Systematic Synthesis Design: the SYNGEN Program

James. B. Hendrickson
Department of Chemistry
Brandeis University
Waltham, MA 02254-9110

Abstract

The aim of this presentation is to show that the design of organic synthesis routes can be systematized and yield reasonable results if: a logical, rigorous and digital schema is used to generalize organic structures and reactions; and a clear and stringent criterion is applied to severely restrict the output.

Introduction

Although the synthesis of organic compounds is as old as the science itself, there has not been until very recently any rigorous protocol for synthesis design - for formulating the best routes to a given target molecule. A synthesis route is a sequence of reaction steps from available starting materials to the target compound, and the number of possible routes is usually in the millions for targets of interest. The traditional approach to synthesis design was to work backwards from the target structure, deriving sequences of reactions and intermediate structures stepwise backwards until available starting materials had been found.

Until the advent of the computer, synthesis design was viewed as an art form, derived more from inspiration than from system, since the number of possible routes is so enormous. Thus, the variety of organic reactions available is such that there may easily be 30 "last reactions" which result in the target molecule from intermediates one step back. Then there will be 30 more possible reactions to make each of those 30 intermediate molecules, so that in five steps back from the target there are 30^5 = 24 million routes possible, almost none yet complete back to real starting materials! Many syntheses are 10-20 steps long.

The first computer approaches simply applied the huge work capacity of computers to growing these routes stepwise backwards from the target, but left the choice of the best routes to the operator. To avoid overload in practice the operator must delete almost all growing routes early in the search before he knows whether they might later prove to be superior. The routes produced will then owe more to the preconceptions of the user than to the computer. This mechanical approach basically leaves the intellectual work of decision-making entirely to the operator, i.e., an interactive mode commonly labeled CAOS (computer-assisted organic synthesis). This still remains the central basis of almost all other programs to date such as LHASA (Corey *et al.* 1985), SECS (Wipke *et al.* 1977), SYNCHEM (Gelernter *et al.* 1990), and others (Barone & Chanon 1986).

Our basis instead (Hendrickson 1971, 1986, 1990; Hendrickson & Toczko 1989) was to ask the computer to do most of the intellectual work, independent of the operator. Because of the huge search space involved, any development of a logical protocol for systematic synthesis design demands first two basic definitions:

(a) a rigorous system of description for molecules and reactions to simplify and generalize the vast field of chemical information, and to systematize it compactly and digitally for facile computer manipulation;

(b) focused and very stringent criteria to severely restrict the selection of only the best syntheses from the enormous number of possibilities.

These preparations are not trivial, but they have not been applied in the other synthesis design programs.

System of Characterization

In our SYNGEN program the system of description for molecular structures is first defined by a dichotomy between the skeleton and the functional groups: the primary or more general part of the molecule is the underlying backbone skeleton (usually linked carbons atoms); the functional groups are then the reactive centers (π–bonds and attached heteroatoms) located at specific sites on that skeleton.

A compact digital format to generalize descriptions is then applied. Each skeletal carbon is defined as having four synthetically important kinds of attachments, skeletal and functional. The numbers of these on each carbon are then used to describe any molecule numerically. For fast comparison and numerical ordering of compounds we derived a unique binary number identification (Hendrickson & Toczko 1983, 1984) for the skeleton first, followed by a listing of the numbers for functional group attachments on each atom of the skeleton. This generalization allows simple, rough descriptions of generalized families of compounds to be used rapidly first, refining only the few selected ones later to finer chemical detail.

The same system describes all possible reaction changes by defining a *unit reaction* as a unit exchange of these attachment types at each changing skeletal carbon. The possible unit reactions can all be generated *de novo* and they provide a complete roster of the net structural changes that may be applied to any molecule to generate its precursor (or product) in any reaction (Hendrickson & Sander 1995).

Guiding Criteria

The central criterion for the best routes is *economy:* the shortest and most efficient with lowest overall cost. The shortest routes are those which start with real, available starting materials and have the fewest reaction steps. Since the most efficient routes are convergent (Hendrickson 1977), only convergent syntheses were sought, i.e., those in which two molecules are joined in the last step and these two are themselves each separately constructed by joining two smaller ones.

The number of reaction steps is minimized by requiring that each step create a skeletal bond, i.e., using only skeletal construction reactions, with no extra steps taken to repair incorrect functional groups, since skeletal construction is the essential part of any synthesis.

Protocol

(1) Skeletal Dissection

These definitions and criteria direct the formulation of a systematic protocol for quickly pruning down the huge search space and generating the best syntheses. The first simplification for pruning is to focus only on the skeletons of the involved compounds, ignoring the functional groups and so vastly reducing the detail that must be considered. This reduces the problem to finding the shortest way to assemble the target skeleton from

the skeletons of available starting materials kept in a catalog.

This major simplification derives from the observation that syntheses proceed from small starting molecules to larger target molecules. Hence it is obligatory that skeletal bonds be constructed in the assembly of the target. The key to any synthesis route then is its *bondset,* the set of skeletal bonds that are made in assembling the target from its starting materials. We may generate all possible bondsets by systematically cutting the target skeleton, but we limit them to the convergency criterion above.

To find the most efficient, convergent routes, the target skeleton is cut into two pieces, and each of these two into two again. This affords four starting skeletons to assemble the target. The steps are minimized by stopping at these two levels of dissection and demanding that the four starting skeletons so derived must then be found in the catalog of available starting materials. For a C_{20} target the four starting materials will average five carbons; as skeletons of this size are abundant in the catalog we will find many acceptable bondsets in just these two levels of skeleton cuts. Further dissection implies more steps, hence less desirable routes.

This initial focus on the primacy of the skeleton greatly delimits the millions of possible routes by finding first only the shortest skeletal assemblies (bondsets) of the target from real starting skeletons. However, each of these bondsets will subsequently expand to many routes differing in the detailed chemistry of the functional groups; these routes may then be generated in a second pass through each allowed bondset. Computer implementation is simplified since each bondset represents an independent subdivision of the overall synthesis tree and so can be expanded separately.

(2) Functional Groups

The required functional groups may now be generated successively around the construction of each skeletal bond of a bondset, taken in order. To do this, the roster of possible construction reactions is applied in reverse to the target and then on to successively generated intermediates, converting the functional groups at each defined bond to those of its reaction precursor. This process will end with the generation of the actual starting materials, now with their functional groups as well as their skeletons. These actual starting materials must all be found in the catalog or the generated route is rejected. In this way we will generate only the shortest routes, as sequences of construction reactions only with no extra steps to repair functional groups. These then must be the "ideal syntheses" sought; it often surprises chemists how many there are!

These concepts afford a rapid and massive pruning of the huge field of possible syntheses. The search is focused at the outset on a catalog of real starting compounds and so avoids searching blind as in the other programs. The routes generated will be only the shortest convergent ones from no more than four real starting compounds available in the catalog. They will use no more steps than just the sequential construction reactions necessary to join the skeletal bonds of the bondset and also arrive at the final functional groups of the target. Such stringent criteria result in a relatively small number of ideal synthesis routes.

The Process in SYNGEN

The protocol for SYNGEN following these tenets is conceptually simple and has two parts as described, the first part being the skeletal dissection. To dissect for convergent synthesis, the program cuts the target skeleton into two intermediate skeleton pieces all ways, cutting one acyclic bond or two cyclic bonds in the same ring. This is the first level of dissection. For the second level the two pieces are then each cut in two again all ways, yielding sets of four starting skeletons; the routes so generated are only retained if all four starting skeletons are found in the catalog. Therefore, for each bondset no more than six bonds will be cut, implying a maximum of six construction reaction steps in each synthesis.

Having defined with these bondsets which skeletal bonds are to be constructed, the second part of the protocol generates the functional groups. Starting with the functional groups of the target, SYNGEN works backwards through each bondset separately, applying to each successive dissected bond the roster of all possible construction reaction changes. At first level this then generates the functionalized intermediates; at second level it generates the actual starting materials; all are then compared with the catalog. From this generation process, the complete routes from real starting materials may be assembled. By definition they must be the shortest routes, and they are generated independent of operator intervention. The only database employed here is the starting materials catalog (presently some 6000 compounds); no internal reaction database is needed for the generation of reactions.

The construction reactions so generated may however be tested for chemical viability by direct comparison with literature precedents. Using the same generalized format for reaction description, we have also created a program named COGNOS for retrieval from reaction databases of reactions matching an input query (Hendrickson & Sander 1995). We can now tie the index for the half million reaction entries in COGNOS to SYNGEN, finding both the frequency and average yield of the closest matches to each construction reaction as it is generated by SYNGEN. This has the effect of validating these generated reactions with practical references.

In actual use SYNGEN starts with a facile input drawing of the target on the screen, and then proceeds to process it in batch mode while other targets are being entered or results examined. It usually takes less than five minutes to generate all viable reactions and starting materials, combining them into complete routes. The output then shows a summary screen with the numbers of viable bondsets, starting materials, intermediates and reactions for each level of dissection, as well as the total number of whole routes generated. These may then be separately examined. The whole routes themselves are displayed, one per screen, arranged in order of overall cost. The cost is derived from starting material costs amplified by the yield loss at each step, and adding also a cost for doing each step so that the number of steps is also important in the overall cost figure.

Conclusion

The unique distinction of SYNGEN from other synthesis design programs, such as LHASA, SYNCHEM, etc., lies in this use of a logical protocol which strikes back into the synthesis tree to find and focus on those real starting materials which are the fewest steps from the target, thereby insuring the shortest and most efficient syntheses.

The protocol works effectively because of its system of generalized description which allows massive preliminary pruning by skeleton before dealing with functionality details. Furthermore, the protocol is not interactive; it does not depend on user choices and so systematically applies consistent criteria throughout to generate all the best routes. Of course the user will ultimately interact with the output to select a few best routes, but only after a full but relatively small set has been generated from these stringent criteria.

Acknowledgments. This work has been made possible for some years by the support of the National Science Foundation.

References

Barone, R.; Chanon, M. **1986.** Computer-aided Organic Synthesis. In "Computer Aids to Chemistry", Vernin, G.; Chanon, M., eds., Ellis Horwood, Chicester, chap. 1: a general review with almost 50 programs, but not all are directly involved with synthesis design .

Corey, E. J.; Long, A. K.; Rubinstein, S. D. **1985.** Computer-assisted Analysis in Organic Synthesis. *Science* 228: 408.

Gelernter, H.; Rose, J. R.; Chen, C. **1990.** Building and Refining a Knowledge Base for Synthetic Organic Chemistry.. *J .Chem. Inf. & Comp. Sci.* 30: 492.

Hendrickson, J. B. **1971.** A Systematic Characterization of Structures and Reactions for Use in Organic Synthesis. *J. Am. Chem. Soc.* 93: 6847.

Hendrickson, J. B. **1977.** Yield Analysis and Convergency. *J. Am. Chem. Soc.* 99: 5439

Hendrickson, J. B. and Toczko, A. G. **1983.** Unique Numbering and Cataloguing of Molecular Structures. *J. Chem. Inf. & Comp. Sci.* 23: 171.

Hendrickson, J. B.; Grier, D. L.; Toczko, A. G. **1984.** Condensed Structure Identification and Ring Perception. *J. Chem. Inf. & Comp. Sci.* 24: 295.

Hendrickson, J. B. **1986.** Approaching the Logic of Synthesis Design. *Accts. Chem. Res.* 19: 274.

Hendrickson, J. B. and Toczko, A. G. **1989.** Systematic Synthesis Design: The SYNGEN Program. *J. Chem. Inf. & Comp. Sci* 29: 137.

Hendrickson, J. B. **1990.** Organic Synthesis in the Age of Computers. *Angew. Chem. Intl. Ed.* 29: 186.

Hendrickson, J. B. and Sander, T. **1995.** COGNOS: A Beilstein-type System for Organizing Organic Reactions. *J. Chem. Inf. & Comp. Sci.* accepted for publication.

Wipke, W. T.; Braun, H.; Smith, G.; Choplin, F.; Sieber, W. **1977.** SECS-Simulation and Evaluation of Chemical Synthesis: Strategy and Planning. *ACS Symposium Series* #61:97.