Tetrahedron, Vol. 37, Supplement No. 1, pp. 359 to 370, 1981
Printed in Great Britain

0040-4020/81/Suppl. 1-0359-12 $02.00/0
Pergamon Press Ltd.

A LOGIC FOR SYNTHESIS DESIGN[1]

JAMES B. HENDRICKSON, ELAINE BRAUN-KELLER and GLENN A. TOCZKO
Edison-Lecks Chemistry Laboratories, Brandeis University, Waltham, MA 02254, U.S.A.

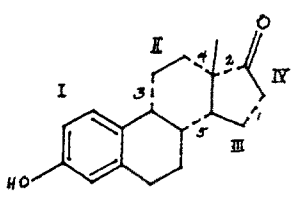
(*Received in U.S.A.* 17 *June* 1980)

Abstract—The background for synthesis design logic is presented, followed by the development of a procedure aimed at assessing all and locating the best synthetic routes. The procedure has two stages, first a dissection of skeleton, then a generation of necessary functionality on it to afford successive construction reactions. The implementation of this plan with computer programs is described, with some results. Reduction to computer has in turn served to clarify the overall logic, which is accordingly somewhat different from previous descriptions.

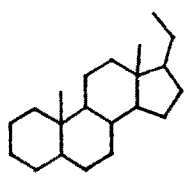
Synthesis of organic molecules has a long history but the conceptual act of design, of selecting a synthetic route or sequence of reactions, remains undefined, an art in the midst of a science. Before 1967 it was not even addressed in the literature.[2] Our minds, our knowledge, our literature, even our starting material catalogs are not organized for it. We speak of a synthesis as "elegant" but no one can define this elegance or say whether another route might have been shorter or simpler.

Size of the problem

Synthesis design is difficult because there are a vast number of synthetic routes possible to any given target from tens of thousands of available starting materials. This results from a combinatorial monster of what pieces to use, what order to take them in, and what reactions to link them with, and because the scope of possibilities is so vast it is probably not generally appreciated. In order to see this, we may consider just the possibilities of putting the target skeleton together from starting material pieces, ignoring the choice of reactions necessary to do it. This skeleton is a graph and there are a number of ways to dissect it.[4] If it has b bonds or links and we decide to make λ of them in the synthesis there are $\binom{b}{\lambda}$ combinations possible of bonds to make, or sets of starting material pieces to use. The simplest steroid synthesis is that of estrone (**1**) in which 5 out of 21 skeletal bonds are made (dotted in **1**).[5]



(**1**)



(**2**)

There are 20,349 combinations of five bonds possible, of which this is only one. Since the average starting material piece used in synthesis has three skeletal carbons,[6] a dissection of the C_{21} cortical steroid skeleton (**2**) should yield seven pieces and so $\lambda = 10$; 10 of 24 bonds to make offers almost two million possible combinations of skeletal dissection alone. Then the order in which any one of these combinations of pieces is put together is still unspecified, and there are $\lambda!$ possible orders for any starting material set. The five dotted bonds in **1** were made in the numbered order shown.[5] The combinations of pieces and orders must be multiplied, giving 2.5×10^6 routes for assembling estrone by making five skeletal bonds, or 7×10^{12} routes to the C_{21} steroid,[7] still irrespective of the chemistry, i.e. the functional groups and the reactions involved. The number of routes to create the C_{21} steroid skeleton from one-carbon units, i.e. total synthesis, is 6.2×10^{23}, more than Avogadro's number! If we made each molecule of **2** a different way we would make a mole of steroid.

The idea of routes to the target is often expressed graphically as a synthesis tree (Fig. 1) with lines indicating the reactions and points the intermediates;[8] the circled points are starting materials and two synthetic routes are marked as heavy lines from starting material to target (T). The tree shown is very deceptive in being so very small a part of the whole. The problem then is to find a systematic way to assess all the routes and locate the best ones through this enormous tree, within clearly defined constraints.

Traditional approach

The original computer procedure, used first by Corey and Wipke[8a] and since by others,[9,10] was to follow systematically the presumptive reasoning of chemists, back from the target functionality stepwise, using a built-in library to find all last reactions and their substrates and then repeating this for each intermediate so derived. This procedure immediately generates a large number of intermediates at the first level and, as each intermediate becomes target in turn, an almost exponential increase in numbers thereafter. The need to prune the tree here is met by prediction of reaction yields at each step and, in the interactive programs, by allowing the chemist to select favorites. There are several intrinsic weaknesses in this approach.

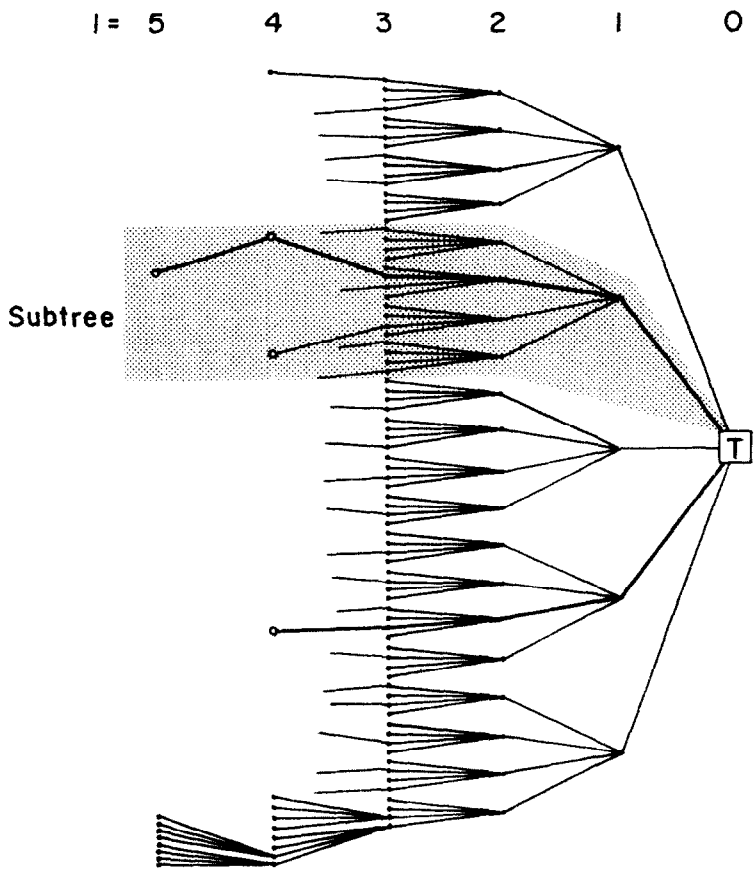


Fig. 1. The synthesis tree.

(a) The prediction of yields is notoriously imprecise and so a poor basis for comparison among thousands of routes assessed.

(b) The procedure has no direction: it ignores the available starting materials until found at the end instead of actively forcing the search to converge on them. Similarly the process is not actively directed to seek economy.

(c) Directed by functionality and reactions dictated by it, the procedure cannot synthesize unfunctionalized targets without incorporating dummy functional groups and there is no heuristic to indicate where to locate these.[8b] Furthermore, such dummy functional groups, removed before the target is reached and so leaving no trace in its structure, are not uncommon in real syntheses of functionalized targets as well.

(d) At the mercy of a library of known reactions there is no opportunity to generate new synthetic reactions.

Finally, however, while synthetic chemists rarely describe the conceptual basis for a route chosen, it seems unlikely that they have employed such a mechanical procedure. A broader grasp of the whole structure is implicit in their work and some relation to a special starting material or key assembly reaction is often apparent.

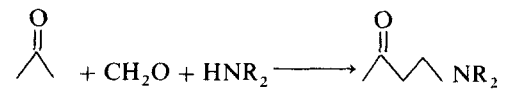
Tree searches

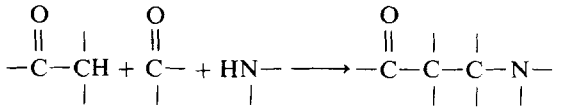
We can approach the synthesis design problem as a huge tree search, comparable to such searches in other fields. The size of the tree demands a reduction and four ways of doing this are common.

(a) *Systematize.* It is common in tree searches in other fields to define the units in the search space numerically since digital expression allows all possibilities to be simply mathematical combinations. This provides both rigorous definition and also the confidence that all combinations can be found mathematically. Also there is the advantage in digital description that it affords easy and rapid computer manipulation.

(b) *Simplify.* Here we define the material in the tree more broadly, in effect coalescing trivial distinctions so as to manipulate fewer items. Then after selection of a few optimal choices, roughly defined, these may then be refined back to normal detail. *Simplification* is analogous to map-making: the larger the area mapped, the fewer the distinctions made, the greater the level of abstraction in representation. Our common representation of molecules used in synthesis is a pictorial one, a graph of connected atoms, and at a level of abstraction that omits much information (interatomic distances and angles, congestion, charge distribution, etc.). Even this level of abstraction is too

(1) $CH_3COCH_3 + CH_2O + HN(CH_3)_2 \xrightarrow[25°]{H_2O} CH_3COCH_2CH_2N(CH_3)_2$

(2) 

(3) 

(4) **20 + 2 ⟶ 20·1**

detailed to map all possible involved molecules in a synthesis tree. A next level of abstraction, with more severe generalization, would reduce molecular description to a simple, linear digital one, capable of encompassing a much larger search space practically. Such a description is offered below and may be illustrated here with four levels of increasing abstraction for the Mannich reaction, each with more severe generalization than the last, each increasingly omitting more detail that is left implicit for the chemist to fill in from his general knowledge.

(c) *Subdivide.* If the tree is still too large to handle, it is reasonable to subdivide it into subtrees and so to operate on them one at a time sequentially. In order to do this the subtrees must be independent of each other, with no crossovers between them. Such a subtree is shown shaded in Fig. 1.

(d) *Select.* The crucial operation, of course, is to select a few optimal routes from so many in the tree. This requires criteria to direct the search, to reject some solutions and to assign priorities to the rest. Furthermore, the criteria must be very stringent ones if only a few routes are to be selected as optimal. The basic criterion is one of economy of time and materials. If yield prediction is to be abandoned, then the basic criterion becomes that of shortest routes, with fewest steps. This means the search will be speeded by actively directing it toward the nearest available starting materials, thus converging onto the shortest possible sequences.

Overview

Synthesis is a skeletal concept. The aim is to create a large target molecule from small starting material pieces. To put together these pieces is fundamentally to assemble the skeleton. In the combination of *pieces* × *order* × *reactions* the traditional focus has been on the *reactions*, but it is the *pieces* × *order*, i.e. the skeleton and its dissection, which is more important in taking a broader view of the problem. We can discern a dichotomy in molecular structure between the skeleton and the functionality on it. There is a parallel distinction in reactions, i.e. those that build skeleton—*constructions*—and those that alter functionality (*refunctionalizations*) with no change in skeleton.[11] Hence our first major simplification of the problem will be to consider only the skeleton. This simplification is enormous: the thousands of acyclic starting materials with six linked carbons or less are represented by only 13 skeletons. The main consequence of this focus on the skeleton is that only construction reactions are obligatory in any synthesis. It follows from this that the shortest, most economical synthetic route is a sequence of constructions only with no intervening refunctionalizations. Thus an ideal synthesis may be defined as one in which the starting materials come correctly functionalized to initiate their constructions and that the functional groups remaining after one construction are exactly those required to initiate the next. Also, at the end when the target skeleton has been fully assembled the functional groups remaining from the last construction are exactly those of the target.

Such an ideal synthetic route is a sequence of constructions only, with no refunctionalizations, and is very rare in practice;[12] the average synthesis contains twice as many refunctionalization steps as construction steps.[6] However, such a synthesis must be the shortest and constitutes a goal to aim for in seeking economy. It is also a very stringent criterion for selection since it lays such heavy demands on the overlapping functionality directing successive constructions.

This concept then dictates a procedure. We should dissect the target skeleton first in the best ways to sets of starting material skeletons and check a catalog for their availability. Only then, for the best sets found, will we generate the functionality necessary for self-consistent sequences,[12] working sequentially back from target functionality until the required functional groups on the starting materials are revealed. These are then again sought in a full starting material catalog for their availability. Hence there are two stages, the first a major simplification of the target to skeleton only and dissections of that skeleton into optimal pieces and orders. Second then comes the generation of the reactions to link them, a sequence only of constructions with the necessary functionality to drive them. In this way we must find the shortest routes from real starting materials to the target.

Skeletal dissection

The simplest gross description of any synthesis is the *bondset.* This is just the set of skeletal bonds (λ in number) which are constructed in the synthesis. The bondset may be further defined by the order in which

they are constructed (*ordered* bondset). The set of dotted bonds in **1** is the bondset for that synthesis,[5] and it is numbered in the order they were made. (Common practice is to make one in four of all skeletal bonds.[6]) The bondset directly shows the starting material skeletons and the carbon sites on each one at which construction occurs, usually 2–3 sites each. The four pieces used to make **1** are labeled with Roman numerals. The bondset defines a single independent subtree of the whole synthesis tree, containing all the reactions used to create the target from one set of starting material skeletons in all possible ways. The subtree for the bondset may therefore be fully explored without interference from other parts of the tree. In view of the many possible bondsets noted above, we shall need very stringent criteria for dissecting the target skeleton into a few best possible ones. We shall also need a basis to put these bondsets into priority order.

Synthesis plans

A basis for ranking synthetic routes can be made from the *synthesis plan*, a graph of the sequence of operations in the synthesis.[13] The synthesis plan is a small section from the whole synthesis tree, representing a single synthesis, as in the plan in Fig. 2

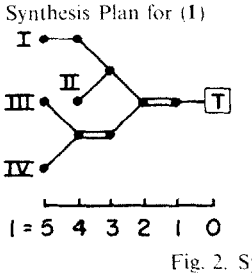


Fig. 2. Synthesis Plans.

for synthesis of estrone (**1**). The lines represent reactions, as vectors to the right. The points are compounds, those at left with one line out being starting materials, the one at far right with one line the target, and those between with 2–3 lines joining being intermediates. The horizontal lines are refunctionalizations and construction *cyclizations* (double lines). The other constructions, which join two pieces, are called *affixations* and represented by a pair of angled lines meeting at the affixation product. The starting materials are Roman numbered as in **1**. Figure 2 shows piece I being refunctionalized once, then affixed to piece II, while III and IV are separately joined and cyclized, then finally affixed to I–II, cyclized and refunctionalized to the target (T). The path-length (l) for each piece can be easily seen, none more than five. The five constructions are numbered in sequential order matching the order of the dotted bonds in structure **1**. Most syntheses are more extended, having more refunctionalization reactions. Removal of the horizontal refunctionalization lines coalesces the whole synthesis plan to a construction plan, showing construction reactions only, and a further coalescence of the cyclizations, just to circles around the intermediates which cyclize, affords the *affixation plan*, also shown in Fig. 2 for the estrone synthesis (**1**).

Overall yields can be calculated directly from the synthesis plan if all the individual step yields are known. These cannot be known in advance at the planning stage, however, but it is reasonable to assume that any mechanistically sound reaction can be optimized to a similar good yield. Hence in calculating overall yields of sequences we simply assume ~80% for each reaction. However, the more convergent a synthesis is the less meaningful is the overall yield. A more accurate measure of overall yield is the total weight of starting materials required. This is simply $W_{sm} = \Sigma_i M_i x^{l_i}$, where M_i is the molecular weight of starting material i and l_i is its path-length, the number of steps piece i passes through to the target. The term x is the inverse of the average yield (for 80% yield, x = 1.25). Values of W_{sm} can be used to compare the relative efficiency of different routes. However, at the planning stage the detailed molecular weight (M_i) will not be known and so we can substitute for it just the number of skeletal carbons (n_i) in the starting material piece. Hence we define $W = \Sigma_i n_i x^{l_i}$ as a measure of total starting material weight for purposes of comparing planned routes. It turns out that even if all reactions have the same yield the value of W is lower and hence better for some routes than others, even with the same number of steps. This is a function of the form of the synthesis plan. Therefore, we can rank various plans in a priority order, first by the number of steps and then for those with the same number of steps by total starting material weight (W).

Examination of various synthesis plans for different routes allows us to draw up some rules about the efficiency of plans, seeking minimal values of W.

(a) Obviously economy demands minimizing the number of refunctionalizations. It may be noted especially that protecting groups require two refunctionalizations; hence protecting groups may be seen as good chemistry but bad synthesis. Only if the protecting group is present in the starting material and comes off automatically in another reaction (no separate removal operation) does it represent a good plan. Most of the refunctionalizations in published syntheses are protecting group manipulations.

(b) Refunctionalizations are least damaging to efficiency and W if done early, preferably on a starting material prior to its incorporation in constructions.

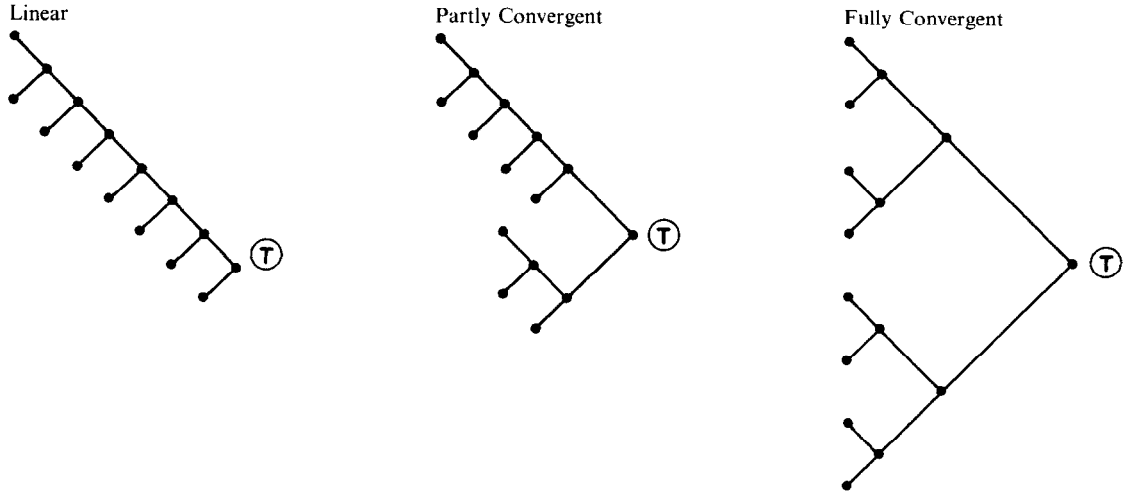
(c) Resolution to chiral intermediates is especially severe since it must be less than 50% yield. Hence a resolution is equivalent to about four refunctionalization steps since four steps at 80% is 41% overall. It is important, therefore, to resolve as early as possible on the smallest possible intermediate, if it must be done at all.

(d) The calculations of W dictate that cyclizations should occur as early in the plan as possible, before other pieces are added to the cyclizing unit.

(e) Large starting materials should be added as late in the sequence as possible to minimize W.

Convergency

The main variable affecting economy in synthesis plans is convergency, a concept first expressed by Velluz *et al.*[14] In a convergent plan the pieces are assembled separately and independently, then linked together afterwards near the end of the synthesis. This may be seen by comparing the affixation plans for three cases of affixing eight pieces (Fig. 3). There is a continuum of partially convergent sequences between a strictly linear one, at left, and a fully convergent one, at right (there are 23 possible affixation plans of



$\Sigma l_i = 35$	27	24
$\Sigma x^{1_i} = 24$	18	16

Fig. 3. Affixation plans for eight pieces.

varying convergency for eight pieces). The most convergent possible has the lowest Σl_i and also the lowest Σx^{1_i} as illustrated in Fig. 3. The extent of possible convergency is related to the bondset, as many bondsets do not allow a fully convergent sequence. Also convergency defines the order of bond making in a bondset.

The best bondsets are dictated by full convergency. These result from truncating the affixation plans for fully convergent total syntheses (i.e. from one-carbon pieces) to those with starting material pieces of 2–4 carbons already linked.[13] Thus the affixation plans for fully convergent bondsets can always be mapped onto the plans for fully convergent total syntheses.

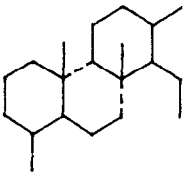
In practice the convergent bondsets are found by dividing the target skeleton first into two roughly equal parts, cutting the fewest rings. Then each part is cut again the same way until pieces of 2–4 carbons are found. The resulting ordered bondsets may then be ranked by calculating W, which will be essentially the same for all and minimal. Convergency not only defines both starting material pieces and order of assembly but is also a stringent selector of bondsets. The C_{21} steroid (**2**) cut into eight acyclic pieces affords two million bondsets but only 45 of these can have fully convergent plans from pieces of 2–4 carbons.[13] The convergent routes are only about 50% better in W than linear ones on the affixation plans but when refunctionalizations are added to string out the whole synthesis plan the preference rises sharply and can favor the convergent plans by a factor of 5–10 times.

Other dissection modes

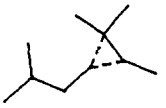
There are other heuristic bases for dissecting the target skeleton for economy, i.e. for fewest steps. All the following ideas can be sought during the search for convergency.

(a) Minimizing steps includes minimizing λ and this implies seeking the largest possible starting material pieces. Many of these can be discovered by comparing the two parts found at each cut with a catalog of starting material skeletons.

(b) Two equivalent halves or nearly equivalent halves generated by any cut can allow one half to be made from the other and so delete the steps necessary for its separate construction. The skeletons of the quassinoid family of natural products (**3**) and that of chrysanthemic acid (**4**) illustrate this idea.[15] If one part dissected by a cut is not identical to the other but contains its skeleton, one may also be made from the other more quickly than by separate construction. Again these identities can be discovered while making cuts toward convergency if the skeletons of cut parts are compared with each other.



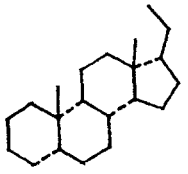
(**3**)



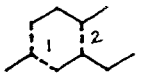
(**4**)

(c) Multiple constructions are more economical since they construct several bonds in one operation. For two constructions these can be double affixations, annelations (i.e. affixation + cyclization), or double cyclizations. Multiple cyclization is the basis for economy in the Johnson steroid syntheses,[17] one of which is shown (**5**) with the partial bondset of its multiple cyclization dotted. Such multiple cyclizations will occur last in the sequence. Annelations, including the Diels–Alder and Robinson methods, are all located when a skeleton is cut in two through a ring. The double affixation requires three pieces, two of them identical and joined to the third in a single reaction. The idea may be illustrated by the C_{10} skeleton (**6**) which has 720 possible ordered bondsets of $\lambda = 3$ but

only ten which meet this condition (with the identical pieces $\geq C_2$). One of these is shown in (**6**), with the bondset order numbered to show the double affixation as the first step.



(**5**)



(**6**)

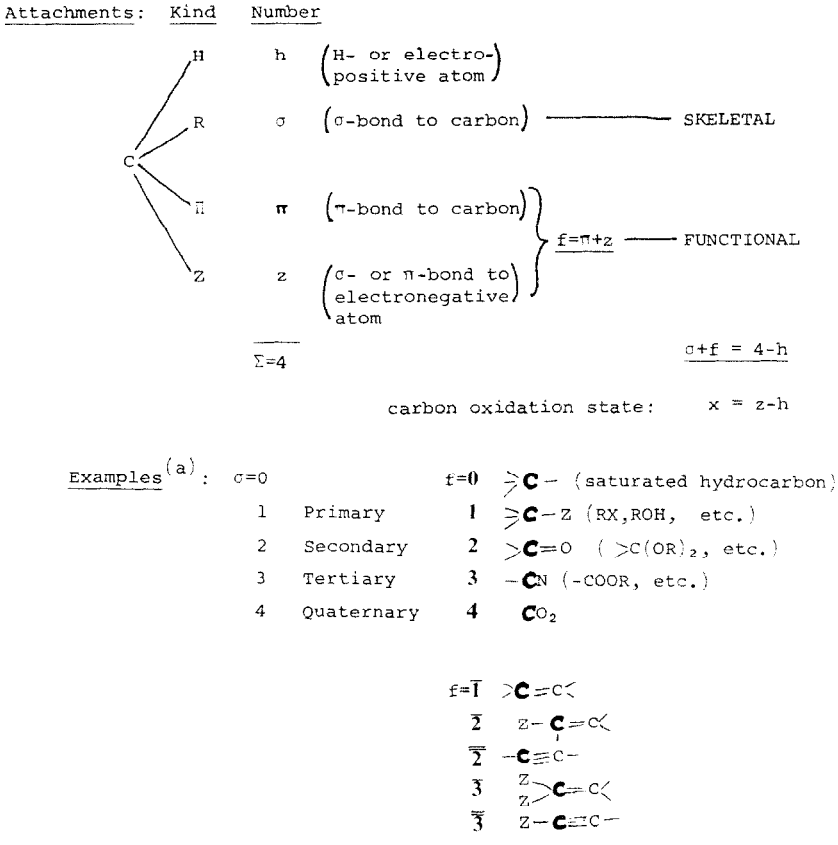
With these heuristics optimal bondsets may be found, and ranked by number of steps and weight of starting materials, via stepwise cuts of the target skeleton toward full convergency, while seeking the other conditions at each cut. The entire dissecting operation is capable of a simple and systematic treatment by the computer. It is in fact the very kind of operation that computers handle much better than people.

Reactions

Having examined the pieces and their order of assembly, we move on in the second stage to the actual reactions. Here we need to generate the functional groups necessary to direct selfconsistent sequences[12] for each bondset selected above. At this point we require a simple, rigorous definition of functionality for the computer to manipulate numerically. The point of digital description here is to coalesce trivial distinctions of functionality, to enable rapid computer handling, and especially to convert all possible variants into simple mathematical combinations so that all can be systematically treated and found. Furthermore, a really basic description will encompass all possible functionalities and reactions, including those not currently chemically feasible.

The numerical characterization of structures[18,19] starts from a definition of four kinds of attachment any carbon may have, and then assigns to each carbon the number of each kind of attachment it has, as summarized in Fig. 4. Thus any structural carbon is definable by two numbers, σ and f, denoting its skeletal level and its functional level,[20] and examples are shown. The value of x accurately gives the oxidation state of any carbon, and so $\Sigma\Delta x$ for any conversion shows the overall change in carbon oxidation state (see bottom of Fig. 5).

Any reaction is now simply defined by the *net structural change* from substrate to product, or equally in the reverse direction. A single reaction step is rigorously defined as a unit exchange of attachments,



(a) In the functionality examples the f-value refers to the boldface carbon and unspecified bonds are nonfunctional (R or H). The overbar specifies the π-value.[20]

Fig. 4. Numerical characterization of Structure.

	Oxidative	Reductive	Isohypsic	
Construction	RH	RZ	RΠ	RR
Fragmentation	ZR	HR	ΠR	
Addition	ZΠ	HΠ	RΠ	ΠΠ
Elimination	ΠH	ΠZ	ΠR	
Substitution	ZH	HZ	ZZ	HH

Relations in Reaction Changes:

$$\Delta f + \Delta\sigma + \Delta h = 0$$

$$\Sigma\Delta x = 2\Sigma\Delta f - \Sigma\Delta\pi + \Sigma\Delta\sigma$$

Fig. 5. Reaction labels. The label is the change at one carbon in one reaction step. Unit exchange of attachments = bond made; bond broken.

and so a reaction as commonly understood may sometimes (though seldom) consist of more than one successive reaction steps. This unit exchange is labeled by two letters, first the kind of attachment made, then the one lost; the 16 possible exchanges at one carbon are listed and described in Fig. 5. Thus reductions of ketone to alcohol or halide to hydrocarbon are equally HZ reaction steps, and $\Delta x = -2$ since all the changes in numerical values (f, σ, x, etc.) for a carbon are exactly implicit in the reaction label. A single reaction step will involve more than one carbon if R or Π exchanges occur since these demand the same change on an adjacent carbon. This system allows a simple but rigorous organization of all possible organic reactions just as the Beilstein system organizes all possible structures.[19]

Two examples of the notation are presented in eqns (5) and (6). The values of f, x and σ are shown for each specified carbon as are the letter symbols on the arrows for each reaction step.

Reaction generation

Once the ordered bondset is defined, the skeleton at each point in the sequence is specified. Hence for functionality generation we need only manipulate f-values for each carbon since the σ-values are determined. Any structure is represented by an ordered list of f-values for the list of numbered carbons

$$\mathrm{R{-}CH_2{-}Cl} \xrightarrow[(HZ)]{Mg} \mathrm{R{-}CH_2{-}MgCl} + \mathrm{\overset{O}{\overset{\|}{C}}(H){-}R'} \xrightarrow[(RH\cdot RZ)]{} \mathrm{R{-}CH_2{-}\overset{OH}{\overset{|}{C}}(H){-}R'} \qquad (5)$$

f = 1	0	2	0	1
x = −1	−3	+1	−2	0
σ = 1	1	1	2	2
ΣΔx = −2		0		
ΣΔσ = 0		+2		

$$\mathrm{R{-}\overset{\ominus}{\overset{\cdot\cdot}{C}}H{-}PO(OR)_2} + \mathrm{\overset{O}{\overset{\|}{C}}(R){-}R} \xrightarrow[(RH\cdot RZ)]{} \left[\mathrm{R{-}\overset{(RO)_2PO}{\overset{\diagdown}{C}}H{-}\overset{O^{\ominus}}{\overset{|}{C}}(R){-}R}\right] \xrightarrow[(\pi Z\cdot\pi Z)]{} \mathrm{R{-}CH{=}\overset{}{\overset{|}{C}}(R){-}R} \qquad (6)$$

f = 1	2	1	1	$\bar{1}$	$\bar{1}$
x = −1	+2	0	+1	−1	0
σ = 1	2	2	3	2	3
ΣΔx = 0			−2		
ΣΔσ = +2			0		

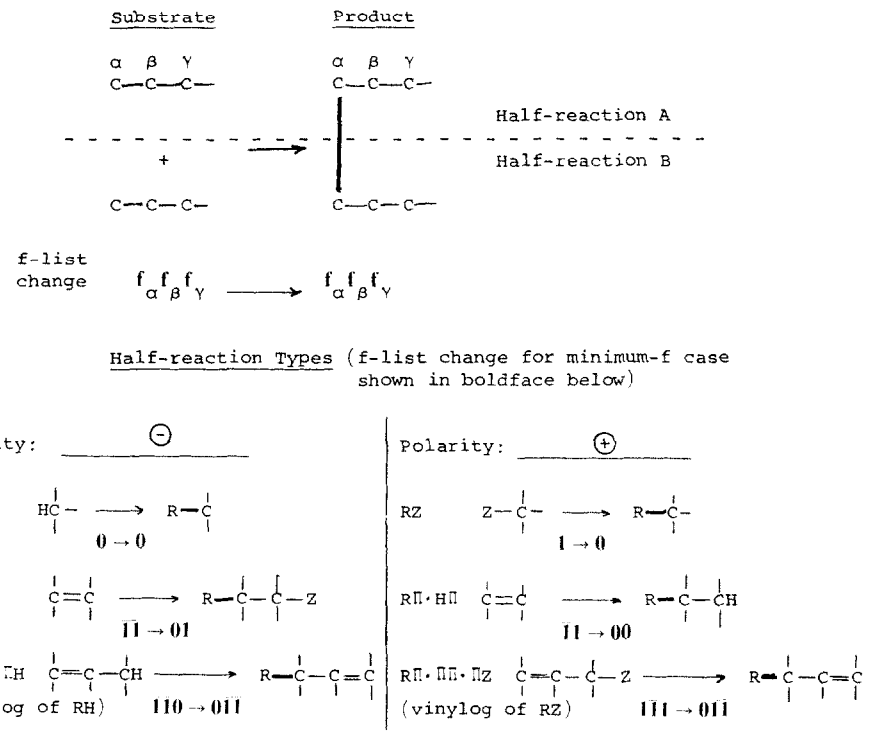


Fig. 6. Construction reactions.

in the skeleton. This *f-list* for any substrate changes in a defined way into the f-list of the product for any given reaction, and so in reverse the f-list for any substrate may be generated for a particular reaction from the f-list of the product.

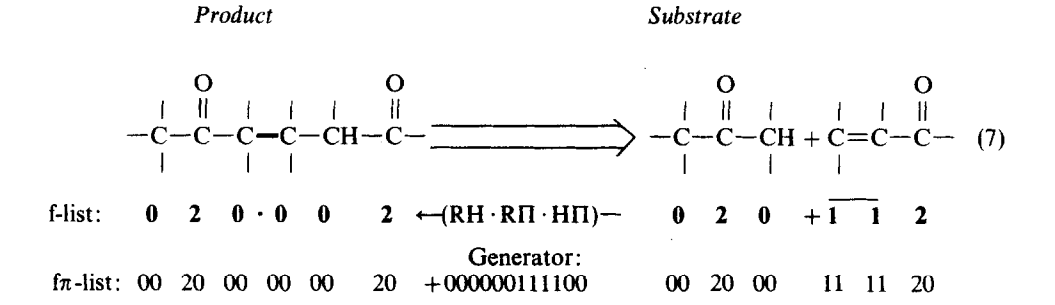
Construction reactions may be generalized with part structures as shown at the top of Fig. 6, labeling the carbons α, β, γ on each piece out from the carbon (α-carbon) forming the new bond. Each piece then displays a half-reaction which may be described by the change in the f-list ($\mathbf{f_\alpha f_\beta f_\gamma}$) from substrate to product (or vice-versa). The changes at the α-carbon must be RH, RZ or RΠ (RR is ignored as also involving a C–C bond fragmentation). With the first two the functionality at carbons β and γ does not change, but RΠ requires a loss of Π at the β-carbon also, i.e. HΠ, ZΠ or ΠΠ. Of these the first two require attachment changes at both α and β, the third a change at γ as well. Vinylogous reactions require three carbons since they exhibit a ΠΠ change at the β-carbon and so demand ΠH or ΠZ at the γ-carbon. Therefore, the six formal possibilities for construction half-reactions on up to three carbons are shown at the bottom of Fig. 6, and necessarily include all possible half-reaction steps.[21]

In these generalized structures in Fig. 6 the unspecified attachments may be made to H, R or Z. The case of minimum necessary functionality implies these attachments only to R or H and these are the f-lists shown. If one or more of these unspecified attachments is to Z it is a parallel example at higher functional level but the net change is the same,[22] as in the simple RZ half-reaction, which can be alkylation (**1 → 0**), the minimum-f case shown, or carbonyl addition (**2 → 1**), or acylation (**3 → 2**), or carbonylation (**4 → 3**).

Finally, the f-list change in a half-reaction is either oxidative ($\Delta x = +1$) or reductive ($\Delta x = -1$). The former are designated as ⊖-polarity since they are nucleophile half-reactions, and the reductive ones as ⊕-polarity since they are electrophiles. A full construction combines two half-reactions, and if one is ⊖-polarity and the other ⊕-polarity this is a common isohypsic (− + or + −) construction, i.e. one with no overall oxidation state change ($\Sigma\Delta x = 0$). These are the only full constructions used here.[23]

The 18 possible full construction combinations (9 + − and 9 − + from Fig. 6) may be generated along any strand of six carbons surrounding a designated construction bond of the bondset. This is done by adding an f-list generator corresponding to the net change in f-list, to the product f-list of that strand in order to create the substrate f-list.[24] In the computer the Π-overbars must be separated from the f-values as separate digits so that an fπ-list is used. For each of the 18 full constructions this is a 12-digit number for the six carbons across the bond formed. The generator for each of the 18 reactions is simply derived by subtracting the minimal fπ-lists of product from substrate for each combination in Fig. 6, and stored in core for use. Their use is illustrated with the Michael addition in eqn (7).

By way of expansion of this procedure to approximate real chemistry more closely, we may briefly note some options. First many practical constructions are in reality a construction together with a refunctionalization. This is true in many reductions of halides to Grignard reagents and other high-energy carbanions, also in elimination following construction (Wittig, aldol, etc.). The examples of eqns (5) and (6) show two such common synthetic



(7)

operations, each of which is two reaction steps (construction and refunctionalization) in the definition here, but accepted for our self-consistent sequences as equivalent to single steps in the synthesis. These automatic refunctionalization steps may then be added to the construction generators to create new generators for these two-reaction-step operations. With the present added refunctionalizations we have expanded the list of full constructions from 9 to 32. Secondly, restrictions may be included on the use of the construction generators, i.e. restriction on inadequate activation, incorrect regiospecificity, or preference for another reaction over construction. These restrictions may be applied as tests of the values of f, π, σ, etc. at the carbons proximal to the bond constructed and may be used to invalidate reactions which, though formally correct from Fig. 6, cannot occur. This eliminates much chemically unreasonable output, such as RH constructions at unactivated sites, etc. Finally, numerical tests can be devised to delete reactions with incompatible functional groups elsewhere in the molecule.

Program

The overall logic resolves itself into a procedure for the computer.

(a) The target molecule is entered graphically as a skeleton with f-values at functionalized sites as shown at the top of Fig. 7. This is resolved by the computer into a simple carbon connectivity table or adjacency matrix of the carbon skeleton, i.e. a symmetrical n × n matrix of 1's and 0's, with the n carbons numbered. Then the f-values of the carbons are carried in the diagonal of this connectivity matrix.[29]

(b) The skeleton is cut in two all possible ways and the parts searched in a catalog of starting material skeletons. Then the procedure is repeated for those pieces not found in the catalog. The program seeks matches not only with the catalog but also with all other cut parts, to locate identical pieces for common synthesis or double affixations. It can also find all ring cuts in the whole target skeleton for multiple cyclization. When all cut parts are identified in the catalog, ordered bondsets are put in a priority order according to minimum number of steps and minimum W, and these may be displayed. The C_{10} molecule in Fig. 7 is skeletally dissected in the top row and one of the ten bondsets for double affixation (**6**) is illustrated.

(c) For each bondset the functionality generators are all applied to each cut bond successively in reverse order. The six carbons around the last bond constructed in the target are taken first and the generators added successively to the target f-list for those six carbons to generate the f-lists of last-reaction substrates; most will not give viable substrates[24] and restrictions applied before generating will eliminate more. The same procedure is then applied to the next bond back in the bondset order, using the new intermediate functionalities generated in the first pass, and so on until all bonds are cut and the resultant required functionality on the starting material skeletons is revealed. These are then searched in the full starting material catalog and the routes from available starting materials are recorded.[25]

(d) In principle the output can be displayed several ways. The starting material skeletons, oriented as they appear in the target, can be graphically displayed with f-values entered on functionalized carbons, as shown in Figs. 7 and 8. Further, each intermediate in a synthesis could be similarly laid out in sequence. For many syntheses this is too much output. More compactly, for each bondset we can list on one line the f-lists of the

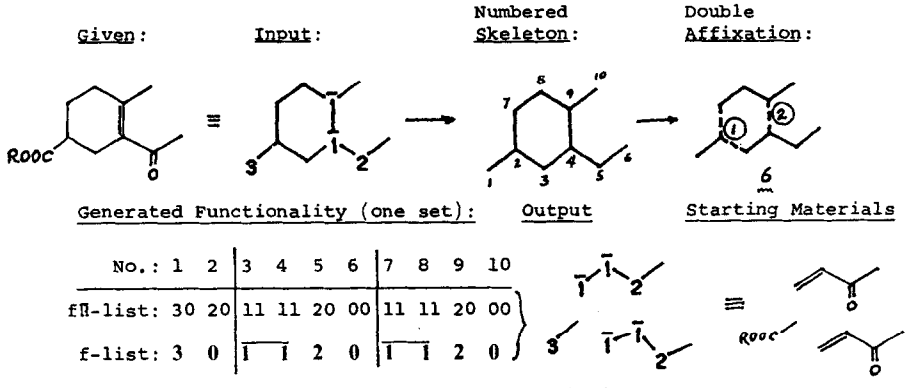


Fig. 7. Dissection of a C_{10} molecule.

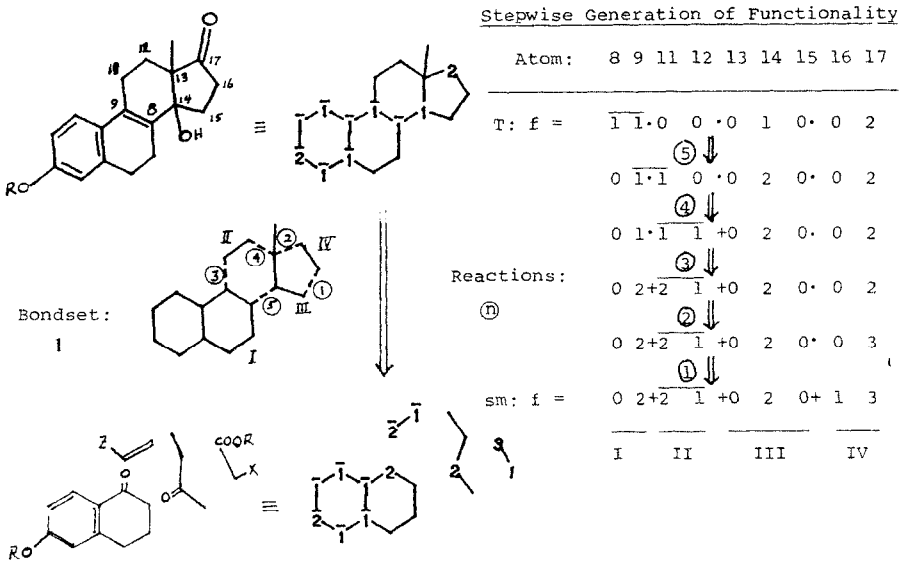


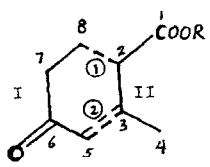
Fig. 8. Generation of estrone synthesis

starting materials followed by simple symbols denoting the nature of the two half-reactions in each successive construction. This allows the chemist to perceive immediately the net structural change and implied chemistry. Such output is shown directly as received from the computer in Fig. 9. Definition of the detailed construction labels has been omitted as trivial in this discussion,[26] but will generally be apparent in all three figures if the f-lists are all refined to normal structures and their successive conversions examined.

One double affixation route to **6** is recorded in the lower half of Fig. 7. In Fig. 8 is an illustration of the stepwise generation of one self-consistent sequence to the intermediate in the estrone (**1**) synthesis;[5] the last three reactions are those in the published synthesis, the first two are reasonable variants. The five constructions are circled in the ordered bondset and in the generated self-consistent sequence.

While the logic is basically simple, the programming details are extensive.[26] At the present time parts of the overall program have been written as separate modules for separate testing, and are not all yet complete nor tied together to make a single overall executive program. The graphics input has been created to accept a drawn structure on an ordinary CRT screen and then to normalize automatically a poorly drawn one, equalizing bond lengths and angles (Figs. 6 and 7). The present starting material catalog is about 8500 compounds.[27] The adjacency matrices of these compounds are ordered for quick search by row-column inversions such that the binary list obtained by stringing out the matrix entries is a maximum number. Then the skeleton catalog is simply a listing of these numbers in numerical order for easy searching. When a skeleton found by the program is to be looked up its matrix binary list is similarly maximized and that number sought in the catalog.[26]

Another program is being created to locate starting materials in the catalog which are one or two steps away from a structure generated by the main



STARTING MATERIALS								CONSTRUCTIONS	
1	2	3	4	5	6	7	8	2- 8	3- 5
30	10	20	00	00	20	00	10	11.R1	*21.E1
30	10	20	00	00	20	11	11	11.R2	
30	00	20	00	00	20	00	10	*A1.11	
30	00	20	00	00	20	11	11	*A1.12	
30	10	20	00	00	20	00	10	R1.11	
30	10	20	00	00	20	11	11	R1.12	
31	11	20	00	00	20	00	10	B2.11	
31	11	20	00	00	20	11	11	B2.12	
32	22	20	00	00	20	00	10	C2.11	
32	22	20	00	00	20	11	11	C2.12	
30	11	21	00	00	20	00	10	B2.11	
30	11	21	00	00	20	11	11	B2.12	
30	22	22	00	00	20	00	10	C2.11	
30	22	22	00	00	20	11	11	C2.12	
30	10	10	00	20	20	00	10	11.R1	*Γ1.21
30	10	10	00	20	20	11	11	11.R2	
32	22	10	00	20	20	00	10	C2.11	
32	22	10	00	20	20	11	11	C2.12	
30	11	21	00	20	20	00	10	12:R1	
30	11	21	00	20	20	11	11	12.R2	

I II

Fig. 9. Self-consistent sequences generated.

procedure. This allows locating not only exact matches in the starting material catalog but also starting materials which might be converted to a desired compound by refunctionalization, rearrangement or fragmentation. This is possible because the numerical change in f-values between any two structures can be used to calculate the number of reaction steps separating them.[28]

The convergency dissection module is currently being built, but a separate program has been written to find all double affixation opportunities (see Fig. 7), i.e. all skeletal dissections into three pieces with two identical and each of these connected to the third from the same site.[26] Another program has been created to derive all multiple cyclizations with alternating bonds in the bondset as in **5**. The bondset in **5** is only one of 34 possible bondsets of $\lambda = 4$ with alternating bonds.

The functionality generator module is now operative and produced the functionality results in Figs. 7–9. At present its output is in the form shown in Fig. 9, generating starting materials as fπ-lists and listing the successive construction reactions used.[26] There are reproduced in Fig. 9 only 20 sequences of the 32 which were generated, but a close examination shows how closely related many of these sequences are.

CONCLUSION

The overall aim of the project is to find all the *shortest routes* of successive constructions to a given target from given starting materials. The basis is a major simplification first to skeleton only, in order to encompass the whole synthesis tree and find optimal bondsets. Then, second, for each bondset there is a simplification of functionality to numerical f-values to encompass the whole subtree for that bondset. Selection is based on fewest steps and highest yields (expressed as minimum W), selfconsistent sequences and available starting materials. The system is intended to be rigorous and clearly defined so that the chemist may know that it produces all possible syntheses within these constraints, including some with presently unfeasible chemistry which may challenge the chemist to invent new reactions. In this way we hope to produce a set of optimal syntheses of any target which may serve as a standard for comparison with synthetic plans invented by practicing chemists. R. B. Woodward saw himself as an artist in synthesis design and indeed his best syntheses pointed to elements incorporated in the logic here. The intent of this project is not to replace art in organic synthesis but to show where real art lies.

Acknowledgement—We are grateful to the National Science Foundation for support of this work through a grant, CHE–7712267.

REFERENCES

[1]This article is dedicated to the memory of R. B. Woodward, the master of synthesis who had so much to teach us. The senior author was his student from 1950 to 1957.

[2]E. J. Corey, *Pure Appl. Chem.* **14**, 19 (1967). An earlier article by Woodward[3] extolls the art of synthesis but says little about the conceptual process of design.

[3]R. B. Woodward, *Perspectives in Organic Chemistry* (Edited by A. Todd) p. 155. Interscience (1956).

[4]J. B. Hendrickson, *J. Am. Chem. Soc.* **97**, 5763 (1975).

[5]S. N. Ananchenko and I. V. Torgov, *Tetrahedron Letters* 1553 (1963); H. Smith *et al.*, *Experientia* **19**, 394 (1963); *J. Chem. Soc.* 5072 (1963).

[6]Taken from a survey of syntheses. In *Art in Organic Synthesis* (Edited by N. Anand, J. S. Bindra and S. Ranganathan), Holden-Day, New York (1970). The word "pieces" here is roughly the same as "synthons"[2] but that term has been confused by several different usages and so is avoided here.

[7]For a target of n skeletal atoms and r rings the number of skeletal bonds is $b = n + r - 1$. If it is cut into k pieces, the number of bonds cut is $\lambda = k + \Delta r - 1$, or if all pieces are acyclic, $\lambda = k + r - 1$. If the average piece is three carbons, $k = n/3$ and $\lambda = n/3 + r - 1$. The combinations of pieces × orders equal $\lambda! \times \binom{b}{\lambda} = \frac{b!}{(b-\lambda)!}$.

[8a]E. J. Corey and W. T. Wipke, *Science* **166**, 78 (1969). For more recent descriptions of their programs see *Computer-Assisted Organic Synthesis* (Edited by W. T. Wipke and W. J. Howe). ACS Symposium Series 61 (1977); [b]In their first paper Corey and Wipke drove a search for Diels–Alder reactions just from a six-ring in the skeleton; the idea was further developed by: E. J. Corey, W. J. Howe and D. A. Pensak, *J. Am. Chem. Soc.* **96**, 7724 (1974).

[9]K. K. Agarwal, D. L. Larsen and H. L. Gelernter, *Computers & Chemistry* **2**, 75 (1978).

[10]M. Bersohn, A. Esack and J. Luchini, *Ibid.* **2**, 105 (1978); see also *Chem. Rev.* **76**, 269 (1976).

[11]The term skeleton here is used strictly to mean the carbon skeleton. It will be possible, however, at a later date to incorporate heteroatoms in rings into the definition of skeleton. Construction reactions are defined as those which create C–C σ-bonds.

[12]A sequence of constructions with no refunctionalizations is defined as a *self-consistent sequence* since the functional groups remaining after each construction are consistent with the functional requirements for the next.

[13]J. B. Hendrickson, *J. Am. Chem. Soc.* **99**, 5439 (1977).

[14]L. Velluz, G. Valls and J. Mathieu, *Angew. Chem.* Intl. Ed., **6**, 778 (1967).

[15]A synthesis of the C_{20} quassinoids from two identical C_{10} molecules via the partial bondset of **3** is currently underway in our laboratories and the synthesis of chrysanthemic acid has been economically achieved[16] by the bondset in **4**.

[16]J. Martel and C. Huynh, *Bull. Soc. Chim. Fr.*, 985 (1967).

[17]W. S. Johnson, *Biorganic Chem.* **5**, 51 (1976); *Angew. Chem.* Intl. Ed., **15**, 9 (1976).

[18]J. B. Hendrickson, *J. Am. Chem. Soc.* **93**, 6847 (1971); *J. Chem. Educ.* **55**, 216 (1978).

[19]J. B. Hendrickson, *J. Chem. Inf. Comput. Sci.* **3**, 129 (1979).

[20]In order to distinguish the separate functionalities of Π and Z, the value of f is used with one or two overbars to separately indicate $\pi = 1$ or 2, respectively.

[21]Half-reactions of larger span can be comparably derived, as with vinylogous addition ($R\Pi \cdot \Pi\Pi \cdot \Pi\Pi \cdot \Pi H$) by inserting four more Π symbols into the string. However, we limit ourselves to half-reactions involving up to only three carbons, as the longer spans are so uncommon.

[22]There may also be unchanging functionality on adjacent carbons, usually activators like the carbonyl at β for the RH reaction α- to ketones. Although they do not themselves change and are not part of the active f-list, these adjacent functional groups often have a strong effect on reaction course and must be recognized.

[23]Oxidative $(--)$ and reductive $(++)$ full constructions can of course be used for affixation only if the two halves are identical and so have little synthetic value, although they can be used in cyclizations.

[24]The substrate(s) generated are then tested mathematically for structural reality: $0 \leq f \leq 4$ and $(4 - \sigma) \geq f \geq \pi$ and $\pi \geq 2$, with all values positive.

[25]The program described here first finds all skeletal dissections followed by all functionality generations. While this logically follows from the prior discussion, from the

standpoint of computing efficiency it is not yet clear whether a hybrid algorithm may not be faster. We are considering alternatives in which one skeletal cut is in some cases followed directly by functionality generation before further skeletal cuts. Some exploration of these hybrid alternatives will be required to clarify their relative efficiency.

[26]Two papers describing the details of the computer operations have been submitted for publication.

[27]The starting material catalog consists of those compounds in the Aldrich catalog which are also stored in connectivity table form in the data base of the National Institutes of Health/Environmental Protection Agency. We thank Dr. G. W. A. Milne of the NIH for providing these and Dr. A. E. Fein of Fein–Marquart Associates for providing the data base search.

[28]J. B. Hendrickson and E. Braun-Keller, *J. Comp. Chem.* in press.

[29]L. Spialter, *J. Am. Chem. Soc.* **85**, 2012 (1963); *J. Chem. Doc.* **4**, 261 (1964).