Pure & Appl. Chem., Vol. 61, No. 3, pp. 589–592, 1989.
Printed in Great Britain.

Systematic synthesis design: the SYNGEN program

James B. Hendrickson* and A. Glenn Toczko

Department of Chemistry, Brandeis University, Waltham, MA 02254-9110

Abstract - We develop here a systematic approach to locating within the vast "synthesis tree" of possible routes to any target molecule just those few syntheses which are optimal with respect to efficiency of assembly. The approach has two phases. In the first the target skeleton is dissected all ways which produce convergent assembly plans from starting skeletons available in a catalog. In the second the chemistry is generated to produce all paths of construction reactions only, from real starting materials, for each such plan. The SYNGEN program to execute this search is described.

The theory of organic synthesis design has had little systematic development; it remains in practice an art in the midst of a science. There is not even any clear definition of a "best synthesis plan". In order to examine the problem systematically we have taken *economy* to be the criterion and measure of a good synthesis, economy defined as minimizing cost and time (with time equated to number of steps). Designing optimal syntheses is a problem simply because there exists such an enormous number of possible synthesis plans or routes for any given target molecule. It is conceptually a simple matter to generate backwards from the target all possible intermediates one step back, and for each of these all intermediates one more step back, and so to continue back until the intermediates have become simple enough molecules to be called starting materials. However, this generation produces huge numbers of routes: if there are 30 possible different reactions which result in the target, and 30 each for its immediate precursors, there are 30^n possible routes of n steps, i.e., about 24 million 5-step syntheses. Hence the necessity for selection is paramount, and evidently very stringent criteria will be required to reduce these combinations to a few best synthesis plans. Our intent is to take a fresh look at this "synthesis tree" of choices and to define ways to order and simplify this search tree to find all the shortest, most efficient synthesis plans from available (and inexpensive) starting materials, thus following the criterion of economy.

There is a simple dichotomy in molecular structures: of the skeleton (carbon framework) on one hand; and the appended functionality on the other. This dichotomy is reflected in reactions also: between the construction reactions, which create skeletal bonds; and refunctionalizations, which transform the functionality without altering the skeleton. In the broadest view synthesis is a skeletal concept since it commonly creates a large, complex target molecule from small simple starting material molecules. In a survey of syntheses we find the average size starting material to be only three carbons (incorporated into the target); hence an average synthesis plan will create one in every three or four of the target skeletal bonds. It is therefore evident that construction reactions are central to synthesis, are indeed obligatory. Refunctionalizations in principle are not, although the average synthesis includes twice as many refunctionalizations as constructions.

From this perspective, then, we should first examine the most efficient ways to assemble the target skeleton from available starting material skeletons. Hence we examine ways to dissect the given target skeleton, cutting the fewest bonds, into skeletons to be found in a catalog. It may be noted that this first step greatly simplifies the search tree by considering only molecular skeletons (there are only 13 acyclic skeletons of six carbons or less but thousands of available starting materials as their functionalized variants).

We define that set of skeletal bonds that require construction in any synthesis as a *bondset*. Removal of the bonds of any bondset from the target skeleton defines the starting material skeletons. An ordered bondset is one for which the order of constructions of its bonds is also defined. An ordered bondset in effect defines the *assembly plan* for the target skeleton, i.e., the sequence of constructions to assemble it. The simplest overall description of any synthesis is its ordered bondset. The

number of possible bondsets, however, is not trivial. For a skeleton of b bonds with λ bonds cut there are possible bondsets, and the number of orders for each one is $\lambda!$ Thus the number of ordered bondsets possible is $b!/(b-\lambda)!$. In our illustration, the Torgov-Smith synthesis of estrone (with b=21 bonds) the bondset has λ=5, which means there are 20,349 bondsets of 120 orders each, or almost 2½ million possible assembly plans. An "average" synthesis of estrone (C_{18}) would require six average three-carbon starting materials. For this combination there are over 100 billion possible assembly plans. A dramatic illustration of the vast variety of possible synthesis routes is that of the "total" synthesis of cortical steroids, from "coal, air and water," i.e., one-carbon starting materials. Here there is only one bondset, that with all bonds cut: $\lambda = b = 24$. The number of possible assembly plans for this skeleton is then $24! = 6\times10^{23}$, which implies that, to make a mole of cortisone, each molecule may be made by a different route! And this is without consideration of any chemical reaction detail.

Fortunately, not all skeletal assembly plans are equally good, and the most efficient are relatively few, especially when we focus the search on available starting material skeletons. The simplest, and most common, assembly plan is a linear one, in which starting materials are sequentially, i.e., serially, linked to the growing skeleton. By contrast a fully convergent plan is a parallel sequence, in which the starting materials are all first linked together in separate pairs and then these paired intermediates are themselves joined pairwise, and so on until the target is finally made by linking the penultimate pair of intermediates. There are hybrid plans between linear and convergent and they grade from the least efficient linear plans to the most efficient convergent ones. This efficiency can be measured either by the sum (S) of the steps each starting material unit must pass through (taking a loss at each step) or by the total weight (W) of all starting materials required, at some average yield per step. For K starting materials there will be $(K-1)$ steps necessary to join them in any plan. For K=8 (7 steps), S=35 steps for linear and only 24 steps for convergent plans, while W=24 for linear and only 16 for convergent plans (for unit-weight starting materials). Thus the convergent plans for the same number of steps are more efficient than the linear ones by about a third; when the plans are extended to 20-30 steps with refunctionalizations the discrepancy increases considerably, to factors of five or more.

The fully convergent plans are few and easy to derive. The procedure is to cut the target skeleton into two pieces and then cut each piece into two pieces again, all possible ways. This can be repeated, but if we elect to stop there at two levels of cuts, each second cut will have produced a set of four starting material skeletons and a corresponding ordered bondset and assembly plan. For a C_{20} target these starting material skeletons will average five carbons each, and the experience with our organized starting material catalog is that a wide variety of functionality is still available for five-carbon skeletons, but falls off rapidly for larger ones. Hence we have a basis to expect a fair set of viable syntheses from real starting materials with assembly plans so generated. These fully convergent assembly plans will have bondsets of no more than λ=6 and must be the most efficient plans for economy. Now instead of over 40 million plans for the estrone skeleton at $\lambda \leq 6$ (all of which can be linear), the fully convergent plans, with all starting material skeletons available from our catalog (~6000 compounds) number only 875, and for many of these the necessary chemistry will not be found. Accordingly, this represents a very stringent selection criterion.

In this way we create all possible fully convergent assembly plans (ordered bondsets) and must now generate the necessary chemistry to construct these bondset bonds in the right order from real starting materials. Since economy is our criterion and since only construction reactions are truly obligatory, then the shortest synthesis will be a sequence of contructions only, without any refunctionalization reactions. Such a sequence is labelled an "ideal synthesis". These are quite rare in practice but constitute a search goal incorporating the very stringent criterion which is required if we are to have only a few routes selected from the tree. An ideal synthesis can be described as one wherein the chemist takes from the shelf two starting materials bearing the right functionality to initiate a construction reaction to join them; that product in turn has the right functionality to initiate construction of the next bondset bond, and so on through the construction of each defined bond, until the target skeleton is all constructed and the correct target functionality is the natural consequence of the construction sequence. We shall look to define such ideal syntheses in the retro direction back from the target, through each convergent bondset in turn, generating the chemistry from both the bond to be constructed and its surrounding functionality. This process will ultimately arrive at the necessary starting material functionality. The starting materials so generated are looked up in the catalog, where they must be found in order to validate any route.

A computer can seek the requisite chemistry in two ways: either look it up in a comprehensive database of reactions; or generate all possible reactions and test them mechanistically. Both have problems. A database must be huge and so is cumbersome, difficult to assemble, variable in quality and never complete, and also it will create

no new chemistry. The generation requires no database but must derive from a system of reaction description which can create every possible net structural change in some abstract or generalized format. This procedure will produce everything including new chemistry, but will produce too much output including a fairly high proportion of unrealistic or unacceptable reactions. General mechanistic rules can much reduce this non-viable output but will also delete a small number of unusual but viable reactions.

We elected to use the second approach for its generality and simplicity: to generate all reactions and then test their mechanisms. This requires a system of description for structures and reactions which is rigorous, general and fundamental and which abstracts structural data, coalescing trivial distinctions. For the computer it must also be digital. Our system describes four synthetically significant kinds of attachment for any carbon: H for hydrogen (or other electropositive elements), R for σ-bonds to other carbons, Π for π-bonds to carbons and Z for bonds (σ- or π-) to electronegative heteroatoms (e.g., N,O,X,S,P). The number of each kind of attachment then is denoted by h,σ,π and z, respectively, with a sum of 4. That this system is fundamentally sound is shown by calculation of the oxidation state at each carbon as $x = z-h$; the sum of such calculations over the involved carbons in any reaction gives the correct oxidation state change.

Structures are simply described: the σ-values of the carbons represent their skeletal level; their functionality is then given by the two digits, $z\pi$, and h is found by subtraction $h = 4-(\sigma+z+\pi)$. A molecule of numbered carbons is then simply annotated by a $z\pi$-list of the carbons in order. Reactions are described in terms of unit reactions at each carbon, i.e., unit exchanges of attachments. These may each be described with two letters: the first the attachment bond made; the second the bond broken. With four kinds of attachment there are 4x4=16 unit reactions at one carbon, e.g., HZ a simple reduction, ZH an oxidation, ΠH an elimination of H to form a π-bond, etc. The reactions of interest here are the constructions, RH, RZ and RΠ; and any unit reaction involving ±R or ±Π at one carbon must have a coupled unit reaction at an adjacent carbon. This abstract description of reactions allows the generation of all possible net structural changes and indeed can be applied to create a "Beilstein system" for cataloguing reactions, i.e., with a place in the catalog for any reaction whether known or not.

In order to generate a construction reaction for a bondset-designated bond, the strands of carbons out from that bond at each end are described with $z\pi$-lists of their functionality and to this can be applied $\Delta z\pi$-lists characteristic of each possible construction mode, in order to generate the substrate $z\pi$-list from that of the product (or vice-versa). Furthermore, rough, general mechanistic tests for each construction reaction can be simply applied as tests of the digital values of h,σ,π or z on each of the carbons proximal to the construction bonds. These tests can be used to delete the most obviously non-viable reactions when generated, or to mark others as mechanistically uncertain for later inspection by the chemist.

The SYNGEN (Synthesis Generation) computer program was designed to apply the foregoing logic to any input target. In the first phase, bondset generation, only the target skeleton is examined. For the first level, this is cut into two skeletons. For the second level, each one is cut again into two more. All four starting material skeletons so generated must now be found by consulting the catalog. Each successful operation creates a convergent ordered bondset. In the second phase SYNGEN goes through each bondset successively, generating intermediates (as $z\pi$-lists) back from the target for each bond in order, using every possible construction change (as $\Delta z\pi$-lists) and deleting nonviable cases by mechanism tests for each reaction. When all the bonds in any bondset have been passed through in order, the program will have generated the $z\pi$-lists on the starting material skeletons, thus defining real starting materials. A valid synthesis plan, or sequence, is generated if those starting materials are found in the catalog.

The SYNGEN program is written in about 50,000 lines of FORTRAN with a catalog of about 6000 starting materials as data. Target structures are input via a very facile, rough drawing module which then normalizes them to neatly drawn structures. The program then processes the target, by the protocol described above, without operator intervention, and stores its output for later viewing. On our MicroVAX computer this usually requires less than three minutes operation. The output is summarized in terms of four categories: bondsets, starting materials, intermediates and reactions, for each of the two levels of cuts. These may total as little as ten reactions or as many as a thousand. For this reason there is a flexible menu of options for selecting from the output, as necessary to focus on only a few routes. Each category may be examined separately, fully displayed graphically. Selections (retain/delete) can be made from any category and further in terms of other bases such as starting material cost, number of reactions in sequence, uncertain mechanistic viability or regiochemistry, removal of chemically equivalent reactions, etc. In this way it is possible to prune down the output, select the best options or focus on particular variants.

The vast number of possible synthetic routes in the theoretical synthesis tree clearly demands stringent criteria of selection as the central focus of any design program which aims to assess all possibilities and to locate only a manageably small set of optimal routes. While the SYNGEN program certainly creates short, decisive syntheses with its tight protocol, sometimes these are not chemically satisfactory or for various reasons good routes are missed. In order to loosen somewhat these restrictions we are reexamining refunctionalization reactions. In practice these are twice as common as constructions, whereas SYNGEN in effect excludes them completely in order to minimize steps. It is true that the abstracted nature of functionality description contains some implicit refunctionalizations, apparent to the chemist, such as heteroatom group changes, protection, chiral activation, etc. However, refunctionalization may be valuable before the construction sequence, to repair a large key starting material available in the catalog, or during or after the construction sequence to alter the final functionality to that of the target. The latter case is most apparent when dummy functional groups are used to initiate skeletal construction and then removed at the end, leaving no trace in the target, most obviously in the synthesis of saturated hydrocarbons or targets with large central hydrocarbon regions.

Our digital description allows a simple calculation of the number of steps of refunctionalization, the "reaction distance", between any two structures. Therefore, starting material repair is possible by calculating the distance between available catalog compounds of the same skeleton as a generated starting material. For valuable starting materials of large skeleton, 1-2 steps of refunctionalization is presently allowed by SYNGEN. For alterations during and after construction we are creating a new program to generate syntheses accepting 1-2 steps of refunctionalization, in a *forward* rather than retro direction. The bondsets are created as before and then all the catalog starting materials are examined for their reaction distance to target and coupled pairwise all ways in the forward direction, deleting all combinations which diverge in distance after construction. By continuing to keep only those which on construction move toward the target functionality, we should avoid the otherwise explosively unfocussed combinatorics of such forward generation.

Finally in order to enhance its practical value we are currently incorporating into SYNGEN a procedure whereby it can look up and display on request any literature precedents for the reactions it generates, from reaction databases such as SYNLIB, REACSS or ORAC. This should serve both to demonstrate that the generated chemistry is realistic and also to lead the operator to procedures for practical execution in the laboratory.

Basically the SYNGEN output shows synthetic routes, from available starting materials to the input target. The program creates many sensible, viable routes; it succeeds in generating known syntheses and also finds new routes of equal efficiency, not precedented in the literature. The strength of the program lies in the fact that it does produce all possible routes within specifically defined criteria: all convergent skeletal assemblies from two levels of cuts using a sequence of construction reactions only from available starting materials. Thus the operator knows exactly what kinds of routes SYNGEN will produce, that it systematically finds all within this definition, and that they are in principle the shortest and most efficient syntheses. Therefore, the program can provide an optimal set of routes against which to compare other synthetic ideas, in effect a set of standards for practical synthesis planning.

The authors wish gratefully to acknowledge the dedicated effort and considerable computer expertise of our associates on the synthesis design project: Dr. David L. Grier, Dr. Zmira Bernstein, Ms. Ping Huang, Mr. Todd Miller and Mr. Camden Parks. We also thank the National Science Foundation and Eastman Kodak Company for financial support.

REFERENCES

References and full discussion of these developments will be found in:
J.B. Hendrickson, Accts. Chem. Research, 19, 274 (1986).