retrieved by CAS ONLINE, a new search service.

SUMMARY

Because development of rigid formatting rules for generating graphical representations of chemical structures would be impractical, a set of comprehensive guidelines has been developed at CAS. The guidelines provide preferred formats for acyclic structures, rings and ring systems, and representation of stereochemistry. They also describe a number of methods for reducing crowding in diagrams. The guidelines, which have been used for CAS publications for about 10 years, were originally developed for manually drawn diagrams, but have proved to be of equal value for subsequent computer-assisted generation of diagrams. More general use of these guidelines (such as in chemical education, research and development, and information storage and retrieval) would facilitate communication of chemical structure information among scientists.

REFERENCES AND NOTES

(1) Crosland, M. P. "Historical Studies in The Language of Chemistry"; Dover: New York, 1978.

(2) For example: Soltzberg, L. J. "Computer Graphics for Chemical Education", *J. Chem. Educ.* **1979**, *56*, 644-9.

(3) For example: Corey, E. J.; Wipke, W. T. "Computer-Assisted Design of Complex Organic Syntheses", *Science* 1969, *166*, 178-92; "Computer-Assisted Drug Design", *ACS Symp. Ser.* **1979**, *No.* 112.

(4) Rush, J. E. "Handling Chemical Structure Information"; *Annu. Rev. Inf. Sci. Technol.* **1978**, *13*, 209-62.

(5) (a) Patterson, A. M.; Capell, L. T.; Walker, D. F. "The Ring Index", 2nd ed.; American Chemical Society: Washington, DC, 1960. (b) See pp xii–xiii. See also ref. 10.

(6) Blake, J. E.; Brown, S. M.; Ebe, T.; Goodson, A. L.; Skevington, J. H.; Watson, C. E. "Parent Compound Handbook—Successor to the Ring Index", *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 162–7.

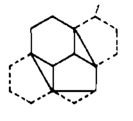
(7) "Chemical Abstracts Index Guide": (a) Vol. 76, 1972, Introduction, Paragraphs 8, 15, 140–63, and 202–12. (b) Vol. 76–85, Cumulative, 1972–6, Appendixes, Paragraph 15.

(8) Blake, J. E.; Farmer, N. A.; Haines, R. C. "An Interactive Computer Graphics System for Processing Chemical Structure Diagrams"; *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 223–8.

(9) Dittmar, P. G.; Mockus, J.; Couvreur, K. M. "An Algorithmic Computer Graphics Program for Generating Chemical Structure Diagrams"; *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 186–92.

(10) The wording in this rule is a brief generalization of orientation and numbering rules A-22, A-31.2, A-32.23, A-34.2, A-41.2, B-1.5, B-3.4, B-10, C-12.7, C-12.8, and C-15 in: IUPAC. "Nomenclature of Organic Chemistry: Sections A, B, C, D, E, F and H"; Pergamon: Oxford, 1979. The rules in this reference supersede those in ref 5.

(11) Figure 13c appears to violate the orientation and numbering rules of ref 10. However, this ring system is drawn for orientation purposes as though every component ring were six membered, as shown, i.e., as if the ring system were pyrene. While pyrene numbering begins at position 1 in the accompanying diagram, locant 1 of Figure 13c is assigned to the first available position according to rule A-22 of ref 10.



(12) These terms are used as in rule A-21.5 of ref 10.

Method for Generating a Chemical Reaction Index for Storage and Retrieval of Information

MARGARET A. MOSBY*

Tompkins-McCaw Library, Medical College of Virginia, Virginia Commonwealth University,
Richmond, Virginia 23298

LEMONT B. KIER

Department of Pharmaceutical Chemistry, Medical College of Virginia, Virginia Commonwealth University,
Richmond, Virginia 23298

Received March 6, 1980

A new method for indexing chemical reactions is described. The calculation of a reaction connectivity index results in a unique number. This number does not provide hierarchical or relational information. It encodes the concept of the reaction process in a unique identifier which is suggested for use, much as the CAS Registry number is used, to optimize ease of storage, manipulation, and retrieval from large computer files.

INTRODUCTION

The retrieval of information in chemistry has become increasingly complex. The total amount of literature to be covered in an exhaustive search is now so vast as to preclude a systematic manual examination of the available sources. It is, however, possible to use a computer for this task. The Chemical Abstracts Service files, Derwent's files, and the American Petroleum Institute's file, as examples, are all available for computer access in some form through one or more data-base vendors. The major benefit, of course, is the speed with which a body of relevant material can be extracted from a much larger file of bibliographic citations or chemical data. The primary drawbacks with this approach are associated with the forms in which data is put into the computer files, with the processing capabilities of the computer, and with the software commands available to search the files.

The main objective of the information scientist performing a literature search is to maximize relevant retrieval without sacrificing completeness. This objective is most easily achieved when searching files that are structured by the use of a controlled vocabulary and/or other formal systems. Hence, the first recognized need in chemical information retrieval was a unique way to identify chemical compounds that would eliminate the confusion of multiple names used for the same structure. The Chemical Abstracts Service Chemical Registry System has done this since 1965, based on a system first developed by Gluck.[1] The registry number is based on the representation of the chemical structures in the form of topological tables and is unique and unambiguous. Now it is at least as important to the chemical community to have

0095-2338/80/1620-0217$01.00/0 © 1980 American Chemical Society

consistent ways to index and retrieve descriptions of the reactions in which these molecules may be used. Beach et al.[2] describe at length all the approaches which must be considered when searching for reaction information manually. At the present time, all these approaches must be considered for a thorough computer search. The resulting string of synonyms, eponymous terms, and key words complicates strategy formulation. Often in an attempt at thoroughness the rate of relevant return must be sacrificed.

Schemes for indexing reactions have been available for many years. Valls[3] gives a very thorough history of the development of reaction indexing. He breaks down the approaches to the compound oriented (e.g., Lynch[4] and Gelberg[5]) and transformation oriented (e.g., Weygand,[6] Theilheimer,[7] and Vleduts[8]).

Since that time, extensive work has been done in this area by Hudrlik,[9] Hendrickson,[10] Zeigler,[11] Bersohn,[12] and Willett,[13] among others. These systems have chiefly been developed to answer the specific questions of the synthetic chemist. They are meant to be access systems to complex reaction information in, and of, themselves. In fact, Willett[13] echoes Valls and the others when he points out that a reaction does not lend itself to listing "in a canonical form, such as via the CAS Registry System". By using a calculation based on molecular connectivity, however, it is possible to generate an index of unique numbers for chemical reactions. This number should stand, as the CAS Registry number does, only as a unique identifier. It is not meant to contain any hierarchical or relational information. This is the major difference between this index and every other reaction indexing system that has appeared in the literature. The reaction connectivity index is not meant to contain information which will, by itself, answer complex reaction questions. It encodes the concept of a process in a unique number. This unique identifier can be used to optimize the ease of storage, manipulation (e.g., coordination in a computer search), and retrieval. It is meant to be used as part of a coordinated search strategy and to enable both the indexer and the retriever to be free from the ambiguities and complexities of natural language.

DISCUSSION OF METHOD

A chemical reaction, in contrast to a molecule, is a dynamic event, proceeding from reactants to products. A molecule is a single entity which can be described by using numerical descriptors. Several of these are in use. A chemical reaction involves the change of one molecule to another, each with a possible codified index. The process then is a change in molecular structure.

Information associated with any aspect of a chemical reaction ideally could be stored under some numerical index derived from a general model characteristic of all examples of a molecular change associated with a particular reaction. The index should be different from an index associated with some other reaction; it should be easily derived for a general class of reactions irrespective of nonparticipating variations in the structure; it should be clearly related to the reaction in some way so that chemists anywhere use the same index.

An answer to this problem lies in the structural description of molecules developed by Kier and Hall called *molecular connectivity*. Molecular connectivity is a description of molecular structure based upon the adjacency or connectivity of atoms forming bonds and ultimately molecules. The method leads to numerical indexes which encode information about the number of atoms, branching, cycles, unsaturation, and heteroatom content. The indexes have been demonstrated to relate closely to calculated molecular volumes and surface areas, to numerous physicochemical properties, and to the biological activity of many series of molecules.[14] The index is easily calculated and needs no empirical assignments or assumptions.

Table I. Connectivity δ Values of Atoms in Molecules

atom	situation in molecule	δ^V
R		1
C	primary	1
	secondary	2
	tertiary	3
	quaternary	4
	in $=CH_2$	2
	in $-CH=$, benzene	3
	\\ in C= /	4
	in $-CH=$	3
	in $-C\equiv$	4
N	in $-NH_2$	3
	in $-NH-$	4
	\\ in N– /	5
	in nitrile, pyridine, nitro	5
O	in hydroxyl	5
	in ether, carbonyl	6
F	fluoro	7
Cl	chloro	0.7
Br	bromo	0.25
I	iodo	0.15
S	$-SH$	0.5
	$-S-$	0.6

CALCULATION OF THE MOLECULAR CONNECTIVITY INDEX, ${}^1X^V$

The calculation of the ${}^1\chi^V$ index begins with writing the nonhydrogen skeleton for reactant and product. A δ^V value or connectivity degree is assigned to each atom. This is the number of bonds to nonhydrogen atoms plus the number of nonbonding and π electrons. The δ^V value can be expressed as

$$\delta^V = Z^V - h$$

where Z^V is the number of valence electrons on the atom and h is the number of hydrogens on the atom.

Halogens above fluorine are treated in a more elaborate way, since each has seven valence electrons. The $\delta^V = (Z^V - h)/(Z - Z^V)$ where Z is the total number of electrons. Table I shows the δ^V values for several atoms in different hybrid states.

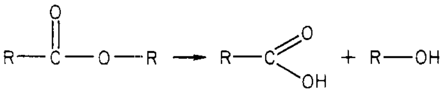
For each bond in the molecule, the product of the δ^V values forming the bond is calculated, and then the reciprocal square root is computed to give a bond index $(\delta^V_i\delta^V_j)^{-1/2}$. These are summed over the entire molecule to give ${}^1\chi^V$:

$${}^1\chi^V = \sum(\delta^V_i\delta^V_j)^{-1/2}$$

CALCULATION OF THE REACTION CONNECTIVITY INDEX, Δ^1X^V

Recognizing that a chemical reaction is a process of conversion of one molecular structure to another, all reactions can be classified by identifying the general structure of reactant and product. A chemical reaction occurs at a part of a molecule, leaving another part unchanged in the conversion. Thus, a general model for a reaction need only consider the structural fragment in which change occurs, in both the product and reactant. A simple, useful reaction index may be designed from a numerical index encoding the general structure of a reactant and of a product in a particular transformation.

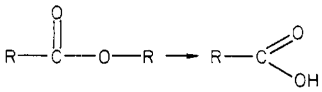
Hydrolysis of an ester illustrates this point. It is possible to write a general structure for the initial and final structures in the process or reaction called ester hydrolysis, with the following symbols:



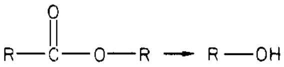
This symbolic statement of the general reaction is familiar to all organic chemists, a fact which makes any reaction index derived from it of great general value. While it is true that the carbonyl group, C=O, remains unchanged in the process, it is explicitly stated in the symbolic statement since it contributes to the chemical identity of the reactant and one product. Its inclusion or omission in the symbolic statement is immaterial to the numerical value of the index calculated. This becomes obvious in the following development of the reaction index.

The use of the symbol R reflects all structural features of the molecules not undergoing a change in the process. Chemically we recognize that there are various levels of participation and influence of R, but the qualitative results are correctly symbolized above for the reaction.

The hydrolysis reaction modeled above may be viewed as a process of acid synthesis from an ester. Interest is thus focused upon the molecular participants, symbolized by



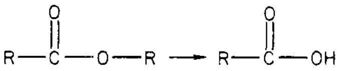
By use of equivalent reasoning, the process of interest may be the synthesis of alcohols by ester hydrolysis:



Depending upon the focus of attention of the investigator, any of these models may typify the general reaction under which he seeks to encode, store, and retrieve information. The essential aspect of the encoding scheme is the unambiguous and universally recognizable choice of a general model symbolizing the process or reaction of interest. For this reason, the following rules are adopted, based on familiar chemical logic and intuition, for the description of a general model for a chemical reaction.

GENERAL MODEL FOR A REACTION

The first step in the derivation of a reaction index is the choice of the general reaction model, as the focus of attention, which an investigator wishes to encode, store, and retrieve information. If the synthesis of acids from esters by hydrolysis is the principle interest, then the reaction is modeled thus:

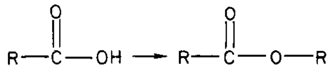


The rule is adopted that all parts of the molecule not in the transformation of ester to acid are designated by a general symbol R. The carbonyl oxygen is retained in the general model since it is, by definition, a part of the ester and carboxylic acid groups involved in the transformation. Only the organic molecule participating in the reaction and the product of interest are included in the reaction model. Other reagents, solvents, catalysts, etc., are not included. The process of interest is the generalized structure of an ester being transformed to an acid.

This general model may also symbolize the hydrolysis of a lactone to an acid, realizing that an alcohol moiety is present in the same product. Nevertheless, if the principle interest is in the conversion of ester to acid, then this model is symbolic of the general reaction of lactone hydrolysis to an acid.

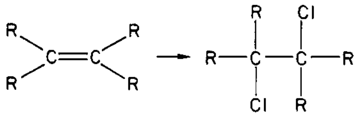
It should be noted that in this or in any general model, the reactants and products are considered in a neutral state; thus, the acid is not described as an ion since that is a variable controlled by reaction conditions, not the structure transformation inherent in the reaction.

A corollary to this model would be the case where the principle interest for encoding a reaction index would be the reaction forming an ester from an acid. This would be symbolized as



the reversed model of ester hydrolysis to an acid.

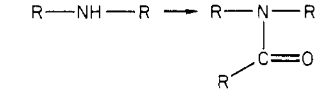
Another example of the designation of a general reaction model is the case of the chlorination of an alkene:



The model depicts only the functional part of the alkene, the double-bonded pair of carbons. The remainder of the molecule does not participate in the changed structure represented by the product. The general symbols R are thus used to satisfy the valencies of the double-bonded carbons and reflect the remainder of the molecule which is immaterial in the general reaction model. The product is symbolized with two chlorine atoms. There is no explicit inclusion of chlorine or chlorinating agent as a reactant. The presence of chlorine in the product reflects the general reaction of interest, the chlorination of a double bond.

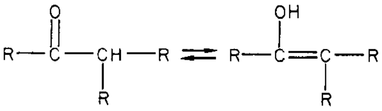
The general reaction model carries no symbolism reflecting mechanism, intermediates, catalysts, or special conditions. It symbolizes only the beginning and the end of a transformation. The product representation in the case above carries no information that the addition results in a trans arrangement of chlorine atoms. The symbols convey only atoms and their neighbors. Geometry, stereochemical arrangement, bond length, etc., are not part of the model. This, in fact, is the kind of information likely to be found stored under this reaction connectivity index.

Another example of the designation of a general reaction model can be built around the acylation of amines. The reactant amine and the product, an amide, are symbolized by



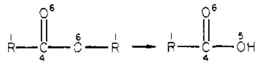
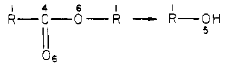
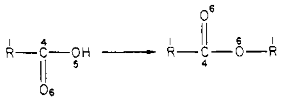
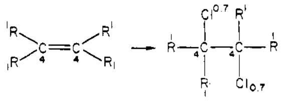
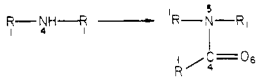
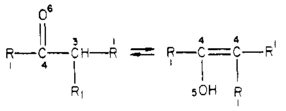
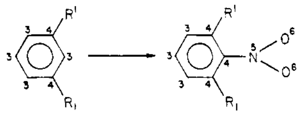
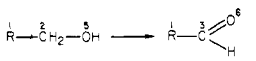
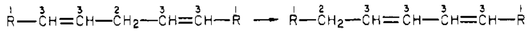
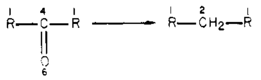
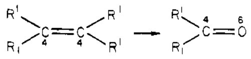
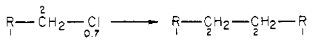
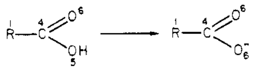
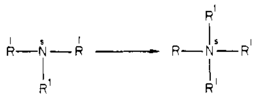
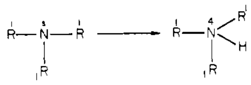
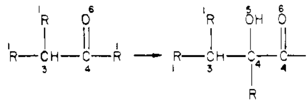
The nature of the acylating agent, catalysts, conditions, water, and other products is not part of the model. Only the amine bearing a replaceable hydrogen and the amide product are represented. The amine represents any primary or secondary amine or ammonia. Chemical intuition tells us that tertiary amines will not acylate this way. The acyl portion of the product is described in its most general form, with an R symbol and the carbonyl group.

A fourth example is the keto–enol tautomerism. The general model shows explicitly only those atoms involved in the transformation:



The molecule on the left, the reactant, is shown as a general ketone with one α-hydrogen and its bonded carbon atom. This hydrogen migrates in the tautomerism to the oxygen, forming a carbon–carbon double bond. Only these salient features are depicted in the general reaction model. Since this may be a highly reversible reaction, an index can be calculated reflecting

Table II. General Reaction Models and Reaction Connectivity Indexes

no.	description of reaction	general reaction model with δ^V values	$\Delta^1\chi^V$
1	ester hydrolysis to acid		−0.388
2	ester hydrolysis to alcohol		−0.869
3	esterification of an acid		0.388
4	chlorination of alkene		1.195
5	acylation of amines		0.822
6	keto–enol tautomerism		±0.173
7	aryl substitution (nitration)		0.512
8	alcohol oxidation		−0.210
9	bond rearrangement		0.055
10	carbonyl reduction		0.210
11	ozonolysis		−1.046
12	ring formation		0.362
13	acid dissociation		−0.019
14	amine quaternization		0.447
15	amine protonation		0.158
16	aldol condensation		0.974

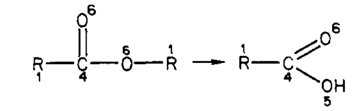
either the forward or reverse action.

Substitution reactions on aromatic rings are modeled as shown by example 7 in Table II. The benzene ring with the two R substituents is the model for any aromatic hydrocarbon undergoing a substitution reaction. The R groups flanking the carbon undergoing substitution symbolize the most general case where none, one, or two substituents are ortho to the entering nitro group. The remaining atoms of the ring shown in the general model for reactant and product do not contribute to the numerical value of the index. They are included in order to aid the chemist in identifying the chemical nature of the process and the types of molecules undergoing change.

It is obviously impossible to describe the general reaction model for every known reaction in this short space. Adhering to the rules laid down in this section should make it possible to treat every reaction in a uniform, unambiguous manner, in preparation for the computation of the reaction connectivity index. A number of examples are illustrated in Table II.

CALCULATION OF THE REACTION CONNECTIVITY INDEX

The calculation of the reaction connectivity index follows from the designation of the general reaction model. This can be illustrated by considering the first entry in Table II, the

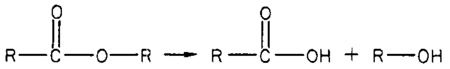


model for the hydrolysis of an ester to an acid. The valence connectivity values, δ^V, are assigned to each atom in the model structures (see Table I). By convention, the δ^V value for all R symbols is unity. The bond index values, $(\delta^V_i\delta^V_j)^{-1/2}$, are computed for each atom pair, i and j, forming a σ bond. The bond indexes are summed over all bonds in a molecule to give the molecular connectivity index, $^1\chi^V$. The reaction connectivity index, $\Delta^1\chi^V$, is $^1\chi^V$(product) minus $^1\chi^V$(reactant).

It is apparent that inclusion or omission of the carbonyl group in both structures of the model is immaterial, since the index value would be the sa.ne. It is desirable to include the carbonyl group since the chemist associates this feature with an ester and an acid.

By convention, the bond indexes are computed to four decimal places and summed to three places. The $\Delta^1\chi^V$ values are computed to three places with a plus or minus sign.

It is possible to calculate a reaction index from a general reaction model including both products of ester hydrolysis:



The same index of opposite sign would encode the general reaction of ester formation from alcohol and acid. This single number, computed from the general reaction model, is the code under which information may be stored or retrieved for the event described by the model.

It is not possible to claim that every general reaction will have a unique reaction connectivity index. Our intuition is that redundancies would be rare and that "false drops" in the retrieval process would probably be minimized when terms appropriate to the query were coordinated with the connectivity value.

Variations in the adoption of the general reaction model could be introduced to permit a more detailed refinement of reaction classification. Thus, entry 5 in Table II describes acylation of any amine. The amine is nonspecified by using two R symbols for carbon or hydrogen atoms. A more specific index may be calculated for an amine acylation reaction involving only primary amines. This would have a calculated reaction connectivity index of 0.877. This could index a separate file or could be a subfile under the general acylation index of 0.822.

Another variant might be the use of a reaction connectivity index computed from a general reaction model reflecting two products, as, for example, ester hydrolysis to both acid and alcohol. From such a general reaction model, a $\Delta^1\chi^V$ would be computed.

CONCLUSION

As described above, the value for a specific reaction type can be quickly calculated; it is entirely dependent upon the structure; it is universal for all reactions of the same type; it is sufficiently different from the values of other reactions to provide an unambiguous identification.

We believe the generation of a reaction index file, for use with large chemical information and/or data files, would greatly enhance the ability of the information specialist to refine retrieval. The appropriate value(s) for different reaction(s) could be incorporated as a searchable field in computer-accessible files just as the CAS Registry number has been ıncorporated.

REFERENCES AND NOTES

(1) Gluck, D. J. "A Chemical Structure Storage and Search System Developed at Du Pont", *J. Chem. Doc.* **1965**, *5*, 43–51.
(2) Beach, A. J.; Dabek, H. F., Jr.; Hosansky, N. L. "Chemical Reactions Information Retrieval from Chemical Abstracts Service Publications and Services", *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 149–55.
(3) Valls, J. In "Computer Representation and Manipulation of Chemical Information"; Wipke, W. T., et al., Ed.; Wiley, New York, 1974.
(4) Lynch, M. F.; Harrison, J. M.; Town, W. G.; Ash, J. E., Eds. "Computer Handling of Chemical Structure Information"; Macdonald and American Elsevier, New York, 1971.
(5) Gelberg, A. J. "Rapid Structure Searches via Permuted Chemical Line Notations. IV. A Reactant Index", *J. Chem. Doc.* **1966**, *6*, 60–1.
(6) Weygand, C. "Organisch-chemische Experimentierkunst", 4th ed.; Barth, Leipzig, 1970.
(7) Theilheimer, W. "Synthetic Methods of Organic Chemistry", 26 volumes; Karger, Basel and New York, 1946–1972.
(8) Vleduts, G. E.; Mishchenko, G. L. *Tr. Vses Konf. Inf. Poisk. Sist. Autom. Obrab. Nauchno-Tekh. Inf., 3rd, 1966*; *Chem. Abstr.* **1969**, *70*, 43991m.
(9) Hudrlik, P. F. "Reaction Index" in "Survey of Organic Synthesis", Buehler, C. A.; Pearson, D. E., Eds.; Wiley: New York, 1977; pp 1001–18.
(10) Hendrickson, J. B. "A Systematic Organization of Synthetic Reactions", *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 129–36.
(11) Ziegler, H. J. "Roche Integrated Reaction System (RIRS). A New Documentation System for Organic Reactions", *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 141–9.
(12) Bersohn, M.; MacKay, K. "Steps toward the Automatic Compilation of Synthetic Organic Reactions", *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 137–41.
(13) Willett, P. "Computer Techniques for the Indexing of Chemical Reaction Information", *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 156–8.
(14) Kier, L. B.; Hall, L. H. "Molecular Connectivity in Chemistry and Drug Research"; Academic Press, New York, 1976.

Computer-Assisted Synthetic Analysis. Long-Range Search Procedures for Antithetic Simplification of Complex Targets by Application of the Halolactonization Transform

E. J. COREY,* ALAN K. LONG, JOHANN MULZER, HARRY W. ORF, A. PETER JOHNSON,[†] and ALAN P. W. HEWETT

Department of Chemistry, Harvard University, Cambridge, Massachusetts 02138

Received July 23, 1980

A major problem in computer-assisted synthetic analysis is the development of techniques for long-range searches directed toward retrosynthetic simplification of a complex target. The approach used in the Harvard program, LHASA, combines multistep antithetic analysis with an efficient method for prescreening targets to eliminate less promising routes and to limit the number of precursors generated. These techniques are illustrated for the halolactonization search in LHASA, and the method for preevaluation of pathways is described in detail. A number of chemical examples are included.

The most elegant and simplest chemical routes for the synthesis of complex molecules are to be found by approaches which combine several lines of analysis so as to allow the concurrent application of several powerful strategies. This collection of strategies generally includes the overarching and most general principles of synthetic design such as the rigorous application of antithetic (retrosynthetic) search, coupled antithetic–synthetic tree generation, the use of self-reinforcing cycles of perception and analysis, and the characterization of unique features of the problem or target structure. In addition,

[†] Department of Organic Chemistry, The University, Leeds LS2 9JT, England.

0095-2338/80/1620-0221$01.00/0 © 1980 American Chemical Society