SYNGEN Program for Synthesis Design: Basic Computing Techniques

JAMES B. HENDRICKSON* and A. GLENN TOCZKO

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

Received February 7, 1989

The central importance of optimal route selection in synthesis design is emphasized. The basis for such selection in the SYNGEN program is outlined: prior selection of convergent assembly bondsets for the skeleton, followed by development of abstracted functionality requirements for sequential construction of such convergent bondsets from actual available starting materials. The detailed computing methods currently used in SYNGEN are delineated.

In the major computer programs intended to design organic syntheses[1-6] the enormous size of the "synthesis tree" (all possible paths) is always emphasized, and the major effort has usually been devoted to the computational work of generating all the possible reactions and intermediates retrosynthetically from the target. This in effect aims at generating the whole tree. Once granting that this can be done, however, immediately points up the central importance of the *selection* of a few optimal routes, since the whole tree of choices is so huge. Indeed, the keys to such stringent selection must be central to the initial evolution of any approach to mechanized synthesis design if it is not to be compromised by excessive output. The present paper summarizes the reasoning for this selection which lies at the base of the SYNGEN program[6] and describes the details of the computing methods in current use in that program.

We require first a definition of optimal routes, one that can be stringently applied as a basic criterion for their selection. We chose the criterion of economy: to seek the shortest and most cost-effective routes. Then we require a basis for a massive simplification and subdivision of the huge search space in the synthesis tree. The first simplification was to examine only the skeleton of the target to discern the most efficient plan for its assembly from small pieces, i.e., starting material skeletons. Even this is not simple, for a huge number of dissections for any target skeleton is possible, generated by the combinatorics implicit in sequential cutting of skeletal bonds. We chose to generate only convergent plans of assembly since they were shown to be more cost efficient than linear plans; the weight of starting materials required can be 1.5–5 times more in a linear than in a convergent plan of the same number of steps.[7] The convergent plan is easy to generate by cutting the target skeleton into two pieces and each piece into two again, etc. In SYNGEN this is currently limited to no more than two bonds per cut and to only two levels of cutting, i.e., the target skeleton into two intermediates and these into two starting skeletons each. The plans are further limited to those with all four starting skeletons found in a catalog of available starting compounds. Finally, the first level cut, on the target, is limited to cuts with the smaller piece containing at least one-fourth of the target carbons, for combinatoric reasons, although this limit may be altered by the operator. In our experience with cuts into more disparate pieces the larger piece can usually not be made from real starting skeletons in one more cut.

This limiting procedure generates bondsets of up to six target skeleton bonds cut and prunes the tree considerably: the total number of assembly plans possible for the estrone skeleton is 41 million, but the convergent plans limited as above, with our catalog of about 6000 compounds, is only 1432 plans. Thus, the skeletal focus vastly simplifies the whole tree, and the convergent criterion then stringently selects only an optimal few. Even these 1432 assembly plans are easier to comprehend when broken down as only six bondsets at the first level, the rest being cuts at the second level that feed those six. Furthermore, each bondset represents a separate, independent subtree of the whole synthesis tree, and this subdividing of the search space now allows the separate development of the possible chemistry for each bondset, one after the other.

The second phase of SYNGEN is concerned with this development of the chemistry: of the functionality on the skeletons and their reaction changes. This requires a further simplification of the tree. The compounds and reactions in the search space are described by an abstract system of functionability representation, which is digital for rapid computer manipulaton, and also is generalized to drastically reduce the search space by coalescing trivial distinctions of detail. By use of this system, all possible reactions can be generated for a compound from the digital representation of its functionality. Stringent selection is applied here from the initial criterion of shortest route: the only routes sought are those that sequentially construct the skeletal bonds of any bond set, with no intervening refunctionalization steps, from real starting materials to the target. These must be the shortest routes. Finally the generated reactions must be tested on mechanistic grounds to delete chemically nonviable ones before a final optimal set of all the shortest, most efficient synthetic routes from the synthesis tree is presented.

DESCRIPTIVE SYSTEM FOR COMPOUNDS AND REACTIONS

The system begins by describing four synthetically important kinds of attachments to any carbon atom: H for attachment of hydrogen, or other electropositive atoms; R for σ-bond to another carbon; Π for π-bond to another carbon; and Z for a bond (σ or π) to an electronegative heteroatom (N, O, S, X). The numbers of each kind of attachment on any carbon are then h, σ, π, z, respectively, and they add up to four. This system, summarized in Figure 1, allows generalized descriptions of both structures and reactions that cover all possible variations. For any structure the functional groups on each carbon of the skeleton may be abstracted with two digits, z (=0–4) and π (=0–2). With the skeletal carbons numbered, the structure is defined by a $z\pi$-list, the $z\pi$ values of the carbons in numerical order. The oxidation state (x) of each carbon is also simply defined as $x = z - h$, which allows oxidation state changes in reactions to be easily calculated.

Reactions can all be described in terms of *unit reactions*, which are unit exchanges of one kind of attachment for another on each involved carbon. The unit reactions at any one carbon

0095-2338/89/1629-0137$01.50/0

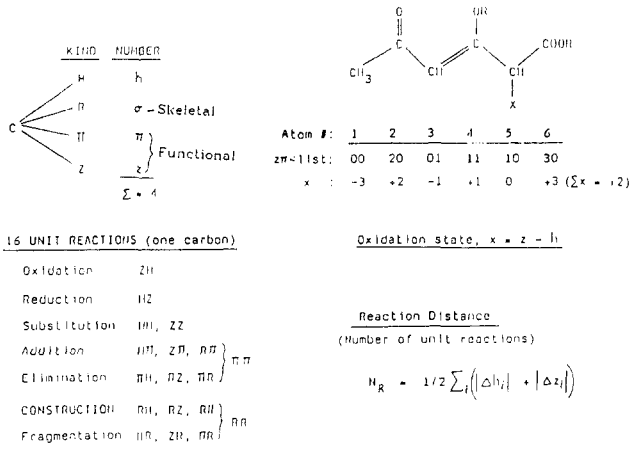


Figure 1. Digital characterization of structures and reactions.

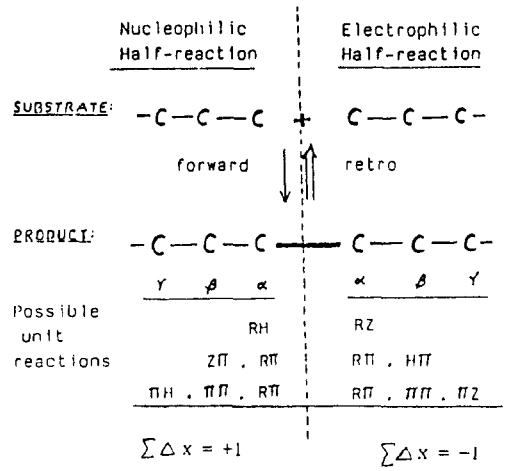


Figure 2. Generalized form of construction reactions.

are designated by two letters, the first for the attachment bond made, the second for that broken. Thus, there are 16 (=4 × 4) two-letter descriptors for all possible unit reactions at any one carbon. Reactions involving R or Π at a carbon must have an adjacent carbon also with R or Π. Thus, construction reactions are shown by RH, RZ, or RΠ at each of the two carbons being joined in the construction. It is these construction reactions that are sought by SYNGEN across each sequential bondset bond to build the skeleton of the target with its functional groups in place. These construction reactions exhibit functionality changes on a strand of carbons out from the bond constructed on both sides (Figure 2). Each side is separately characterized as a half-reaction—one nucleophilic and one electrophilic. The functionality changes in each half-reaction occur on a reactive strand of up to three carbons, labeled α, β, and γ, out from the bond constructed. All possible unit half-reactions on three carbons are listed below in Figure 2. SYNGEN then can generate all constructions of any bondset-designated skeletal bond. The program will recognize the functional groups on each strand out from the bond in the product as a $z\pi$-list and then generate the substrate by applying each unit reaction set in turn, altering the $z\pi$ values of each involved carbon as directed.

Half-Reactions. In the development of SYNGEN we originally used only the simple unit construction half-reactions defined from RH, RZ, and RΠ changes at the α-carbon in the six ways shown in Figure 2. This revealed two shortcomings in practice. In the first place, a number of good, practical, one-step constructions did not turn up. These proved to be *composite constructions*, composed of a construction unit reaction coupled (before or after) with a refunctionalization unit reaction in situ. For these cases the list of viable half-reactions had to be expanded to include these composites. In the second place, many constructions were generated that were chemically unreasonable; these are discussed below.

The composite constructions are of three kinds: prior reduction to form a nucleophilic carbanion to be used in construction; subsequent elimination to form a double bond across the bond constructed (α–α') or the α–β bond; bond tautomerization either before or after construction, either allylic or keto-enol. The overall net structural changes for these as well as for the six unit half-reactions of Figure 2 were expressed as $z\pi$-lists for substrate and product on the $\alpha\beta\gamma$ strand of reacting carbons, and therefore the change in $z\pi$-list on these atoms is characteristic of each half-reaction.

The basic half-reactions were further subdivided in terms of the substrate functionality level ($z + \pi$) at the α-carbon. The resulting list of 25 half-reactions used in SYNGEN is shown in Figure 3. Here the $\alpha\beta\gamma$ strand for the minimal requisite functionality is shown for each as partial structures of substrate and product, as well as the corresponding $z\pi$ values at each carbon. There are 16 nucleophilic and 9 electrophilic half-reactions that can combine to create 114 full constructions. Although 16 × 9 = 144 reactions, three (E1, F1, 2E) can only combine with each other, forming α–α' double bonds, in two full constructions (E1·2E and F1·2E), so there are 2 + (14 × 8) = 114 full constructions.

Each half-reaction is labeled with a two-character notation. The first character usually represents the minimum required functionality level ($z + \pi$) of the substrate α-carbon, a letter identifying nucleophiles or a number identifying electrophiles. The exceptions are the reductively formed carbanion nucleophiles (R1, R2, R3, RT), the elimination composites E1 and F1, and the π-nucleophiles (and aromatics), P1. The second character usually represents the half-reaction *span*, i.e., the number of carbons (1–3) that change attachments in the $\alpha\beta\gamma$ strand. The exceptions are RT, an allylic tautomerization, and 2E and 2F, the electrophiles (aldehyde/ketone of $z = 2$) that eliminate after construction. The span number specifically denotes only those atoms that change attachments; it may be noted that the A1, C1, and P1 half-reactions require the presence of a β-atom, though only the α-atom changes (i.e., span = 1). Furthermore, the P3 reaction is intended to imply involvement of the γ-H in an allylic carbanion with a β'-withdrawing group so that it tautomerizes into conjugation. The result is the same overall change in $z\pi$-list as P1, so that the two are identical except for tests of mechanism in the next section.

These half-reactions should be generally recognizable as familiar chemical families from the partial structures shown in Figure 3. The brief formal descriptions appended to each may often be extended to common names. R1 is the standard Grignard or organometallic reagent and R3 its allylic variant, RT if the double bond tautomerizes into conjugation to an electron-withdrawing group. Such a group can also direct metal reduction of a double bond at α–β to form an α-carbanion (R2). E1·2E represents dehydrating aldol reactions to form α–α' double bonds, and F1·2E is the Wittig reaction. Where P1 is electrophilic substitution on alkene or aromatic, B2 is the addition counterpart on alkenes, and C2 represents electrophilic addition on an acetylene followed by tautomerization to the keto form. The B3 half-reaction is half of the ene reaction or of nucleophiles created from allylic silanes or stannanes where H = Si or Sn, etc. Among electrophiles, 11 is an alkylation, 21 carbonyl addition, 31 is acylation, and 41 is carboxylation, while 12 is conjugate addition. The program also provides on demand a help screen with these definitions displayed, but we have found that the two-character descriptors generally become quite familiar very quickly. In any case, for reasons of space, the display of reactions from SYNGEN is

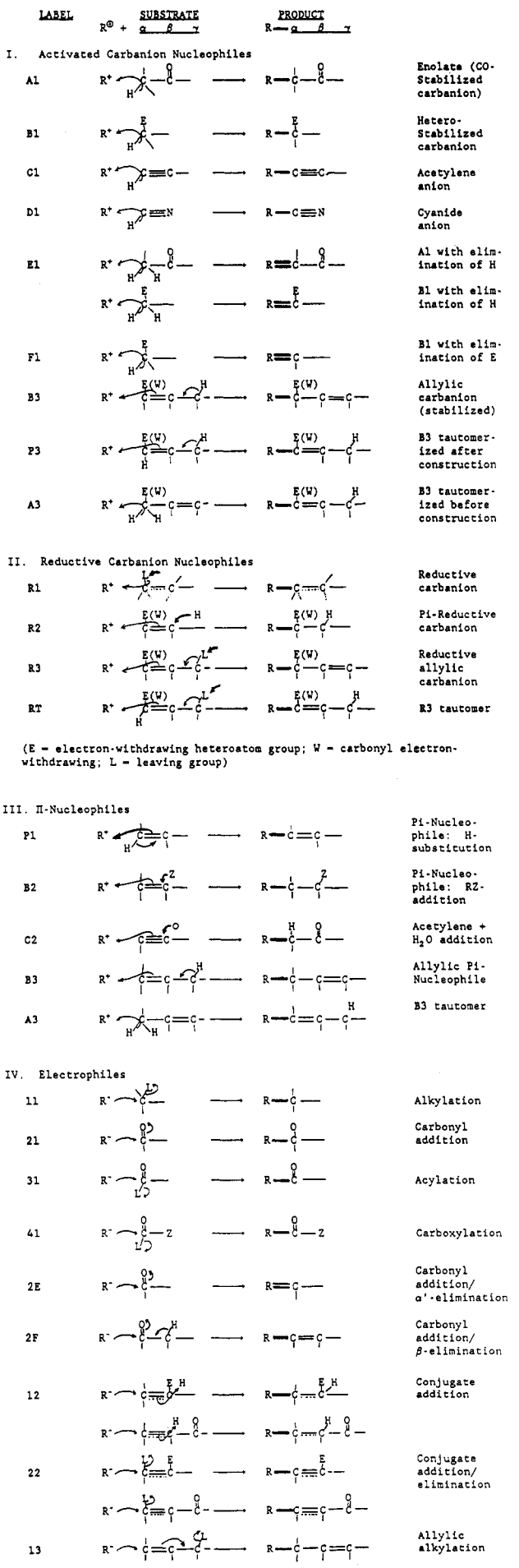


Figure 3. Construction of half-reactions used in SYNGEN.

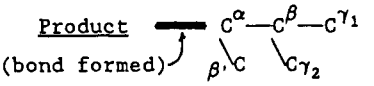
annotated with this shorthand.

Mechanistic Tests. As noted in the previous section, simple application of all half-reactions to generate constructions and substrates from a product and a designated bond very often produced chemically unrealistic reactions. The reasons for this were twofold. First, many reactions require activating functional groups that do not alter in the construction. Second, many desired constructions may take another, unwanted, course owing to the presence of nearby functional groups that deflect it by elimination from carbanions, incorrect π-bond regiochemistry, etc. To reduce the output to more manageable size by eliminating at least grossly incorrect constructions, we elected to incorporate these rough mechanism tests. The problem here of course is that too much mechanistic restriction will curtail the possibility of turning up new and interesting chemistry, while having no restrictions leads to excessive output of unrealistic chemistry.

We have sought to balance these extremes in a flexible way by incorporating mechanism tests for each half-reaction in the form of modules that may easily be separately replaced in the program as experience clarifies their value. Furthermore, while these tests disallow the generation of some unacceptable half-reactions, others of borderline acceptability are retained but are flagged with a series of notations marking their problems. The operator may then identify such problem reactions when examining output to further eliminate such reactions selectively. These are discussed below.

The mechanism tests are of two kinds: "require" and "reject". The former tests for the presence of required activation and regioselectivity on the atoms near the construction bond; the latter searches for the presence of incompatible functional groups and interfering side reactions. To do this we recognized that merging all electronegative heteroatom attachments as z (Figure 1) was too severe a generalization for these mechanistic tests. Hence, we defined a subset of z to indicate not simply the *kind* of attached atom, as N, O, S, or X, but rather its *mechanistic function*. These *z-functions* are labeled: E for electron-withdrawing; O for electron-donating; and L for leaving group, to define groups attached by one bond, i.e., for $z = 1$. For $z = 2$ or 3, separate singly bonded attachments can also be E, O, or L, but multiply bonded attachments are described as W for electron-withdrawing (C=O or C=N). When a skeletal atom is a heteroatom (N, O, S), it is treated as a special carbon, responding just like carbon to all the same half-reaction changes in $z\pi$ on construction but utilizing mechanistic requirements and restrictions that are different from carbon.

Accordingly, a string of computer bits is assembled to describe the skeletal heteroatom nature and z-function for each of the atoms of the reactive strand in the product structure; the four atoms described are α, β, β', and any γ attached to β (β' implies an atom attached to α but off the reactive strand):



The checklist string for each atom is then S, N, O, L, E, O, W, one bit (1 or 0) to indicate presence or absence of each feature on that atom. Thus, the 4-atom checklist is a 32-bit word. This product checklist is then compared against the same form of checklist for each possible half-reaction for which the product is a candidate. Note that the product must already contain the minimum functionality of Figure 3 to be a candidate for any particular half-reaction. The checklist for the whole product strand (four atoms) is compared, all at once in one AND operation, with each of the two half-reaction checklists (require and reject). A product containing no required features for a given half-reaction is disqualified and that half-reaction is not generated or recorded. If the product

Table I. Mechanism Test Checklists[a]

	α							β							γ							β′						
no π	S	N	O	L	E	O	W	S	N	O	L	E	O	W	S	N	O	L	E	O	W	S	N	O	L	E	O	W
A1	R	R	R	–	–	–	X	R	–	–	X	–	–	R	–	–	–	–	–	–	–	R	–	–	X	–	–	R
B1	X	X	X	–	R	–	X	–	–	–	X	–	–	–	–	–	–	–	–	–	–	–	–	–	X	–	–	–
D1	R	X	X	X	X	X	R	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
R1	X	X	X	–	–	–	X	–	X	–	X	–	–	–	–	–	–	–	–	–	–	–	X	–	X	–	–	–
11	–	X	X	X	X	X	X	–	X	–	–	–	–	–	–	–	–	–	–	–	–	–	X	–	–	–	–	–
21	X	X	X	R	R	R	X	X	X	X	–	–	–	–	–	–	–	–	–	–	–	X	X	X	–	–	–	–
31	–	X	X	X	X	X	R	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
41	–	X	X	X	X	X	R	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
B2	X	X	X	X	–	X	X	X	X	X	R	X	R	X	X	X	X	–	–	–	X	X	X	X	–	–	–	–
B2H	X	R	–	X	–	X	X	X	X	X	R	X	R	R	R	R	R	–	–	–	X	X	X	X	–	–	–	–
C2	X	–	X	X	–	X	X	X	X	X	X	X	X	R	–	–	–	–	–	–	X	X	X	X	–	–	–	–
R2	X	R	R	–	R	–	X	X	X	X	X	X	–	X	–	–	–	X	–	–	X	–	X	X	X	–	–	R
12	X	X	X	X	X	X	X	X	X	X	–	R	–	X	–	X	X	–	–	–	R	X	–	X	–	–	–	X
12H	X	X	X	X	X	X	X	X	R	R	X	X	X	X	–	X	X	–	–	–	–	X	–	X	–	–	–	X
12C	X	X	X	X	X	X	X	X	X	X	X	–	X	X	–	X	X	–	–	–	–	X	–	X	–	–	–	–
π(αα′)																												
E1	–	R	X	–	R	–	X	–	–	–	X	–	–	R	–	–	–	–	–	–	–	–	–	–	X	–	–	R
F1	X	X	X	–	–	–	X	–	–	–	X	–	–	–	–	–	–	–	–	–	–	–	–	–	X	–	–	–
2E	X	X	X	X	X	X	X	X	X	X	–	–	–	–	–	–	–	–	–	–	–	X	X	X	–	–	–	–
π(αβ)																												
P1	X	X	X	X	–	X	X	X	X	X	X	X	X	X	X	X	X	–	–	–	X	X	X	X	–	–	–	–
P1H	X	R	X	X	X	X	X	X	X	X	–	–	–	X	–	–	–	–	–	–	–	X	X	X	X	–	–	–
R1	X	X	X	–	–	–	X	X	X	X	X	X	–	X	–	X	X	X	–	–	X	–	X	–	X	–	–	–
R2	X	R	X	X	R	X	X	X	X	X	X	X	–	X	X	–	–	X	–	–	X	–	X	–	X	–	–	R
B2	X	–	X	X	–	X	X	X	X	X	R	X	R	X	–	–	–	–	–	–	X	X	X	X	–	–	–	–
12	X	X	X	X	X	X	X	X	X	X	X	R	X	X	–	X	X	X	X	X	R	–	–	–	–	–	–	X
12H	X	X	X	X	X	X	X	X	R	X	X	X	X	X	–	X	X	X	X	X	–	–	–	–	–	–	–	–
12C	X	X	X	X	X	X	X	X	–	X	X	–	X	X	–	X	X	–	–	–	–	–	–	–	–	–	–	X
22	X	X	X	–	–	–	X	X	R	R	–	R	X	X	–	X	X	X	–	–	R	–	X	–	–	–	–	X
2F	X	X	X	X	X	X	X	X	–	X	–	–	X	X	–	–	–	–	–	–	–	X	X	X	–	X	–	–
P3	X	X	X	X	R	X	X	X	X	X	X	X	–	X	R	R	R	X	X	–	X	X	X	X	X	X	–	R
A3	X	X	X	X	R	X	X	X	X	X	X	X	–	X	X	R	X	X	X	–	X	–	X	X	X	X	–	R
A3C	X	X	X	X	–	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	–	X	X	–	–	–	–
RT	X	X	X	X	R	X	X	X	X	X	X	X	–	X	X	X	X	X	X	–	X	–	X	X	X	X	–	R
π(βγ)																												
B3	X	X	X	X	X	X	X	X	X	X	–	X	–	X	X	X	X	–	–	–	X	X	X	X	–	–	–	X
B3H	X	R	R	–	R	–	X	X	X	X	–	X	–	X	X	R	R	–	X	–	X	–	X	X	X	–	–	R
R3	X	R	R	–	R	–	X	X	X	X	–	X	–	X	X	X	X	–	X	–	X	–	X	X	X	–	–	R
A1H	X	X	X	X	X	X	X	X	X	X	–	–	–	X	X	R	R	X	X	X	X	X	–	–	X	–	–	X
13	X	X	X	–	X	–	X	X	–	X	–	–	–	X	X	X	X	–	–	–	X	–	X	–	–	–	–	–

[a] No restrictions or requirements on triple bonds at αβ = half-reactions C1, R1, 22. The designations H and C refer to skeletal heteroatom present and cyclization, respectively.

qualifies for required activation, it is then compared against the reject checklist, and if it contains any reject features, it is also disallowed. An example is shown in Figure 4 and discussed further below.

The two kinds of checklists for each half-reaction are combined in the tabulation of Table I, which divides up the half-reactions according to the presence of π-bonds left in the product (compare with Figure 3). Any feature marked R is a requirement for the half-reaction; any feature marked X causes rejection of that half-reaction for the candidate structure. Features marked with a dash are of no consequence to either check. The data in Table I can be read with a little practice, comparing the entry in Table I with the corresponding part-structure in Figure 3. Thus, the A1 (enolate carbanion) half-reaction requires a β-carbonyl, or heteroatoms (S, N, O) as the α-atom; a sulfur at β is also accepted. However, a leaving group on any β-atom, or a carbonyl at α, will reject its generation. The β-leaving group in fact rejects all carbanion nucleophiles (A1, B1, E1, F1, P3, A3, R1, R2, R3, RT). The conjugate addition, 12, leaving no π-bond in the product (in group "no π", Table I), requires an electron-withdrawing heteroatom (cf. $-NO_2$, $-SO_2R$, etc.) on the β-carbon or a γ-carbonyl but rejects them on α or β′. The 12H entry is the same reaction with a skeletal heteroatom at β instead, but without functionality attached (cf. no nitrone substrates). Different entries are occasionally provided for different demands on cyclization: the 12 half-reaction on cyclization is allowed without electron-withdrawing groups, if Markovnikov regiocontrol is provided (below). The π-bond additions are repeated in the "no π" and the π(αβ) groups since addition to double bonds affords the former products, addition to triple bonds the latter.

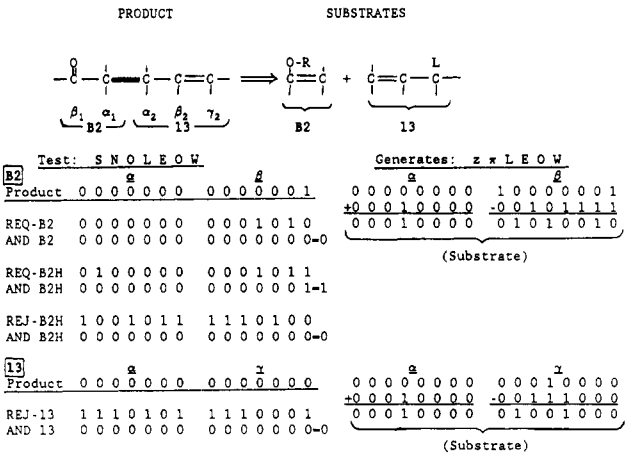


Figure 4. Sample test and substrate generation for a reaction.

Most half-reactions have requirements intrinsic to their (mechanistic) descriptions in Figure 3, and these are checked

in the generating program apart from the requirements labeled R in Table I. These requirements include the presence of π-bonds as used in Table I to categorize the half-reactions, as well as the required z_β for B2, C2, or H at β for 12 and R2 or at γ for A3, RT, P3, etc. Some other requirements are also tested separately from the R entries in Table I. The 11 (alkylation) half-reaction requires either $h_\alpha > 0$ or else cyclization. Products with triple bonds generated for substrates [B2, R2, or 12 in $\pi(\alpha\beta)$, or C2] will not generate if that triple bond would be in a ring. Allylic half-reactions with spans of three (A3, RT, P3, B3H, R3) demand δ-atoms to be unfunctionalized ($z\pi = 0$), to validate regiochemistry. The most important such separate check is the regiochemical test of Markovnikov's rule, which allows unactivated π-nucleophile additions (P1, B2, C2) only if $\sigma_\alpha \geq \sigma_\beta + 1$. Here the regiochemically uncertain ones ($\sigma_\alpha = \sigma_\beta + 1$) are allowed but flagged for operator examination later. Electrophilic additions (12) that are unactivated but cyclizing are similarly flagged. A similar regiochemical test (*o*, *m*, *p*) of aromatic substitution has not yet been incorporated, and so all are allowed.

Logic Flow in SYNGEN. In terms of computing detail, the ordered bondsets are first generated from the skeleton, the target bonds are numbered and tabulated as an array identifying each with its two input-numbered atoms, and the smallest set of smallest rings,[8] identified from the connectivity, is tabulated with included bonds. The bonds are sequentially cut, and then a determination is made of whether they are ring bonds; if so, all other larger numbered bonds in the same ring only are also cut to deliver two fragments for the first level. At the second level these fragments are similarly cut again, defining two further fragments each. All fragments produced are then compared with the skeletons of the starting materials in the catalog. For every second-level dissection yielding found skeletons in the catalog, the bondsets are recorded for sequential application of the reaction generators in the second phase of the program.

The second phase examines these dissected bonds sequentially in a nest of DO loops, first of cut level, then of the target or subtarget compound, then of bondsets in the level, and then of the individual bonds in each bondset. Where a ring is cut, cutting two bonds, each of the two orders of construction is examined independently. When a given construction bond is set to examine in the innermost loop, the several $\alpha\beta\gamma$ strands out from that bond on each side are identified, and checklists are assembled for each strand in the form of a four-byte string for α, β, γ, β', corresponding to Table I: a 1 is set in each corresponding bit if the atom is S, N, or O and/or the *z*-function is L, E, O, or W.

The strands on each end of the bond are examined separately, in another DO loop, for generation and testing of half-reactions. The course of the program here is diagrammed in Figure 5. The strands are sorted for π-bonds at $\alpha\beta$ and at $\beta\gamma$ to divide the flow into the relevant half-reactions. When these have been tested and new substrate $z\pi$-lists generated for the active strands at each end of the bond, the two halves are matched for polarity (one nucleophilic and one electrophilic half-reaction) and the successful new $z\pi$-lists recorded. After cutting into two fragments—one acyclic bond or two ring bonds—these new $z\pi$-lists now record the functionality on each generated substrate. The program now compares these $z\pi$-lists with those in the starting material catalog for the same skeleton. All generated starting materials must be found by the second level for their reactions to be recorded.

Catalog Comparisons. A standard format for recording compounds is employed both in the starting material catalog and for the retrosynthetically generated substrates from the construction reactions. This takes the form of a linear string composed of *skeleton–heteroatoms* (*SNO*)*–zπ-list–z-function*

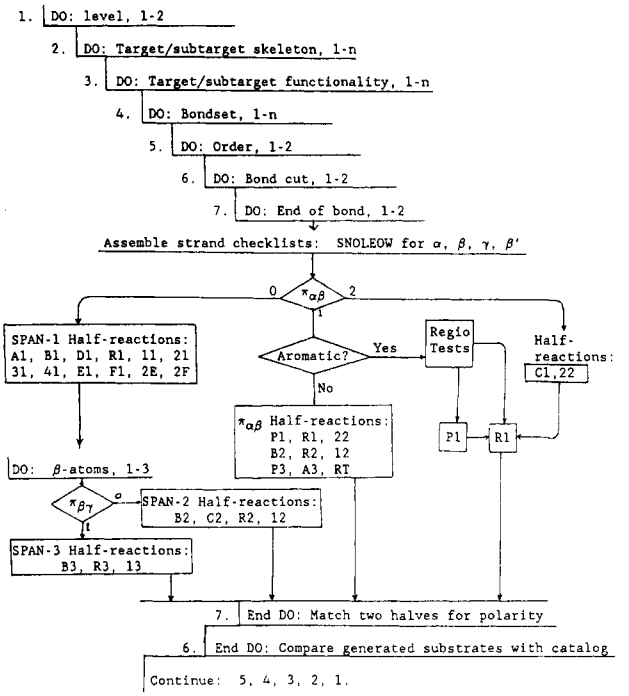


Figure 5. Logic flow for generating constructions and substrates.

(*LEOW*) *list* in that order. Thus, comparisons can be made of skeleton only from the front end of the list, of skeleton plus $z\pi$-list, or of the whole string with functionality more defined by *z*-function. The atoms in any skeleton are first canonically numbered by maximizing the adjacency matrix to a maximal bit-string of connectivity.[9] The catalog compounds are then listed in numerical order, first of the skeleton connectivity string, then of the remaining lists (heteroatom and functionality) in the string. With the catalog entries in numerical order, comparison with generated compounds is a very fast operation.[9]

A number of catalog compounds may have the same identity string when they vary only in trivial ways that are not recorded, and only one is kept in the comparison catalog; the duplicates may be called up on demand at output time. Thus, the carboxyl carbon is entered as $z\pi = 30$ (*z*-function = W) whether it is an acid, acid chloride, nitrile, or any ester. Nitriles and esters are entered both as carbon skeleton only ($z\pi = 30$) and as nitrogen- or oxygen-containing skeletons, for which the carbon atom becomes $z\pi = 02$ and 20, respectively. When esters are entered as carbon skeleton only, the alcohol moiety is separately cataloged. The catalog presently contains 6086 compounds of which 4782 have different identity strings; the skeletal size limits of the catalog are from 3 to 16 skeletal atoms.

When a generated substrate has the same skeleton as some in the catalog but different functionality, the program calculates how many unit reactions of refunctionalization are required to create it from each catalog entry (reaction distance, N_R, in Figure 1).[10] If it is only one or two steps away from some starting material, and if its skeleton is large enough to qualify for this repair, the generated substrate will be allowed and recorded with its reaction distance ($N_R = 1$ or 2 steps). The required minimum size to allow refunctionalization may be preset by the operator.

There is one exception to the requirement that all starting materials must be found by level 2. For large targets with a reasonably balanced size distribution of the four starting materials, some may not be found. In our experience starting materials of C_6–C_9 may not be in the catalog but can always be made by one further dissection, i.e., a third level of construction. While the combinatorics do not allow separate

Table II. Half-Reaction Generators: $\Delta z\pi$LEOW (Product → Substrate)

	α						β						γ					
half-reaction	Δz	Δπ	ΔL	ΔE	ΔO	ΔW	Δz	Δπ	ΔL	ΔE	ΔO	ΔW	Δz	Δπ	ΔL	ΔE	ΔO	ΔW
A1, B1, C1, D1, P1, P3	0	0	0	0	0	0												
E1	0	−1	0	0	0	0												
F1	+1	−1	0	+1	0	0												
2E	+2	−1	0	0	0	+1												
R1, 11, 22	+1	0	+1	0	0	0												
21	+1	0	0	0	−1	+1												
31	+1	0	0	0	0	0												
41[a]	0	+3	0	0	0	0												
R2, 12	0	+1	0	0	0	0	0	+1	0	0	0	0						
B2/z = 1	0	+1	0	0	0	0	−1	+1	−1	0	0	0						
B2/z = 2	0	+1	0	0	0	0	−2	+1	0	0	+1	−1						
C2	0	+2	0	0	0	0	−2	+2	0	0	0	−1						
2F	+2	−1	0	0	0	+1	0	+1	0	0	0	0						
A3	0	−1	0	0	0	0							0	+1	0	0	0	0
B3	0	+1	0	0	0	0							0	−1	0	0	0	0
R3, 13	0	+1	0	0	0	0							+1	−1	+1	0	0	0
RT	0	0	0	0	0	0							+1	0	+1	0	0	0

				in HEX numbers		
Δzπ LEOW as one byte	α	β	γ	α	β	γ
A1, B1, C1, D1, P1, P3	00000000			0	0	0
E1	−00010000			−10	0	0
F1	+00110100			+34	0	0
2E	+01110001			+71	0	0
R1, 11, 22	+01001000			+48	0	0
21	+00111111			+3F	0	0
31	+01000000			+40	0	0
41[a]	+00110000			+30	0	0
R2, 12	+00010000	+00010000		+10	+10	0
B2/z = 1	+00010000	−00111000		+10	−38	0
B2/z = 2	+00010000	−00101111		+10	−2F	0
C2	+00100000	−01100001		+20	−61	0
2F	+01110000	−00010000		+71	−10	0
A3	−00010000		+00010000	−10	0	+10
B3	+00010000		−00010000	+10	0	−10
R3, 13	+00010000		+00111000	+10	0	+38
RT	00000000		+01001000	0	0	+48

[a] Substrates of $z = 4$ from 41 are kept as $z = 3$, $\pi = 3$.

delineation of all such third-level routes, the program allows some routes with these indicated as level 2 *intermediates*. This implies the understanding that they can be synthesized but will require a longer overall route. These may be rejected or examined separately at the operator's choice when the output is viewed, as shown below.

Generation of Substrates. In the flow chart (Figure 5) each active strand out from each end of a bond is sequentially examined for each possible half-reaction, shown in several groups depending on π-bond presence in the product being examined. For each half-reaction the required and reject strings of Table I are first applied to ascertain if the reaction is at all viable. If so, selection flags are then set as discussed below, and then the substrate $z\pi$LEOW for each changing strand atom is computed. A working record of all atoms is maintained in the form of $z\pi$LEOW for each atom, and the changes for each half-reaction (product → substrate) for atoms α, β, and γ are tabulated in Table II and may be compared with the pictorial representation of Figure 3.

An example of the treatment of one construction is shown in Figure 4 for finding a successful B3·13 reaction, drawn as partial structures at top. The left strand with $z_\beta = 2$ is accepted for B2 tests but fails the require test as the AND operation equals 0; it is then tried with B2H (see Table I) and passes with a nonzero require test. The reject test is also acceptable (AND = 0). Only two of the four atoms (α,β) are shown. Similarly, the 13 test is accepted for the right-hand strand (no require/Table I and reject AND = 0). For substrate generation, the $z\pi$LEOW array is used for αβγ since skeletal atom type (S, N, O) will not change but z and π will. This array is one byte per atom with two bits each for z and π (=0–3). The generating bytes from Table II are added as shown, and the resultant $z\pi$LEOW values for the substrate atoms are recorded. These generated substrates are now converted to the standard compound format with maximized connectivity, for search in the catalog as discussed above. They are also similarly compared with previously generated substrates to remove duplication in the growing record of intermediates and starting materials generated.

Output and Operator Selection Modes. When all successful reactions have been generated, the successful bondsets and their intermediates, starting materials, and reactions are all tabulated and cross-indexed for output display purposes. The output is all graphical, each compound and reaction formulated onto the original, normalized drawing of the target with appropriate excision of atoms and bonds and with the generated changes in functionality. The new functional attachments are kept in the program only as values of z and LEOW. They are appended to the skeleton for graphical output of substrate compounds by attaching a letter and bond at a computed angle.

For output the operator may separately examine four categories—bondsets, starting materials, intermediates, or reactions, at either level—and make *retain* or *delete* selections of items in any category to prune excessive output or closely examine a small subset of the output. The numbers of items remaining in the subset in each category after selection are then computed and displayed. A representative page of reactions from a steroid synthesis is shown in Figure 6 with the remaining totals displayed at the top and a menu for deletion and page viewing at the right.

Because output may be large, a further flexible mode of

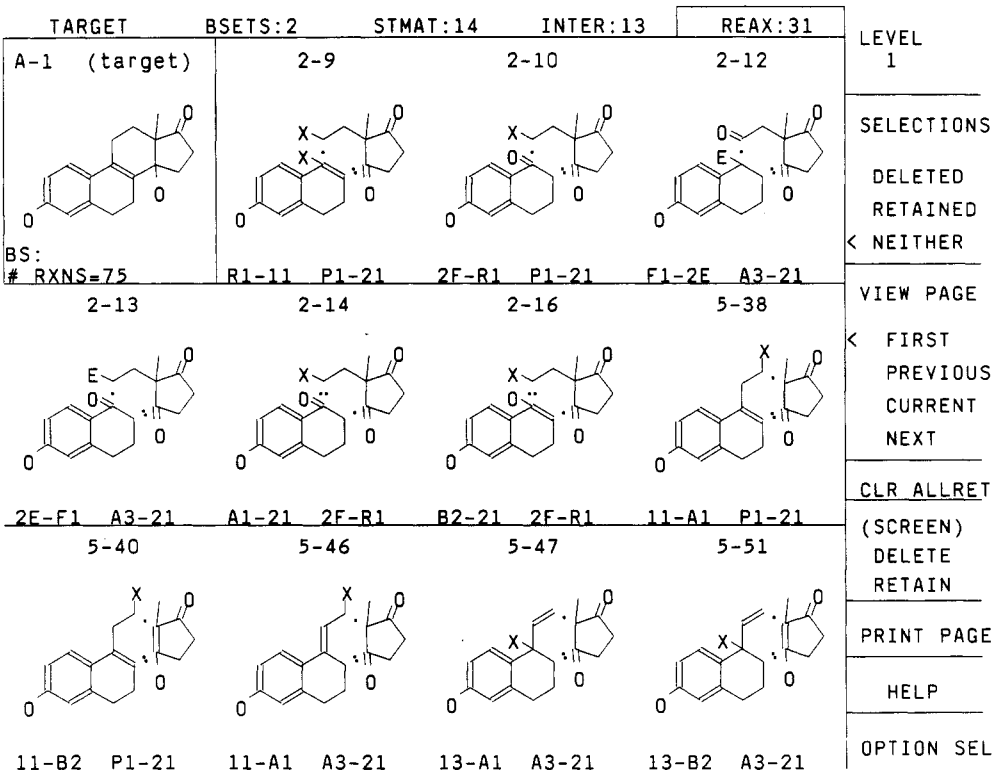


Figure 6. Sample output of reactions.

COMPOUND/SELECTIONS	SM-1-	SM-2-	RXN-1-	RXN-2-
COST LIMIT none$(1) none$(2)	9	143	37	391
SM NOT REFUNCTIONALIZED	4	45	21	59
SM REFUNCTIONALIZED	5	98	16	332
INTERMEDIATES	32	3	71	17

REACTION/SELECTIONS	RXN-1-	RXN-2-	
UNCLEAR REGIO:KETONE	6	20	
UNCLEAR REGIO:PI-BOND	2	14	MOVE CURSOR TO
MIXED MODE CONSTRUCTION	8	84	LEVEL -1-
TWO-STEP ANNEL: ACID/BASE	39	206	LEVEL -2-
TWO-STEP ANNEL: RED.CYC.	33	182	DESC. FOR 1&2
BALDWIN RULES VIOLATION	13	76	ENTER
CHEM. EQUIVALENT REACTIONS	30	55	R FOR RETAIN
CHEM. EQUIV.: REVERSE-POLARITY	5	35	D FOR DELETE
CHEM. EQUIV.: REVERSE-ORDER	0	13	C FOR CLEAR
COMPATIBILITY: LEAVING GROUP	24	261	Clear All Flags
COMPATIBILITY: KETONE	50	59	HELP
COMPATIBILITY: ALDEHYDE	0	6	EXIT

Figure 7. Indirect selections menu.

selection is also made available to apply to compounds and reactions. Here a number of choices are offered to apply across the whole output. The screen of choices is illustrated in Figure 7 with the opportunity to retain or delete one or more choices at either level 1 or 2, and the numbers then display which compounds and reactions remain after the selection. Returning to the four categories then allows the remaining items to be displayed, as in Figure 6 for reactions.

The starting material selections incorporate various features about starting materials that may recommend the sequences in which they are involved. First of these is their cost, kept in the catalog as 1986 Aldrich price in cents/mole. The catalog can be updated with current Aldrich prices tapes, but this is probably not often necessary since the *relative* price positions are likely to be maintained over long periods. Here it is possible to enter from the keyboard a maximum allowed starting material cost. The screen will then show the remaining number of reactions using only this group of starting materials. Here again, as with all individual or group selections, the deletion of some starting materials will cause implicit deletions of others as well as of some reactions, intermediates, and bondsets.

The other starting material group selections in Figure 7 include the option of deleting (or retaining) sequences from starting materials that need prior refunctionalization. One can also focus especially on routes that involve at least one starting material found at the first level, i.e., routes that are shorter and so more efficient because one compound at the first level does not need to be made. Similarly, it is possible to delete or separately examine sequences depending on a third-level construction of an intermediate (C_6–C_9) at level 2, syntheses that are longer than the norm.

The reaction group selections in Figure 7 are designed to isolate or remove reactions of dubious practicality or simply to prune out excessive output that is essentially chemically equivalent. The former are restrictions thought to be not severe enough for total rejection in the SYNGEN mechanism tests. Hence, instead they are flagged when generated for assessment here by the operator. The following groups are currently installed.

(1) *Regiochemistry* is defined for two situations. A ketone acting as an enolate carbanion may have a hydrogen on its other side (i.e., h_γ) so that the regiochemistry for A1 or E1 is unclear. This is flagged in cases in which the α-atom is not N or O, there is not a second carbonyl or π-activation at β', or the construction is not a cyclization of ring size equal to 5 or 6. The second situation is for π-nucleophiles in B2, C2, or P1 half-reactions or π-additions (12) and invokes Markovnikov regioselectivity, as noted above under Mechanistic Tests. The examples that will turn up in this group are those with equal substitution on both double (or triple) bond atoms in the substrate, i.e., *products* with $\sigma_\alpha = \sigma_\beta + 1$ and otherwise unactivated. Constructions with these features will be displayed selectively if retain is chosen or deleted from the display on delete command.

(2) *Mixed-mode* refers to constructions in which the two half-reactions are normally used under different conditions of acid or base. When a half-reaction is generated, a flag is set for those that are so characterized, as follows; the other

half-reactions are regarded as valid under either acid or base conditions and are not flagged. Those belonging to the "acid only" group are π-nucleophiles; those in the "base only" group are several electrophiles only flagged when *not* in a cyclization.

acid only: P1, B2, C2, B3, and A3, when activated

base only: 11, 12, 22, 13, when not cyclizing

The group selection for mixed-mode constructions then displays, or deletes, on command those constructions composed of a nucleophile in the first list with an electrophile in the second.

(3) *Annelations* are a pair of sequential constructions creating two bonds in one ring. The first construction is intermolecular, joining two molecules, while the second is intramolecular, cyclizing the intermediate formed in the first. The two constructions taken together are commonly displayed as one in the reactions screen, as shown in Figure 6. From the point of view of synthesis efficiency, the preference is for two constructions that may proceed in one pot, i.e., as if they were one reaction, like the Robinson annelation. If two displayed constructions cannot be executed in one step, then they are flagged as two-step annelations for either of two selections: TWO-STEP ANNEL:ACID/BASE implies that one construction (or both) is mixed-mode as in (2) above or that one is nominally under basic conditions and the other acidic; TWO-STEP ANNEL:RED.CYC. implies that the second construction is a reductive one (R1, R2, R3, RT) and so cannot be run directly after the first without an intervening isolation of the intermediate product.

(4) *Baldwin's rules* are flagged as violated in the event that a 5-membered or smaller ring is being formed by a reaction that adds to an endo double bond, i.e., between atoms which are in that ring. This affects reactions B2 and 12, adding to double bonds (at α–β), and to B3, R3, and 13, which shift endo α–β double bonds (see Figure 3). The flag is also set for cyclizations forming a 3- or 4-membered ring by reaction to an exo triple bond (α–β exo to the formed ring), affecting reactions C2 and B2 or 12 on triple bonds.

(5) *Equivalent reactions* are those that present only trivial distinctions in chemical sense and so may be ignored in an effort to prune excessive output. The computer applies the half-reaction generators quite mechanically and will often produce several variants for one construction which to the chemist appear conceptually equivalent, such as alkylation to form the same product either with an α,β-unsaturated carbonyl electrophile (12) or with its β-halo-carbonyl equivalent (11). Deletion of these equivalent reactions can more effectively orient the operator to the synthetic "ideas" produced without the burden of excessive minor variations. Thus, the general entry (CHEM.EQUIVALENT REACTIONS) flags as equivalent variants those cases for which the same product is formed via the same bondset bond from a different starting material using a different half-reaction of the same polarity. The primary half-reaction in each case and its corresponding equivalent half-reaction are tabulated in Table III, showing the generated substrate in each case. Thus, the example mentioned above is the alkylation equivalent, for which the 12 half-reaction is taken as primary and its equivalent 11 generation is flagged as equivalent and may be deleted as desired.

Two other kinds of equivalence are also flagged and examined by the next two entries in Figure 7. The entry for reverse polarity refers to cases in which a bond is made from the same substrates by simply reversing the polarity, i.e., reversing which side is nucleophile and which is electrophile. The common case is R1·11 vs 11·R1 since the net generation of R1 and 11 substrates is the same. The equivalents of this type are set out in Table IV.

Table III. Chemically Equivalent Half-Reactions

PRODUCT		SUBSTRATE	PRIMARY REACTION	EQUIVALENT REACTION
CARBANIONS				
R—C—CH	A1 ⟹	CH—CH	A1	R1, R2
	R1 ⟹	CZ—CH	R1	R2
	R2 ⟹	C≡C		
R—C—C≡C	B3 ⟹	C≡C—CH	B3	R3
	R3 ⟹	C≡C—CZ		
R≡C—	E1 ⟹	CH_2—	E1	F1
	F1 ⟹	CHZ—		
ALLYLIC				
R—C—C≡C	A1 ⟹	CH—C≡C	A1	A3
	A3 ⟹	C≡C—CH		
	11 ⟹	CZ—C≡C	11	13
	13 ⟹	C≡C—CZ		
	R1 ⟹	CZ—C≡C	R1	R3
	R3 ⟹	C≡C—CZ		
ENOL TYPE				
R—C—CO	A1 ⟹	CH—CO	A1	B2, C2
	B2 ⟹	C≡CZ	B2	C2
	C2 ⟹	C═CZ		
ALKYLATION				
R—C—CH	12 ⟹	C≡C	12	11
	11 ⟹	CZ—CH		

Table IV. Reverse Polarity Constructions

product	primary	equivalent
$\pi_\alpha = 0$	R1·11	11·R1
	R1·13	11·R3
	R3·11	13·R1
	R3·13	13·R3
$\pi_\alpha = 1$	R1·11	22·R1
	R1·13	11·R3
	R1·22	11·R1
	R3·22	13·R1
$\pi_\alpha = 2$	R1·22	22·R1

The last equivalence is that of reverse order and refers to the order of making the two bonds in an annelation. Here there will be two reaction entries with the same bond set, substrates, and reactions, differing only in the order in which the two bonds are made. One is then relegated to equivalent status and so may be ignored with the delete command.

(6) *Compatibility* refers to the presence of certain functionality that may interfere with the generated construction, but is not found on the reactive $\alpha\beta\gamma$ strand (or at β') and so is not tested in the mechanism tests. The first compatibility check is that of leaving groups found off-strand in either substrate during construction involving basic carbanion conditions. The half-reactions flagged in such a case are A1, B1, C1, E1, F1, R1, R2, R3, and RT.

The other compatibility check is one in which an aldehyde or ketone is found off-strand in either substrate and so may preferentially react instead of the desired carbonyl electrophile. These are flagged as "ketone" when half-reactions are 21, 31, 41, 2E, and 2F. If the off-strand functionality is actually an aldehyde, the aldehyde flag is also set on the grounds that aldehydes are more reactive and so more seriously incompatible. The operator can therefore isolate here those cases in which ketones or aldehydes need protecting groups if the reaction is to be used in practice.

An example of a selected screen of choices is shown in Figure 8, in which starting material costs have been limited to \$50 and \$60, respectively, for the two cut levels, and refunctionalized starting materials have been entirely deleted, as have second-level intermediates. Chemical equivalent reactions of two kinds have also been deleted. The result is a dramatic reduction in reactions at the two levels, from 71 and 391, respectively, to 4 and 15.

Summary. SYNGEN is programmed in about 50 000 lines of FORTRAN for a VAX computer (3000 or 8000 series) and requires 1–2 megabytes of active memory. The input is a

COMPOUND/SELECTIONS	SM-1-	SM-2-	RXN-1-	RXN-2-
COST LIMIT 50$(1) 60$(2)	3	13	2	15
SM NOT REFUNCTIONALIZED	3	13	2	15
SM REFUNCTIONALIZED	0	0	0	0
INTERMEDIATES	4	0	4	0

REACTION/SELECTIONS	RXN-1-	RXN-2-	
UNCLEAR REGIO:KETONE	0	0	
UNCLEAR REGIO:PI-BOND	0	0	MOVE CURSOR TO
MIXED MODE CONSTRUCTION	0	0	LEVEL -1-
TWO-STEP ANNEL: ACID/BASE	4	3	LEVEL -2-
TWO-STEP ANNEL: RED.CYC.	0	5	DESC. FOR 1&2
BALDWIN RULES VIOLATION	0	0	ENTER
CHEM. EQUIVALENT REACTIONS	2	3	R FOR RETAIN
CHEM. EQUIV.: REVERSE-POLARITY	0	0	D FOR DELETE
CHEM. EQUIV.: REVERSE-ORDER	0	0	C FOR CLEAR
COMPATIBILITY: LEAVING GROUP	0	9	Clear All Flags
COMPATIBILITY: KETONE	4	12	HELP
COMPATIBILITY: ALDEHYDE	0	0	EXIT

Figure 8. Indirect selections: sample choices.

rapid, crude drawing of the target structure on a Tektronix mode graphics terminal and proceeds without operator input, as described above, to generate all possible sequential constructions of convergent bond sets of two levels only from real starting materials in its catalog. The program requires about 2 min for a steroid of the complexity in Figure 6. A number of sequential targets may be input with each going to batch mode for its output generation. These are stored when finished in a directory for operator examination later using the delete/retain modes for further selection from large outputs as described above. The reactions produced are generally sensible and constitute all possible combinations within the constraints applied. The results seen to date seem to mirror sensibly the expectations of those constraints. Further expansion of the program is currently under development.

ACKNOWLEDGMENT

We are grateful for support provided by the Eastman Kodak Co. and by a grant (CHE-86-20066) from the National Science Foundation.

REFERENCES

(1) Corey, E. J.; Long, A. K.; Rubinstein, S. D. *Science* **1985**, *228*, 408, and earlier references cited therein.
(2) Wipke, W. T.; Braun, H.; Smith, G.; Choplin, F.; Sieber, W. *ACS Symp. Ser.* **1977**, *61*.
(3) Bersohn, M. *Bull. Chem. Soc. Jpn.* **1972**, *45*, 1897. Bersohn, M.; Esack, M.; Luchini, J. *Comput. Chem.* **1978**, *2*, 105. Bersohn, M.; MacKay, K. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 137.
(4) Gelernter, H.; Sridharan, N. S.; Hart, H. J.; Yen, S. C.; Fowler, F. W.; Shue, H. J. *Top. Curr. Chem.* **1973**, *41*, 113. Gelernter, H.; Sanders, A. F.; Larsen, D. L.; Agarwal, K. K.; Bovie, R. H.; Spritzer, G. A.; Searlemen, J. E. *Science* **1977**, *197*, 1041. Agarwal, K. K.; Larsen, D. L.; Gelernter, H. *J. Comput. Chem.* **1978**, *2*, 75.
(5) Marsili, M.; Gasteiger, J.; Carter, R. E. *Chim. Oggi* **1984**, *9*, 11. Gasteiger, J. In *Computer in der Chemie*; Zeigler, E., Ed.; Springer: Berlin, 1984; p 207.
(6) Hendrickson, J. B. *Acc. Chem. Res.* **1986**, *19*, 274.
(7) Hendrickson, J. B. *J. Am. Chem. Soc.* **1977**, *99*, 5439.
(8) Hendrickson, J. B.; Grier, D. L.; Toczko, A. G. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 195.
(9) Hendrickson, J. B.; Toczko, A. G. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 171.
(10) Hendrickson, J. B.; Braun-Keller, E. *J. Comput. Chem.* **1980**, *1*, 323.

Multiple Constructions in Synthesis Design

JAMES B. HENDRICKSON* and PING HUANG

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

Received February 7, 1989

Multiple construction reactions that may take place in one laboratory operation are argued to be important keys to rapid creation of the target skeleton in synthesis design. The several paths for such constructions, forming two to four skeletal bonds, are logically formulated. Two of the major classes, double affixation and multiple cyclization, are then articulated as computer programs to find all skeletal variants for a given target and then to apply appropriate reactive functionality to the generated starting skeletons to produce viable routes to the target.

In the development of short, efficient synthesis routes to complex target molecules we have argued that the construction of the target skeleton should be the key consideration.[1] This arises from the basic observation that syntheses proceed from small starting molecules (average of only three carbons incorporated into the target) to form large target molecules. Hence, we concluded that only those reactions ("constructions") that link the small starting skeletons into the target skeleton are obligatory for synthesis. Therefore, the shortest syntheses will consist of only construction reactions, or at least will sharply minimize the refunctionalization reactions (those which alter functionality without changing skeletal bonds). To this end we developed the SYNGEN program to generate syntheses consisting of sequential construction reactions only, presumably the shortest routes.[1]

The ideal of the shortest synthesis then becomes one of minimizing refunctionalization reactions, but should also consider minimizing the number of construction steps as well. An average published synthesis constructs about one-third to one-fourth of the target skeletal bonds, i.e., for targets of 10–30 carbons, making about 3–9 skeletal bonds. In the SYNGEN program minimization of the number of constructions arises by dissecting the target skeleton into convergent assemblies of starting skeletons. These are found by cutting the target skeleton into two parts first and then, at the second level, cutting each of those two into two smaller skeletons, finally accepting only those ordered bondsets so derived that result in all four starting skeletons found in a catalog of available starting materials.[2] This convergent dissection procedure allows no more than two bonds cut to divide a skeleton into two pieces and so will result in no more than six bonds (and usually less) cut from the target, i.e., those bonds to construct in the synthesis.

Another way to minimize the number of construction steps is to employ reactions that make more than one construction per step, i.e., multiple construction steps, and this approach

0095-2338/89/1629-0145$01.50/0