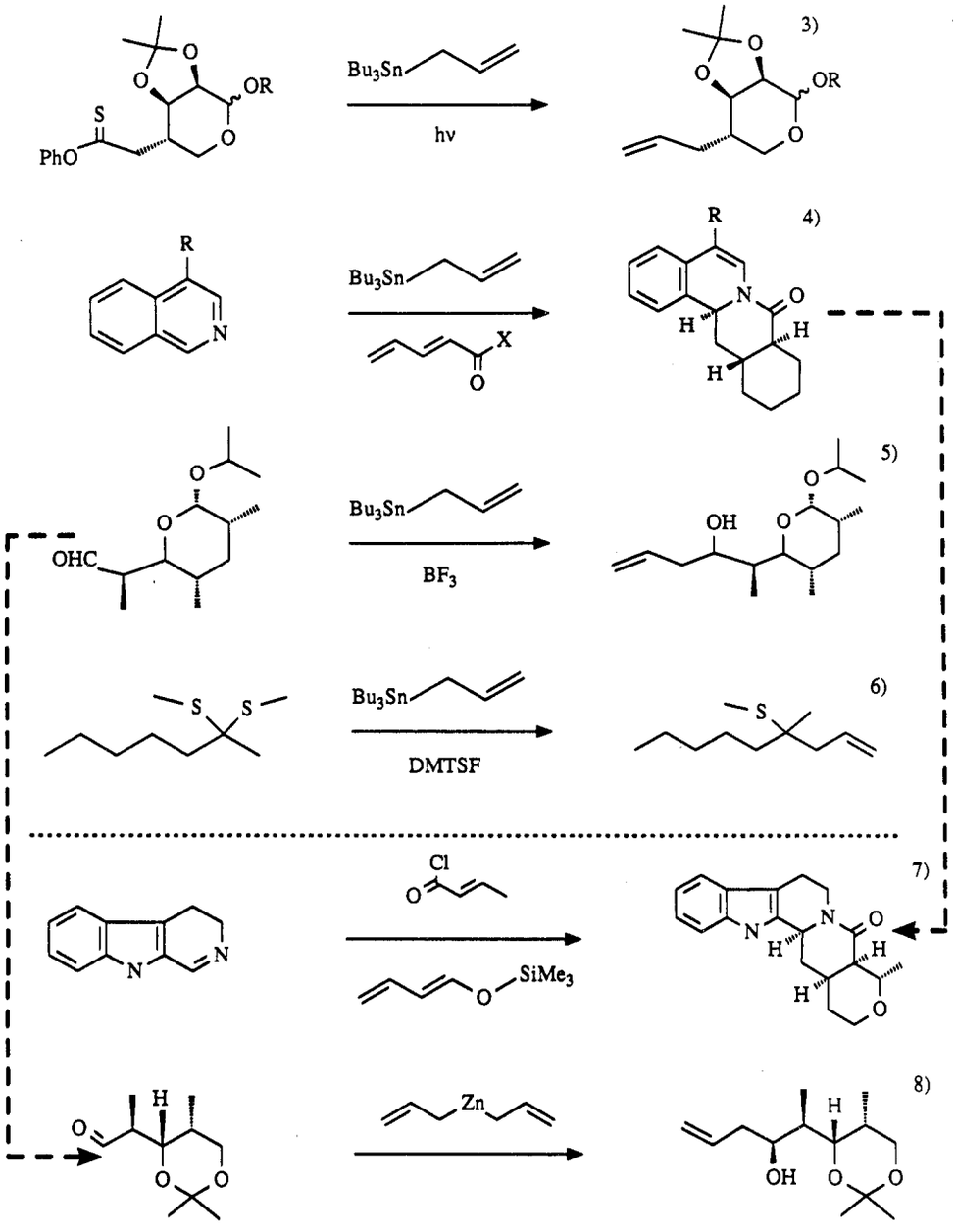


Figure 1. Similarity in the reaction domain.

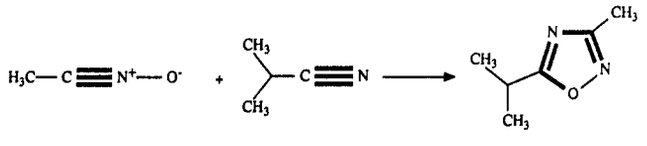


Figure 2. Similarity search query for heterocyclic ring formation.

hetero rings of different size or which did not involve formation of a heterocyclic ring (type C) (Figure 4).

For some selected searches (circled in Figure 3) the percentages of A, B, and C were calculated. As expected, searches with high molecule parameter settings yielded reactions where the similarity of the compounds is emphasized more than the reaction itself. An increase in the reaction parameter values changes the results up to a point where only [3+2] cyclizations are being retrieved. Overall, we were able to extract from the database information on the preparation of 16 different five-membered heterocyclic systems (Figure 5) formed by cyclization of 3- and 2-atom fragments, thus proving the validity of the approach. Some representative reactions are shown in Figure 6.

Michael Reaction. Nonheterocyclic synthetic chemistry classification schemes are often based on reaction type. The common denominator in this case is the changes occurring at the reaction site(s). Most commonly, this information is extracted from a reaction database by well-defined substructure searches. Sometimes substructure queries cannot accommodate all possibilities without making the query too general. A case in point is the well-known Michael reaction. The donor and acceptor molecules can contain a wide variety of functional groups (Figure 7), and a general substructure query based on this information would generate an unreasonable number of unrelated hits in addition to the desired Michael reactions. For instance, a search over the JSM-REACCS[21b] Database with the query shown in Figure 8 resulted in 155 hits, less than half of which were actual Michael reactions. The others included some tandem reactions in which the initial Michael addition is followed by another reaction (e.g., Robinson annelation), but mostly they were false hits. These results were obtained even without considering intramolecular reactions and those

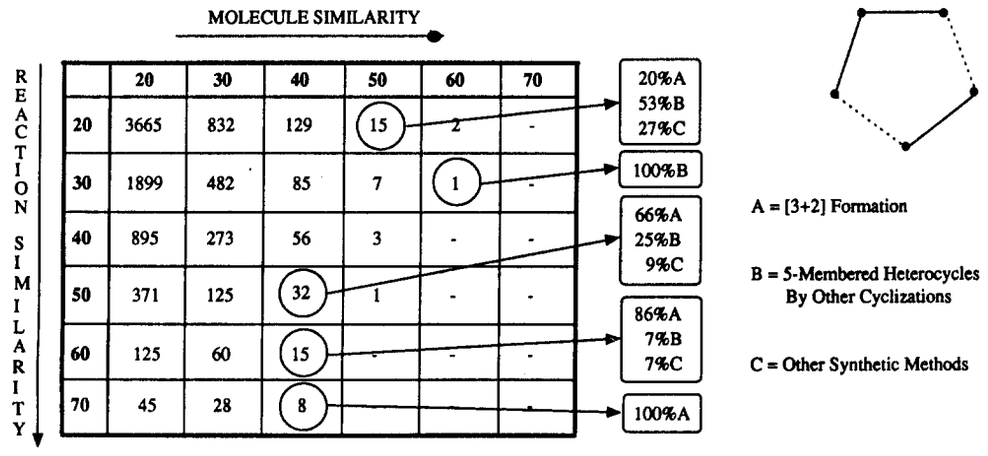


Figure 3. Results of similarity search with query from Figure 2.

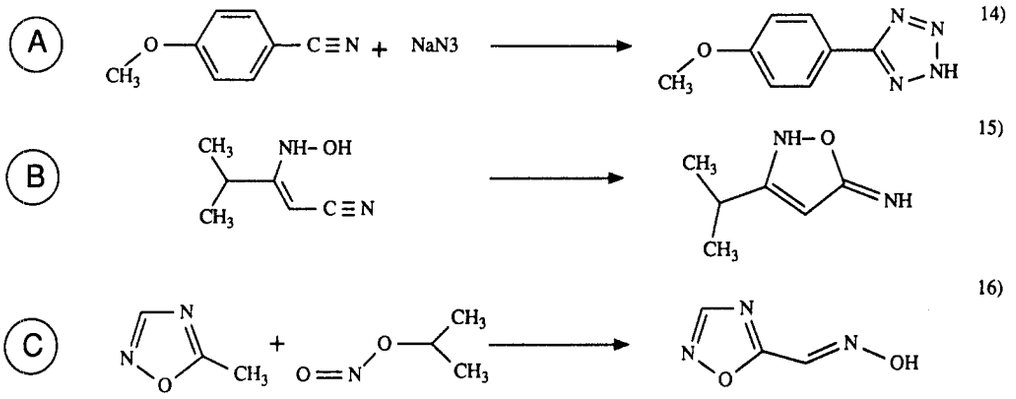


Figure 4. General reaction types found in similarity search with query from Figure 2.

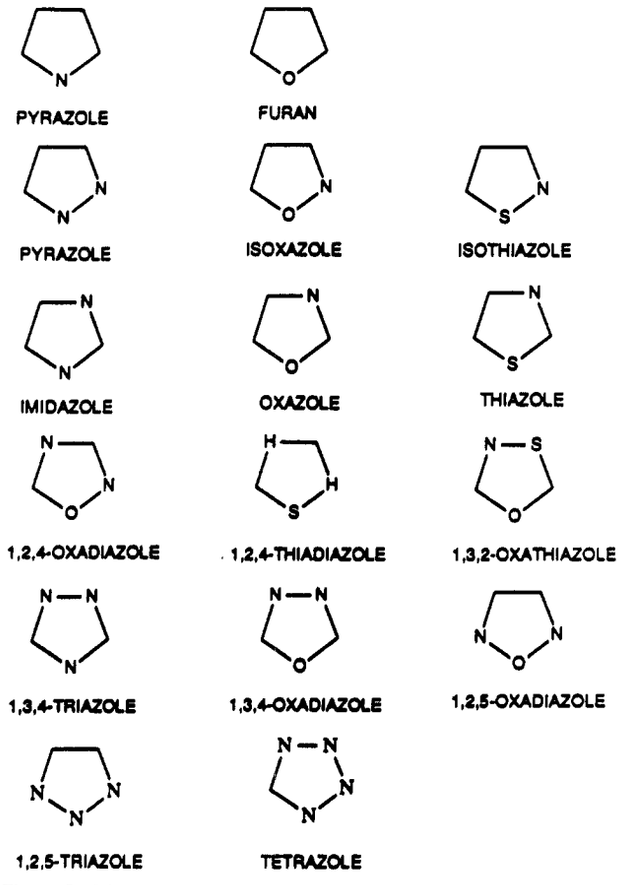


Figure 5. Ring systems formed by cyclization of [3+2] atom fragments (Current Literature Database).

involving functional groups containing triple bonds (e.g., nitriles). A keyword search over the database for titles containing the word "Michael" did not improve the results—only 45 of the previously retrieved Michael reactions fell into that category.

Similarity searching on the other hand provided interesting and useful results. A search over the same database using a typical Michael addition as the query (Figure 9) yielded 65 hits. The parameters were set to 70/20 to emphasize reaction type similarity. The hitlist contained Michael reactions with a variety of differently substituted educts, including intramolecular substrates, and reactions run under quite different conditions (Figure 10). Additionally, several "Michael-like" reactions were found, underscoring the creativity factor and the complementary nature of similarity searching. The combination of substructure, similarity, and keyword searching utilizing standard list handling procedures can provide a useful method of classifying reactions in a large database according to reaction type.

REACTIONS WITH SIMILAR REAGENTS

We have seen that it is possible to combine reactions of similar reaction type, but to cluster reactions that have similar reagents in common is a more difficult problem. For example, Lewis acids have been used as catalysts to promote many reaction types with a wide array of substrates. If the similarity of the reagents is used as the criterion for grouping different reactions, other methods than those described above are needed.

Our approach to this problem makes use of those features in REACCS which allow manipulation of molecules. All molecules which are part of a reaction (starting materials, products, reagents, catalyst, and solvents) are stored in a

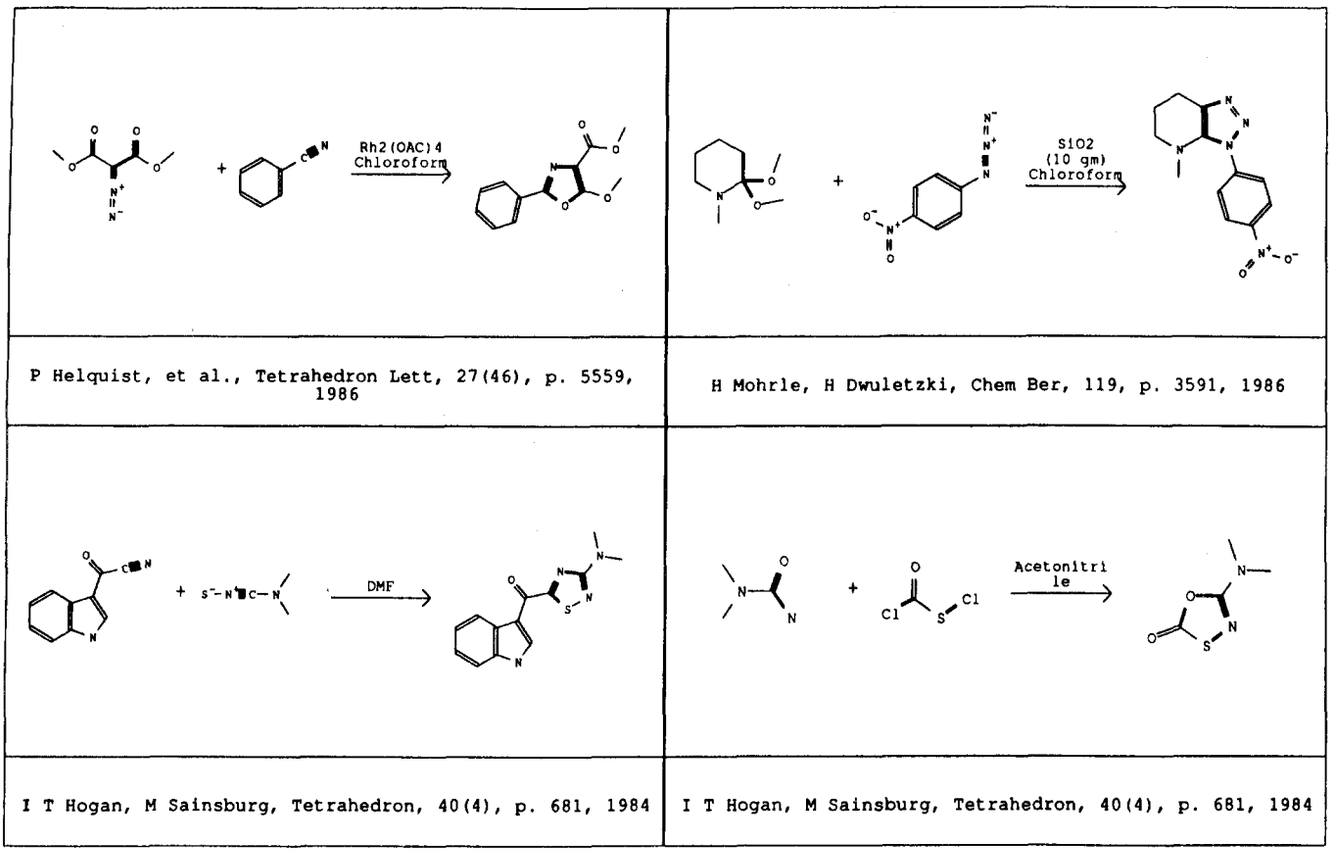


Figure 6. Representative examples from similarity search with query from Figure 2.

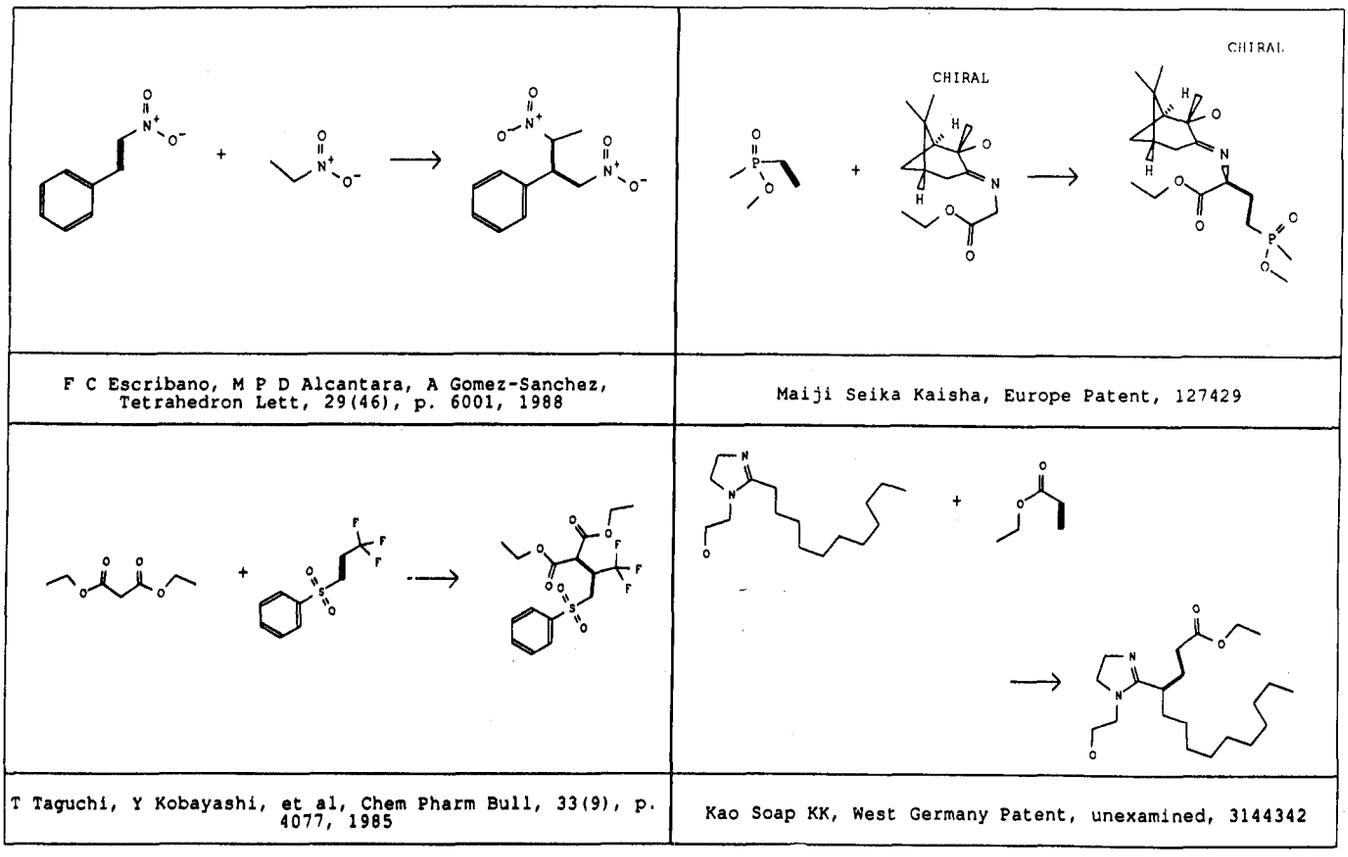


Figure 7. Diversity of substrates in the Michael reaction.

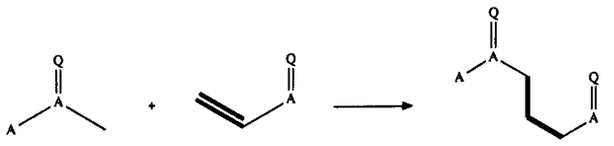


Figure 8. Query for RSS search—Michael reaction.

Figure 9. Query for similarity search (70,20)—Michael reaction.

REACCS molecule database as structures together with pertinent data. Substructure and similarity searches therefore can be carried out easily, and the results (stored as temporary or permanent *lists* of items) can be converted into the corresponding reactions.

Lewis Acids. Lewis acids exist in many forms and, unless one is interested in a closely related family of molecules (e.g., alkylaluminum halides), substructure or formula searches are meaningless. Similarity searching, on the other hand, could be very suitable for this task. We chose three different Lewis acids—$TiCl_4$ (**1**), Et_2AlCl (**2**), and $BF_3{\cdot}O(Et)_2$ (**3**)—for our searches over the Current Literature File (CLF) database (ca. 42 000 molecules and 27 000 reactions). After some experimentation, the molecule similarity threshold for these searches was set to 50. The resulting hitlists, 42 molecules for **1**, 18 for **2**, and 44 for **3**, were combined using Boolean OR. The majority of the 99 molecules in the new lists (L_n) were Lewis acids. By using the REACCS command "L_n AS CATALYST" the list was then converted into a list of reactions (1570 reactions) which contained these molecules as catalysts. This list provided a wealth of information on a wide variety of Lewis acid mediated reactions with many different catalysts; some examples are shown in Figure 11.

While a search of this type can provide chemists with a general overview over a broad area of synthetic chemistry, it is more likely that the search for reactions with similar reagents will be restricted to a more closely defined area. This is exemplified by a search for reactions involving different crown ethers. Though the structures of these compounds may be quite diverse, most reactions can actually be retrieved by a truncated name search. Most reagents of this type will have the fragment "crown" embedded in their name or commonly used abbreviation. Nevertheless, a similarity search might provide additional information and further insight into the use of different crown ethers. The simplicity of the search makes this choice even more viable.

Crown Ethers. The use of crown ethers in synthetic chemistry as phase-transfer catalysts has had an impact on many methodologies. Nucleophilic substitutions in particular benefited from their application. A typical example of this reaction type is shown in Figure 12 This reaction from the CLF database shows only the abbreviation for the crown ether in the diagram, but the full structure of 18-crown-6 (Figure 13) is actually stored in the database. It can be called up easily and, without having to draw the large structure, can be used for subsequent searches.

To gain an overview of the utility of this and other crown ethers a similarity search with this molecule over the same database was carried out as shown in Figure 13.

With the molecule similarity threshold set to 40 the search gave 123 hits, examples of which are shown in Figure 14. This

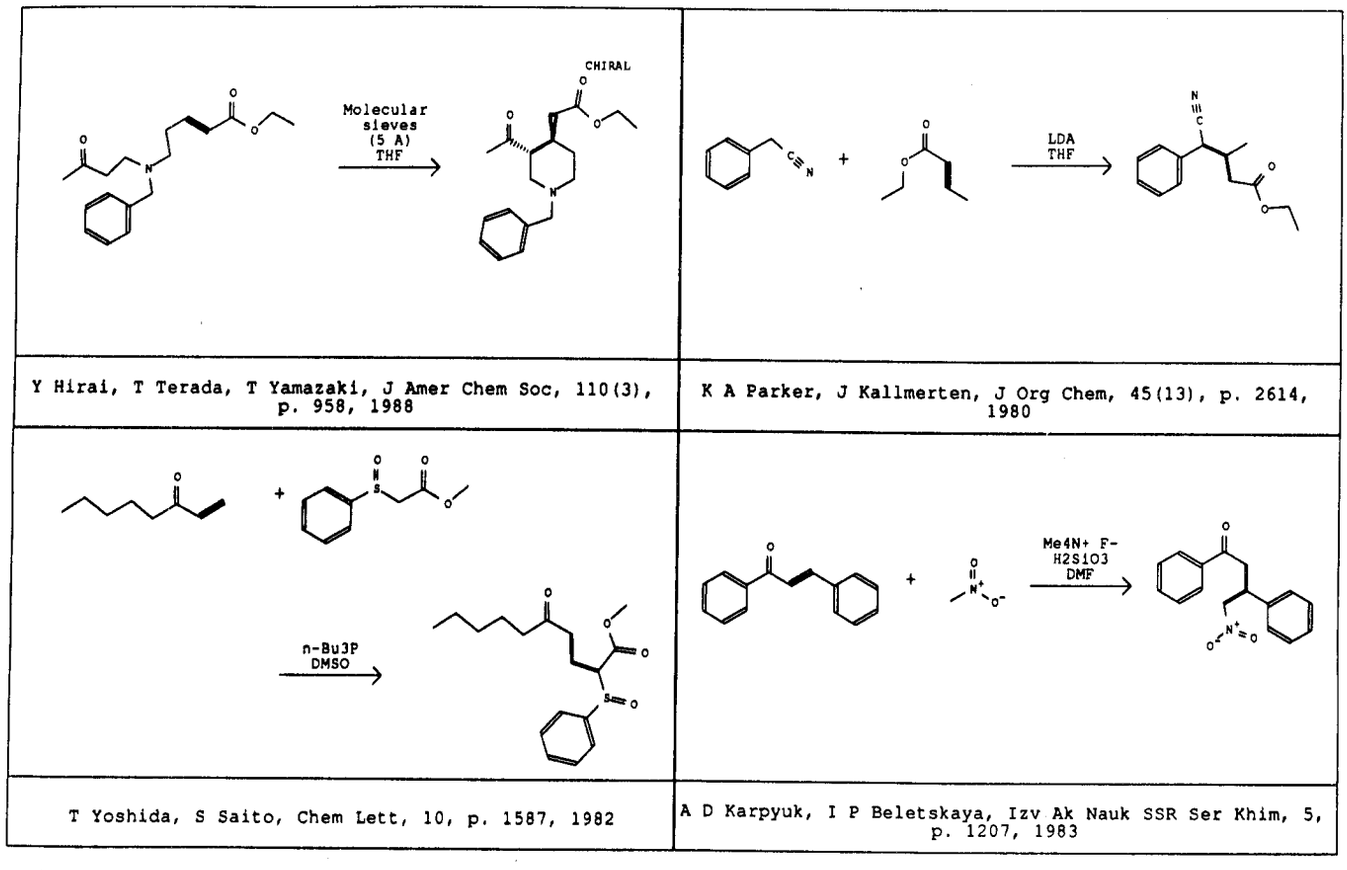


Figure 10. Representative examples of the similarity search for Michael reactions.

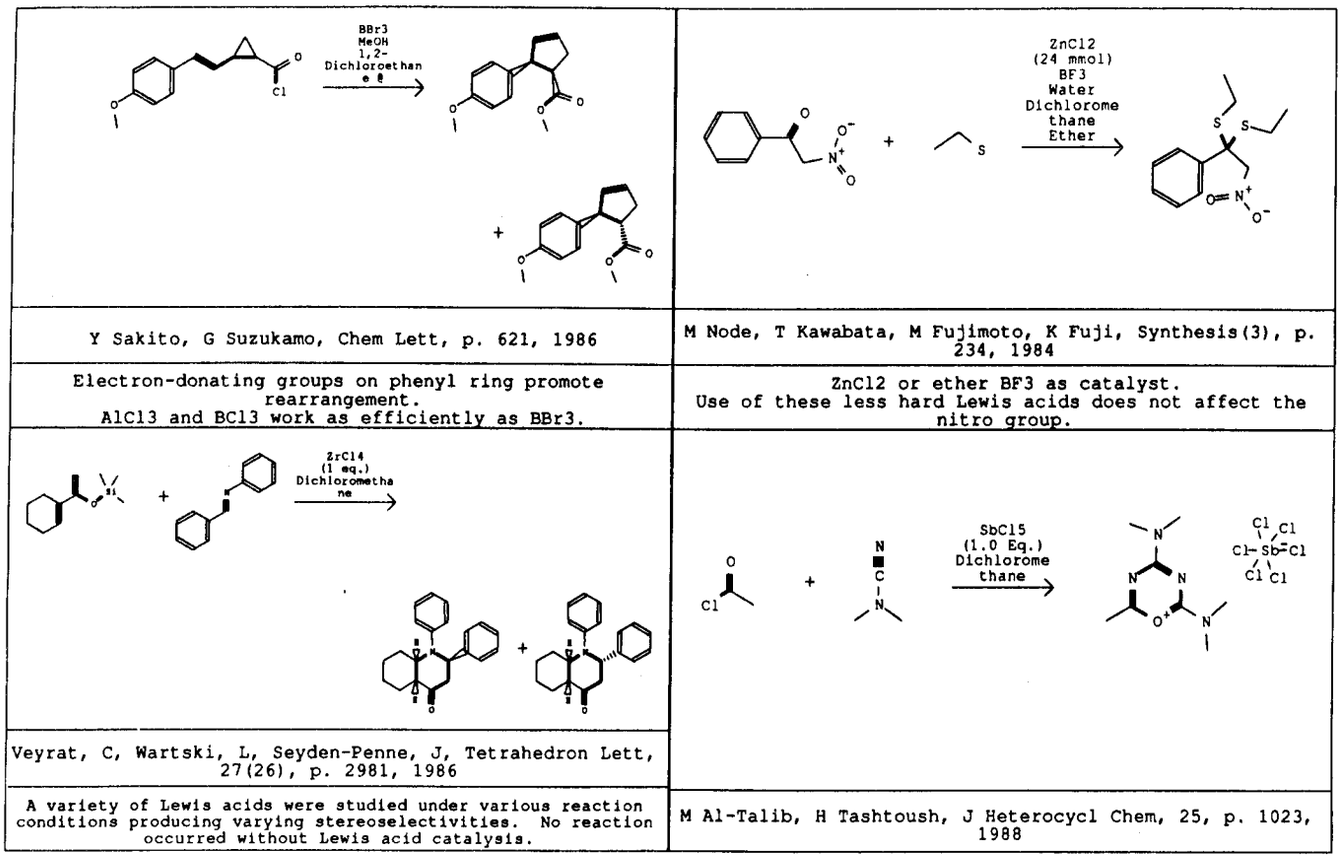


Figure 11. Lewis acid mediated reactions from CLF.

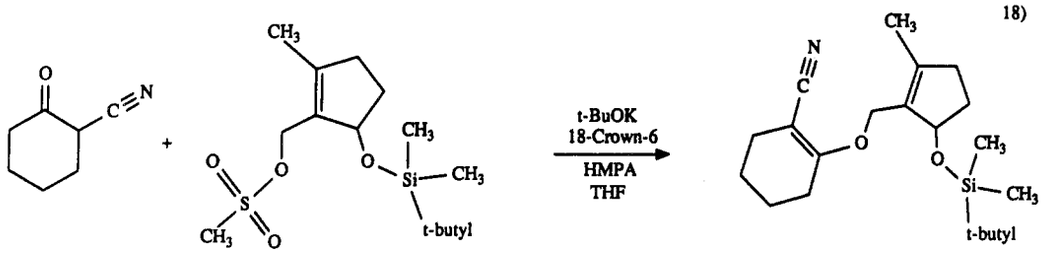
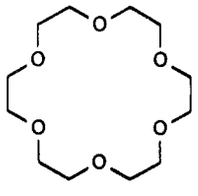


Figure 12. Reaction from CLF with crown ether as catalyst.



SEARCH: SIMILAR AS CATALYST

Figure 13. Query for similarity search for reactions using crown ethers as catalysts.

sample listing shows that many ways in which crown ethers can be used effectively with different reaction types.

These searches which employed molecule similarity searching and subsequent molecule–reaction conversions again showed the complementary nature and usefulness of similarity searches, particularly if a broad overview over different reactions using the same type of reagent or catalyst is required.

SYNTHESIS PLANNING

Since publication of the pioneering paper on logic-based synthesis planning by Corey and Wipke[19] two decades ago, much progress has been made in this area and the programs have grown in sophistication. Despite these efforts, the programs, with a few exceptions, have not yet become standard tools in the laboratory. Lack of general availability and the experimental stage of development of most of these programs are probably the main reasons that they are not used more widely. Much more important, though, is the impact these programs had on the strategies chemists apply in developing new syntheses; the retrosynthetic, analytical approach propagated by most of the programs has become the standard in synthesis planning. Utilizing this methodology, chemists rely mostly on their knowledge of synthetic chemistry which, by nature, is biased toward familiar reactions. The problem is compounded by the large number of new synthetic methods which are published at an ever increasing rate. Reaction indexing systems which access large databases on new synthetic

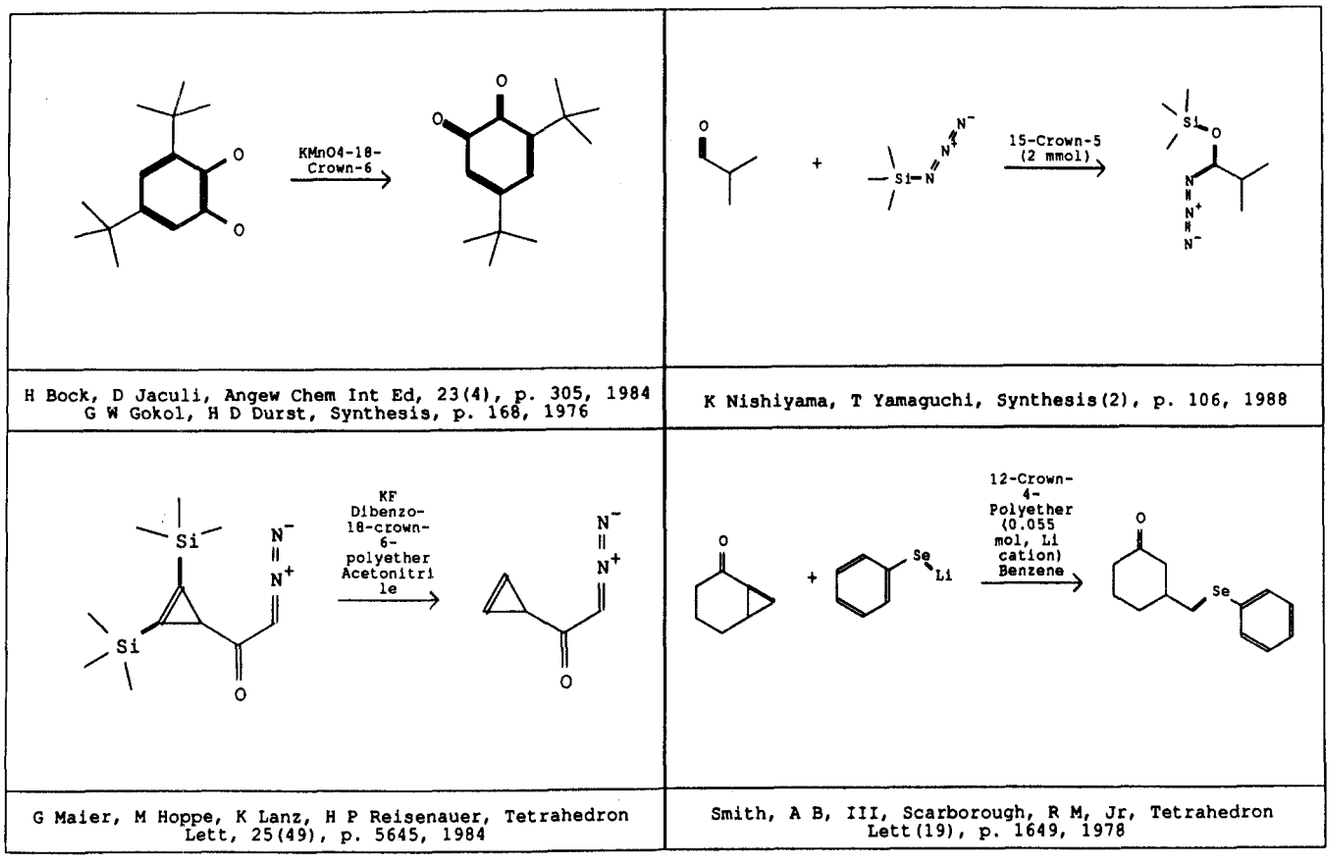


Figure 14. Examples from similarity/conversion search with 18-crown-6 as query.

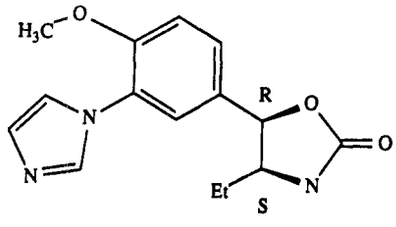
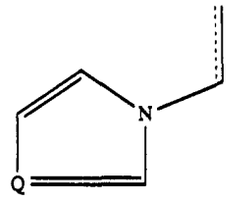
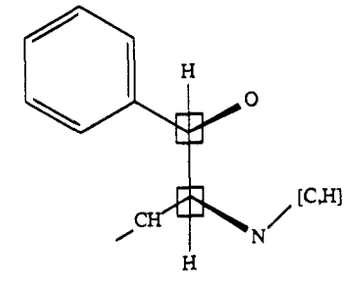


Figure 15. Target for synthesis planning.



SEARCH: SSS AS PRODUCT NOT SSS AS REACTANT

Figure 16. Substructure/conversion query for the formation of the heterocyclic phenyl fragment.



SEARCH: SSS AS PRODUCT NOT SSS AS REACTANT

Figure 17. Substructure/conversion query for the formation of erythro amino alcohols.

methodologies can make significant contributions in the planning of multistep syntheses of new compounds.

To prove this point we chose the compound shown in Figure 15 as our synthetic target. This structure was designed based on topological pharmacophores claimed to be responsible for the biological activity in phosphodiesterase-inhibiting cardiotonic agents.[20]

To find the most effective synthetic sequence it is good practice to develop and analyze as many reasonable pathways as possible, i.e., to develop a synthesis tree with many branches. Similarity, substructure, and conversion searches were used to generate these sequences. We started out with a similarity/conversion search with the structure in Figure 15 as our query and the search command "SIMILAR AS PRODUCT NOT SIMILAR AS REACTANT". This command will retrieve reactions in which the product(s) are similar to our target but the starting material(s) are not.

To obtain the maximum number of reasonable suggestions, a global search was conducted over 10 REACCS databases with a combined total of ca. 225 000 reactions.[22] With the molecule similarity set to 60, a total of 92 reactions were retrieved. The resulting hitlists were inspected and 13 reactions that bore high relevancy to our problem were chosen to represent the first level of the tree, i.e., potential starting materials for one-step reactions leading to the target. These new compounds then in turn became new targets for the same type of similarity search. Depending on the target structure the molecule similarity threshold was varied between 50 and 60;

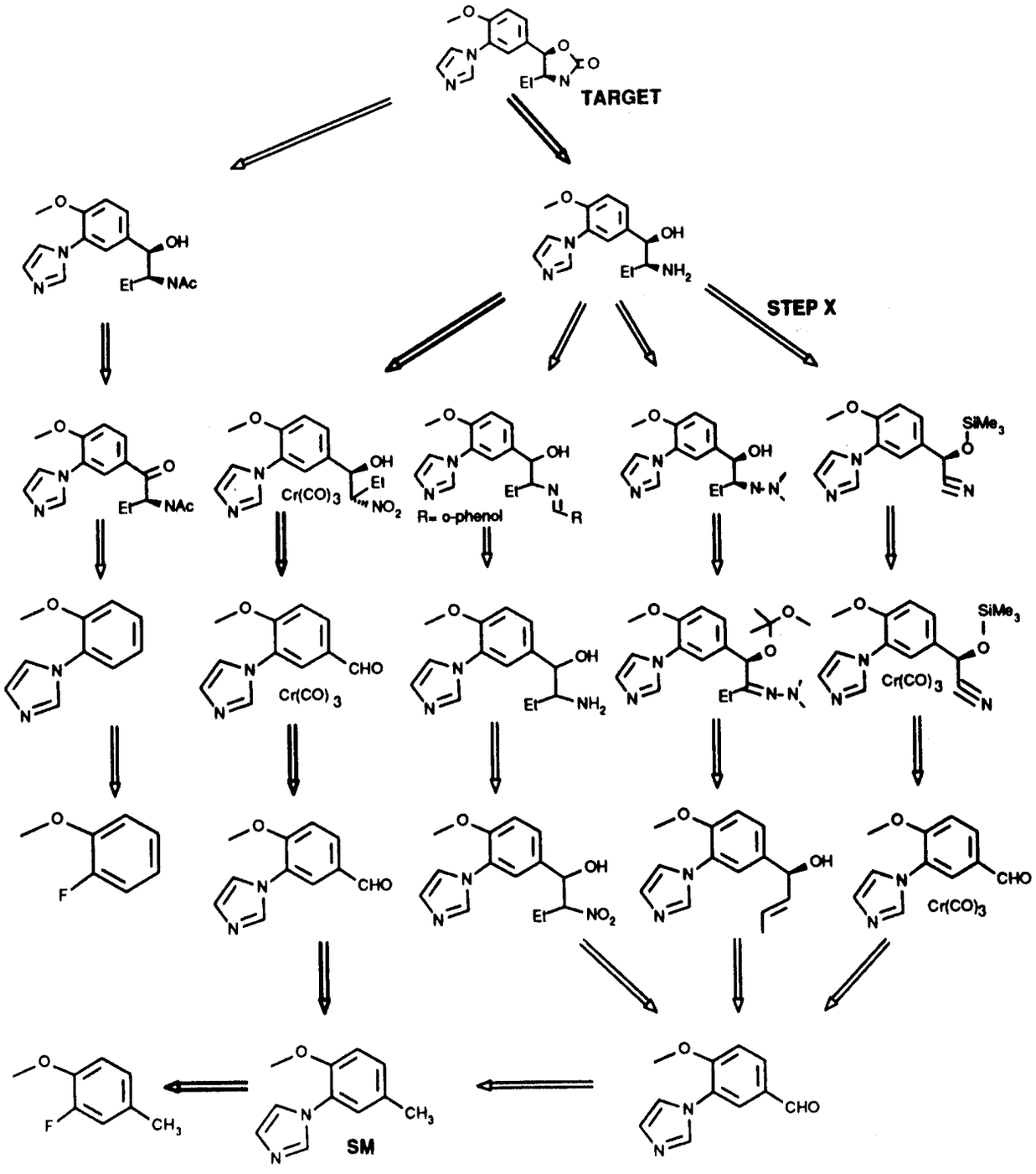


Figure 18. Synthetic sequences for the preparation of "TARGET".

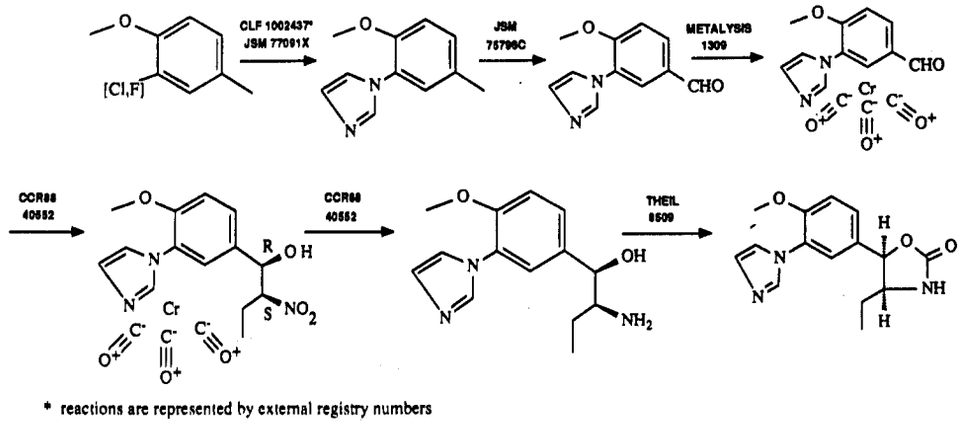


Figure 19. Synthetic sequence from the synthesis tree shown in Figure 18.

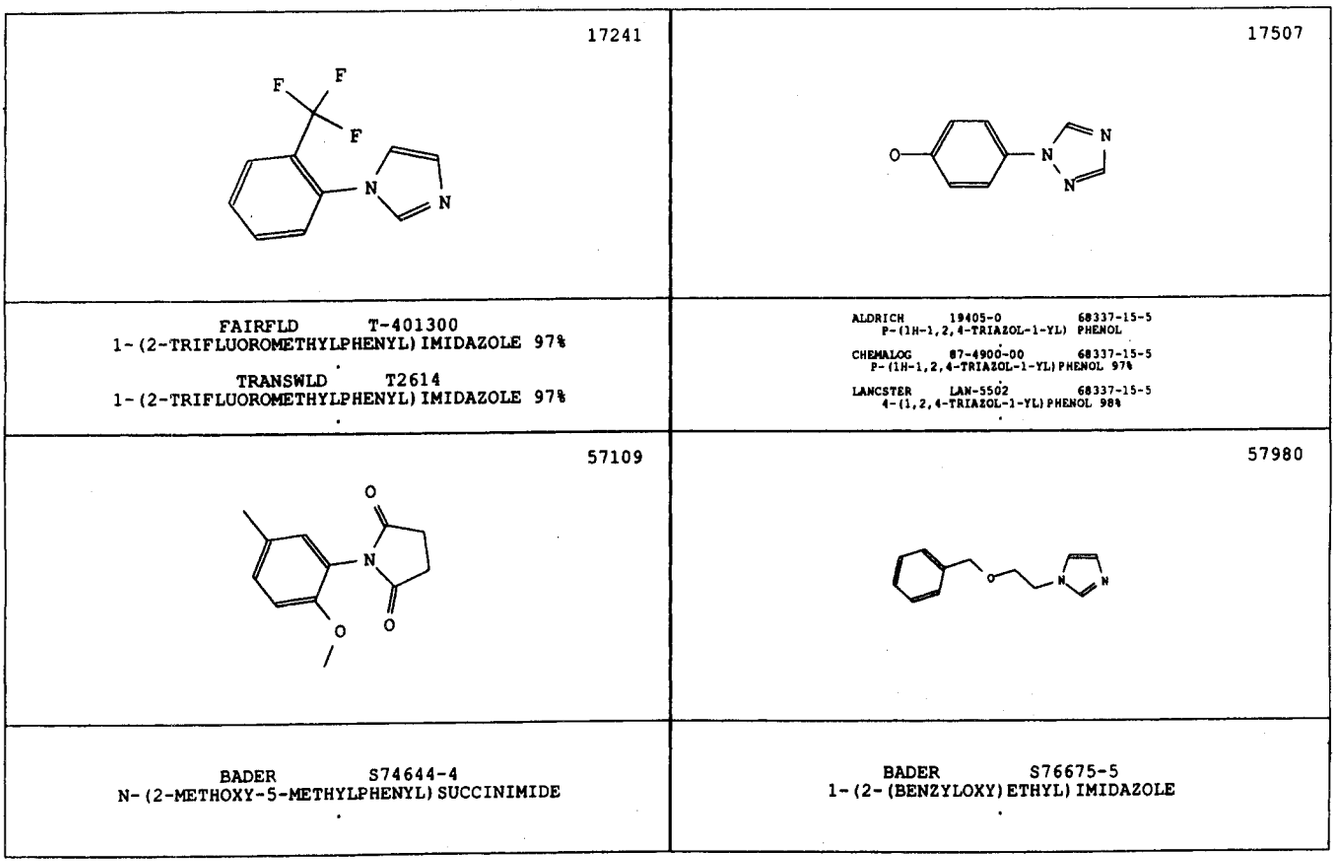


Figure 20. Potential starting materials from FCD for the preparation of derivatives.

the results (total hits vs relevant reactions) mirrored those obtained in the first search. Repeating this procedure for several levels, always choosing the most relevant examples, we created our first approximate synthesis tree. The results from the similarity searches were then augmented by more specific substructure and conversion searches to support some of the suggestions. For example, methods for the preparation of the imidazole-phenyl fragment were retrieved from the databases with the query shown in Figure 16. It is important to note that the procedure is highly interactive and that the intellectual input by the synthetic chemist will supplement the machine-generated results.

Although a specific isomer is shown in Figure 15, we were looking for methodologies which would allow the preparation of any desired stereoisomer, preferably directed by the choice of starting material. The search for reactions leading to the erythro isomer, for example, was carried out with the query in Figure 17 and gave among other reactions "Step X" highlighted in Figure 18.

By fine-tuning the results of the similarity search with more specific queries like those outlined above, the initially large number of reactions was reduced to a manageable size. Applying this procedure we neither lost any important information nor did we inject any bias into our searches. The combination of all searches resulted in many reasonable pathways which together formed the synthesis tree, an important part of which is shown in Figure 18. One sequence, very likely to succeed, is highlighted in Figure 18 and shown in detail in Figure 19. The individual steps in this sequence are solidly supported by analogous reactions retrieved from one or more of the databases used for this application.

The target compound often becomes the lead compound in a new project, and the synthetic scheme is only useful if the original starting material(s) can easily be substituted by derivatives or similar compounds to prepare a series of compounds for testing purposes. In our case this meant finding viable substitutes for the compound labeled SM in Figure 18. Therefore, this compound became the query structure in a molecule similarity search over the Fine Chemicals Directory. The search resulted in a wide variety of potential starting materials, some examples are shown in Figure 20.

CONCLUSION

We have shown in these studies that similarity searching in REACCS can be successfully applied, alone or in combination with other available methods, to the many challenges synthetic chemists face in their daily work. The problems to which solutions were suggested include reaction classifications and synthesis planning. The results indicate that similarity searching in the reaction domain is a powerful tool complementing other important searching procedures.

REFERENCES AND NOTES

(1) Willett, P. *Similarity and Clustering in Chemical Information Systems*; Research Studies Press: Letchworth, Herfordshire, England, 1987.
(2) (a) Moock, T. E.; Grier, D. L.; Hounshell, W. D.; Grethe, G.; Cronin, K.; Nourse, J. G.; Theodosiou, J. *Tetrahedron Comput. Methodol.* **1988**, *1*, 117 and references therein. (b) We are grateful to a referee for pointing out that similarity searching has also been implemented recently in ORAC (unpublished information).
(3) Keck, G. E.; Kachensky, D. F.; Enholm, E. J. *J. Org. Chem.* **1984**, *49*, 1462.
(4) Yamaguchi, R.; Otsuji, A.; Utimoto, K. *J. Am. Chem. Soc.* **1988**, *110*, 2186.
(5) Nakajima, N.; Hamada, T.; Tanaka, T.; Oikawa, Y.; Yonemitsu, O. *J. Am. Chem. Soc.* **1986**, *108*, 4645.
(6) Trost, B. M.; Sato, T. *J. Am. Chem. Soc.* **1985**, *107*, 719.
(7) Martin, S. F.; Benage, B.; Hunter, J. E. *J. Am. Chem. Soc.* **1988**, *110*, 5925.
(8) Yamamoto, Y.; Maruyama, K. *Heterocycles* **1982**, *20*, 357.
(9) Ref 1; p 54.
(10) Vladutz, G. Do we still need a classification of reactions. In *Modern Approaches to Chemical Reaction Searching*; Willett, P., Ed.; Gower Publishing Company: Brookfield, VT, 1986; p 202.

(11) Herger, R. Reaction Planning (Computer-Aided Reaction Design). In *Chemical Structures, The International Language of Chemistry*; Warr, W. A., Ed.; Springer Verlag: Berlin, 1988; p 385.
(12) Hendrickson, J. B.; Miller, T. M. Submitted for publication in *J. Am. Chem. Soc.* We thank Professor Hendrickson for a preprint of this paper.
(13) Moock, T. E.; Nourse, J. G.; Grier, D. L.; Hounshell, W. D. The Implementation of Atom–Atom Mapping and Related Features in the Reaction Access System (REACCS). In *Chemical Structures, The International Language of Chemistry*; Warr, W. A., Ed.; Springer Verlag: Berlin, 1988; p 303.
(14) Maxwell, J. R.; Wasdahl, D. A.; Wolfson, A. C.; Stenberg, V. I. *J. Med. Chem.* **1984**, *27*, 1565.
(15) Fomum, Z. T.; Asobo, P. F.; Landor, S. R.; Landor, P. R. *J. Chem. Soc., Perkin Trans. 1* **1984**, 1079.
(16) Kenley, R. A.; Bedford, C. D.; Dailey, O. D., Jr.; Howa, R. A.; Miller, A. *J. Med. Chem.* **1984**, *27*, 1201.
(17) Olsen, R. S.; Fataftah, Z.; Rathke, M. W. *Synth. Commun.* **1986**, *16*, 1133.
(18) Ziegler, F. E.; Nangia, A.; Schulte, G. *J. Am. Chem. Soc.* **1987**, *109*, 3987.
(19) Corey, E. J.; Wipke, W. T. *Science* **1969**, *166*, 178.
(20) Erhardt, P. W. *J. Med. Chem.* **1987**, *30*, 231. Erhardt, P. W.; Hagedorn, A. A., III; Davey, D. D.; Pease, C. A.; Venepalli, B. R.; Griffen, C. W.; Gomez, R. P.; Wiggins, J. R.; Ingebretsen, V. R.; Pang, D.; Cantor, E. *J. Med. Chem.* **1989**, *32*, 1173.
(21) All databases mentioned are REACCS-readable files of synthetically important reactions and commercially available molecules: (a) Theilheimer, over 46 000 reactions. *Synthetic Methods of Organic Chemistry*; Theilheimer, W., Ed.; 1946–1980 Vols. 1–35 (Karger). (b) REACCS-JSM, over 51 000 reactions. *Journal of Synthetic Methods*, 1980–1988, Vols. 6–14 (Derwent). (c) Current Literature File (CLF), over 44 000 reactions. Abstracted from the current synthetic literature (~35 journals from 1983 to present) by a consortium of university researchers. (d) ORGSYN, over 5 000 reactions. *Organic Syntheses*, 1921–1988, Vols. 1–68 (Wiley & Sons). (e) CHIRAS, over 14 000 reactions. Abstracted from the chemical literature from 1975 to present. Scope is limited to methods for asymmetric synthesis.
(22) The following public databases were used: MDL's THEIL, ORGSYN,[21d] JSM-REACCS, CLF, CHIRAS,[21e] METALYSIS, and the CCR databases (1986–1989) from the Institute of Scientific Information (ISI), Philadelphia.