Pure & Appl. Chem., Vol. 60, No. 11, pp. 1563–1572, 1988.
Printed in Great Britain.

Synthesis design logic and the SYNGEN (synthesis generation) 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, as well as some current developments for program expansion.

The literature of organic chemistry contains virtually nothing on the logic involved in the design of a synthesis. In a science as logical as organic chemistry this is surprising: synthesis design in practice is an art in the midst of the science. Against what theoretical background can one assess how good a synthesis is, whether another route might be shorter or cheaper? Indeed, what is the test of a good synthesis, or even the definition of a good synthesis? And yet most chemists would probably agree that there is some intrinsic underlying logic possible for the selection of an optimal synthetic route to any target molecule. Our intent here is to take a fresh look at this problem of synthesis design[1].

THE SYNTHESIS TREE

As a first step we look at the "synthesis tree", a graph of the process, of the successive steps in all possible synthesis routes for a single target (Figure 1). In the graph the lines are reaction steps, proceeding from left to right, the points are compounds: starting materials (heavy dots); intermediates; and the target itself (T). In general molecular complexity diminishes to the left away from the target, as for the classical "total synthesis from coal, air and water". Since nature and the chemical industry provide many syntheses of relatively simple compounds in abundant quantity out from the left side of the tree, we are thereby supplied with many substances marked as starting materials. The *levels* of the tree, shown at the top, represent the *reaction distance*, the number of steps of an intermediate or starting material from the target on any synthetic sequence or route. The decreasing yields (at 80% per step) with reaction distance are shown below. Two particular simple routes from starting materials to the target are picked out as heavy lines. Such depictions of synthesis routes are called *synthesis plans*.

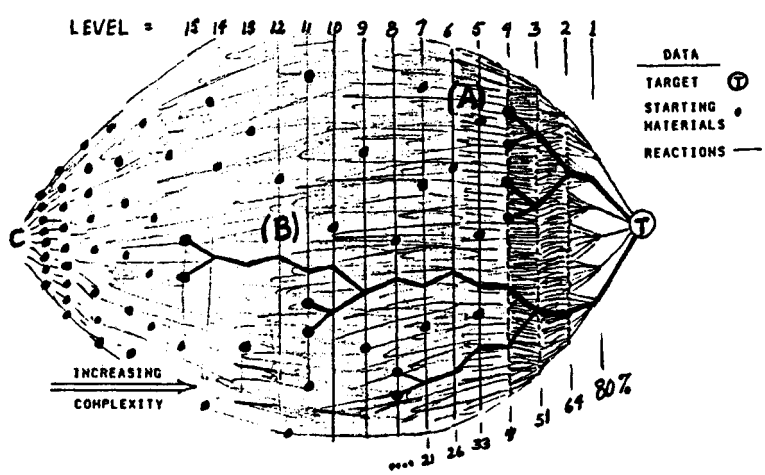


Figure 1. The Synthesis Tree

For any target of reasonable complexity the synthesis tree is certainly far larger than is generally appreciated, and this is probably the underlying reason why the logic of route selection, of synthesis design, has been so little addressed. The size of the tree can be indicated by observing that, if there are 30 possible last reactions to the target and also to each of its near precursors, still similarly complex, there will be 30^n possible routes from a level n steps back, i.e., 30^5 = 24 million routes from only five steps back. Furthermore, these combinations will vastly increase for longer routes, and most syntheses are much longer. The data available to construct and assess the synthesis tree for any target are: the structure of the target; a catalog of the structures of all available starting materials; and a catalog of all organic reactions.

It is conceptually relatively simple to create the tree for a given target by generating backwards all possible reactions to intermediates one step back (level 1), and for each of these one more step back, and so to continue back until all the generated precursors have become simple enough to be found as starting materials. The millions of routes this procedure will generate make it clear that the essential point of synthesis design is not so much generation of routes as the selection, from so many, of the few optimal routes one might want for the laboratory. Indeed, whatever the selection process, there is serious concern as to whether all meaningful precursors would be generated by the stepwise backwards approach. In many syntheses, dummy functional groups are used to assist in constructing the target and then removed, leaving no trace in the target structure. This is obvious in the synthesis of saturated hydrocarbon targets, less apparent elsewhere, as in the last step of estrone synthesis in Figure 2, which might not be found by generation backwards. Indeed, complete retro generation of the full tree from the target nominally demands reinstatement of all possible functional groups at all hydrocarbon sites, thus vastly increasing the already enormous size of the tree.

Instead of generating the whole synthesis tree and then making selections, our approach will be to define first the criteria for optimal routes and let these serve to select first only a few small segments of the tree to examine, those that contain the defined optimal routes. These can then realistically be refined and assessed completely. To do this we must first simplify the tree drastically to its essential outlines, so as to see clearly which segments suit the criteria. There is an analogy with maps here: the larger the mapped space, the more details are coalesced or omitted - whole towns become dots. But when a desired place is located, a more detailed map of that area restores the detail. With full structures and reaction descriptions our synthesis tree is too detailed and complex a map for rational exploration. It is also so large that very stringent constraints will be required to simplify it and select only a few best synthesis plans.

CRITERIA AND COMPARISON

Hence, we must start by defining the criteria for the optimal desired synthesis, in order to focus the selection process from the tree. It will be clear directly that not all chemists will agree on the criteria, but we must assign some in order to provide a basis. We opted for a criterion of economy: of time, effort and cost of materials. Time and effort are reflected in the number of steps, so that the primary basis will be minimum steps. The cost is not only that of the starting materials (and reagents) but also a function of the order in which they are assembled. The available data do not include reaction yields since most of the specific reaction steps in the synthesis tree have never been carried out, and the possible accuracy of yield prediction is far too imprecise to be meaningful for selection. The result is that we will focus on finding all the *shortest* routes with the most efficient synthesis plans. Looking at the synthesis tree this means locating all the routes with the nearest lowest-level starting materials, i.e., finding synthesis plans like (A) in Figure 1, avoiding those like (B). Instead of just generating precursors stepwise backwards we need a logic basis to strike back into the tree to focus on the nearest available starting materials for the target.

In order to compare various synthesis plans, especially those with the fewest steps, we can calculate the total weight of all starting materials (W_{SM}) required to make a mole of target[2], assuming a common or average yield per step, by the formula $W_{SM} = \sum_i \Pi_i \times^{l_i}$ where Π_i is the molecular weight of the i^{th} starting material and l_i is its level on the synthesis tree; $\chi = 1/y$, where y is the yield per step. The synthesis plans (A) and (B) in figure 1 represent particular real syntheses of estrone taken from ref. 3. The shorter plan (A), that of Torgov and Smith, is outlined in Figure 2 with structures as well as the corresponding 5-step synthesis plan, extracted from the synthesis tree; starting materials are lettered, intermediates numbered. The synthesis requires, at 80% yield per step, about a kilogram of starting materials to make a mole of estrone. For plan (B) in Figure 1, that of Velluz, the calculated W_{SM} is 8.6 kilograms. With these guiding criteria and the calculation for comparison of plans, we can now look to simplify the enormous synthesis tree.

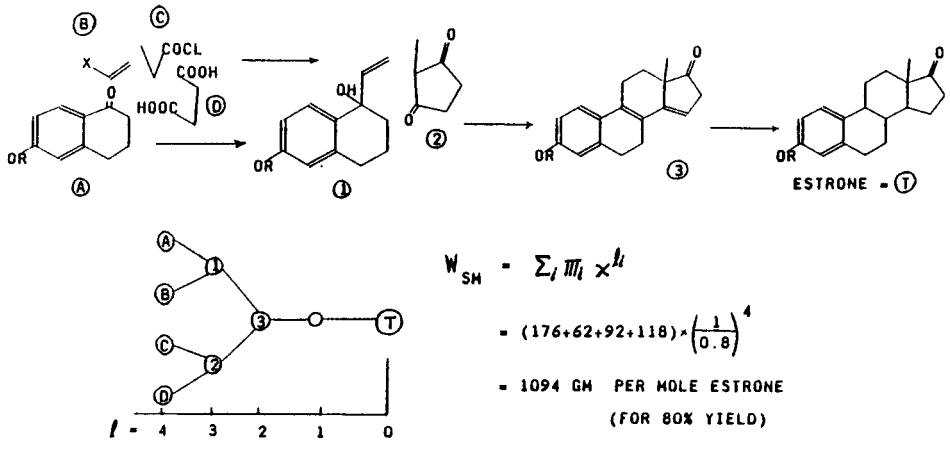


Figure 2. Synthesis Plan for Estrone

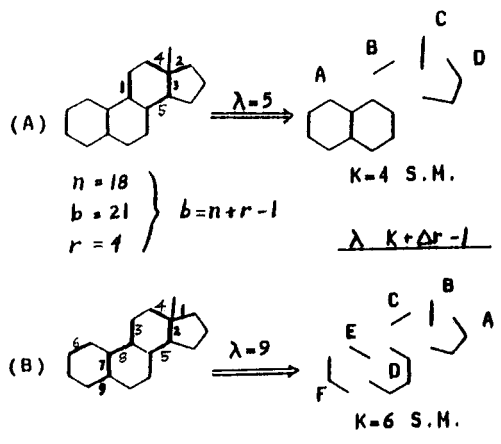
There is a simple dichotomy in molecular structures: of the skeleton (carbon framework) on one hand; and the appended functionality on the other. This dichotomy is reflected in reactions also: between the construction reactions, which create skeletal bonds; and refunctionalizations, which transform the functionality without altering the skeleton. With this distinction in mind, then, in its essence *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); an average synthesis plan will construct one in every three or four of the target skeletal bonds. It is therefore evident that these construction reactions to link small starting units are central to synthesis, are indeed obligatory. Refunctionalizations in principle are not, although the average synthesis includes twice as many refunctionalizations as constructions.

SKELETAL ANALYSIS

By looking first only at the skeletons we greatly simplify the synthesis tree: note for example that there are only 13 acyclic skeletons of six carbons or less but thousands of available starting materials as their functionalized variants. From this skeletal perspective, then, synthesis design becomes a search for the most efficient ways to assemble the target skeleton from available starting skeletons. Hence, the procedure will be to dissect the target skeleton, cutting the fewest bonds, into starting skeletons to be found in a catalog. In this way, we simplify the tree to allow a focus on the best routes for assembling the skeleton; the few best can then be refined and elaborated with chemical detail.

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. These are shown for syntheses (A) and (B) of estrone in Figure 3. An ordered bondset in effect defines the *assembly plan* for the target skeleton, i.e., the sequence of constructions necessary to assemble it. The simplest overall description of any synthesis is its ordered bondset, or its equivalent, the assembly plan, as seen by comparing with the details in Figure 2. The number of bonds in a bondset (λ) is a function of the number of starting skeletons (K) formed by its dissection and the number of rings cut (Δr), as $\lambda = K + \Delta r - 1$. The corresponding assembly plans are shown at the bottom of Figure 3. A weight, W, may be calculated for such plans but the number of carbons (n_i) is used in place of molecular weight since in this skeletal planning phase nothing is known about the functional groups. These assembly plans are basically the synthesis plans without the refunctionalization steps, i.e., only λ steps are involved. A difference in weights in assembly plans will be much magnified when the plan is expanded to include refunctionalizations, as seen with (A) and (B) whose assembly plans differ in weight by less than a factor of two while the full synthesis plans differ by eight times.[2]

Despite this simplification of the synthesis to register only skeletal assembly, the number of possible bondsets is not trivial. For a skeleton of b bonds with λ bonds cut there are $\binom{b}{\lambda}$ 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. Since estrone has n = 18 carbons, an "average" synthesis would require six average three-carbon starting materials, and so need λ = 9 bonds in its bondset. This is the case for the Velluz synthesis (B), for which we can calculate that there are over 100 billion possible ordered bondsets to assemble that skeleton in nine constructions.



Number Possible: Bondsets		Orders		Ordered Bondsets
$\binom{b}{\lambda}$	x	$\lambda!$	=	$b!/(b-\lambda)!$
20,349	x	120	=	2.44×10^6
293,930	x	362,880	=	1.07×10^{11}

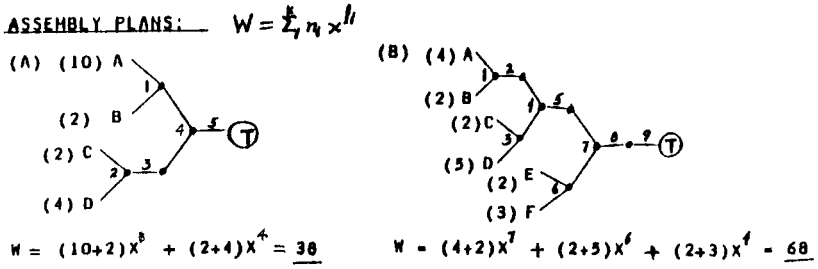


Figure 3. Bondsets and Assembly Plans

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 $24! = 6 \times 10^{23}$, 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.

EFFICIENCY IN PLANS

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.

The relative efficiencies can be assessed in two ways,[2] summarized in Figure 4. For assembling K starting skeletons there will be $(K-1)$ steps necessary to join them by any plan. For $K=8$ (7 steps) the weights (W) for each plan can be calculated as in Figure 3, using unit-weight starting materials and construction yields of 80%. Alternatively, the sum (S) of the steps each starting material must pass through (taking a yield loss at each step) is another measure of efficiency. Calculated in Figure 4, these show that the linear plan is half again more inefficient than the convergent, and the discrepancy swells to several times when the plans are extended to 20-30 steps to include refunctionalizations (compare the assembly plan to the full synthesis plan differences for estrone syntheses (A) and (B) in Figures 2 and 3). Accordingly, our economy criterion suggests that we narrow the selection of bondsets/assembly plans to fully convergent plans only.

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 to minimize steps we elect to stop there at two levels of cuts, so that 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

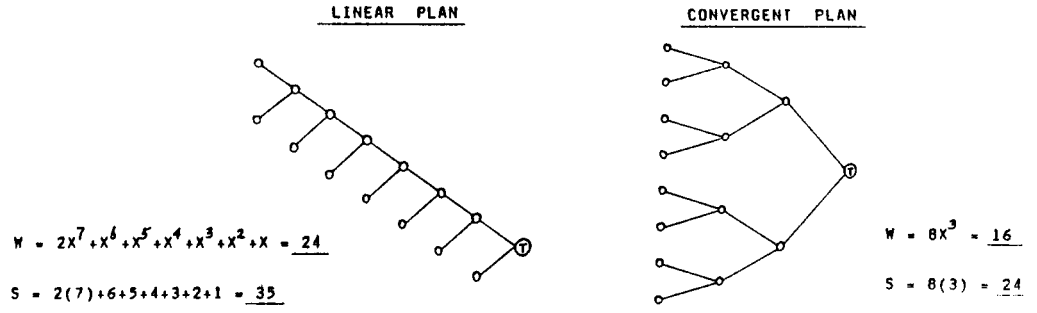


Figure 4. Comparison of Assembly Plans

fully convergent assembly plans will have bondsets of no more than λ=6 and must be the shortest and 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 with ~400 skeletons) number only 875, and for many of these the necessary chemistry will not be found. Accordingly, this represents a very stringent selection criterion.

Thus far, we have drastically simplified the synthesis tree to its bare outlines of skeletal assembly and narrowed the choices to a relatively few fully convergent assemblies from available starting skeletons found in just two levels of dissection. Now we must expand these few skeletal modes to incorporate functionality so as to assess the chemistry involved in constructing the bondset bonds in their defined order. In effect, we expand the several assembly plans to full synthesis plans. Since economy is our criterion and since only construction reactions are truly obligatory, then the shortest synthesis will be a sequence of constructions 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 also the natural consequence of the construction sequence.

This logic for parsing and selecting from the synthesis tree via criteria of economy can now be summarized in a protocol for a synthesis generation program which has two successive phases of tree simplification and selection.
(1) The skeletal phase consists of dissecting the target skeleton all ways at first level into two pieces, and each of these cut in a second level to two more. If the four starting skeletons so produced are found in a catalog, this defines a valid ordered bondset with minimal construction steps (λ), which is a fully convergent assembly plan.
(2) The functional phase now takes each ordered bondset in turn and generates in a retro direction from the target functionality the necessary functional groups of the successive intermediates, for construction of each bondset bond in order, back to starting materials. Each starting material functionality is generated in this way and looked up in the full catalog (skeleton & functional groups). Only routes with all four starting materials found are validated. It should be noted that the starting material data is important in serving first to focus the search for best routes via skeletons and also, second, to prune the generated output of shortest syntheses by demanding the correct functionality as well.

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 cope with the considerable variety of possible functional groups and reactions, we must follow our analogy with maps and generalize or abstract structural data to coalesce trivial distinctions. This requires a system of description for structures and reactions which is rigorous, general and fundamental, to simplify and systematize the search space.

DESCRIPTIVE TERMS FOR STRUCTURES AND REACTIONS

Our system[1,4] describes four synthetically important 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 structure is then simply annotated as a compact list of simple numbers, a $z\pi$-list of its carbons ordered by their skeletal numbering. Thus, linearly numbered crotonic acid is 30.01.01.00 and all acyl variants (esters, nitriles, etc.) are generalized in the same notation; acetoacetates are 30.00.20.00 and their enol ethers are 30.01.11.00.

Reactions are characterized clearly and simply[1,5]. A *unit reaction* is defined as a unit exchange of attachments on each carbon. This can be expressed with two lettes: the first being the attachment bond made; the second the bond broken. Thus, there are 16 possible unit reactions at any carbon, all combinations of the four kinds of attachment made or lost. The reduction of alkyl halide to alkane is an HZ unit reaction, as is reduction of ketone to alcohol. The oxidation state change is found from $\Delta h = +1$ and $\Delta z = -1$ so $\Delta x = \Delta z - \Delta h = -2$. Simple oxidations are ZH reactions, $\Delta x = +2$, ΠH represents an elimination of H to form a π-bond, HΠ the reverse, addition reaction, etc. Any unit reaction involving ± Π or ±R at one carbon must have a coupled unit reaction at an adjacent carbon.

The abstract description of reactions allows the generation of all possible *net structural changes* as attachment changes at the involved carbons. This can be applied to characterize and catalog all organic reactions[5], analogous to the Beilstein system for organizing structures in that all possible reactions, presently known or unknown, have a defined place in the catalog. This becomes a very useful basis for organizing and searching any compendium or database of reactions.

GENERATION OF CONSTRUCTION REACTIONS

The construction reactions, RH, RZ, RΠ, are central to our procedure. A construction reaction is a combination of two half-reactions on each side of the constructed bond - one half nucleophilic (oxidative, $\Sigma\Delta x = +1$), the other electrophilic (reductive, $\Sigma\Delta x = -1$). In general a strand of up to six carbons spanning the constructed bond (up to three for each half-reaction labeled δ,β,γ out from the bond formed) contains all the carbons undergoing attachment changes and all possibilities for construction half-reactions can be defined as outlined in Figure 5.

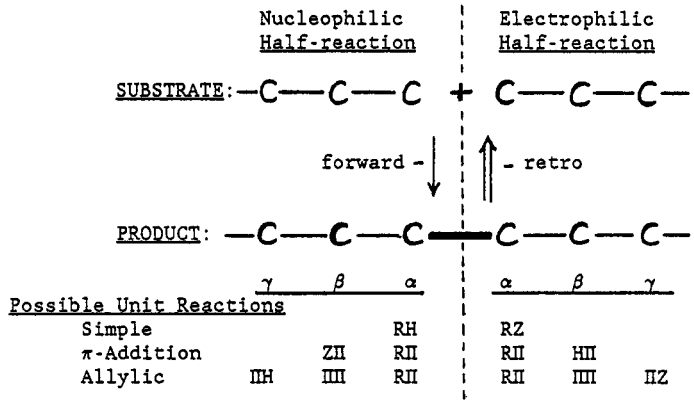


Figure 5. Generalized View of Construction Reactions

In order to generate a construction reaction across a bondset-designated bond, the strands of carbons out from each end of that bond are described by a $z\pi$-list and each defined construction reaction is described by an analogous single number list, i.e., its characteristic $\Delta z\pi$-list, which added to the product $z\pi$-list will generate the substrate $z\pi$-list (or vice-versa). Each reaction list represents its net structural change as $\Delta z\pi$, the

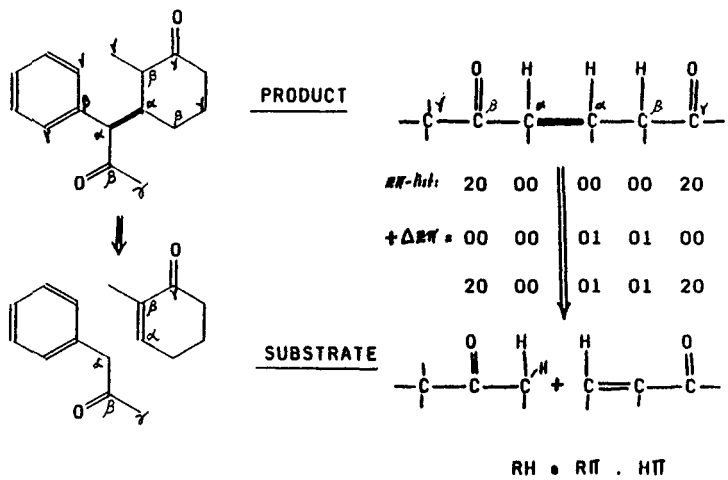


Figure 6. Example of Reaction Generation

arithmetic difference between substrate and product lists. Practical use of the generator is illustrated in Figure 6. Expression of the substrates and products as simple digital lists and this numerized generation of one from the other have several advantages: it is a very fast process for the computer; all possible conversions become simply a set of mathematical combinations so that none will be missed; the digital expression generalizes minor distinctions in functionality; no library database is necessary; presently unknown conversions are generated as potential new chemistry; and finally mechanistic tests can be made by quick numerical checks of functionality lists.

Generation of constructions with the $\Delta z\pi$-lists corresponding to the six basic half-reactions misses some real-life constructions for which the overall change is a composite construction and refunctionalization, as in the Wittig reaction = construction + elimination, or the Grignard reaction = reduction of RX + construction. When these composite half-reactions are added, and some subdivided to reflect common chemical divisions, the program contains 24 construction half-reactions, each characterized by its own $\Delta z\pi$-list generator. It should be noted that these numerical formulations involve surprisingly little information loss from normal reaction descriptions, which are easily reconstituted from these number lists.

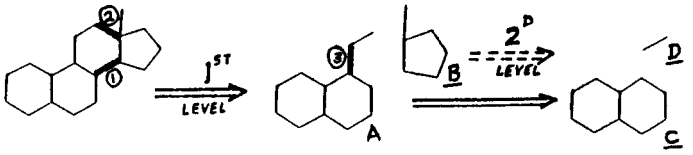
In practice, however, computer generation of all possible constructions for each bond in a bondset usually produces a high proportion of non-viable results (such as RH + R'X ⟶R-R') and so we can prune much of this using quick numerical checks of h,σ,π,z on the carbons near the construction bond to test for mechanistic viability. For each half-reaction we test for required activation, regioselectivity, interfering side reactions, presence of incompatible functional groups, etc. Here we recognize that merging all heteroatom types into the number z is too severe a generalization for mechanistic discrimination. Therefore, we define a subset of z to indicate the *function* of the heteroatom as electron-donating, electron-withdrawing, or leaving group. The use of mechanism tests for each generated reaction also allows the expansion of the defined skeleton to include N,O,S atoms as well as carbon, since they undergo the same construction half-reactions as carbon but with different mechanistic requirements. These mechanistic tests are all collected into two characteristic lists ("require" and "reject") for each half-reaction, each applied across all relevant skeletal atoms at once in a quick AND operation which checks h,σ,π,z,z-function, and skeletal atom type. These reaction checklists are also autonomous and may be replaced at will to reflect looser or tighter mechanistic requirements as desired.

THE SYNGEN PROGRAM

The SYNGEN (Synthesis Generation) computer program was designed to apply the foregoing logic to any input target[6]. The program consists of about 50,000 lines of FORTRAN with a catalog of about 6000 starting materials as data, and uses about a megabyte of active memory on a MicroVAX computer. Target structures are input crudely with a very fast, facile drawing module which then normalizes them to neatly drawn structures. The program then processes the target without operator intervention, usually in less than three minutes, and stores its output for later viewing.

An overview of its procedure is shown in Figure 7, illustrated with an estrone precursor known to form in an ideal synthesis[3]. This is shown as one of the generated routes from available starting materials. The generated skeletons are lettered and generated functionality on them numbered for each one, with found catalog entries underlined. The generated sequence of $z\pi$-lists for the changing atoms, through the three constructions from target to starting materials, is appended at the right; the program records the particular half-reactions used each time.

1. SKELETON - Dissect to all fully convergent ordered bondsets which result in all starting skeletons found by second level.



Functionality Generation

Atoms:	8	9	11	12	13	14
	01	01•00	00•00	10		
$z\pi$-lists:	⇓①					
	00	01•01	00•00	20		
	⇓②					
	00	10•01	01+00	20		
	⇓③					
	00	20+11	01+00	20		
	C-3	D-5	B-2			

Starting Materials

2. FUNCTIONALITY - From target functionality and each bondset generate sequential constructions through the sequence of ordered bonds to available starting materials.

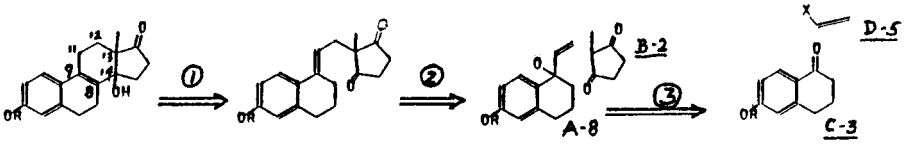
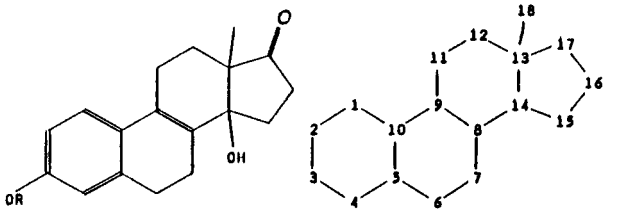


Figure 7. Overview of SYNGEN Procedure

The output for the same target is summarized as shown in Figure 8 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 depending on the nature of the target. 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.



LEVEL	B'SETS	ST. MAT.	INTERMED.	REACTIONS
1	4	7	26	61
2	34	169	3	393

Figure 8. SYNGEN Summary Screen

Selections (retain/delete) can be made from any category, and also 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. A typical page of reactions is shown, for two selected bondsets, in Figure 9 along with the screen menu. Each entry, numbered for bondset and reaction, shows the two substrates to join at sites indicated with one or two dots for first and second constructions respectively. The shorthand notation below shows the two pairs of half-reactions for the two constructions (there are 24 2-character notations of half-reactions easily learned or expanded on a help-screen). The first level construction of Figure 7 is to be found as reaction 4-41. For the same estrone precursor *SYNGEN* found several other, new syntheses, summarized in traditional form in Figure 10.

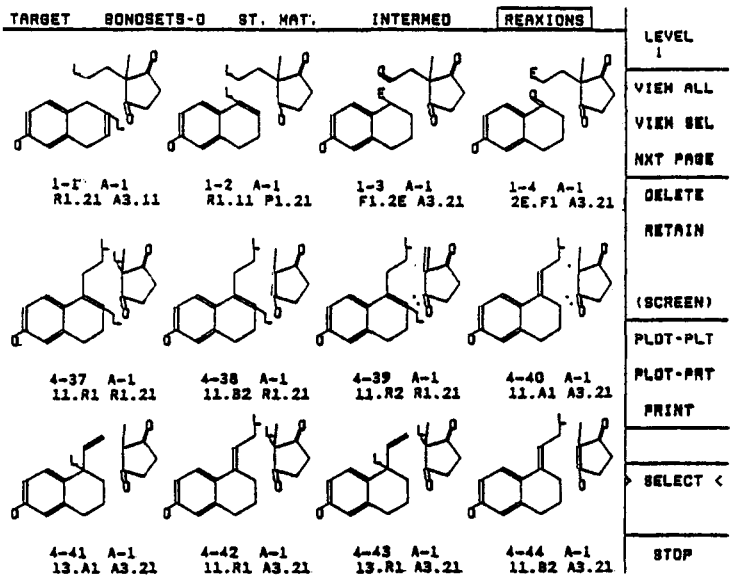


Figure 9. SYNGEN Reactions Output

DEVELOPMENTS FOR PROGRAM EXPANSION

Although the entries generated by the program do not derive from a reaction library or database, there will be considerable practical value in using such a database to discover literature examples as close precedents to the reactions generated by *SYNGEN*. We are currently applying our system of reaction organization as an overlay to reaction catalogs to allow faster search for the catalog entries that best match a generated reaction. Both the REACSS and SYNLIB catalogs are under investigation. Such precedent searches should serve both to demonstrate the validity of the generated chemistry and also to lead the operator to procedures for practical execution in the laboratory.

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,

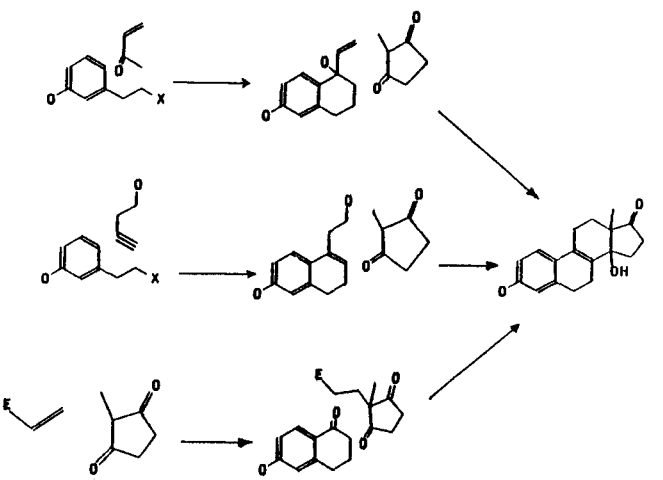


Figure 10. Selected New Syntheses From SYNGEN

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, i.e., 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 for the required skeletons are examined for their reaction distance to target. All those not too distant are coupled pairwise all ways in the forward direction, deleting combinations for which constructions are not viable, as well as those which diverge in distance after construction. By continuing to keep only those which after each construction move forward toward the target functionality, we should avoid the otherwise explosively unfocussed combinatorics of such forward generation. In this way we should create paths to estrone via more functionalized precursors of the same skeleton, as in Figure 2, and also of course routes to saturated hydrocarbon targets.

CONCLUSION

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.

Acknowledgements

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

1. J. B. Hendrickson, Accts. Chem. Res. 19, 274, (1986).

2. J. B. Hendrickson, J. Am. Chem. Soc. 99, 5439 (1977).

3. D. Taub in "The Total Synthesis of Natural Products", ed. J. Apsimon, Wiley-Interscience, 2, 641 (1973); 6,1 (1984).

4. J. B. Hendrickson, J. Am. Chem. Soc. 93, 6847 (1971); J. Chem. Ed. 55, 216 (1978).

5. J. B. Hendrickson, J. Chem Inf. & Comp. Sci. 3, 129 (1979).

6. J. B. Hendrickson, D. L. Grier and A. G. Toczko, J. Am. Chem. Soc. 107, 5228 (1985).