The Knowledge Engineering Review, Vol. 12:4, 1997, 369–386

The application of computers to generate organic syntheses

JAMES B. HENDRICKSON

Department of Chemistry, Brandeis University, Waltham, MA 02254–9110, USA

Abstract

A major task of organic chemistry is the synthesis of compounds of interest (''targets'') from available starting compounds. We address here the generation by computer of the best synthesis routes, or sequences of reactions, to such compounds. The number of possible routes in the search space is enormous and so the reasoning is examined in some detail, leading to a protocol for the operation of our program, SYNGEN, which has been under development for some years. It requires a new representation of the data – molecular structures and their reactions – which is digital in format and quite unlike the usual graphical depictions of organic chemistry. This is presented as a problem in artificial intelligence, and every effort has been made to present it with a minimum background of organic chemistry required of the reader.

1 Introduction and background

The overlap of organic synthesis and computers has traditionally been small. The reason lies in the fact that organic chemistry does not deal with mathematics, and although it possesses its own rigorous logic, it rarely requires number-crunching for its problem solutions. It deals instead with molecular structures presented visually as stylized pictures, their transformations not easily or naturally reduced to mathematical, hence computer, formats. Accordingly, there has been very little connection between organic synthesis and artificial intelligence, although it would appear to be a natural one, as this review intends to show.

Organic chemistry deals with molecules which arise from the unique ability of carbon atoms to bond to each other into extended chains and rings, and also to append bonds to other atoms, usually hydrogen, oxygen, nitrogen and halogens (chlorine, bromine, iodine). The number of possible organic compounds is essentially unlimited, and the nature of organic chemistry is to understand the transformations (reactions) these compounds can undergo when ''reagents'' (other reactive compounds) are applied to them.

Reactions are general in that they alter the bonds only at a particular site of a few atoms in the molecule as a whole, and so may be generalized for use in many molecules which have the same defining group of reacting atoms and bonds, while their other atoms remain unchanged. The many molecules which make up any observable quantity of a compound are all the same, and so it is common for chemists to speak of the compounds in terms of their individual molecules, as defining their characteristic behaviour. Indeed, chemists tend to speak of compounds and their molecules interchangeably.

Organic chemistry is also the basis for understanding the organic compounds which make up the substance of living matter and its reactions in cell metabolism. The study of biochemistry is a subset

of organic chemistry; its study involves the synthesis of active biochemicals and medicines. The chemical industry exists to put the reactions of organic chemistry to work to make (i.e., synthesize), in large quantity, compounds of interest to society, such as plastics, dyes and pharmaceuticals.

The aim of this paper is to review the logic of the problem of designing synthesis "routes", i.e. sequences of reactions from available "starting materials", which might be used to synthesize a desired molecule (compound). These starting materials are compounds with relatively small molecules, already available for sale; thousands of such compounds are manufactured for the market, many in tank-car lots.

I propose to outline our development of a logic for synthesis design, to create a basis for the computer to derive the best syntheses for any given "target" molecule. Unlike areas of chemistry like quantitative analysis, there has traditionally been no rigorous, programmable protocol to follow in designing a synthesis, hence little attention to linking this problem with the computer. For some years my research has been focused on developing and programming this logic (Hendrickson, 1971), and my coworkers have been implementing it in the computer program SYNGEN[1]

While there are a handful of other synthesis programs in development worldwide (Barone & Chanon, 1986), each has taken a different approach and there has been little interaction among them. A comparative survey of these would be much longer, so I have limited this article just to the logical base of our effort with a minimum of chemistry detail. The intent here is to present this logic for a broad audience who may be interested in what kinds of computational methods can be used to automate this sort of mainstream task from natural science.

A synthesis route consists of a sequence of reactions proceeding from starting materials (SM) via intermediate compounds (I_n) to the target (T), a sequence of steps as in Figure 1(a), in which each step ($\rightarrow$) is a reaction. Since we know the target, we can reason backwards by applying *transforms*, reactions described in the reverse direction, as in Figure 1(b). The linear sequence here implies one compound changed into another one at each step. However, there are usually several starting material compounds with smaller molecules than the target molecule. Hence many of the reactions will join two molecules to make a bigger one.

This results in the elaboration in Figure 1(c), which is then reformulated in Figure 1(d) as a *synthesis plan*, a graph of the *process* of synthesis in the forward direction (Hendrickson, 1977), in which the lines represent reactions and the molecules of the compounds are lettered as intermediates (I_n) or starting materials (SM_n). The pairs of angled lines represent a single reaction which joins two molecules, read from left to right.

The complexity of the problem arises from the fact that there are thousands of starting materials available and literally millions of possible routes to any target from them, a combinatorics problem

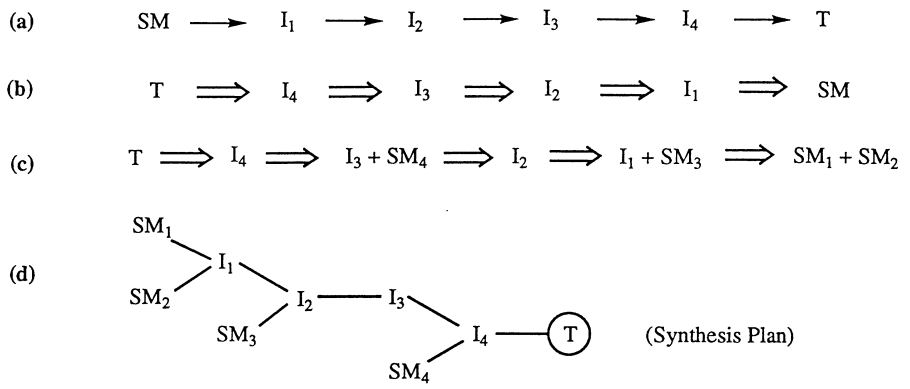


Figure 1 Generalizations of synthesis routes

[1]This article was taken from a lecture given at the AAAI Spring Symposium on Systematic Methods of Scientific Discovery in March 1995, and adapted from a lecture at the BASF Jubilee Symposium (Hendrickson, 1990).

of vast scale – its size is discussed further below. To solve the problem, we are given the molecular structure of the target (T) and a catalogue of starting material structures (SM). We also have the whole text of organic reactions, which may be expressed as transforms in reverse, and so generate these synthesis plans back from the target molecule.

We may start by applying transforms to the target molecule, different sites on which will transform variously into different intermediates. Each of these is in turn a subtarget for further transform application, and soon we will generate millions of synthesis plans (i.e. routes) of varying length. In effect, we are generating the plans of Figure 1(d) stepwise backwards from the target (T). These sequential generations of synthesis plans will each be terminated when all the last molecules generated are starting materials (SM) found in the catalogue.

We may generalize this procedure for solving the problem as: a starting state, the target molecule (T); new states (I_n) generated by recursively applying transforms as operators; and a test of success in matching starting materials found in the catalogue. The problem is parallel to that of Valdes-Perez (1995) to elucidate all multistep pathways for the stepwise *mechanism* of a single complex chemical reaction in terms of the involved atoms and the bonds which are made or broken. As he points out, "The decision about when to apply each operator is guided by heuristics. In real-life tasks, problem spaces are often too large to be searched unselectively and so the selectivity is provided by the heuristics." Thus, it is these heuristics which we must seek to apply with the computer.

A fuller understanding of the process may be gained by clarifying the minimum essentials of reactions. Organic reactions are basically seen as the overall breaking and making of particular bonds in the involved molecules, using selective reagents (often simple inorganic compounds: $NaBH_4$; CrO_3; $AlCl_3$, etc.). Many general reactions have been discovered or invented over a century and a half, and more continue to be developed each year. Figure 2(a) illustrates a simple example, which combines compounds **A** and **B** to form **C**. The compounds are represented by their individual molecules. **A** + **B** are first combined to form an unstable entity (molecule with an O-Li bond) which in the laboratory is directly converted to product **C** by adding water. This is considered as a single step and would commonly be generalized for synthesis, as in Figure 2(b), showing just the overall change in compounds **A** and **B** to form **C** and generalizing the iodine (I) as any halogen (X).

Further chemist shorthand is shown in Figure 2(c). The lines are bonds between the atoms or groups at each end, with carbon understood as being present at the end of any otherwise unlabelled

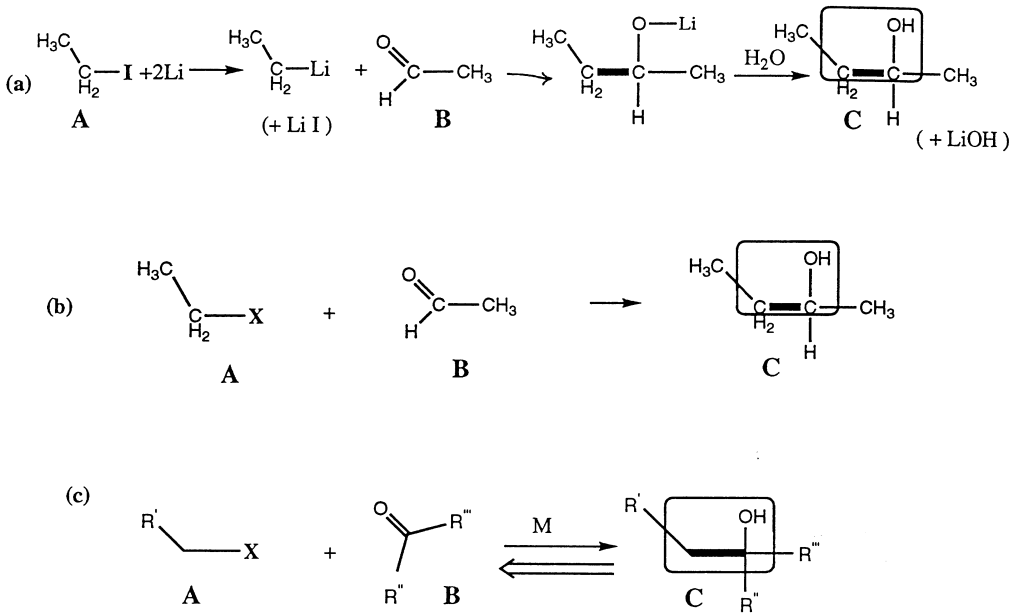


Figure 2 Sample organic reaction

bond, hence at the discontinuity joining two bond lines. The hydrogens on each carbon are only implied; there are always enough to make up four bonds to every carbon (there are two hydrogens bonded to the carbon at the left end of the bold bond line in **C**, none at the right end which already has four bonds).

The groups R in Figure 2(c) just generalize the rest of the molecule which does not change. M is used to generalize some metal, as Li or Mg here, but the actual inorganic compounds (LiI, LiOH) are left out as extraneous to the synthesis focus. The reaction as executed in the laboratory proceeds to the right, from reactants **A** and **B** to product **C**, as shown with a simple arrow. The corresponding transform is indicated with a double arrow and shows in reverse the precursors **A** and **B** which will make the product **C** with this reaction. The atoms which change their bonds make up the reaction center and are circled in the three representations in Figure 2.

The traditional approach asked the computer to emulate the moves a chemist would make (Corey & Wipke, 1969; Wipke, 1978), starting with the structure of the target, T, and applying transforms to generate the possible precursors, then taking each of these as a new target to generate its precursors, and continuing stepwise backwards until the user arrives at starting materials which he knows he can buy at the store.

This process simply offers to show the user all possible alternatives, and allows him to pick what he likes. For this reason it is called Computer-Assisted Organic Synthesis (CAOS), and is set up as interactive with the user. There is in this simple procedure no recognition of the huge size of the generated output, and so the critical aspect of judging which routes are best is left to the user, not winnowed by the computer.

A number of criteria to reduce the output by user-chosen constraints have since been overlaid in the major program of this kind, LHASA (Corey et al., 1985), and a focus on recognition of available starting materials has recently been undertaken (Johnson et al., 1992). However, the output remains swollen in part because full chemical detail is employed to describe reactions, and so many routes appear which differ only in trivial ways – no generalization or abstraction to essential detail is applied. The other programs under construction (Barone & Chanon, 1986) share these problems, and none are sharply focused on a criterion defining the best syntheses and applied by the computer.

This procedure also depends upon the presence of a library of known transforms, kept as a database in the computer and keyed to be applied wherever the nature of a product under consideration matches that of a transform in the library. The transforms are equipped with crude heuristic caveats (“no good if ‘X’ is present in molecule, etc.”) and estimations of the reaction yields, i.e. how much product should be formed in the reaction (some of the compound is always destroyed in a reaction). These yield estimations are notoriously inexact and not accurately generalizable, since real examples differ a lot from one to another. Furthermore, a full library of transforms can never be complete and must be constantly upgraded to incorporate new reaction types as they are discovered.

2 Logical basis for synthesis generation

We undertook to apply artificial intelligence reasoning to the problem (Hendrickson, 1986). The search space is clearly illustrated by the *Synthesis Tree,* a graph of the process of synthesis, shown in Figure 3. The lines are reactions, as in Figure 1(d), directed to the right (their transforms to the left); the many points at the ends of the lines are the intermediate compounds, and the larger dots signify available starting materials, diminishing in size to single-carbon materials, **C**, at the far left in Figure 3.

Two representative synthesis routes are shown as heavy lines, sequences of reactions from starting materials to the target (T) via intermediates. These are individual synthesis plans (as in Figure 1(d)) embedded in the tree. The levels shown represent the number of steps to the target, and below are shown the overall yields if each step has a yield of 80%. These mean that the farthest starting material in the lower synthesis, at level 7, will be converted to the target in seven steps with 21% overall yield (i.e. $0.87 = 0.21$); the 21% yield tells us that about five units of that starting material will be required to make one unit of target and four units will be successively lost in the sequence.

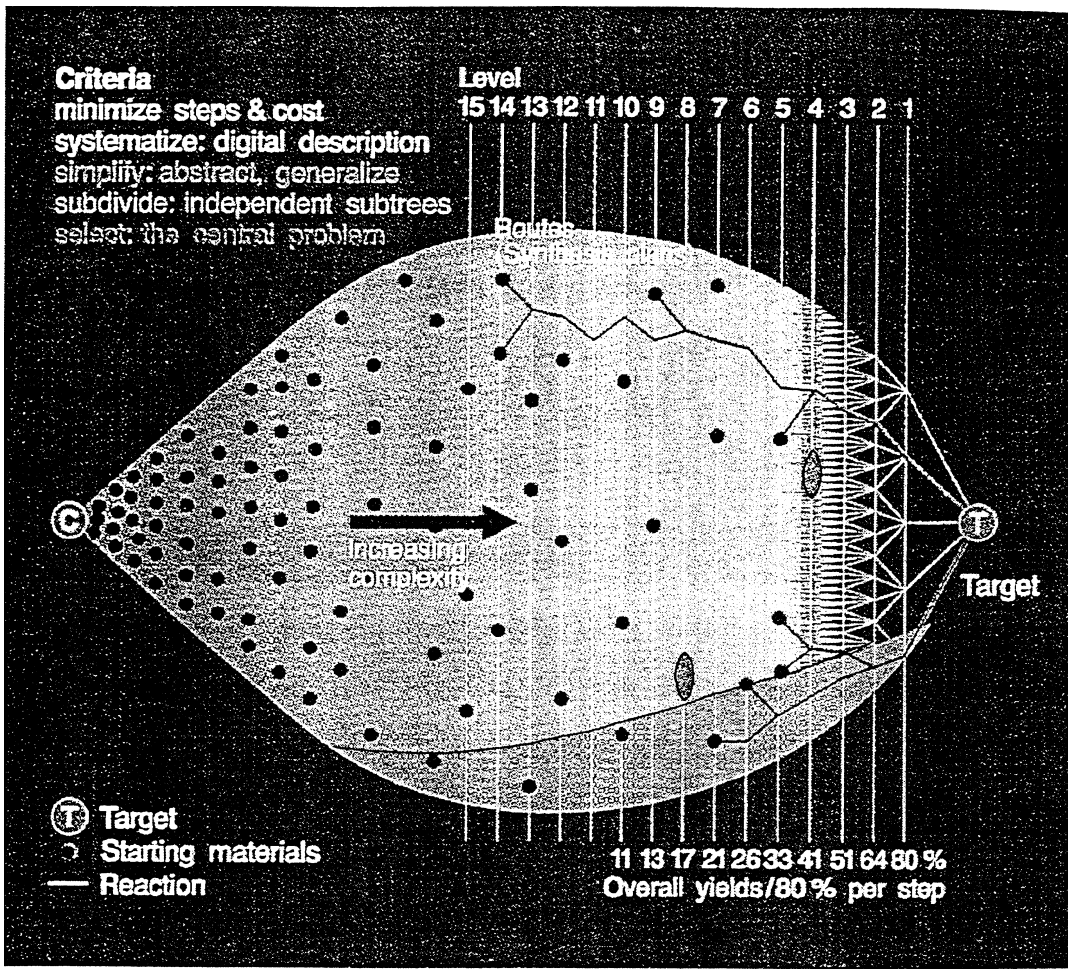


Figure 3 The synthesis tree

For the computer it is simple in principle to generate the whole tree for a given target; the real problem is to select the optimal routes. We can follow the traditional, stepwise backwards approach on Figure 3. It would not be unreasonable to expect 30 different reactions which would convert intermediates at the first level into the target (only five are drawn in Figure 3). Since each of these 30 intermediates will still be of comparable complexity, each will also have some 30 reactions from intermediates two steps (levels) back. To proceed back five levels will therefore produce 30^5, or 24 million, possible routes, and most syntheses are longer than five steps. Now the question is, which one is the best route to try out in the laboratory? This question unhappily is just left to the user in the traditional approach. In actual use, the user must drastically prune the tree right away at the first levels in order not to be swamped by alternatives. Hence, he may easily miss good routes long before they are fully generated.

2.1 Simplify, systematize, subdivide, select

The synthesis tree is obviously enormous, much larger than chemists intuitively appreciate. This tree of many synthesis plans is at once the source of the problem and the key to its solution. Because of its huge size, it is imperative first to *simplify* it, to generalize severely the representation of the data – the compounds and reactions. We must abstract the general chemical significance, coalescing trivial detail, as implied by the oval areas in Figure 3 which group together many intermediates and their reactions and make a much simpler tree.

This simplification is analogous to maps in that a map of a small area shows far more detail that

that of a larger region. In practice, we use the latter to find our way to the desired locale and the former to refine the detail as we arrive there. The same principle, to generalize first and refine later on only a few remaining best routes, is used here to seek optimal plans from the synthesis tree. This technique is familiar in artificial intelligence as abstraction spaces, in which the problem solving proceeds first with an abstracted representation of states.

Further, we must *systematize* the data in the search space in a digital format, not only for rapid computer manipulation, but also to ensure that all possibilities can be derived as simple mathematical combinations, so that none will be missed. This is quite impossible with the extensive notation required in the traditional approach for each structural representation of each compound in the tree.

Finally the simplified, generalized tree is still enormous, but it can be made more manageable if we can *subdivide* it into subtrees to allow their exploration one at a time. Such a subtree is shaded in at the bottom of Figure 3. This subdivision requires that the subtrees must be separate and independent of each other.

The central problem remaining is to *select* the few best routes. This requires defining criteria for "optimality", stringent enough to select only a few routes. We opted here for a *criterion of economy* – the shortest routes with the most efficient assembly from the cheapest starting materials. This focuses on minimizing the number of steps to maximize the overall yield. A look at the tree shows that this requires that we focus our search on the nearest starting materials in the tree instead of generating blindly back from the target as in the traditional approach. Further, we wanted all these best routes to be found by the computer through consistent application of the criteria, rather than leaving the process interactive, at the whim of the user and his preconceptions.

3 Focus on skeleton

An important dichotomy in examining molecular structures is that between their *skeleton* and their *functionality*. The skeleton is the backbone of linked carbon atoms, connected by single bonds. The functionality consists either of double (or triple) bonds between carbons on the carbon skeleton, or of bonds from the skeletal carbons to other kinds of atoms, usually oxygen, nitrogen and halogens. The sites of functionality on the carbon skeleton constitute the sites of chemical reactions. This is illustrated in Figure 4 for the cortisone molecule, with the skeleton shown in boldface lines; for the double bond within the leftmost ring, one bond is skeletal, the second is functional. Reactions can occur there, but not in the next ring, which is unfunctionalized.

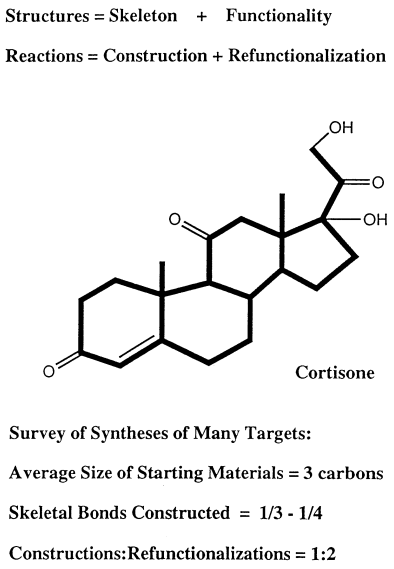


Figure 4 Dichotomy of structures and reactions

This dichotomy extends to reactions as well, by distinguishing *construction* reactions from *refunctionalization* reactions. The former are reactions that create skeletal bonds, those between carbon atoms, as in the example of Figure 2(c). The refunctionalizations simply alter the functionality without affecting the carbon skeleton.

The primary importance of the skeleton and its assembly in synthesis is philosophical: *synthesis itself is a skeletal concept*. Synthesis consists of joining small starting molecules to make a large target; the average size of starting materials in a survey of many published syntheses is only three carbons, while the target molecules are commonly much larger. Therefore, the central, *obligatory* reactions are the skeletal constructions, which join these small molecule pieces. In principle, refunctionalizations are not needed at all, despite the observation that the average synthesis has twice as many refunctionalizations as constructions. Such reactions may be seen as repair jobs to fix the functionality resulting from one skeleton-building construction reaction, in order to prepare the intermediate molecule for the next one. These refunctionalizations interject extra steps into the skeleton-building sequence, and this is not economical of time and effort.

This skeletal focus leads to a stringent criterion for economy: to seek only *ideal syntheses*, those consisting *only* of construction reactions, since such routes, though rare, must be the shortest ones. In an ideal synthesis, the chosen starting materials have the correct functionality to initiate the constructions joining them; these construction reactions make a skeletal bond and also change the functionality on the active carbons. The subsequent constructions must each be initiated by the functionality left by the last, and lead successively to the target skeleton with all its correct functionality in place, as a natural consequence of the chemistry of the construction reaction sequence. Such an ideal synthesis route then will need no refunctionalization reactions at all in its direct sequence of constructions.

These considerations imply that we need to systematize the molecular structures of the compounds in the synthesis tree to deal efficiently but separately with the skeleton and the functionality on the skeleton, both for starting materials and for intermediate structures generated *en route*. We need to determine which skeleton bonds of the target will be the best ones to construct, and then what starting material functionality is required to sequentially construct those bonds, with viable chemical reactions, so as to result in the final target functionality directly.

The focus on skeleton drastically simplifies the tree. While there are perhaps 50,000 available starting materials, there are only 13 acyclic starting skeletons of six carbons or less. Nevertheless, there remains an enormous number of possible ways just to assemble a target skeleton from starting skeletons of only about three carbons. This arises from the "combinatorial monster" of what pieces to use and in what order to join them.

We may designate a *bondset* in a target skeleton as a set of skeletal bonds (λ in number) which are the ones to construct, as in Figure 5 for the skeleton of the steroid estrone. If we remove all the bondset bonds from the skeleton we are left with the several smaller skeletons of the starting materials. Thus, the bondset defines a family of syntheses which all make the same set of skeletal bonds from the same starting skeletons. These syntheses will differ only in the nature and sites of the functionality on the skeletons in the reaction sequences, and in the successive reactions which construct the skeleton and alter the functionality.

For any bondset, the order, or sequence, in which the bonds are made defines a generalized synthesis route, i.e. a plan of assembly of the starting skeletons into the target skeleton. This assembly plan derives from the ordered bondset, as in Figure 5. This is just a generalized route, or synthesis plan (Figure 1(d)), from the synthesis tree, but composed only of constructions. Indeed, the simplest gross description of any synthesis route – of its essentials – is this ordered bondset, which shows the starting material skeletons and the sequence of target bonds which are constructed to link them.

The number of possible ordered bondsets is the number of ways to dissect the skeleton in a certain order: for a target of b bonds this is $(\lambda!b!)/(\lambda!(b-\lambda)!) = b!/(b-\lambda)!$ – the classic number of permutations. To dissect the estrone skeleton ($b = 21$) by cutting five bonds ($\lambda = 5$, as in Figure 5) in any order would result in 2.44 million plans, of which only one is shown. This assembly plan is the

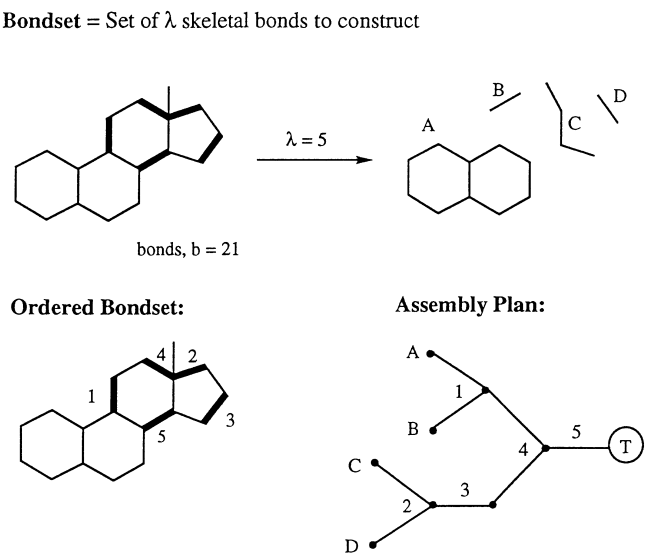


Figure 5 Skeletal bondsets and assembly plans

equivalent of the synthesis plan of Figure 1(d) but representing only the skeletal constructions in the route. It shows pairs of angled lines joining two skeletons, left to right (cf. A + B in construction #1), while horizontal lines imply a construction joining two carbons in one skeleton into a ring.

An assembly of estrone from starting pieces averaging three carbons would require a bondset of $\lambda = 9$ bonds and affords 10^{11} assembly plans (ordered bondsets), i.e. 100 billion routes, even without considering the chemistry involved. A dramatic realization of these combinatorics may be seen in the total synthesis of cortisone (Figure 4) from one-carbon units – constructing all 24 skeletal bonds – for which there are 6×10^{23} assembly plans: this means that to make a mole of cortisone (about a pound) one can make *every molecule* by a different route!

There are, however, two powerful criteria which allow a severe reduction in these numbers. The first is the choice of a *convergent* assembly plan, which can be shown to be significantly more efficient than a linear plan in terms of the total weight of starting materials required (Hendrickson, 1977). The convergent plans can be created very simply by cutting the target skeleton into two pieces (intermediate skeletons) at first level, then cutting each of these again into two more at the second level, and so on as needed. Most plans from random cutting are linear like the assembly plan in Figure 1(d); that in Figure 5 is convergent.

The second criterion is to cut no further than two levels and require that the four starting skeletons so formed in these two successive levels of dissection be found in a catalogue of actual, available starting materials. A large target for synthesis would have 20 carbons and so will result in four starting skeletons averaging five skeletal carbons each. Our experience with such a starting materials catalogue shows that available starting materials are plentiful in functional variety on skeletons of up through five carbons, but rapidly diminish above.

This suggests that by applying these two criteria we should be able to find realistic syntheses with at least some of these plans, and find only the shortest ones, with the fewest skeletal bonds to construct. Not only are these convergent plans the shortest and most efficient syntheses, they are also quite rare and so afford the needed severe reduction in the total bondsets that must be examined.

A sampling of various convergent bondsets from estrone by the two levels of cuts described above is shown in Figure 6. The ordered bondset from cuts 6, 8, 10 is the same as that in Figure 5, and represents the shortest commercial synthesis of estrone from Wyeth laboratories, while that from 6, 9, 10 corresponds to an experimental synthesis realized at Hoffmann-LaRoche but never commercialized.

The convergent bondsets created as in Figure 6 cut no more than two bonds per level (to cut a single ring in two), hence six bonds ($\lambda = 6$) or less in total. For estrone ($b = 21$) with $\lambda = 6$ there are

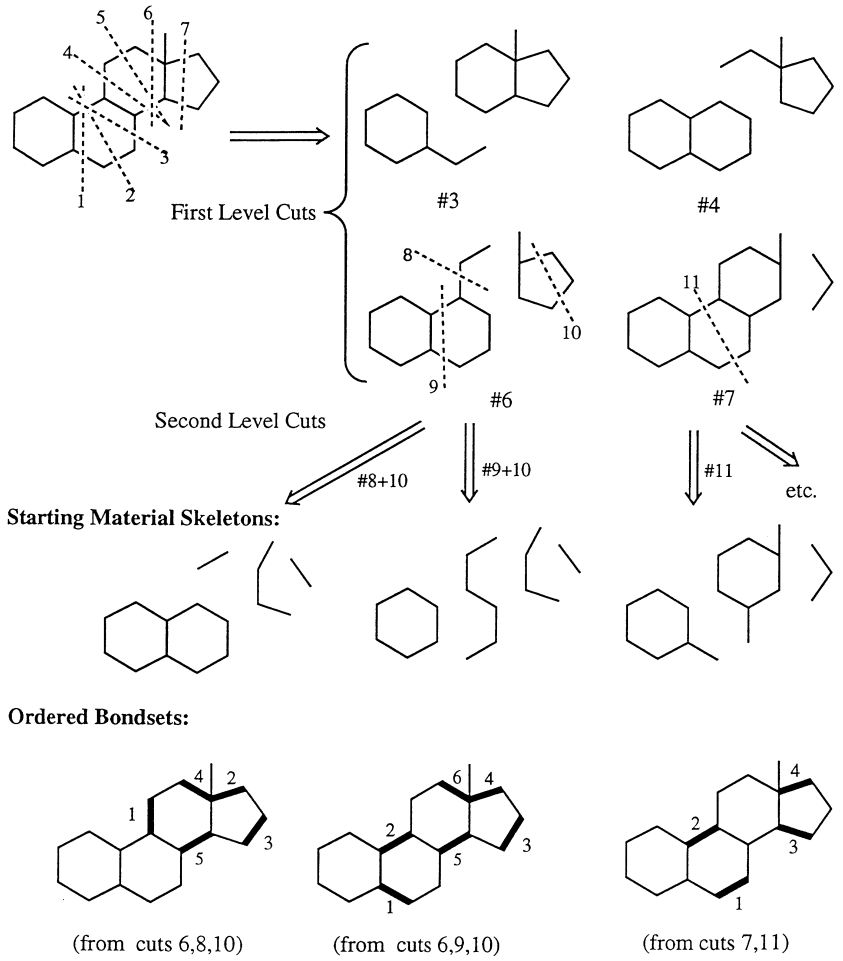


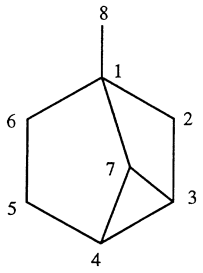
Figure 6 Skeletal dissection into several bondsets

39 million ordered bondsets possible, but the number of fully convergent ones from catalogue skeletons is less than 900, and most of these, examined with real functionalized starting materials and reactions, cannot be expanded into ideal syntheses. Hence we realize the required stringent selection of routes from the whole tree with these criteria: convergent dissection of the skeleton to real starting skeletons available in a catalog, and expansion into ideal syntheses (section 4) from real functionalized starting materials bearing those skeletons.

To implement this dissection process, the representation of any skeleton in the computer must be rigorously unique and compact. Figure 7(A) shows a sample skeleton of eight carbons, arbitrarily numbered. This is considered as a graph, and next to it is the adjacency matrix of its connectivity which fully describes the structure. This matrix may be strung out to form a linear binary string; this is now just a binary number and can be used as an identification number for this skeleton. However, for this eight-carbon structure there are 8! ways to number it, i.e. 40,320 equivalent numberings of the skeleton, from which we must find one which will always be unique.

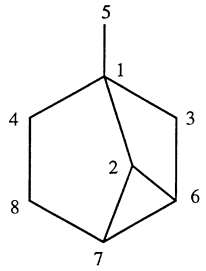
In fact this can be done by finding the one numbering for which its binary string will be the *maximum binary number* (Hendrickson & Toczko, 1983). Such a maximal numbering is shown in Figure 7(B), and the two corresponding binary strings are shown below. The maximal number is a unique canonical number which may be used as an identification for the skeleton, used to order ID numbers for an index like the starting materials catalogue, and used to compare generated skeletons for their identity with any in the catalogue. The catalogue is just a list of maximal binary ID numbers of its contained skeletons, arranged in numerical order so that any search for a match with a generated skeleton becomes a simple binary search, very fast in the computer. We have also created a program to redraw a skeleton graphically from its maximal binary number.

A. Sample Numbering



	1	2	3	4	5	6	7	8
1		1	0	0	0	1	1	1
2			1	0	0	0	0	0
3				1	0	0	1	0
4					1	0	1	0
5						1	0	0
6							0	0
7								0
8								

B. Maximal Numbering



	1	2	3	4	5	6	7	8
1		1	1	1	1	0	0	0
2			0	0	0	1	1	0
3				0	0	1	0	0
4					0	0	0	1
5						0	0	0
6							1	0
7								1
8								

Binary Identification Numbers

A.: 10001111000001001010101000000

B.: 11110000001100010000001000101

Figure 7 Maximal canonical numbering

4 Systematizing reactions and functionality

To proceed now to the chemistry for synthesis routes, we must first represent the functionality on the skeletons and the changes it can undergo in chemical reactions. The functionality will consist of multiple bonds between skeletal carbons, and attached groups of non-carbon atoms (excluding hydrogen atoms, which are normally not reaction sites). Such a system must be digital for the computer, simple but rigorous in its description of chemistry and capable of generalizing and simplifying by coalescing trivial structural detail.

4.1 Representation of structure

The system summarized in Figure 8 (Hendrickson, 1971, 1978) designates four synthetically important kinds of bonds which any skeletal carbon may have. The first kind, labelled **R**, is a single bond to another carbon, i.e. the skeletal connection of any carbon. (The label **R** is common in organic chemistry for groups attached by a carbon-carbon bond, as in Figure 2(c).) The two kinds of functionality mentioned above are: the extra (multiple) bonds between carbons, commonly called π-bonds and labeled as Π here; and the functional bonds to other atoms (commonly nitrogen, oxygen, halogen, etc.), which chemically have a lot in common and are grouped together labelled **Z**. This leaves the remaining bonds to hydrogens, labelled **H**. A more general definition of **Z** and **H** for chemists is that the former are attached atoms more electronegative than carbon, the latter more electropositive.

The number of each kind of bond to any carbon is designated respectively as σ, π, z and h with values of 0–4, and as these must add up to four, the valence of carbon, there are only three variables. We may therefore define the *character* of any carbon by its value of $\sigma z\pi$, since h may be found by the difference of the others from four. The values of the four variables are all 0–4 (π is only 0–2). The character of any carbon is then just a three-digit number. Since the skeletal representation above

implicitly defines σ, only two digits, z and π, serve to represent the functionality of any carbon. The whole molecule may then be digitally and compactly described by a zπ-list, over all the skeletal carbons numbered in the canonically unique maximal order. For the computer only four bits are required per carbon to represent zπ. The whole generalized molecule is now represented first by the maximal binary identification number for its skeleton (from which the σ value may be obtained), followed by the zπ-list of the carbons, also in maximal order, for its functionality.

The importance of this compact, digital representation lies in the ease with which any newly generated compound may be compared with others, either from the synthesis generation or in the starting materials catalog. Furthermore, this comparison may be made by skeleton alone or by the full skeleton-functionality notation, in either case a simple binary number.

The functional notation is shown in Figure 8 for an example molecule, which is just numbered here linearly, not maximally. The values of z and π in the zπ-list are easily understood by reference to the carbons in the molecule above. The relation of the system to real chemistry is emphasized by the fact that it correctly affords the oxidation state (x) of any carbon (and so of a whole molecule as Σx_i) by the simple equation: $x = z - h$, and these values of x are also illustrated in Figure 8. This calculation of x provides as well for the recognition of the important oxidation state change in any reaction converting compound A to compound B, i.e. $\Sigma\Delta x_B - \Sigma\Delta x_A$. It is important to know this oxidation state change in a reaction to understand the kinds of reagents needed to effect it. A negative $\Sigma\Delta x$ implies a reduction and requires a reducing reagent; a positive $\Sigma\Delta x$ needs an oxidizing agent.

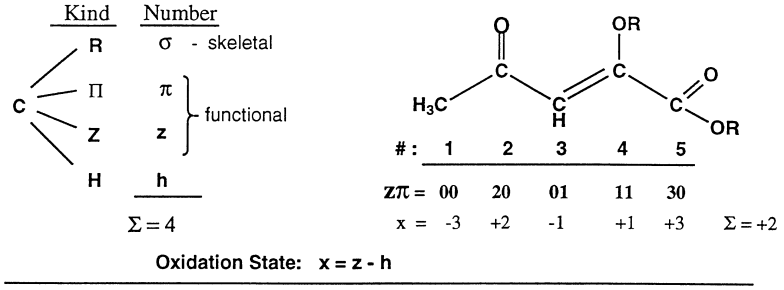
4.2 Representation of reactions

Reactions may also be described using the system presented in Figure 8, for any reaction may be described by the net change, from the reactant to the product, in the character, or bonding, at each skeletal carbon. The most basic reaction change is a simple exchange of one kind of bond for another at any carbon. For the four kinds of bonds, there will be 16 possible combinations at any one carbon, expressed by two letters, the first for the new bond made, the second for the old bond broken.

These combinations are listed below in Figure 8 and named in common general chemical terms. Thus, a reaction which serves to replace a bond to oxygen, halogen, etc., by a bond to hydrogen is generalized as **HZ**. This is a reduction since the change in z is –1 and the addition of h is +1, so the change in the oxidation state is $\Delta x = \Delta z\text{-}\Delta h = -2$. All of the 16 simple exchanges at carbon in Figure 8 show an associated oxidation state change, which can be similarly seen to be equal to –2, –1, 0, +1 or +2. These oxidation state changes are implicit in the 16 single atom exchange notations.

In many reactions there may be several carbons changing their bonds at once. Obviously, if one carbon gains an **R** at the expense of some other bond, another carbon, linked to it in the product, must also gain an **R**, thus forming a carbon–carbon bond between the two. This will describe a construction reaction overall and serves to build the skeleton. Similarly, a gain of Π on one carbon must also result in a gain of Π on an adjacent carbon (an elimination reaction in Figure 8, implying elimination of either a **Z** or an **H** on each of two adjacent carbons to form a double bond between the carbons). This is just a refunctionalization reaction.

Thus it is not uncommon to involve more than one carbon in a reaction, with simultaneous exchanges of attachments on several adjacent carbons. A *unit reaction* is defined as a single exchange of attachments on one or more linked carbons, rarely more than six carbons and usually only three or less. The circled reaction centre in the product in Figure 2(c) may now be simply described as **RZ • RZ** at the two carbons which exchange bonds, i.e. a reductive construction ($\Sigma\Delta x = -2$). The unit reaction for a simple elimination reaction with $\Sigma\Delta x = 0$ would similarly be labelled with the simple exchange on each carbon as $\Pi Z.\Pi H$, the oxidation state change at the first carbon being –1, the second +1. A dot is used to separate the notation for each changing carbon; if the reaction forms a skeletal bond (**R**) the dot is distinguished in boldface as with **RZ • RZ** for the changes at the joining



16 Reaction Exchanges at any one carbon:

Substitution	**ZZ HH** (0)	**Addition**	**HΠ** (-1) **ZΠ** (+1)
Oxidation	**ZH** (+2)	**Elimination**	**ΠH** (+1) **ΠZ** (-1)
Reduction	**HZ** (-2)	**Construction**	**RH** (+1) **RZ** (-1) **RΠ** (0)
Rearrangement	**RR ΠΠ** (0)	**Fragmentation**	**HR** (-1) **ZR** (+1) **ΠR** (0)

(The corresponding oxidation state change of each is shown in parentheses)

Figure 8 Generalized characterization of compounds and reactions

two carbons in Figure 2(c). Organic chemists would write the ΠZ.ΠH elimination as a generalized reaction like this: $-CZ-CH- \rightarrow -C{=}C-$.

As indicated below, most actual reactions are unit reactions. An interesting and valuable consequence of this new representation of reactions is that the *reaction distance* which separates any two compounds may be calculated quite simply by equation (1), using the sum of changes in h and z between the two compounds (Hendrickson & Braun-Keller, 1980). This reaction distance tells how many successive unit reactions (N_R) will be necessary to convert one compound into the other. Thus the single reduction step **HZ** in Figure 8 has $\Delta h = +1$ and $\Delta z = -1$, so that $N_R = 1$ step.

$$\text{Reaction Distance, } N_R = 1/2\Sigma_i\{|\Delta h_i| + |\Delta z_i|\} \quad (1)$$

The construction reactions are those with which synthesis design is specially concerned, for building the target skeleton. They are the only reactions in an ideal synthesis. A construction may be seen as the combination of two half-reactions, one on each of the joining carbons. Usually one is oxidative and one reductive, although the example in Figure 2(c) has two reductive half-reactions (each is **RZ**). Chemically, the two half-reactions are usually independent and may be joined in all combinations to make a full construction reaction.

The value of such a system of representation is that it affords an abstract description of every possible change a carbon may undergo. Hence, it forms a basis for a rigorous organization of all possible organic reactions (Hendrickson, 1992, 1995). Surprisingly, there has not heretofore been a recognized logical system of documenting organic reactions, perhaps because organic chemistry has traditionally not been mathematically oriented (Hendrickson, 1997).

Putting together those changes that must involve a neighbouring carbon, i.e. $\pm$**R** and $\pm\Pi$, we can now build a complete table of all possible unit reactions on 1–3 carbons and such a table is presented as Figure 9. It is a kind of "periodic table" of organic reactions, in which every possible unit reaction has a designated place. It may be noted that there are three implicit variables in the reaction notations as there are in the structures. These may be expressed as $\pm\Sigma\Delta\sigma$, $\pm\Sigma\Delta\pi$ and $\pm\Sigma\Delta x$, the latter incorporating Δz and Δh. These are, respectively, skeletal construction/fragmentation, addition/elimination to π-bonds (double or triple bonds between carbons), and oxidation/reduction.

Hence the table of possible changes should be three-dimensional, with planes of $\Sigma\Delta x$ vs. $\Sigma\Delta\pi$ at different, vertically displaced levels of $\Sigma\Delta\sigma$. Reactions with $\Sigma\Delta\sigma = 0$ are refunctionalizations, those with $\Sigma\Delta\sigma = +2$ are constructions (those with $\Sigma\Delta\sigma = -2$ are fragmentations, breaking carbon–

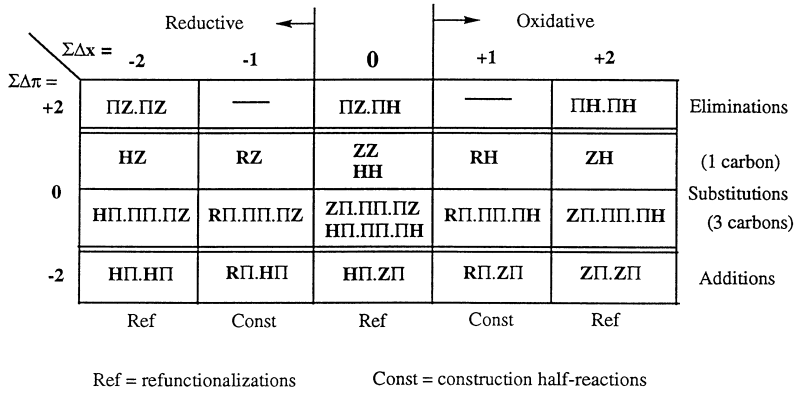


Figure 9 Periodic table of organic unit reactions for one to three skeletal atoms

carbon bonds). The constructions and refunctionalizations are all collected together in Figure 9 in only one plane, with coordinates $\Sigma\Delta x$ and $\Sigma\Delta\pi$. The constructions are shown as half-reactions ($\Delta\sigma = +1$ each), six combinations being possible on 1–3 carbons, and the refunctionalizations as $\Sigma\Delta\sigma = 0$. The fragmentation half-reactions (**–R** in Figure 8, including the rearrangement **RR**) have been left off of the table here as they are not used in the SYNGEN procedure.

Exploring the reaction table in Figure 9, we find the elimination example above, ΠZ.ΠH, in the elimination row and in the column for zero oxidation state change; the other two elimination reactions are either oxidative or reductive with $\Sigma\Delta x = -2$ or $+2$. The corresponding addition reactions are the bottom row with reversed symbols; each reverse pair corresponds across the midpoint of the table. The construction half-reactions all create **R** at the expense of **H**, **Z** or Π.

The two reductive **RZ** construction half-reactions which comprise the construction of Figure 2(c) are a substitution of **R** for **Z**, found in the table of Figure 9 at the left in the substitution row and the oxidative **RH** substitution at the right. The **R**Π construction exchange must have a further –Π exchange on the adjoining carbon, as in the additions row at the bottom, or in the three-carbon complex substitutions. The importance of this rigorous summary of organic reactions lies in its value to apply concisely and digitally a general reaction type to any reaction.

5 Systematic reaction retrieval from databases

The described representation of unit reactions allows for cataloguing the reaction entries in a database of reactions from the chemical literature, arranging them in the families of Figure 9. Their annotation for the computer consists of subtracting the binary $z\pi$-list of the changing carbon string in the product molecule from that in the reactant. This affords a $\Delta z\pi$-list, which may serve as a binary identification number for the reaction family. When this is done, it turns out that each family is characterized by a unique binary number.

Of course, not all observed chemical reactions are just simple unit reactions as in Figure 9, but a computer inspection of over 500,000 different published reactions in several reaction databases shows that about 80% are in fact unit reactions, like that in Figure 2(c). Another 15% are "composite reactions", consisting of two successive unit reactions on overlapping strings of linked carbons (Hendrickson & Miller, 1990). It is possible to systematically derive all possible combinations of two unit reactions on overlapping strings: there are just over 200 of these composites and they all show unique $\Delta z\pi$-lists as well. Hence a catalogue of some 95% of the cases in any database of literature reactions may be assembled for easy retrieval by the computer using the simple notation of the table in Figure 9.

We have now implemented this reaction organization system into a program called COGNOS, which serves to match an input query reaction with the entries in a database, which are themselves indexed by families for quick retrieval (Hendrickson & Sander, 1995). The COGNOS program with

a graphical interface has now been integrated with a database of 400,000 literature reactions and made available on a single CD-ROM for rapid reaction retrieval[2]

This has proved to be a very practical use of the generalizing representation in Figures 8 and 9. The user draws a reaction, as in Figure 2(c), and COGNOS immediately responds with the number of database entries of the same family. Then the number may be reduced substantially by demanding closer matching with the values of σ, z, π on the reaction centre in the query. Its success in finding correct matches lends confidence in this logic for organizing reactions, and the speed with which it finds matches from very large databases makes it very practical to use.

6 The SYNGEN program

Returning to the application of the logic to synthesis generation, we may briefly see how it is applied in the SYNGEN program. To generate reactions from bondsets the program is given a reaction product as its skeletal ID plus zπ-list, as well as a designated bond from the current bondset under examination. Strands of up to three carbons out from each end of the designated bond to construct are assigned labels α, β, γ. SYNGEN then examines the functionality (zπ-list) on these strands and applies the Δzπ-list generator for each possible construction to provide the zπ-lists of the corresponding reactants. As it manipulates bit-strings instead of graphs, the process is very fast for the computer.

An example, illustrated in Figure 10, shows the product graphically at upper left and the zπ-list for one $\alpha\beta\gamma$ strand to the right, each zπ value shown under its carbon atom. The reactant is similarly shown below. Its zπ-list is simply generated by adding the Δzπ-list of this reaction to that of the product. The construction reaction corresponding to this Δzπ-list operator is seen to be **RH•RΠ•HΠ**, composed of two half-reactions from the table in Figure 9.

An overview of the protocol for SYNGEN operation is presented in Figure 11 for a target from the commercial synthesis of the hormone estrone. The protocol has two successive phases. First, only the target skeleton is examined, shown down the left side of Figure 11. It is dissected through

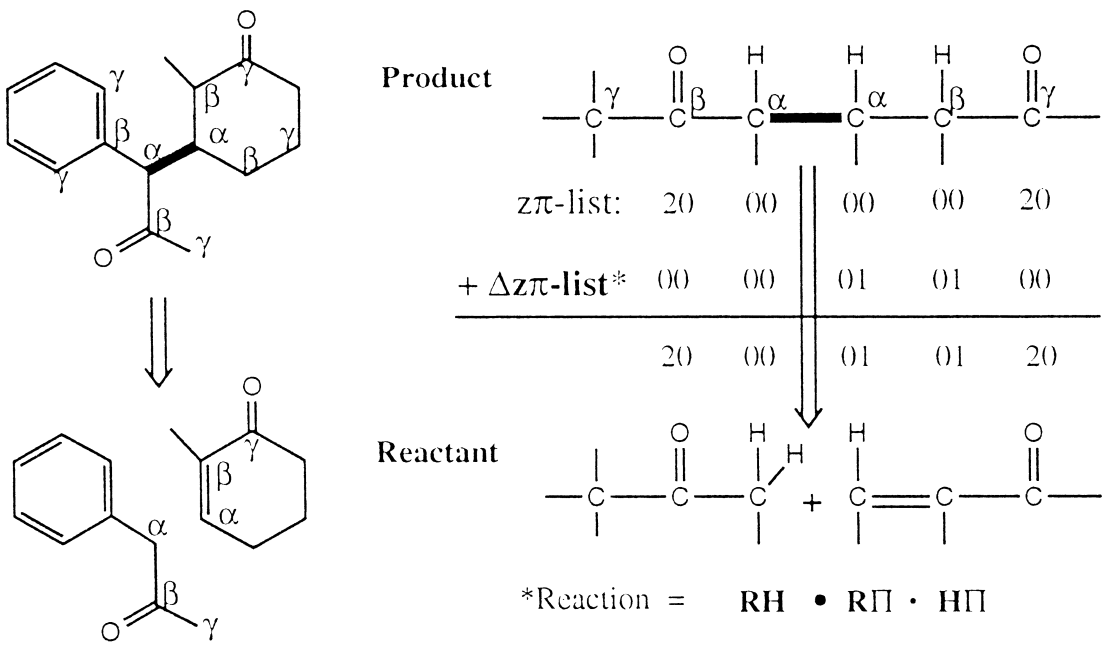


Figure 10 Rapid generation of reactions in SYNGEN

[2]The COGNOS reaction retrieval system with graphic input and display and an archive of 400,000 reactions is available on a CD-ROM from InfoChem GmbH, Landsbergerstrasse 408, 81241 Munich, Germany (tel.: 499–89–58–30–02) or from the publishers Springer-Verlag.

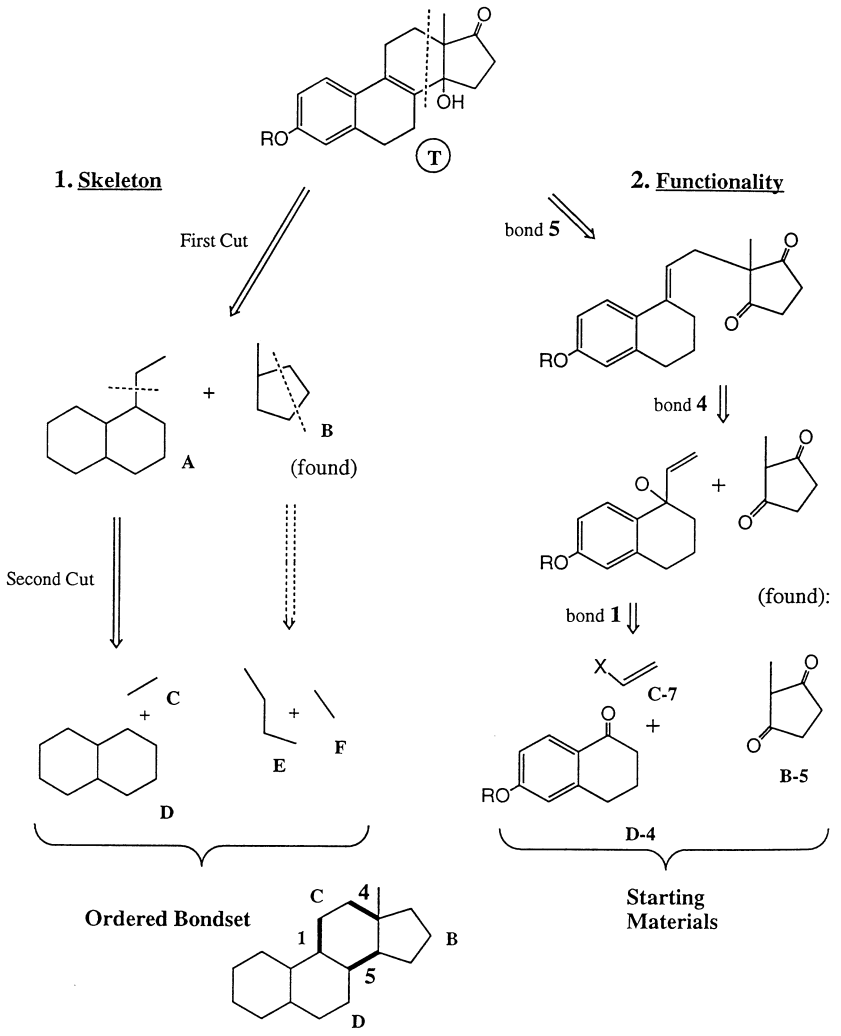


Figure 11 Overview of the SYNGEN protocol

two levels of cuts to four starting material skeletons, and only the resulting ordered bondsets with all four skeletons found in the catalogue are accepted. Only one of the bondsets so derived is shown here (essentially the cuts 6, 8, 10 in Figure 6). Skeleton B is in fact found in the catalogue and so it is flagged not to cut further (dotted arrow) if a correctly functionalized version is found in the second phase.

In the second phase, the process starts again with the target, now with functionality as well, and the designated bondset. Down the right side of Figure 11 is shown a sequence of the reactions successfully derived by the process of Figure 10 applied successively to the intermediates generated one bond at a time from the ordered bondset. By the time the bonds in the ordered bondset have all been successively processed to generate the functionality, many routes on that bondset will have been generated and the full starting materials (functionality as well as skeleton) will have been produced for each. These generated starting materials are now looked up in the catalog. Only if all are found is the sequence accepted as a viable route. In the one route illustrated in Figure 11, the three starting materials, B-5, C-7 and D-4, and the reaction sequence from these to the target duplicates the commercial (and ideal) synthesis route to estrone, as originally developed at Wyeth and generated independently by SYNGEN.

Within any one bondset the chemistry may afford a number of construction reactions to any intermediate. In the convergent model bondset in Figure 5, the assembly plan may include ten reactions to make a single intermediate from A and B, and another ten or so reactions to make an

intermediate from C and D. Combined these already afford 100 routes to the target. Even these shortest routes will therefore combine a lot of varied chemistry. These are generated by SYNGEN from the abstracting of all reactions in the generalized periodic table (Figure 9), and the question which arises now is how good are these generated reactions in practice.

7 Validation of reactions in SYNGEN via COGNOS

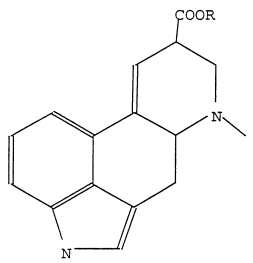
Originally SYNGEN was equipped with an array of broad chemical tests applied to the reactions generated (as in Figure 10), to weed out in practical terms those reactions which were certainly nonviable in terms of general chemical experience. Carried too far, however, this restriction is in danger of eliminating novel chemistry which might be made to work, as well as deleting reactions which might be all right in the particular context. The solution would seem to be to validate the generated reactions by reference to the published literature. This implies assessing the practicality of these reactions by comparison with a reaction database, hence to COGNOS.

Of over a half million reactions available to us, we find an archive of nearly 100,000 constructions. Since these are all formatted in COGNOS by the same $z\pi$-list system used by SYNGEN, it is relatively simple to create a fast look-up table for the computer to match any reaction generated in SYNGEN to the family in COGNOS, further pruned by requiring the same $\sigma z\pi$ character on the changing carbons. This allows labeling each reaction generated with the number of matches found and their observed average yield.

Now the routes created by SYNGEN may be grouped as to precedented practicality, leaving aside those with little literature precedent for separate examination if desired. This use of an installed database is different from that of the other programs mentioned in section 1, since in SYNGEN the database does not control the generation of reactions. All possible reactions are generated first via the "periodic table" of Figure 9, and only then checked against database precedent; those without precedent are still available for scrutiny by the chemist.

8 Practical aspects of synthesis generation

In practice the SYNGEN program is very easy to use. The user simply enters graphically the structure of his target and the program takes over. It generates without any user interaction the shortest ideal syntheses, from real starting materials in the catalog, in just a few minutes. The output is stored in a directory and the results summarized on a screen like that in Figure 12 for the ergot drug lysergic acid. This shows some 905 routes generated and the number of bondsets, starting materials, intermediates and reactions in each of the two levels of dissection. These may be



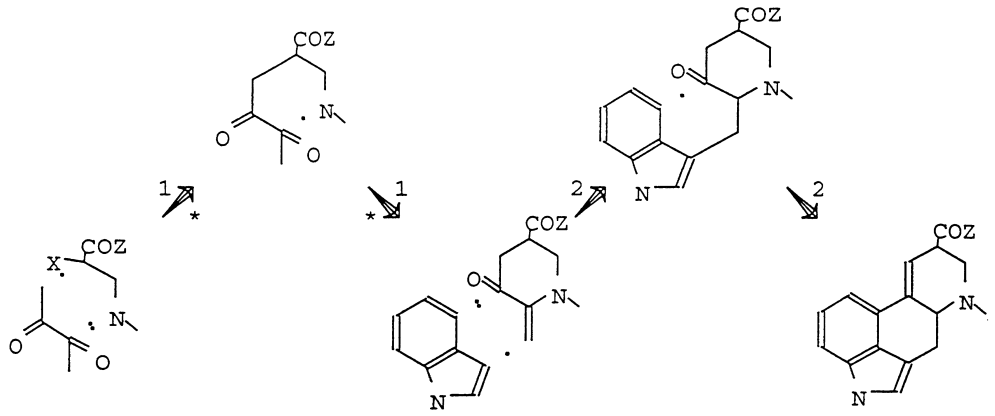
LYSERGIC ACID - 20-MAR-96

MIN SIZE: lev-1 : 3
Δ FG=2 : 8
Δ FG=1 : 5

TOTAL ROUTES
905

LEVEL	BONDSETS	ST. MAT.	INTERMED.	REACTIONS
1	5/ 5	10/ 10	24/ 24	65/ 65
2	26/ 26	71/ 71	0/ 0	132/ 132

Figure 12 Summary output from SYNGEN for lysergic acid synthesis




```
                                        Route #    2
                                    STEPS=            4.
                                    TOTAL SM=        45.
                                    TOTAL COST=     554.
1:11-A1: Alkylation - Enolate
1:A1-2F: N nucleophile - Carbonyl addition/elimination
2:P3-12: Imine-Enamine - Conjugate addition
2:P1-2F: Pi-Nucleophile: H-substitution - Carbonyl addition/elimination
```

Figure 13 Sample synthesis route for lysergic acid

separately viewed and any deleted as desired. The whole routes themselves may be viewed sequentially, ordered by overall cost; one such route is illustrated in Figure 13.

The cost of a route, equation (2), is determined by the cost (Cj) of each starting material (j), amplified by the product of the yields (y_i) of the n reactions each passes through, plus a cost (Q) for each reaction step (for solvents and reagents and for the time spent doing the reaction), with s = number of steps overall.

$$\text{Route cost} = \Sigma_j(\Pi_i^n 1/y_i)_j C_j + sQ \tag{2}$$

The value of SYNGEN lies not only in finding routes to target molecules for the first time, but also in locating all feasible alternative routes which might be environmentally more benign, for the synthesis of existing commercial chemicals. To this end we are currently assembling a quantitative scale of environmental hazard (toxicity, carcinogenicity, etc.) for the starting materials in the catalogue and also appending certain computer estimations of these properties for the intermediates produced, since they will be compounds previously unknown.

Originally written in Fortran over a decade ago, the program has now been upgraded and translated into C, with roughly 40,000 lines of code. In its present state, SYNGEN probably represents over 20 years of development work, during which time the concise description presented here has slowly clarified (cf. Hendrickson, 1986).

In summary, the logic presented here was developed to drastically simplify the synthesis tree of possibilities and further to represent them in a compact, digital format. Then the system in SYNGEN applies the criterion of economy very stringently to locate the shortest routes, requiring them to be convergent assemblies from the closest real starting materials in the synthesis tree, and using only ideal syntheses, containing no refunctionalization reactions. It does all this without interaction from the user. The operation of the program has shown it to be a very facile way of finding all the possible shortest syntheses for any target. It finds known syntheses which meet its criteria of brevity and suggests many other interesting routes. It has the particular advantage of assuring the user that he has not missed some short pathways to his chosen target.

Acknowledgements

I wish to acknowledge the wonderful spirit, devoted diligence, and remarkable computer skills of my colleagues who have made possible the work I have described here: Glenn Toczko, with me since the inception of SYNGEN some fifteen years ago; David Grier; Zmira Bernstein; Ping Huang; Todd Miller; Camden Parks; Thomas Sander; Li Su; and Lingran Chen. I also wish to specially acknowledge the enduring support of the National Science Foundation over these years.

References

Barone R and Chanon M, 1986. "Computer-aided organic synthesis" in Vernin G and Chanon M (Eds), *Computer Aids to Chemistry* Chichester: Ellis Horwood.

Corey EJ and Wipke T, 1969. "Computer-assisted analysis in organic synthesis *Science* **166** 178–192.

Corey EJ, Long AK and Rubenstein SD, 1985. "Computer-assisted analysis in organic synthesis" *Science* **228** 408–418.

Hendrickson JB, 1971. "A systematic characterization of structures and reactions for use in organic synthesis" *J. Am. Chem. Soc.* **93** 6847–6854.

Hendrickson JB, 1977. "Yield analysis and convergency" *J. Am. Chem. Soc.* **99** 5439–5450.

Hendrickson JB and Toczko AG, 1977. "Unique numbering and cataloguing of molecular structures" *J. Chem. Inf. & Comp. Sci* .**23** 171–177.

Hendrickson JB, 1978. "A logical characterization of organic chemistry" *J. Chem. Ed.* **55** 216–220.

Hendrickson JB and Braun-Keller E, 1980. "Generation of reaction sequences" *J. Comp. Chem.* **1** 323–333.

Hendrickson JB, 1986. "Approaching the logic of synthesis design" *Accts. Chem. Res.* **19** 274–281.

Hendrickson JB, 1990. "Organic synthesis in the age of computers" *Angew. Chem. Intl. Ed.* **29** 1286–1295.

Hendrickson JB and Miller TM, 1990. "Reaction indexing for reaction databases" *J. Chem. Inf. & Comp. Sci.* **30** 403–408.

Hendrickson JB, 1992. "Descriptions of reactions: Their logic and applications" *Rec. Trav. Chim. Pays-Bas* **111** 323–334.

Hendrickson JB and Sander T, 1995. "COGNOS: A Beilstein-type system for organizing organic reactions" *J. Chem. Inf. & Comp. Sci.* **35** 251–260; *Eur. J. Chem.* **1** 449–460.

Hendrickson JB, 1997. "Classification of organic reactions" in von Schleyer R (Ed.), *Encyclopedia of Computational Chemistry* Wiley; "Comprehensive System for Classification and Nomenclature of Organic Reactions" *J. Chem. Inf. & Comp. Sci.* **37** 852–860.

Valdes-Perez RE, 1995. "Some recent human-computer discoveries in science and what accounts for them" *AI Magazine* 37–44.

Wipke WT, 1978. "Simulation and evaluation of chemical synthesis – SECS: An application of artificial intelligence techniques" *Artificial Intelligence* **11** 173–193.