Systematic Synthesis Design. III. The Scope of the Problem[1]

James B. Hendrickson

Contribution from the Edison-Lecks Laboratory, Brandeis University, Waltham, Massachusetts 02154. *Received January 22, 1974*

Abstract: The skeletons of organic molecules are viewed as graphs, and a simple graphical procedure for defining and enumerating all contained rings (monocycles) and r-cycles (bicycles, tricycles, etc.) is offered for molecules of $r \leqslant 7$. Also the synthesis tree, for skeletal construction reactions only, is analyzed in a new grid format, applicable to any target molecule which allows enumeration of starting material skeletons and routes to target. The grid is composed of sets of intermediates defined by the number of components and number of rings.

There is a tradition in organic chemistry that there is an infinite number of ways to synthesize any given compound. It is also clear that this is possible by infinite variation of transient groups en route or by aimless refunctionalization. However, a serious approach to the real scope of meaningful variation in synthetic routes has not yet been examined. While it is unlikely that such variations are infinite, it is important to understand the breadth of the possible in synthesis design within the limits of sensible definition so that protocols of systematic synthesis generation can be assessed. It is probably fair to say that there has been to date very little discussion of the allowable restrictions on synthesis formulation or any sharp conceptual definition of the "synthesis tree" of paths and intermediates.[2]

As a step toward clarification of the synthesis tree, we shall focus only on the skeleton building aspect of a synthesis, following the kind of analysis previously applied to the simple model problem of linking substituents to an aromatic ring[1b] and enumerating sets of pathways sharply defined. Furthermore, as a first step, only *direct routes* will be examined, i.e., those which contain only construction steps and no skeletal cleavages.[3] In this way, we may hope to clarify the scope of the synthesis tree by asking such questions as: how many ways may a skeleton be split into 2, 3, 4 ... parts; what sizes of these parts or components (synthons) result, and how many cuts or bond dissections of the skeleton are required;[5] how many synthons of 2, 3, 4 ... carbons exist in the skeleton as potential starting materials of that size; how many ways exist to build the skeleton from given subskeletons; how many annelations or rearrangements[3] are possible, etc.? Can we categorize and count all starting materials and pathways as in ref 1b?

Skeleton building is a problem in graph theory. Mathematically molecular structures as we draw them are simply graphs composed of points (atoms) linked with lines (bonds). The synthesis tree is also a graph, and application of graph theorems allows us to gain insights into synthesis which might not otherwise be apparent, as well as to enumerate synthons and pathways by the use of related combinatorics. Such enumerations will likely be more valuable for defining the various ways in which synthons and pathways are generated and the general size or scope of the problem than for the absolute numbers of possibilities that result. First we shall look at molecular skeletons as graphs and then at systematic ways to disconnect them to build synthesis trees of construction pathways.

The Skeleton as Graph. The carbon skeleton of the target structure[6] may be examined as a connected, labeled graph of points and lines,[7] the carbon atoms (n_0), and their σ bonds (b_0), respectively. The number of (fundamental) rings is $r_0 = b_0 - n_0 + 1$. This graph of the skeleton may be reexpressed mathematically as its adjacency matrix, A (with elements, a_{ij}), in which each row (and column) represents a point (atom), numbered as on the skeleton;[8] thus A is a ($n_0 \times n_0$) square symmetrical matrix. The elements, a_{ij}, are 1 or 0 if the atoms i and j are or are not bonded to each other, respectively, and the diagonal elements, $a_{ii} = 0$. This matrix contains all the information about the skeleton (cf. $b_0 = \frac{1}{2}\Sigma a_{ij}$) and is ideally amenable for computer storage as a Boolean array since it contains only binary information.

In principle all rings can be defined by an algorithm which "walks through" this matrix identifying successive adjacencies until the starting point is reached. Such an algorithm starts at the beginning of each row, passes horizontally to each "1", then drops (or rises) vertically to the diagonal (a_{ii}) entry, then passes horizontally again in either direction to each "1", and repeats, recording each successive atom, i, which is reached until a return to the starting atom is obtained. (The algorithm produces redundancies which must be reduced; i.e., it produces each ring 2ρ times, where ρ = ring size.) This is essentially a simple "tree search" common to many computer operations in data management, and a number of algorithms for defining rings and sets of rings in polycyclic structures have been delineated and computerized.[9]

In order to define a structure fully, each skeletal carbon (i) is separately defined by the number of its bonds to hydrogens (h_i), heteroatoms (z_i), and carbons (σ bonds, σ_i; π bonds, π_i);[1a] this leads to a definition of the *character* of each carbon as $c_i = 10\sigma_i + z_i + \pi_i$. An expanded total structure matrix, S (elements, s_{ij}), can now be made from the adjacency (skeleton) matrix, A, by substituting z_i for $a_{ii} = 0$ in the diagonal and using 2 and 3 instead of 1 for double and triple bonds (between carbons), respectively. The structure matrix, S, defines not only the skeleton (since A may be derived from it) but also the functionality of the structure:

$$\sigma_i = \Sigma_j a_{ij}$$

$$z_i = s_{ii}$$

$$\pi_i = \Sigma_{j \neq i} s_{ij} - \sigma_i$$

$$h_i = 4 - (\sigma_i + z_i + \pi_i) = 4 - \Sigma_j s_{ij}$$

$$c_i = 10\sigma_i + z_i + \pi_i = 9\Sigma_j a_{ij} + \Sigma_j s_{ij}$$

The adjacency and structure matrices for a tricyclic example are shown in Figure 1.

Further matrices can also be defined, such as the incidence matrix ($n_0 \times b_0$) showing the incidence of atoms (n_0) with bonds (b_0), again with 0 and 1 elements, and matrices

Adjacency Matrix (A)

Atoms	1	2	3	4	5	6	7	8	9	10	$\Sigma = \sigma_i$
1	0	1	0	0	0	1	1	0	0	0	3
2	1	0	1	0	0	0	0	0	0	0	2
3	0	1	0	1	1	0	0	0	0	0	3
4	0	0	1	0	1	0	1	0	0	1	4
5	0	0	1	1	0	1	0	0	1	0	4
6	1	0	0	0	1	0	0	1	0	0	3
7	1	0	0	1	0	0	0	0	0	0	2
8	0	0	0	0	0	1	0	0	0	0	1
9	0	0	0	0	1	0	0	0	0	0	1
10	0	0	0	1	0	0	0	0	0	0	1

Structure Matrix (S)

Atoms	1	2	3	4	5	6	7	8	9	10	$\Sigma = (4\text{-}h_i)$
1	1	1	0	0	0	1	1	0	0	0	4
2	1	0	1	0	0	0	0	0	0	0	2
3	0	1	0	1	1	0	0	0	0	0	3
4	0	0	1	0	1	0	1	0	0	1	4
5	0	0	1	1	0	1	0	0	1	0	4
6	1	0	0	0	1	0	0	2	0	0	4
7	1	0	0	1	0	0	2	0	0	0	4
8	0	0	0	0	0	2	0	0	0	0	2
9	0	0	0	0	1	0	0	0	3	0	4
10	0	0	0	1	0	0	0	0	0	0	1

Ring Matrix (R)

Faces	0	1	2	3	$\Sigma = \rho_i$
0	(2)	2	2	1	5(7)
1	2	0	2	1	5
2	2	2	(1)	1	5(6)
3	1	1	1	0	3
					Σ = 18(21)

$$b_o = \frac{21 + (3)}{2} = 12$$

Elements = β_{ij} Ring size = ρ_i

$\alpha = \beta_{ii}$ in parentheses

not included in $\Sigma = \rho_i$

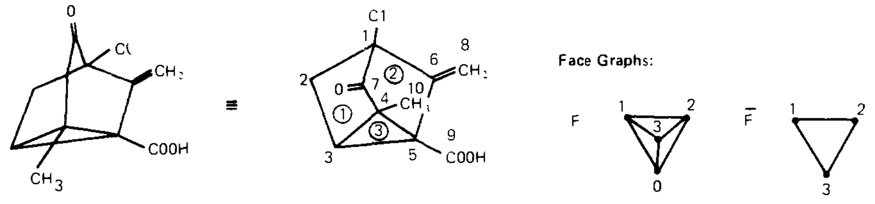


Figure 1. Matrix expressions of a sample structure.

of atoms or bonds with rings, $(n_0 \times r_0)$ or $(b_0 \times r_0)$. One such matrix of interest to us will be the ring matrix, R, in which the rows (and columns) represent the fundamental rings (r_0 in number) of the skeleton. A molecular skeleton can be drawn in many (isomorphic) ways which preserve adjacency. We shall always use a plane graph representation, i.e., one in which no bond lines are crossed. Parenthetically it may be noted that, because of the geometry of carbon and its limitation to $\sigma \leq 4$, nearly all organic structures may be represented as plane graphs.[10] The fundamental rings in a plane graph are the smallest ones, numbering $r_0 = b_0 - n_0 + 1$ (the three rings in Figure 1 are the labeled ones, five-, five-, and three-membered, and do not include four other rings, 123456, 123547, 165347, and 12356, the first three of which are six-membered). The fundamental rings are called *faces* of the graph, defined as bounded by cycles of bonds (lines); the exterior face (face "0") is the outer or unbounded region[11] (and is not counted in r_0).

If we define ring bonds (β) as those lines that are part of a cycle of lines, then all ring bonds in the skeleton are common to two (and only two) faces and may be characterized by those two faces. The other bonds are acyclic bonds (α), parts of (uncyclized) chains. The ring matrix, R, therefore exhibits elements, β_{ij}, which are the number of bonds common to rings i and j. All bonds common to any two given fused rings (including the exterior face) are said to be of the same type. The exterior face is introduced as a zero row (and column) so that β_{oi} (or β_{io}) represents the number of outer bonds of ring i. Hence R is a symmetrical matrix of dimension $(r_0 + 1) \times (r_0 + 1)$, the extra row (and column) for the exterior face. The regular entries, β_{ij}, indicate the number of ring bonds of each type, common to the two fused rings, i and j. The diagonal entries β_{ii} can be used to include all acyclic bonds (α).[12] The row sums in R are then the sizes of the several rings ($\rho_i = \Sigma_{j \neq i}\beta_{ij}$), the zero-row sum being the total of outer bonds on all rings or the monocycle represented by the periphery of the molecule if all rings are fused. The total number of bonds, $b_0 = \alpha + \beta = \Sigma_{i=0}\beta_{ii} + \frac{1}{2}\Sigma_{i \neq j}\beta_{ij}$.

We may now derive from a molecular skeleton (plane graph) a new graph, the *face graph* (F), exhibiting points which correspond to the $(r_0 + 1)$ faces, or rings, of the skeleton (numbered for identification) and lines which correspond to the presence of one or more β bonds common to the faces. The number of lines (e) in the face graph equals the number of types of ring bonds, β_{ij}, and hence half the number of nonzero (and nondiagonal) entries in the matrix R. Alternatively, if all the nonzero β_{ij} entries are changed to "1" (and all β_{ii} to 0), R becomes the adjacency matrix for the face graph. When a ring bond in the skeleton is disconnected, in analyzing for synthesis design, the resultant skeleton will be expressed by a new face graph in which the line corresponding to the type of bond (β_{ij}) cut disappears and its two incident points are coalesced into one, representing a new (enlarged) skeletal ring (or the exterior face extended by the face of the cut ring); each disconnection of bonds of dissimilar type results in the removal of the corresponding line in F. The face graph is valuable both in determining the number of ways of disconnecting (or, in reverse, constructing) the skeleton and in ascertaining the total number of rings in the skeleton. It may be noted here that the face graph is always a connected, plane graph like the skeleton from which it derives since it incorporates the exterior face.

A computer program for determining the total number of monocycles contained in a complex skeleton has recently been offered.[9d] The face graph allows the same determination to be made by hand very simply and quickly for skeletons containing up to seven fundamental rings, a range covering virtually all molecules in normal consideration. To this end consider the face graph with the exterior face (point "0") and its incident lines removed. This may be called the incomplete face graph, labeled $\bar{F}$, containing $\bar{e}$ lines and r_0 points corresponding to faces of the skeleton ($r_0 \leq 7$). Examples are shown in Figure 2. Monocycles larger than simple faces will now be recognized as fusions of two or more adjacent faces (i.e., by disconnection of their common skeletal bonds, β_{ij}). Let the number of monocycles be C_n where n = number of simple faces fused into a larger monocycle ($n \leq r_0$). The total number of rings will then be ΣC_n. Thus, C_1 = number of simple faces, i.e., the traditional description of the molecule as bicyclic ($C_1 = 2$), tricyclic

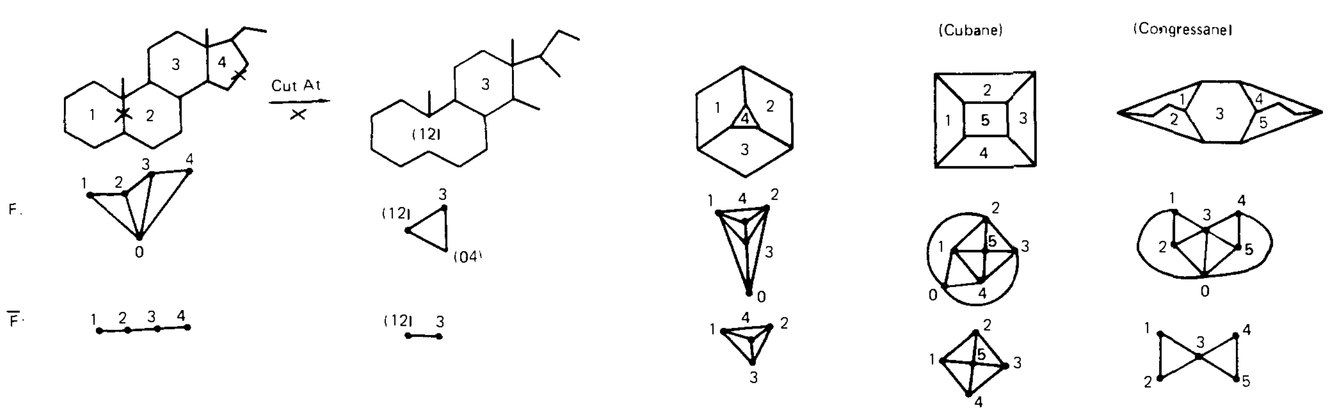


Figure 2. Molecular skeletons and their face graphs.

Table I. Enumeration Formulas for Monocycles[16]

$C_1 = r_0$

$C_2 = \bar{e}$

$C_3 = \ell - 2\Delta = \Sigma_i \binom{d_i}{2} - 2\Delta$

$C_4 = \Sigma_{ij} \binom{d'_{ij}}{2} - 2\Delta' + S$

C_{r_0} = 1 if all skeletal rings are fused[13]

C_{r_0}-1 = number of points removable from $\bar{F}$ to leave a connected graph of (r_0-1) points

C_{r_0}-2 = same for pairs of points to leave a connected graph of (r_0-2) points

Total monocycles for linear $\bar{F}$ = r_0 (r_0+1) /2

(linear $\bar{F}$: 1 2 3 4 5 r_0)
●—●—●—●—●- - - - -●

r_0 = points in $\bar{F}$ (labeled i, j, k, . . .)

$\bar{e}$ = lines in $\bar{F}$ = points in L (labeled ij, ik, jk, . . .)

ℓ = lines in L

d_i = degree of point i in $\bar{F}$

Δ = number of triangles in $\bar{F}$

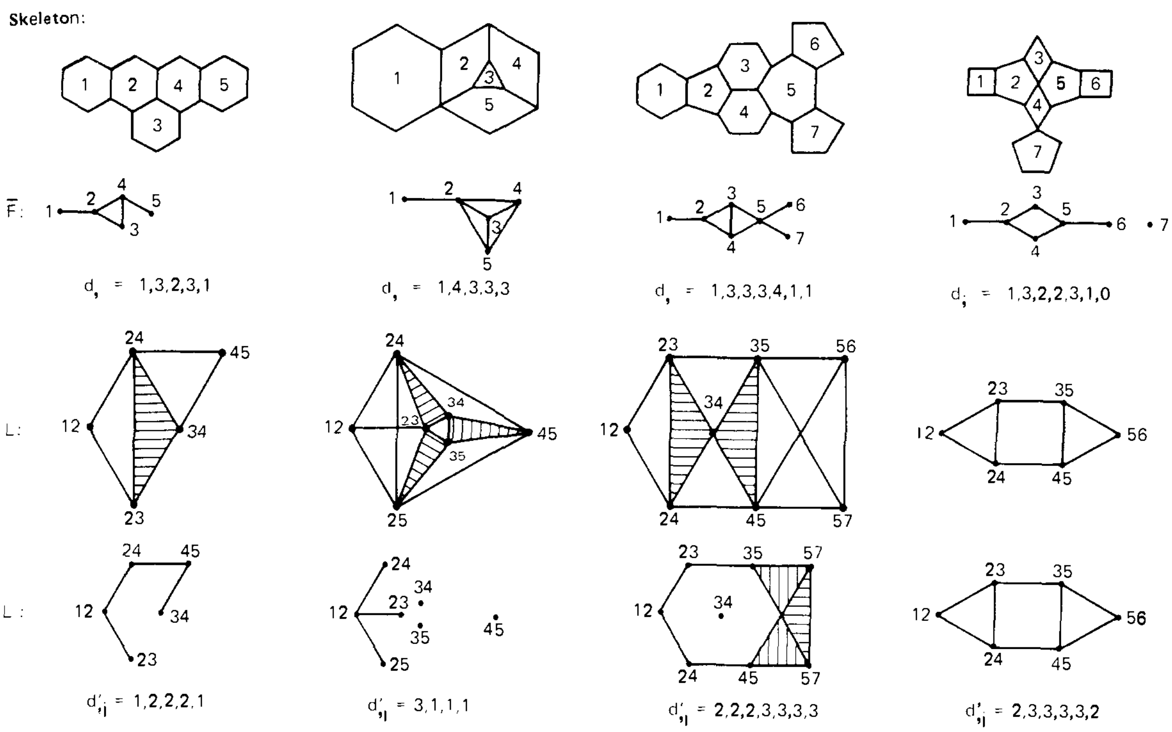
d'_{ij} = degree of point ij in L'

Δ' = number of triangles in L'

S = ☒-3□where ☒ and □ are the numbers of crossed and open squares in $\bar{F}$[14]

(C_1 = 3), etc. ($C_1 = r_0$, the number of points in $\bar{F}$). The number of monocycles made by fusing two faces is then $C_2 = \bar{e}$, the number of lines in $\bar{F}$, since each line in $\bar{F}$ joins two points and so indicates the fusion of two skeletal rings which can be opened into a single larger monocycle. In general, C_n will always be the number of unique combinations of $(n - 1)$ adjacent or linked lines in $\bar{F}$, but simple calculation of $\binom{\bar{e}}{n-1}$ is inadequate since it takes no account of the condition of adjacency in the lines counted, and it will also introduce redundancies to the extent that $\bar{F}$ itself contains cycles. With the simplest cycle, a triangle, in $\bar{F}$ (see example in Figure 1), it is clear that the three points it contains represent three skeletal rings that may be fused into a monocycle (bounded by atoms 12356 in Figure 1), but that combinations of lines, $\bar{e}$, would count $C_3 = 3$ instead of $C_3 = 1$ (there are three two-line combinations in a triangle: 123, 231, 321). The formula for C_3 then derives from counting adjacent pairs of lines and subtracting the redundancy of triangles. In order to count adjacent pairs of lines, we note that such pairs have a central point in common. The *degree* (d_i) of each point in $\bar{F}$ is simply the number of lines contiguous or adjacent to it so that the number of adjacent line pairs is the sum of combinations of each point degree taken twice, or Σ_i ($^{d}_{2}\, i$). Formulas for C_n are listed in Table I.

The potential redundancy problem increases rapidly with



C_1 =	5	5	7	7
C_2 =	5	7	8	6
C_3[16] =	7-2=5	15-8-1=6	15-4=11	8-0=8
C_4 =	3	3+1=4	15-4+1=12	14-4-3=7
C_5 =	1	1	10	4
C_6 =	0	0	5	1
C_7 =	0	0	1	0
Total monocylces =	19	23	54	33

Figure 3. Illustrations of face graphs and monocycle enumeration.

increase in the size and potential number of cycles in F̄ and so limits simple formulas for C_n to C_4, although the principle holds for more polycyclic molecules. However, the counting may be approached from the other end, e.g., $C_{r_0} = 1$.[13] Determination of the (relatively few) large monocycles available by fusing $(r_0 - 1)$ and $(r_0 - 2)$ faces may quickly be made graphically, in the former case by counting the number of single points in F̄ which may be removed and still leave a fully connected graph of $(r_0 - 1)$ points, in the latter case by the number of pair of points (adjacent or not) which can be similarly removed, leaving a connected graph of $(r_0 - 2)$ points.

For the C_4 enumeration (i.e., of monocycles created from four fused faces), we need to count all the four-point connected subgraphs in F̄, of which there are six possible kinds.[14] This is most easily achieved from a reduced line graph of F̄, following this protocol. The line graph L is constructed such that each point in L corresponds to a line (ij) in F̄ and each line in L connecting these points is an indication that these two lines in F̄ are contiguous (meeting at a common point).[15] Following this, those triangles in L which correspond to triangles in F̄ are deleted (lines of the triangle only) to leave a reduced line graph L′, for which the degrees of the points (d'_{ij}) are then listed. The same formula as for C_3, involving combinations of these degrees and subtraction of redundant triangles in L′, is now shown for C_4 in Table I but adds a second redundancy or correction factor based on the number of squares in F̄ (not in L);[14] the factor (S) adds the number of crossed squares[14] and deletes 3 × the number of plain (open) squares.

It is the rapidly increasing number of kinds of n-point connected subgraphs to be counted[14] which renders calculation of C_n above C_4 laborious; while there are, for example, six kinds of connected four-point graphs,[14] there are 20 kinds of connected five-point graphs.[7] The overall procedure for enumerating all the various kinds of monocycles in up to heptacyclic structures is, however, very easy (especially for six or fewer fused rings, not counting C_4 via the reduced line graph). This is summarized in Table I[16] and illustrated in Figure 3. It may be added that, for simple linear face graphs, F̄, such as in the steroid skeleton, the total number of monocycles is given simply by $r_0(r_0 + 1)/2$ (for steroids, $\Sigma C_n = 10$).

Apart from monocycles, it is also important to define and enumerate all other r-cycles (bicycles, tricycles, etc.) contained within the given structure. These will be all combinations of all independent monocycles with each other. For our purposes, the r monocycles which compose an r-cycle (within the r_0-cycle total skeleton) need not all be fused; e.g., ring A and ring C of the steroids constitute one (unfused) bicycle contained in the skeleton and ring A with the perimeter of rings C/D another.

The categories of possible *kinds* of combinations must first be clearly defined since an r-cycle is made from r monocycles which in turn can be severally derived by fusing various numbers of original skeletal faces. The kinds of mono-

Table II. Categories and Enumeration of r-Cycles

Categories[a] for r_0 = 1	2	3	4	5	6	7
Monocycles (r = 1) 1	2	3	4	5	6	7
Bicycles (r = 2)	11	12	13	14	15	16
			22	23	24	25
					33	34
Tricycles (r = 3)		111	112	113	114	115
				122	123	124
					222	223
						133
Tetracycles (r = 4)			1111	1112	1113	1114
					1122	1123
						1222
Pentacycles (r = 5)				11111	11112	11113
						11122
Hexacycles (r = 6)					111111	111112
Heptacycles (r = 7)						1111111
Number of categories: [a]1	3	6	11	18	29	44

Enumerations

$$C_{111\ldots} = \binom{r_0}{r}$$

$$C_{111\ldots n} = C_n \binom{r_0-n}{r-1} \qquad C_{22} = 1/2[\bar{e}(\bar{e}+1)-\Sigma_i d_i^2]$$

$$C_{111\ldots mn} = C_{mn} \binom{r_0-(m+n)}{r-2}$$

C_z = number of r-cycles in category z for a structure with r_0 fundamental rings; z contains r digits

[a] = The categories for any skeleton of r_0 rings include the numbers (z-list) in its r_0 column plus those in all columns to the left.

cycles were defined with a single digit, n, indicating the number of fundamental rings (faces) fused to create them.. Hence in a parallel vein, the r-cycles can be defined by a list, z, of r digits, each of which indicates a monocycle and shows how many faces were fused to create it. The sum of the r digits in the z list must be $\leq r_0$. These z lists are therefore category designations, and the total number of actual r-cycles in any category for a given skeleton is C_z as before, where z is the designating list. Hence C_{13} is the number of bicycles composed of one fundamental ring and one monocycle made up of three fused fundamental rings, as in a steroid with rings A, B, and C opened into one 14-membered ring (C_{13} = 2 for steroid skeletons). A table of possible categories for structures up to heptacyclic is offered in Table II along with combination formulas for enumerating C_z for most of the categories; the numbers of examples (C_z) in larger categories are usually small and easily determined by inspection of $\bar{F}$. In any case, the formulas shown are adequate for enumerating all r-cycles in structures up to pentacyclic. For example, the number of bicycles C_{13} (of one face and one monocycle made of three fused faces) possible in a heptacyclic (r_0 = 7) skeleton is given by $C_{13} = C_3\binom{7-3}{2-1} = 4C_3$, and the number of tetracycles C_{1122} in that skeleton (r_0 = 7) is $C_{1122} = C_{22}\binom{7-(2+2)}{4-2} = C_{22}\binom{3}{2} = 3C_{22}$. A formula for C_{23} is apparently attainable but complex; i.e., the number of bicycles in which one cycle is two fused faces, and the other is three (only possible in skeletons which have five or more faces).

In summary, the basis for r-cycle counting in $\bar{F}$ in the equations described above (and shown in Tables I and II), or by inspection, is one of finding (for C_n) all sets of n-linked points in $\bar{F}$ to represent monocycles of n fused rings, for C_{1n} all combinations of the sets of n-linked points with any other single points, and for C_{mn} all combinations of the sets of m linked points with the sets of n linked points. In the latter cases. the m-linked set and the n-linked set of points may not have points in common.

The individual r-cycles themselves may be designated by a bracketed list of numbers corresponding to the involved fundamental rings or faces; those that are fused into a larger monocycle are enclosed in parentheses.

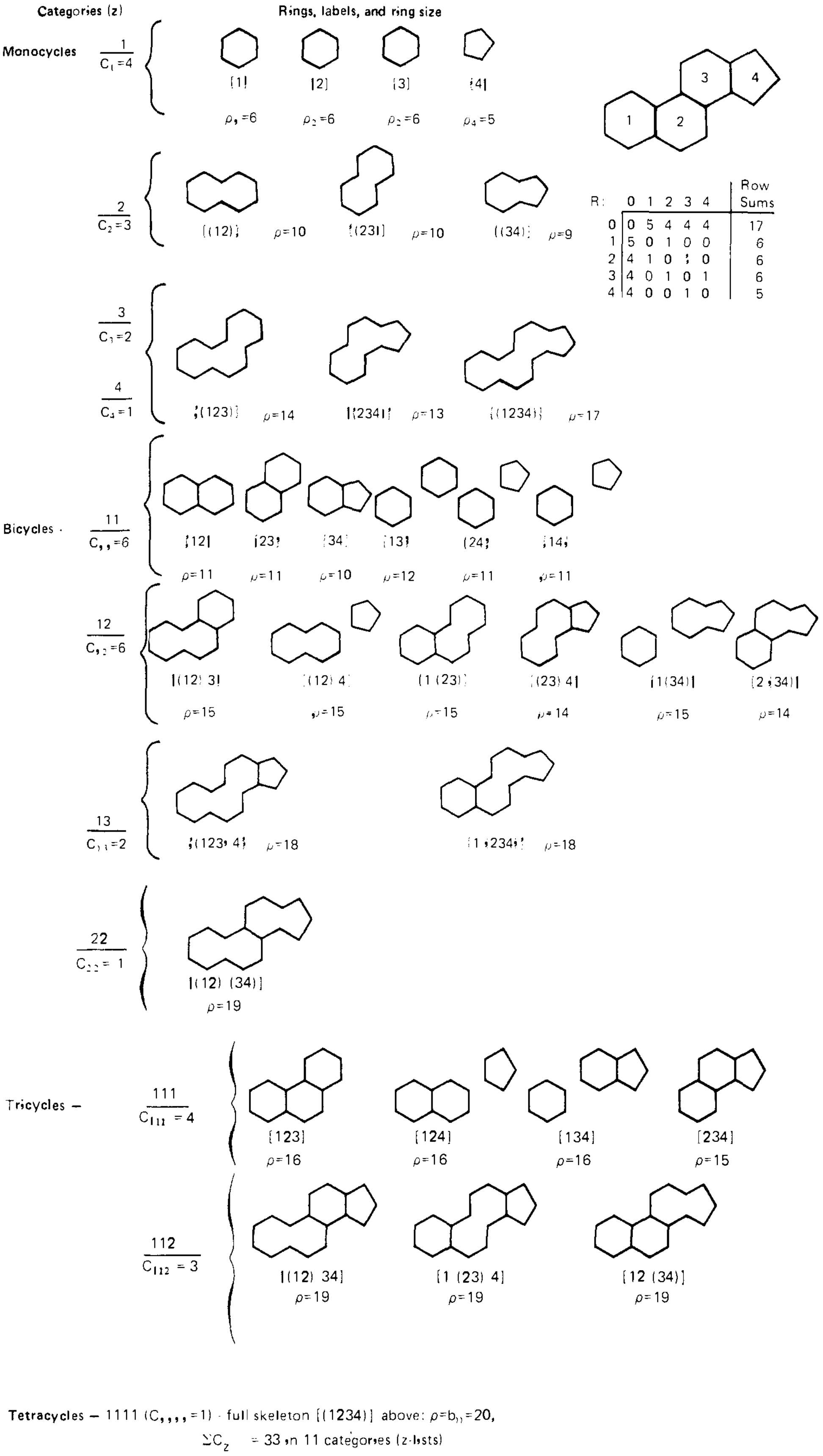
Examples of r-cycle designations are illustrated (see Figure 4):

[(12)3]: bicyclic (r = 2); category: z = 12

[12346]: pentacyclic (r = 5); category: z = 11111

[1(234)6]: tricyclic (r = 3); category: z = 113

[(346)]: monocycle (r = 1); category: z = 3



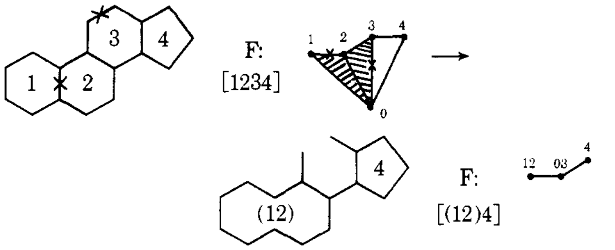
R:	0	1	2	3	4	Row Sums
0	0	5	4	4	4	17
1	5	0	1	0	0	6
2	4	1	0	1	0	6
3	4	0	1	0	1	6
4	4	0	0	1	0	5

Figure 4. The r-cycles contained in the steroid skeleton.

The 33 r-cycles (in 11 categories) in the steroid skeleton are assembled and labeled in Figure 4.

The ring size, ρ, is defined as the number of bonds in the r-cycle, also illustrated in Figure 4. These are obtained by inspection or from the ring matrix, R, in which the row sums are the sizes ($\rho_i = \Sigma_j\beta_{ij}$) of the corresponding fundamental rings. The zero row sum gives the size of the external periphery, i.e., the bonds of the exterior face. The size of other r-cycles is obtained by adding the row sums of all fundamental rings involved in the r-cycle label and substracting all row-column intersections in R that are common to rings in any parentheses. The *complement* of ρ is $(\frac{1}{2}\Sigma_{i \neq j}\beta_{ij} - \rho)$, the remaining number of ring bonds in the skeleton. The complement of ρ corresponds to the cutset subgraph, F′, of F which is disconnected on cutting the parent skeleton down to the particular r-cycle described. The size of the parent ring system (r_0-cycle) is $\rho_0 = b_0 - \alpha$.

When a particular r-cycle is generated by cutting a set of lines (F′) in F, the graphical operation is one of removal of the line (corresponding to a β-type bond) and coalescence of its joined points (i,j) into a single point (skeletal ring), labeled ring (ij). Any pair of lines joining i and j with a third point k (usually "0"), i.e., forming a triangle with the disconnected line in F, is similarly coalesced into a line $(ij) - k$. This new line is characterized by a new β value (number of skeletal bonds of one β type) which is the sum of the two coalesced; i.e., $\beta_{(ij)k} = \beta_{ik} + \beta_{jk}$. When this operation is carried out on **F**, sequentially removing the lines of the cutset F′ and coalescing the remaining points and lines as described, the graph remaining is the face graph of the r-cycle generated, as shown in operation 1 for generating the bicycle [(12)4] of the steroids.



(1)

The graphical procedures outlined here are easy to use and allow definition and enumeration of all the cyclic synthon types in terms of the fundamental rings of the full skeleton which they contain. Hence a catalog of all possible contained rings—as monocycles, bicycles, tricycles, etc.—can be generated (as in Figure 4) with a view to seeing all possible combinations of rings in sets of starting materials. In general this catalog is generated by finding all the categories in Table II and enumerating the number (C_z) of each. The individual r-cycles so defined and enumerated are easily located by inspection of F and the skeleton itself. In principle mechanical generation of the individual r-cycles may proceed by making all possible combinations of c cuts from the e ringbond types (one cut from each type β_{ij}), which means $\Sigma_c\binom{e}{c}$, for $c < e$, generating operations, as in (1). The illustration 1 is one of $\binom{7}{2} = 21$ ways to cut two different ring-bond types, in this case β_{12} and β_{03}, of the $e = 7$ available for the skeleton. However, this procedure is redundant, producing some starting material ring types more than once (e.g., with steroids, cutting β_{03} and β_{04} yields the same bicycle [12] as does cutting β_{03}, β_{04}, and β_{34}). Thus, for steroids, the mechanical generation yields $\Sigma_c\binom{7}{c} = 7 + 21 + 35 + 35 + 21 + 7 = 126$ operations giving only the 33 different r-cycles in the catalog of Figure 4.

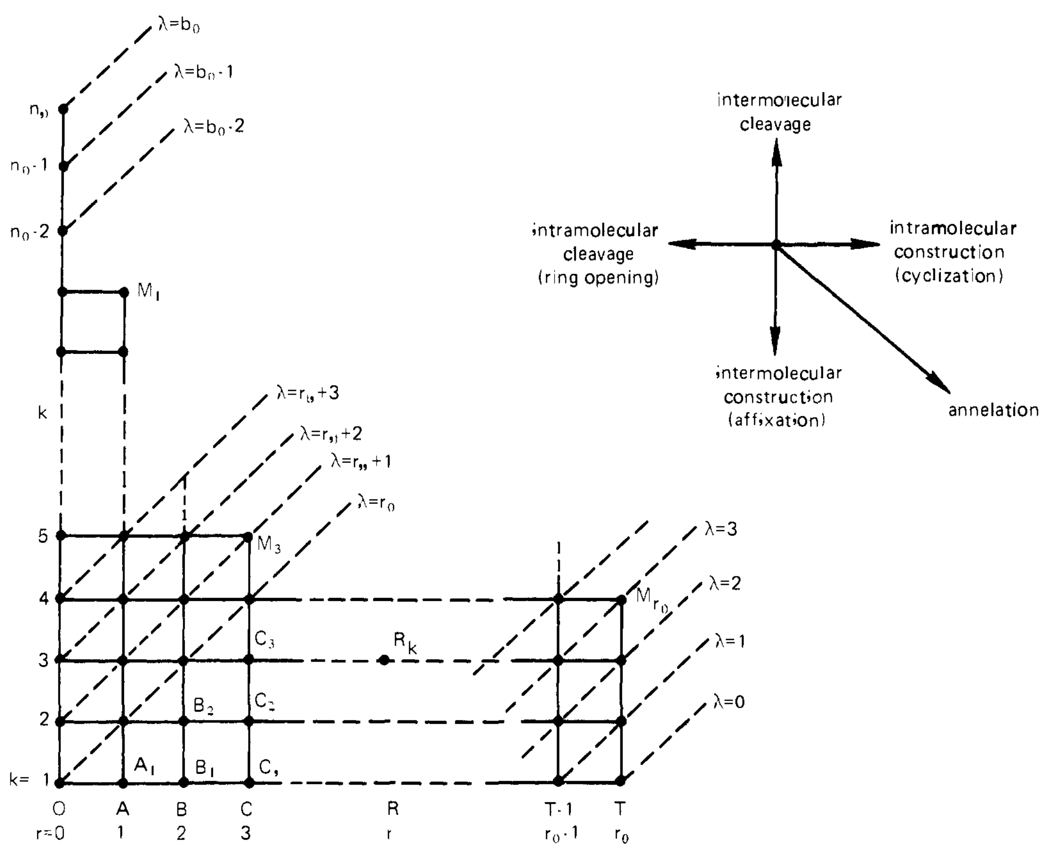
Such a catalog offers insights into the different modes of constructing the target skeleton. However, the catalog only focuses on rings, not on acyclic bonds (or their absence). Therefore, each individual r-cycle defined in the catalog will be a set of starting materials with the defined rings and acyclic atoms variously linked or not. Operation 1 shows a single bicyclic synthon generated, but the [(12)4] definition of it also fits a four-component set of synthons: a cyclodecane, a cyclopentane, and the two remaining atoms of ring 3 as one-carbon components. There are a total of eight sets of [(12)4] synthons further generated by making all possible combinations of cuts in the three acyclic bonds shown in (1). This added variety in the number of components is explored in the next section.

The Construction Tree. A synthesis tree that includes all possible routes is clearly enormous, and its size is even unclear. We may cope with it in a more manageable way by isolating meaningful subsets of the tree; one such subset would be the construction tree, dealing only with reactions creating skeletal bonds and ignoring not only functionalization reactions which do not alter skeleton but also the particular kind of construction reactions used (alkylation, acylation, carbonyl addition, etc.). We can see the levels of the tree as successive construction steps toward the target skeleton. The levels or nodes of the tree may then be grouped according to number (k) of synthon components[17] and number and kind of r-cycles in the synthons.

The target skeleton (T) of n_0 atoms, r_0 rings, and b_0 bonds is considered to be created by sequential construction of all or some of its bonds from all possible sets of precursors. The precursors are themselves sets of smaller synthon components (starting materials) totaling the n_0 atoms of the product; these precursor sets of smaller molecule skeletons are labeled *prestructs*, understood as a total of n_0 atoms, r rings, and k synthon components. A prestruct of tetramethylethylene ($n_0 = 6$) could be the three-component set of ethyl isobutyrate and two molecules of methyllithium ($n_0 = 4 + 1 + 1$); the atoms in the ethanol liberated from the ester are ignored as not appearing in the skeleton of the product. This prestruct would stand as two construction steps (two C–C bond formations) away from the target tetramethylethylene in its construction tree, the functionalization step of dehydration of the initial alcohol formed being also ignored here.

This construction tree can be conveniently expressed as a coordinate system, the *construction grid*, illustrated in Figure 5. The vertical axis of the grid records the number of components (k) in any prestruct at that level, while the horizontal axis records the number of (fundamental) rings (r) in the prestructs with that coordinate. The final target (T_1) is found at $r = r_0$, $k = 1$ at the extreme lower right. The other points represent sets, R_k, of prestructs of r rings and k components (and n_0 atoms in all cases); the number of prestructs in the set R_k is $N_{r,k} = |R_k|$. The horizontal (ring) axis is labeled with letters to distinguish number of rings: A = monocycles ($r = 1$), B = bicycles ($r = 2$), etc., with O used for acyclic ($r = 0$) prestructs.[18] The number of components (k) is appended as a subscript. The prestruct $O_{n_0}(N_{0,n_0} = |O_{n_0}| = 1)$ is a set of n_0 single atom synthons, i.e., the ultimate classical synthetic precursors: coal, air, and water. The component maxima, M_r, represent the most components possible with r rings present ($M_r = \lim k$ for r rings). For the steroid ring system ($n_0 = 17$), $M_1 = 13$ components, i.e., the smallest ($\rho = 5$) monocycle and the remaining 12 atoms as single carbon synthons. Similarly, $M_4 = 1$ (no acyclic bonds to cut into extra components) and $M_3 = 5$ (the smallest tricycle, rings 234, plus the four extra single carbons of ring 1). For a C_{21} steroid-like progesterone, there are four acyclic bonds ($\alpha = 4$) so that $M_4 = 5$, while $M_3 = 9$.

Each line on the grid is a vector indicating one construction step (down or to the right, toward T) or the reverse cleavage (up or to the left, away from T). The horizontal vectors are intramolecular: cyclization (right) and ring opening (left). The vertical vectors are intermolecular, putting two synthons together by forming one bond (down) or cleaving a skeleton into two (up). The term ***affixation*** is offered for the former, the intermolecular reactions linking two separate synthons by formation of one σ bond between them. The four-coordinate directions are then cyclization



T = Target skeleton (r_0 rings, n_0 atoms, b_0 bonds)

λ = Number of bonds disconnected from T_1

A, B, C . . . = monocycles, bicycles, tricycles, etc.; r = 1, 2, 3 . . . respectively (O = acyclic synthons, r = 0)

R = General column of r rings

k = Number of synthon components in each prestruct, with maxima M_r for any r-cycle column (A, B, C, etc.)

R_k = Any intersection, a set of prestructs, all with n_0 atoms, k components and r rings

$|R_k|$ = $N_{r,k}$ = Number of prestructs in set R_k

$\lambda = \Delta r + \Delta k = r_0 - r + k - 1$

$b_0 = n_0 + r_0 - 1$

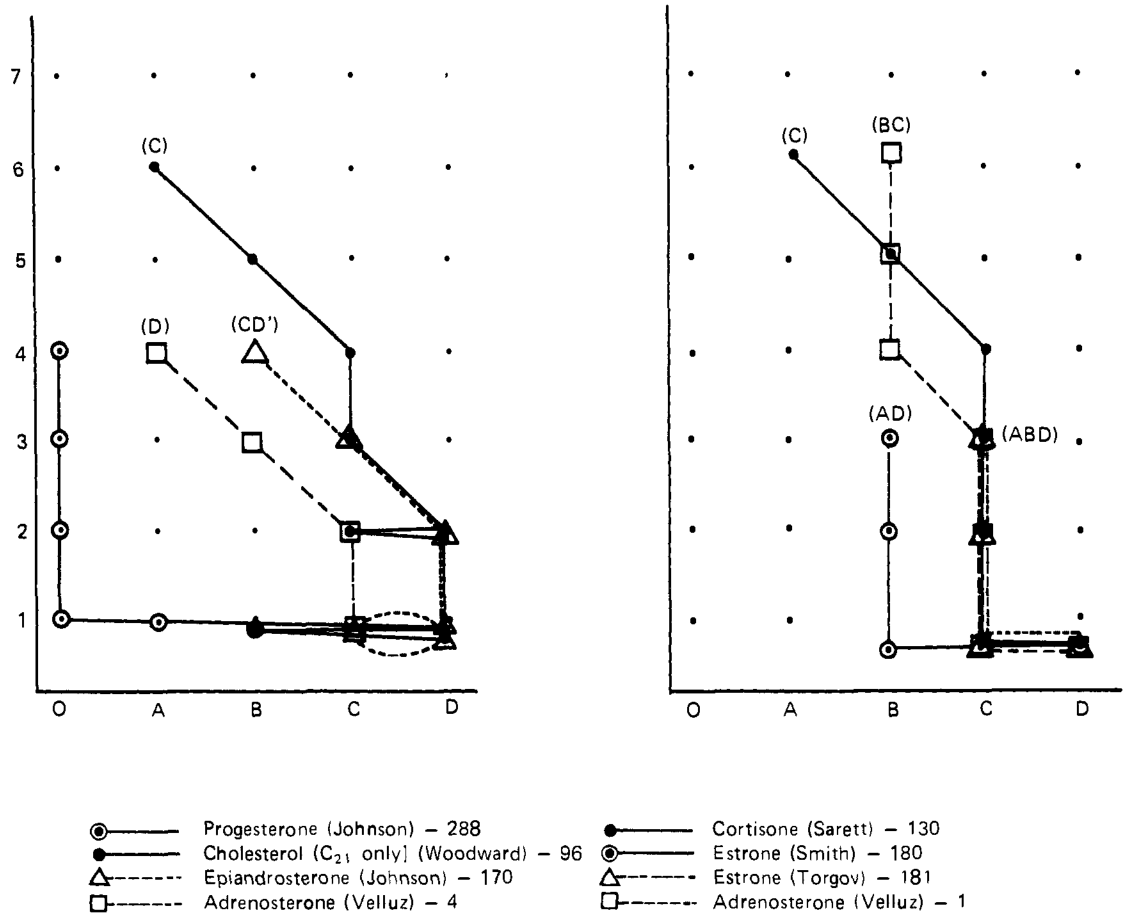
b = bonds in $R_k = n_0 + r - k = b_0 - \lambda$

Figure 5. The construction grid.

($\Delta r = 1$, $\Delta k = 0$), ring opening ($\Delta r = -1$, $\Delta k = 0$), affixation ($\Delta r = 0$, $\Delta k = -1$), and cleavage ($\Delta r = 0$, $\Delta k = 1$). Annelation reactions may be seen as the diagonal conversion ($\Delta r = 1$, $\Delta k = -1$), affixing a synthon and cyclizing it in two steps (two bond constructions), i.e., $R_k \rightarrow (R + 1)_{k-1}$.

The construction tree levels are now the diagonals back to the left from T_1 to O_{n_0}, labeled as λ, the number of links or bond formations required for a prestruct to go to T_1. The family of sets of prestructs at level λ is S_λ and numbers $|S_\lambda| = \binom{b_0}{\lambda}$, i.e., the number of prestructs missing λ bonds of the final skeleton T_1 ($\equiv S_0$). Interrelating equations are given below Figure 5, which summarizes this description of the construction grid.[19] With this framework, the total numbers of prestructs, R_k, for a given skeleton may be calculated and hence the total possible ways to construct the skeleton. For any given synthesis, all starting synthons are assembled as one prestruct of n_0 atoms.[19] k components, and a total of r fundamental rings at level λ. The synthetic sequence, or rather just its construction steps, is then a pathway down and to the right from R_k to T_1. A direct route is a path of λ constructions, consisting only of affixations and cyclizations (including annelations); an indirect route contains one or more cleavage (or ring-opening) steps.

Several pathways representing various successful steroid syntheses[4] are shown on the partial C_{21}-steroid grid in Figure 6. These pathways show quickly the cycles in the prestructs, the length of the route, the construction modes variously elected, including indirect routes, etc. None involve more than six starting material components, and the later syntheses usually involve fewer components and fewer constructions. Several involve ring openings (usually contractions of ring D), shown as horizontal loops. Of those shown only Johnson's biomimetic route builds an acyclic precursor



NOTES: Grid shown is part of that for C_{21}- steroids; r-cycle subsets (labeled with steroid ring letters) shown in parentheses for starting prestructs; references are to page numbers in ref. 4.

Figure 6. Representative steroid syntheses on the grid.

(from O_4) first, and it carries out three cyclizations at once ($A_1 \to D_1'$), followed by two simultaneous ring opening + recyclization steps ($D_1' \to B_1 \to D_1$).

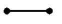
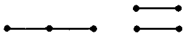
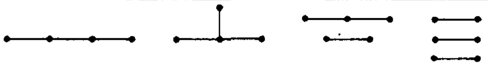
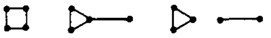
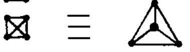
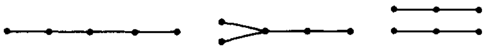
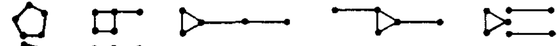
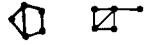
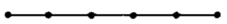
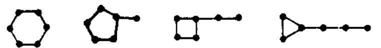
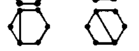
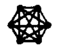
These illustrated pathways in Figure 6 represent families of single synthetic sequences through the construction tree, and the grid is simply a different representation of the tree. Examination of Figure 5 shows that—as in the traditional representation of the tree—the number of intermediates (prestructs) increases as one passes back through the levels (λ) from the target (T_1) and then decreases as intermediates become more primitive until it coalesces at the ultimate one-carbon intermediates, at O_{n_0}. We may now assess the sizes of such trees for particular targets by applying enumeration combinatorics to the construction grid, both for numbers of prestructs at each level and for the number of pathways for each through the grid to T_1.

Enumeration of Direct Routes. In order to appreciate the total numbers of possible prestructs, $|R_k|$, and so the numbers of pathways, $R_k \to T_1$, we need formulas for calculating $|R_k|$, and these may be derived in principle as numbers of combinations of λ bond disconnections on the product skeleton, distinguishing disconnections of rings (Δr) from those of acyclic chains (Δk). Derivation of these formulas is not a trivial nor apparently a previously examined graphical problem.[20] It increases rapidly in complexity with Δr, the number of rings cut, as does the enumeration (above) of monocycles from the face graph. The procedure developed below and summarized in detailed mathematical form in the Appendix is adequate for fully enumerating the construction grid for skeletons which are tetracyclic or less. With a little practice and a desk calculator, such an enumeration may be carried out in an hour or two.[21]

Cuts in α ($=\beta_{ii}$) bonds result in $\Delta k = 1$ per cut, and the number of prestructs with all r_0 rings intact will be the combinations of α bonds cut λ times, or $|T_k| = \binom{\alpha}{\lambda} = \binom{\alpha}{k-1}$. The first cut in one of the set of ring bonds β_{ij} results, however, in $\Delta r = 1$ and $\Delta k = 0$, while subsequent cuts in that same set (i.e., cuts of bonds of the same β type) result in $\Delta k = 1$ per cut since, following the cut in one β_{ij} bond, the others become in effect α bonds. The ways in which they may be cut is still, however, a matter of combinations; e.g., ring 1 of the steroid skeleton has $\beta_{01} = 5$ outer bonds so that the first cut in β_{01} can be made $\binom{5}{1} = 5$ ways, each with $\Delta r = 1$ and $\Delta k = 0$, while two cuts can be made $\binom{5}{2} = 10$ ways with the second cut leading to $k = 2$ components, thus generating ten prestructs of two components each. The number of resultant prestructs with x cuts in one bond type β_{ij} is $N_x = \binom{\beta_{ij}}{x}$. If two independent rings are cut, then all partitions of the x cuts between the two rings must be used, and any single selection of some (u) cuts in one ring allows all combinations of the remaining cuts ($v = x - u$) in the other ring, resulting in a total number of prestructs for x ring-bond cuts (eq 2).

$$N_x = \Sigma_{u,v} \binom{\beta_{ij}}{u} \binom{\beta_{kl}}{v} \text{ where } x = u + v \qquad (2)$$

Table III. Different Ways to Reduce Number of Structure Rings by Multiple Bond Disconnections

Δr	e′	c′ = Δk	Cutset Subgraphs F′
1	1	0	 (K_2)
2	2	0	
	3	1	(K_3)
3	3	0	
	4	1	
	5	2	
	6	3	 (K_4)
4	4	0	 etc.
	5	1	 etc.
	6	2	 etc.
	7	3	⋮
	8	4	⋮
	9	5	⋮
	10	6	(K_5)
5	5	0	 etc.
	6	1	 etc.
	7	2	 etc.
	8	3	⋮
	9	4	⋮
	10	5	⋮
	11	6	⋮
	12	7	⋮
	13	8	⋮
	14	9	⋮
	15	10	 (K_6)

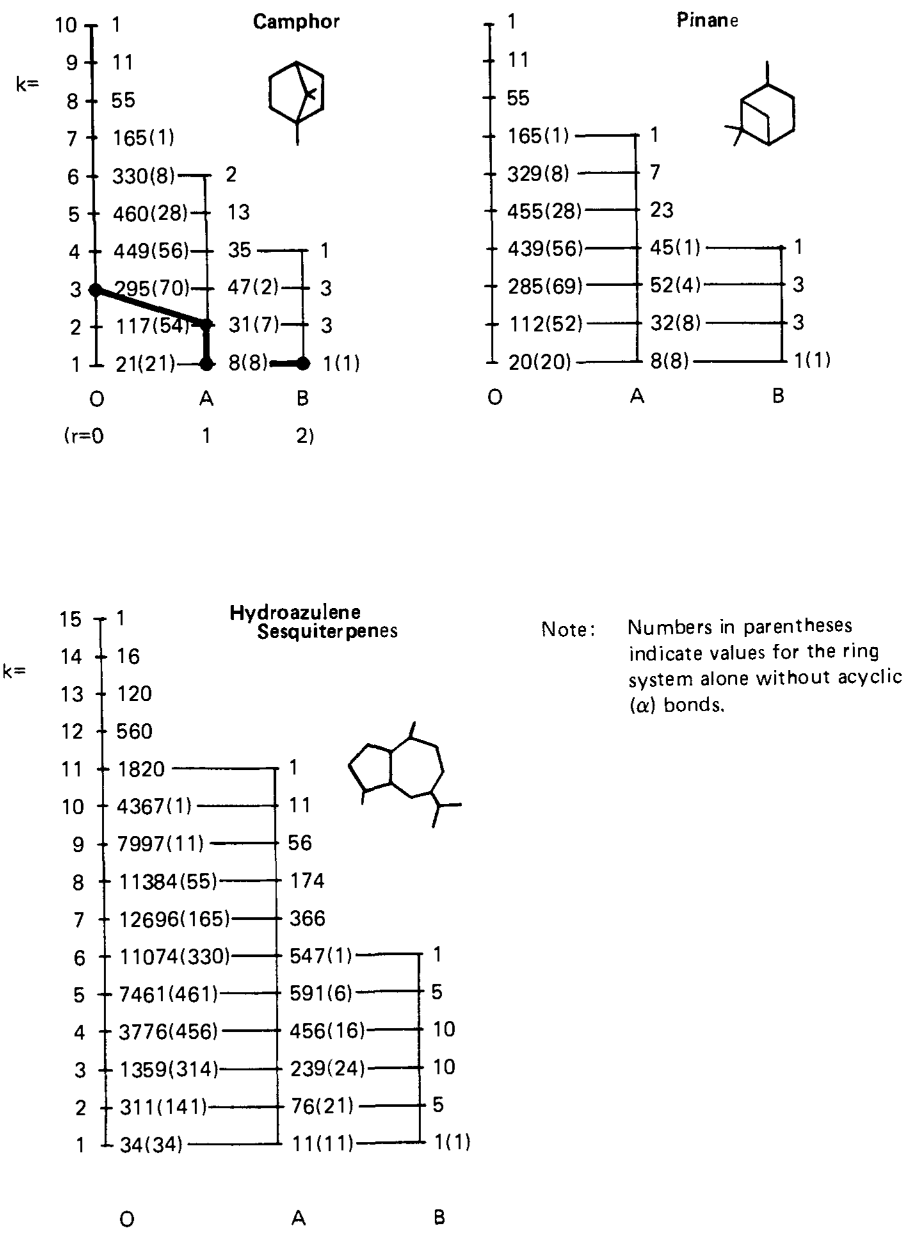
Notes: (a) The cutset subgraph (F′) represents that part of the facegraph which is cut: r′ points, e′ lines, c′ cycles, and k′ components, where e′-c′ = r′-k′ = Δr in the parent skeleton. Representative possible subgraphs are shown with different values of r′, e′, c′, and k′ but common e′-c′ = Δr to suggest the complexity of possibilities. The last subgraph in each Δr group is the maximum or complete graph on r′ points, labeled $(K_{r'})^7$ and defining the maximum number of different β-bonds which may be cut to achieve a given value (Δr) for skeletal ring number reduction. The full variety of cutset subgraphs is illustrated only for Δr ≤ 3.

This leads to an expectation that the general number of prestructs for x β-bond cuts will include summation of all possible partitions of x into u, v, w, etc., with all possible combinations of β_{ij} terms taken together as in eq 3.

$$N_x = \Sigma_{i,j,k\ldots}\ \Sigma_{u,v,w\ldots} \binom{\beta_{ij}}{u} \binom{\beta_{ki}}{v} \binom{\beta_{mn}}{w} \ldots \quad \text{where } x = u + v + w + \ldots \qquad (3)$$

This implies that the number (e') of *types* of β-bonds cut (and the partitioned number of cuts in each, u, v, w, etc.), which is the number of parentheses in the combinations product, will also equal the number of rings cut, Δr, and allow the statement of the number of components per prestruct to be $k = x - e' + 1$. This is, however, necessarily true (i.e., $e' = \Delta r$) for cuts into no more than two β types ($e' \leq 2$), i.e., eq 2. Complexity arises when three or more β types are cut, owing to a feature of the skeleton graph, and this is best understood by examining the possible sets of cuts.

The face graph (F) is a set of ($r_0 + 1$) points, and e lines representing the different types of β bonds. The set of β types which is cut then constitutes a subgraph of F called the cutset subgraph (F′), which is not necessarily a connected graph. The points in this subgraph (r') are the faces of the skeleton which are cut into and are equal in number to the number of different i, j ring numbers in the several β_{ij} which are cut; the number of lines (e') equals the number of different *types* of β bonds cut. Furthermore, each cycle in the cutset subgraph corresponds to an increase in the number of components equal to $\Delta k = 1$ in the resulting prestruct of the product structure. (Alternatively put, whenever the several β-bond types cut constitute a complete



Note: Numbers in parentheses indicate values for the ring system alone without acyclic (α) bonds.

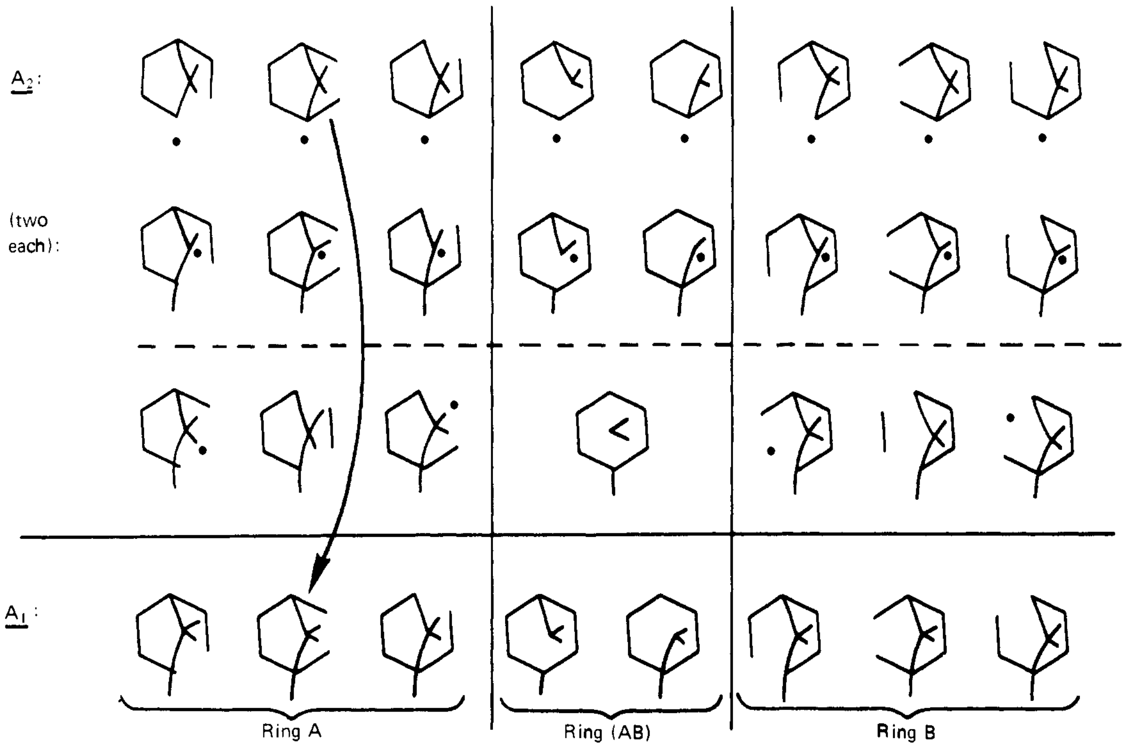
Figure 7. Construction grids for selected terpene skeletons.

cycle in F thus removed, the result is an increase of $\Delta k = 1$). This means that, if the x ring cuts are each into a different β type, i.e., if $x = e'$, then $e' = \Delta r + \Delta k = \Delta r + c'$, where c' is the number of cycles in the cutset F′.

The cutset (F′) thus defines the minimum requisite $\Delta k = c'$. The result is that three ring cuts, if a triangle in F′, only reduce the number of rings by $\Delta r = 2$, where otherwise it would be $\Delta r = 3$. This may be seen in the steroid by comparing three cuts into β_{01}, β_{02}, and β_{03} (with $\Delta r = 3$) and three cuts into β_{01}, β_{02}, and β_{12} (a triangle in F′, removed from F), yielding only $\Delta r = 2$. Similarly $\Delta r = 2$ for cutting β_{23}, β_{24}, and β_{34} or β_{24}, β_{25}, and β_{45} in the second structure of Figure 3. If four rings are arranged (such as 2, 3, 4, 5 in the second structure of Figure 3) so that their face graph is the maximum graph on four points, all six different β-type bonds can be cut while reducing the number of rings by only $\Delta r = 3$ (since $c' = 3$).

The number of ways of cutting a polycycle to result in a given decrease (Δr) in the number of rings is thus a function of the kinds of subgraphs contained in F, and these increase rapidly as Δr increases; this is summarized in Table III and is the reason that the collection of formulas for $|R_k|$ given in the Appendix was not pursued beyond $\Delta r = 3$, which already contains nine possible cutset subgraphs and four ways to achieve $\Delta r = 3$. For $\Delta r = 4$, there are 29 cutset subgraphs and seven ways with 4–10 bond types cut (Table III).

Taking the ring reductions of the cutset subgraph into account for all the kinds of triangle and square subgraph cycles in Table III and multiplying the combinations so obtained for ring cuts (x) with those obtained for cuts in the acyclic α bonds ($\lambda - x$ cuts), one obtains combinatorial formulas for numbers of prestructs ($N_{r,k}$) at any point (r,k) on the grid back to $\Delta r = 3$ from T_1. These general formulas are laid out in the Appendix. The formula of eq 3 is still basic but, as shown in the Appendix, is separately formulated for each of the subgraph-cycle types of Table III, and these totals (N') are added and subtracted as required to obtain particular totals $N_{r,k}$. It is possible to extend them to $\Delta r = 4$, i.e., to serve for tetracyclic skeletons, by invoking the known total prestructs $|S_\lambda| = \binom{b_0}{\lambda}$ at any level λ. For tetracyclics $T_1 = D_1$ on the grid, and columns A, B, C, D may then be enumerated, leaving the acyclic precursors (column O_k) to be obtained by subtraction from $|S_\lambda|$.

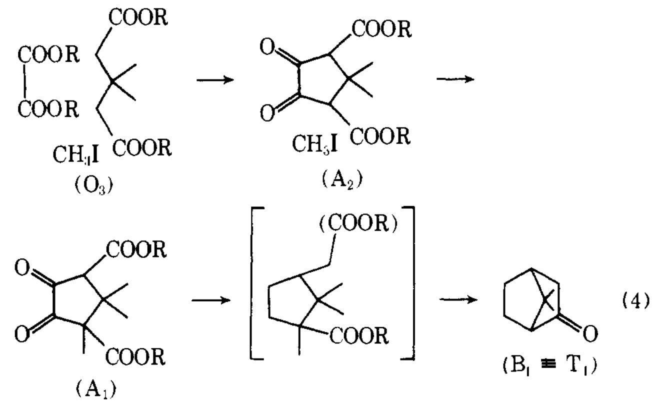


(Actual synthesis of sequence (4) shown with arrow.)

Figure 8. The monocyclic prestructs of camphor ($k < 3$).

Three terpene skeletons are enumerated as illustrations in Figure 7, in each case showing the full skeleton with attached acyclic bonds as well as the numbers of possibilities for the ring system alone ($\alpha = 0$) in parentheses. The latter numbers are much smaller not only because the skeleton has fewer atoms (ring atoms only) but mostly because of the rapid rise in numbers of prestructs when all the combinations of disconnected acyclic bonds are included. With R_1 sets, which are fully connected, however, the numbers are the same with or without inclusion of the α bonds.

By way of illustrating these grids, we may note the classical synthesis of camphor, shown as skeletal construction steps only in sequence 4; the addition and loss of the extraneous carbon (after A_1) does not figure in the grid as the carbon is not one which appears in camphor.[19] This sequence begins with one of the 295 possible three-component acyclic prestructs enumerated for O_3; this is converted by annelation to A_2, methyl affixed to A_1, and cyclized to B_1; the sequence is shown in boldface on the grid in Figure 7. Any examination of just the remaining 294 prestructs to start camphor synthesis from O_3 is already a huge task, and with all the other possible starting points equally considered it is clear that there are a great many possible syntheses of camphor, many of them undoubtedly viable. There are in fact about forty million total *routes* (see below) from single-carbon synthons ("coal, air and water")!



(4)

As a partial expansion of Figure 7 for camphor, Figure 8 shows the monocyclic prestructs, the eight A_1 precursors at the bottom divided into three subsets each characterized by having one of the three monocycles of camphor intact. The lowest A_2 row shows further dissection of β bonds and so constitutes the seven A_2 prestructs of the ring system only (ignoring α bonds), while the 24 other A_2 prestructs in the top two rows correspond to disconnection of the acyclic (α) bonds of camphor. Each of these may either construct an α bond, going to the eight A_1 skeletons, or construct a β bond, going to the three B_2 prestructs (full ring system but two components), 24 routes each, the former vertical on the grid, the latter horizontal. Each of the seven prestructs in the lowest A_2 row also has two routes open, but both are affixations, leading to A_1 structures. The seven possible annelations (see below) from A_2 to B_1 must also occur from the seven A_2 prestructs of this lowest row. The skeletal symmetry of camphor, which makes the vertical sets of ring A and ring B prestructs equivalent, is ignored here. The present treatment deals with labeled graphs, which have no symmetry, and reflects the possibility of asymmetry due to functionalized sites (see sequence 4). The $A_2 \rightarrow A_1$ part of sequence 4 is shown as a arrow in Figure 8. The detailed definition of intermediate prestructs in this manner can often show important avenues of construction which might not be apparent otherwise. This can be most fruitful in the upper levels (low λ, near T_1) where the numbers are not excessive.

The full grid for C_{21} steroids is shown in Figure 9 with the numbers of prestructs calculated for each set R_k and (in parentheses) those for the C_{17} ring system only, without the four α bonds. The numbers are very large indeed, especially for the former. This reflects, of course, the many skeletal combinations of constructing α bonds overlaid on each mode of building the ring system, but the numbers simply

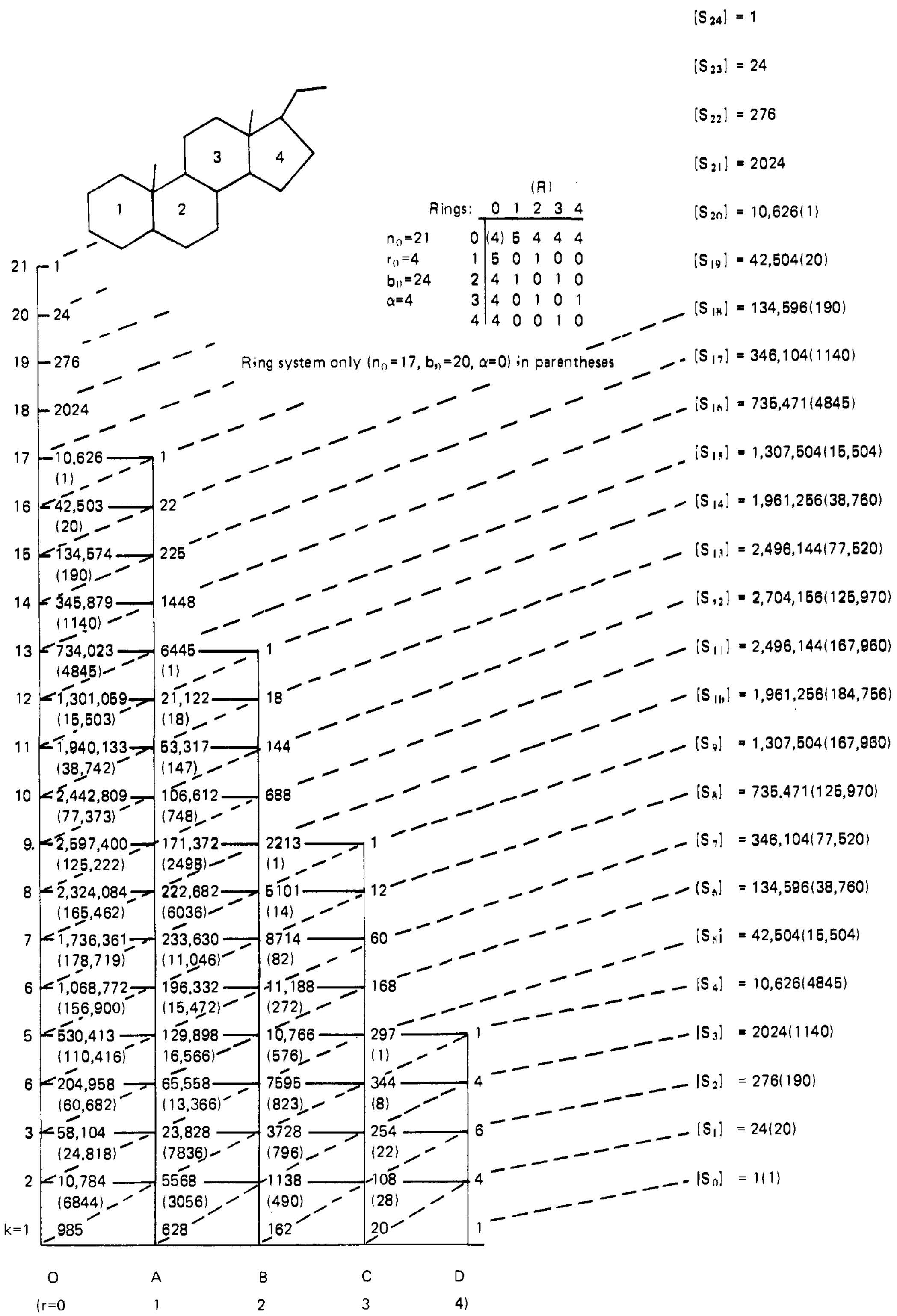


Figure 9. Construction grid for steroids.

for building the ring system are themselves formidable and beyond reasonable manipulation unless drastically reduced by some reductive criteria. The numbers of R_1 precursors are the same for each (C_{17} or C_{21}) since the α bonds are all necessarily connected in the R_1 families, and the upper k values in each R column are not possible for the C_{17} ring system since any C_{21} prestructs can have up to four more components than the C_{17} series. The totals $|S_\lambda|$ for each level are also shown and reflect the expansion, then contraction, of the tree back from the target (T_1) as the increase, then decrease, of combinatorial values of $|S_\lambda| = \binom{b_0}{\lambda}$.

We may note that the set of starting materials (skeletons only) for Johnson's progesterone synthesis at O_4 (Figure 6) is one of 204,958 prestructs formally available for that starting point. In principle these prestructs could all be generated by computer since the formulas developed above represent an algorithm capable of defining what each prestruct actually consists of. Many of these at O_4 will represent a single large synthon ($n_0 < 19$) coupled with three small synthons (C_1 or more), and many of these are only available themselves by synthesis (implying a true starting point higher up the T column). Thus many of these starting pre-

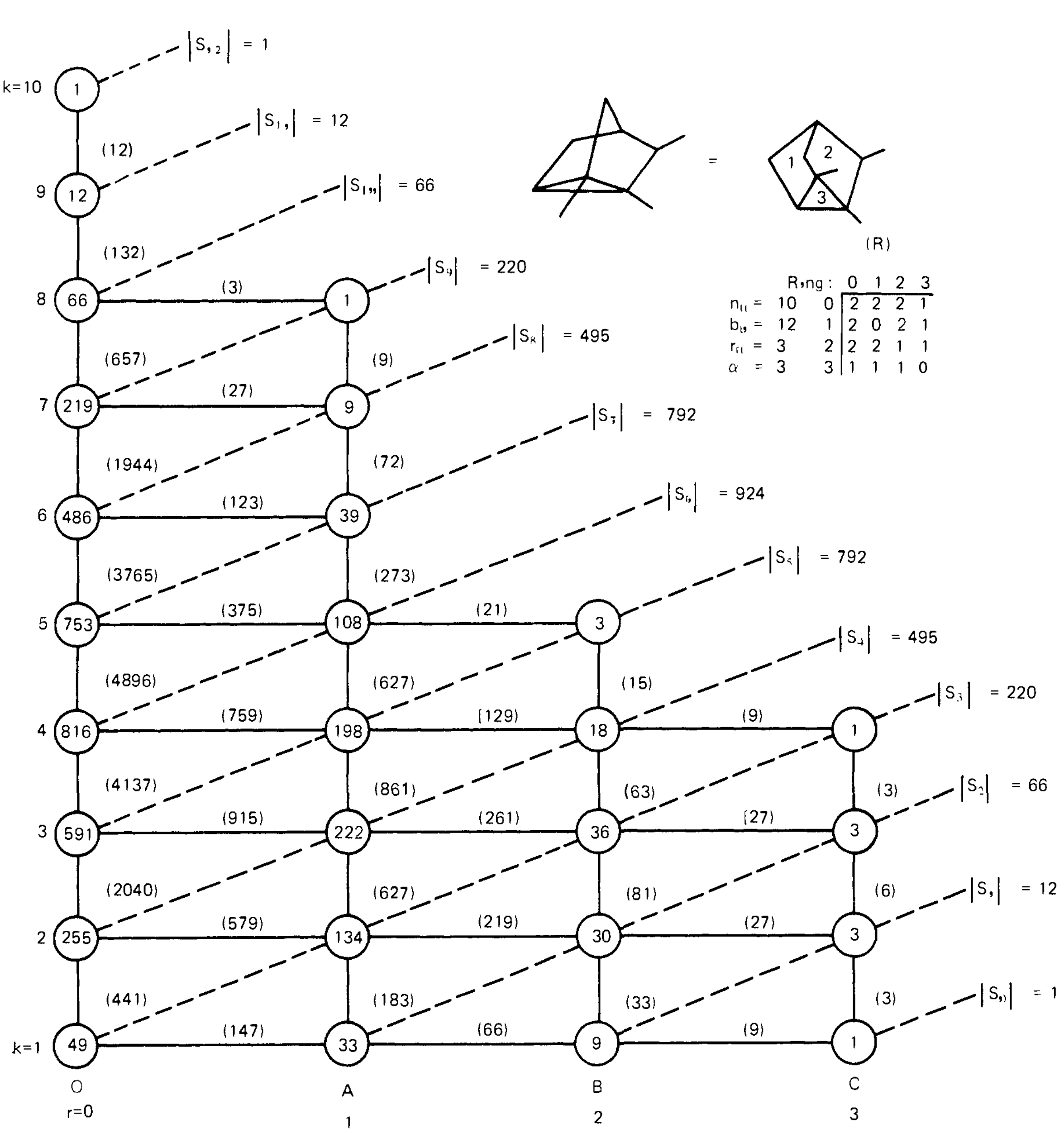


Figure 10. Construction grid for a tricyclene skeleton.

structs could be eliminated and the total reduced if they could first be matched by computer with the skeletons of available starting materials. Alternatively, if we had a smooth procedure for calculating the numbers of prestructs of different partitions by size ($n_1' + n_2' + n_3' + \ldots. = n_O$), we could also eliminate prestructs which involve excessively large synthons. However, this is a very difficult combinatorial problem.[7,17] The adjacency matrix (A) contains the information since A^n is a matrix, each element (i,j) of which shows the number of paths (of length n) through the molecule (graph) from atom i to atom j, and this describes a linear n-bonded synthon linking i and j. The procedure works for enumerating (and defining) the two-carbon synthons from A^2 and the three-carbon synthons from A^3, but the latter requires an adjustment owing to the redundancy of triangles, and the redundancy becomes serious for A^n with $n > 3$.

An alternate, though more limited, approach to enumeration is the matrix-tree theorem in graph theory applied to the product skeleton. This theorem (ref 7, p 152) allows the number of spanning trees,[18] $|O_1|$, to be calculated. A matrix, M, is obtained from $-A$ (the adjacency matrix of the skeleton with nonzero entries made equal to -1) by substituting the degree $(\sigma_i)^{1b}$ of each skeleton atom in the ith diagonal entry. (The row sums of M are therefore zero.) Then, removal of any row and corresponding column yields a determinant, evaluation of which affords the number of spanning trees,[22] i.e., the number of acyclic, one-component (fully connected) prestructs, $|O_1|$. For the tricyclene skeleton in Figure 10, $|O_1| = 49$, obtained either way; for camphor likewise $|O_1| = 21$, for pinane, 20. Appended acyclic bonds do not change the result for $|R_1|$, as Figures 7 and 9 show; hence the matrix-tree approach is more easily applied to the ring system alone.

It has apparently not previously been noted that M may also be used to evaluate $|A_1|$ and indeed any $|R_1|$ for skeletons with all rings fused (connected $\bar{F}$). For each defined monocycle subset of A_1, the rows and columns corresponding to the atoms which constitute that monocycle are simply removed from M, and the remaining determinant is evaluated to yield the number of possible ($k = 1$) prestructs containing that monocycle as the only ring.[23] In Figure 1, the number of monocyclic prestructs containing only ring 2 (atoms 14567) is 5, containing only ring 3 (atoms 345) is

12, containing ring (123), i.e., atoms 12356, is 5. The value of $|A_1|$ is the sum of these values for all monocycles [there are seven: 1, 2, 3, (12), (13), (23), (123)], in this case, $|A_1| = 33$. Since $\bar{F}$ is connected in Figure 1, $|B_1|$ may similarly be found as the sum of the calculations for bicycles.

A more general enumeration of these r-cycle subsets of R_k may be made with the formulas in the Appendix. These formulas are simply applied to the parent skeleton with the characteristic β_{ij} values of the particular r-cycle disallowed in calculation. In this way, one is computing all ways to cut the parent skeleton to R_k without touching any bonds belonging to a particular r-cycle.

When the points (intersections) on the construction grid have been enumerated as $|R_k|$, the number of synthetic pathways is readily computed. Any prestruct in R_k lacks λ bonds from the product. Making some of these may be cyclization (horizontal), others may be affixation (vertical). But the sum of the direct routes leading out from R_k must be $\lambda|R_k|$ as the sum of the routes leading in to R_k must be $(b_0 - \lambda)|R_k|$. This amounts to enumerating the routes corresponding to the lines of the grid. For most points R_k; there is both a horizontal and a vertical way in and a similar pair of ways out; only the sums for these pairs are so enumerated. But for certain points on the periphery, one of the pair is zero. Thus there is only one way in (horizontal) and one way out (vertical) for $T_{k-\max} \equiv T_M$, the highest point on the T column, and the same is true for the highest point on each column (R_M). Also, for O_k, there is only a vertical way in, for T_k only a vertical way out, and for R_1 only a horizontal way out. It is also true that the number of routes out of R_k vertically equals the number of vertical routes into R_{k-1}, i.e., there is only one number enumerating the routes corresponding to a line between two points; if it is evaluated from one point, it holds for the other. The relations are summarized in Table IV, taking $V_{r,k}$ and $H_{r,k}$ to symbolize vertical (intermolecular) and horizontal (intramolecular) routes leading into R_k, and $\bar{V}_{r,k}$ and $\bar{H}_{r,k}$ the corresponding routes leading out.

All the possible routes may be enumerated starting with column T_k and evaluating V for each link down the column from T_M. This leaves H to be evaluated by subtraction, and this in turn represents $\bar{H}$ for the (T − 1) column and following this the $\bar{V}$ and V values for (T − 1), etc., until the grid is complete. The number of overall direct pathways from R_k to T_1 is now the product of all the links along the path chosen, $\bar{H}$ or $\bar{V}$. Alternatively, the number of ways into or out from R_k, or r-cycle subsets of R_k, can also be computed directly by formulas given in the Appendix, analogous to the $|R_k|$ enumeration formulas. It is enough to enumerate horizontal (cyclization) lines $H_{r,k}$ and $\bar{H}_{r,k}$ since vertical ways ($V,\bar{V}$) are obtained by subtraction from $\lambda|R_k|$ or $(b_0 - \lambda)|R_k|$, as in Table IV.

Horizontal (cyclization) paths in to R_k are very easily determined for the individual r-cycle subsets as $\rho = \Sigma\beta_{ij}$ which characterizes any particular r-cycle. The total $H_{r,k}$ for R_k is then the sum of these ρ values for the separate r-cycle subsets of R_k. The horizontal ways out ($\bar{H}_{r,k}$) from R_k may be obtained from the previous enumeration formulas when it is noted that only one cut in a β_{ij} bond of the parent is allowed if the resultant R_k prestruct is to be capable of a one-step cyclization to $(R + 1)_k$. Also it may be noted that the multiple cuts in different β-type bonds which characterize triangle and square subgraphs of F′ (i.e., N'_{21}, N'_{31}, N'_{32}, N'_{33} in the Appendix) result in prestructs in R_k which are incapable of one-step cyclization. The resultant formulas for $\bar{H}_{r,k}$ are included in the Appendix. They may be applied, as with $N_{r,k} = |R_k|$ enumerations, either to the full set R_k, or to its r-cycle subsets of disallowing the $e_i(\beta_{ij})$ values characterizing the r-cycle of the particular subset in the computation.

Table IV. Enumeration of Routs into and out of Prestruct Sets

$$V_{r,k} + H_{r,k} = (b_0 - \lambda)\,|R_k| = (n_0 + r - k)\,N_{r,k}$$

$$\bar{V}_{r,k} + \bar{H}_{r,k} = \lambda\,|R_k| = (r_0 - r + k - 1)\,N_{r,k}$$

$$V_{r,k} = \bar{V}_{r,k+1} \qquad H_{r,M} = (n_0 + r - M)\,N_{r,M}$$

$$\bar{V}_{r,k} = V_{r,k-1} \qquad \bar{H}_{r,1} = (r_0 - r)\,N_{r,1}$$

$$H_{r,k} = \bar{H}_{r-1,k} \qquad V_{0,k} = (n_0 - k)\,N_{0,k}$$

$$\bar{H}_{r,k} = H_{r+1,k} \qquad \bar{V}_{r_0,k} = (k - 1)\,N_{r_0,k}$$

$$H_{0,k} = \bar{H}_{r_0,k} = V_{r,M} = \bar{V}_{r,1} = 0$$

$$V_{r,k} = (r_0 - r + k)\,N_{r,k+1} - H_{r+1,k+1}$$

$$H_{r,k} = (n_0 + r - k)\,N_{r,k} - V_{r,k}$$

The total number of overall construction pathways (direct routes only) may be seen more simply (lumping H and V) as the product of all $\lambda|S_\lambda|$ terms from the starting level (λ) of R_k to $S_0 \equiv T_1$.

Total direct routes from R_k to T_1 = $\lambda!|R_k| = (r_0 - r + k - 1)!|R_k|$.

Total direct routes from $S_{b_0} \equiv O_{n_0}$ ("coal, air and water") = b_0! Thus the number of possible direct routes is enormous, but it is not infinite. The total overall direct routes to camphor number 39,916,800 and for C_{21} steroids, 2.4×10^{18}. It is of course many times more if longer, indirect pathways are used (next section).

An expansion of the construction grid by ring types is possible, as shown in Figure 11, by breaking down each intersection point, R_k, into its r-cycle subsets. These may be enumerated separately, as noted above and in the Appendix. The network of horizontal (cyclization) lines into and out of each such subset will now connect it to a number of different subsets of $(R - 1)_k$ and $(R + 1)_k$ as illustrated in the $k = 2$ row of Figure 11. The sets in columns O and T, however, cannot be so differentiated so that all cyclization ways leading in to the several monocyclic subsets of A_k must originate in O_k and all ways out from the different $(r_0 - 1)$ ring systems of $(T - 1)_k$ must go to T_k. As noted above and in the Appendix, formulas for enumerating H and $\bar{H}$ for the separate subsets are available.

The vertical lines, however, remain simple and parallel, reflecting that, in an affixation of component synthons ($\Delta r = 0$, $\Delta k = -1$), the particular r-cycle ring system necessarily remains intact up any subset column of R_k, unchanged by interconversions vertically. To clarify this, note that each of the ten defined monocycles (such as 1, 2, (12), (1234), etc.) of the steroid (Figure 4) or the three of camphor (Figure 8) represents a single subset of A_k, and that the defined ring is present no matter how divided the rest of the prestruct is into synthons, i.e., no matter whether k = 1, 2, 3, etc. Hence the vertical (affixation–cleavage) interconnections (interconversions) are all single vertical lines, each characteristic of particular single r-cycles. A fully expanded construction grid for the tricyclene ring system (with no α bonds) is illustrated as Figure 12. The intermixed cycliza-

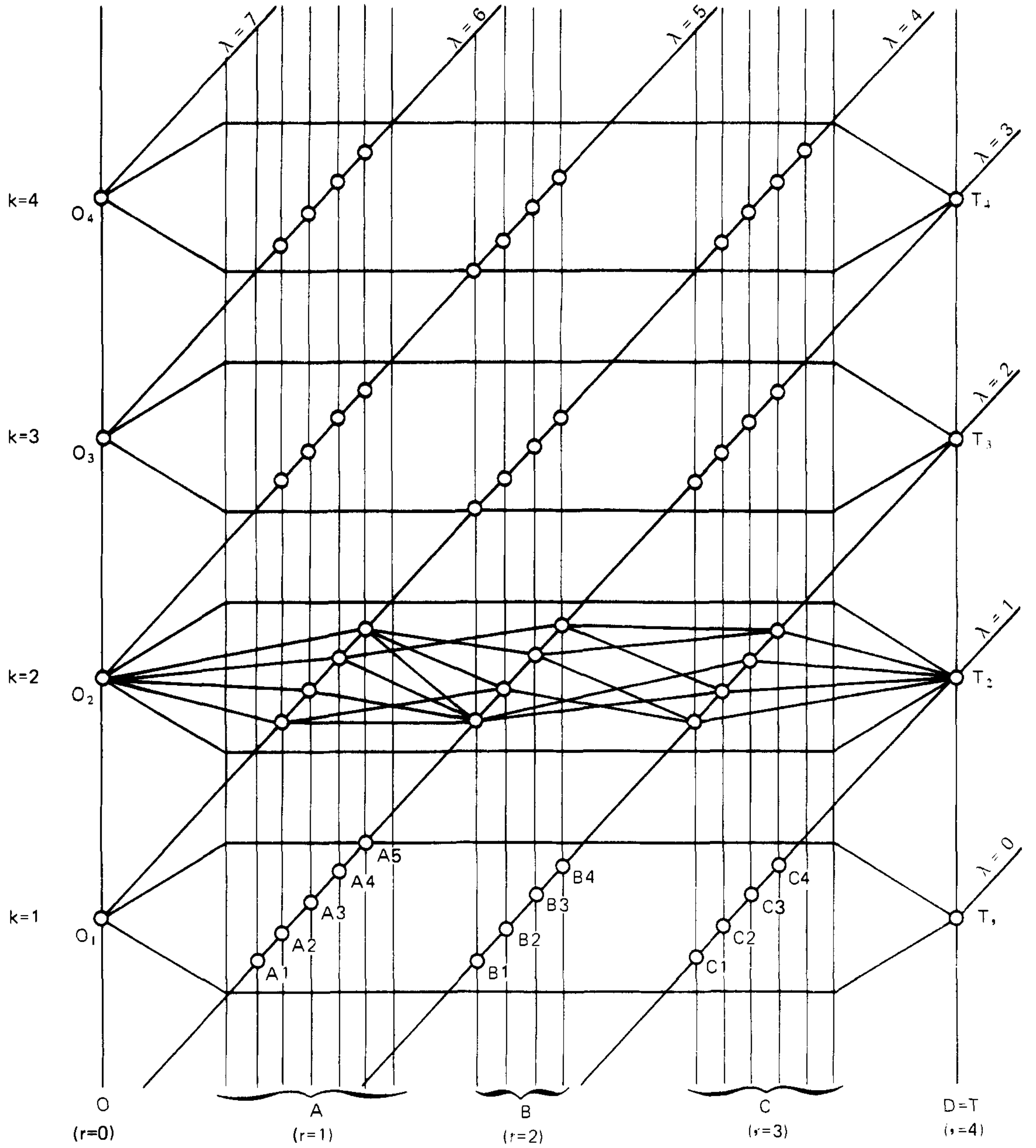


Figure 11. Expanded construction grid with r-cycle subsets.

tion routes from $A_1 \rightarrow B_1$ and $A_2 \rightarrow B_2$ are simply collected and enumerated from each side without showing the detailed links from one subset to another. The numbers here can be expanded into the real skeletons of the various prestructs as done for camphor in Figure 8.

The number of annelations which create any given r-cycle subset of R_k, from $(R - 1)_{k+1}$, may be enumerated also, given the β_{ij} values which characterize the r-cycle of the subset. Annelation is defined as two constructions, one affixing and one cyclizing ($\Delta r = 1$, $\Delta k = -1$), an annelation forming R_k designated as $A_{r,k}$. This requires the formation of two bonds of the same β type. Thus for the target (T_k) column

$$A_{r_0,k} = \sum_{i=1}^{e} \binom{e_i}{2}$$

and is the same for all values of k. For r-cycle subsets in R_k, the same is true as long as the β values used are those for the graph of the r-cycle, generated from $(F - F')$ by line and point coalescence as described before. These new β values for an r-cycle, easily defined by inspection, may be listed serially as e_i' so that $A_{r,k} = \Sigma_i\binom{e_i'}{2}$. For monocycles, the number of annelations if $\binom{\rho}{2}$. Annelations may conveniently be described as $(m + n)$ annelations, in which m and n denote the sizes of the two units combined into a ring of size $\rho = m + n$. Thus carbene addition to olefins is a (1 + 2) annelation, Diels–Alder cycloaddition is (2 + 4) annelation, and Robinson annelation can be either (2 + 4) or (3 + 3) annelation. In Figure 13 are listed the numbers of annelations for the 33 steroid r-cycles. The number of $(m + n)$ kinds of annelations to a monocycle is then the number of partitions of ring size ρ into two parts, $m + n$. The lowest row of A_2 for camphor in Figure 8 represents prestructs for $A_2 \rightarrow B_1$ annelations of the two possible kinds: (1 + 4) and (2 + 3).

Indirect Routes. Indirect routes are longer than direct routes by multiples of two steps.[1b] The length of the route or path is the number of lines traversed on the grid from R_k to T_1. For direct routes, this is $(k - 1) + (r_0 - r) = \lambda$; for indirect routes, the path length must be $(\lambda + 2n)$, where n

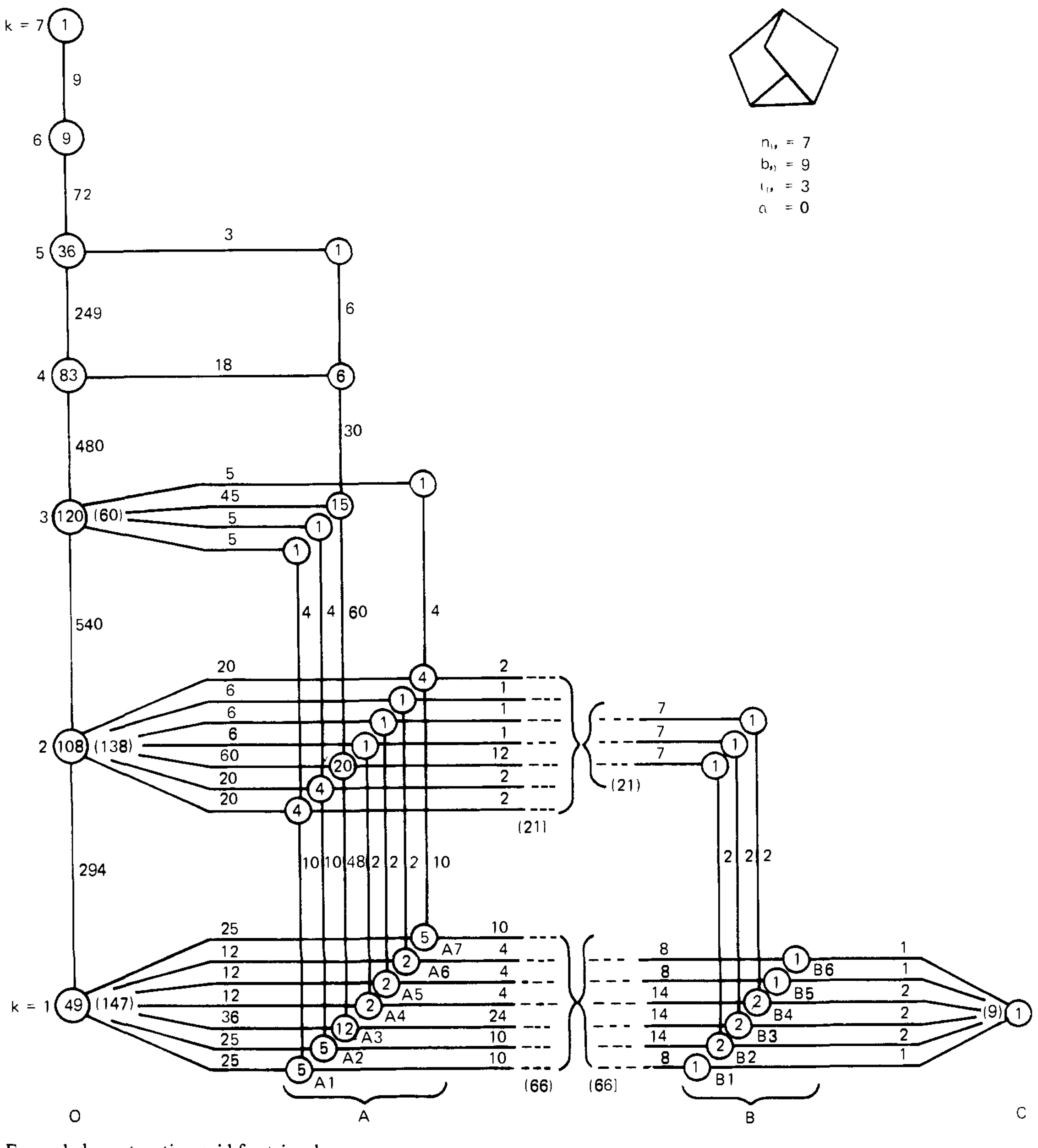


Figure 12. Expanded construction grid for tricyclene.

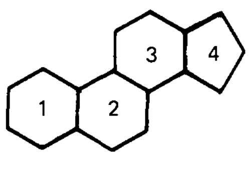
= 1, 2, 3 .... Indirect routes necessarily involve cleavages (or ring opening; i.e., $\Delta k = 1$ or $\Delta r = -1$), the number being n in $(\lambda + 2n)$. Owing to the factor of 2, indirect routes rapidly become inefficient by having too many steps, and most indirect routes used in existent syntheses include $n = 1$ cleavage (or only occasionally two).

Rearrangements may be viewed as equivalent to two-step transformations, one cleavage and one construction, reverting to the same set, R_k. However, the r-cycle subsets of the starting and product R_k prestructs are different. Synthetically the same skeletal change is achieved whether the cleavage + construction are carried out (in either order) in two steps or in one step via rearrangement. Thus the rearrangement of the decalin to hydroazulene skeleton may be achieved either by rearrangement of the central bond through a leaving group in the 1 equatorial position or by ozonolysis of $\Delta^{9,10}$-octalin followed by aldol cyclization.

On the construction grid, enumeration of indirect routes which contain only one cleavage (or rearrangement) and require $(\lambda + 2)$ steps from R_k may be analyzed as follows. The bond cleaved can be a bond between any two atoms of the skeleton which are not actually bonded in the target (T_1).[24] This implies an analysis of a formal $(T + 1)_1$ product, which is a set of all such possible structures. The structures in this set will have n_0 atoms, $(b_0 + 1)$ bonds, and $(r_0 + 1)$ rings, and the number of structures will be $|(T + 1)_1| = \binom{n_0}{2} - b_0$. In principle each structure in the $(T + 1)$ set is now to be enumerated as outlined in the Appendix. Such enumeration is an enormous calculation, even for only one internal cleavage in the direct route, since the number of $(T + 1)_1$ structures is so many; for the 21-carbon steroid skeleton (Figure 2), $|(T + 1)_1| = 186$ structures to be enumerated.

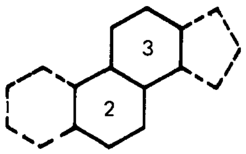
In practice, the only viable possibilities to be considered for $(T + 1)_1$ structures are those in which the extra bond to be cleaved is part of a three- to six-membered ring and perhaps also related by requisite functionality to the final functionality of the product. In any case, if we simply focus on the number of possible extra bonds added to the structure which may be considered for cleavage, the number $|(T + 1)_1|$ above is the maximum number for such cleavable bonds.[25] They need not be cleaved as the last step in the

$$A_{r,k} = \Sigma_i \binom{e_i'}{2}$$



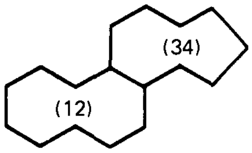
[1234]

$$\begin{aligned} e_1 &= 5 & e_5 &= 1 \\ e_2 &= 4 & e_6 &= 1 \\ e_3 &= 4 & e_7 &= 1 \\ e_4 &= 4 \end{aligned} \qquad A_{4,k} = \binom{5}{2} + 3\binom{4}{2} + 3\binom{1}{2} = 28$$



[23]

$$\begin{aligned} e_1' &= \beta_{02} + \beta_{12} &= 5 \\ e_2' &= \beta_{03} + \beta_{34} &= 5 \\ e_3' &= \beta_{23} &= 1 \end{aligned} \qquad A_{2,k} = 2\binom{5}{2} = 20$$



[(12) (34)]

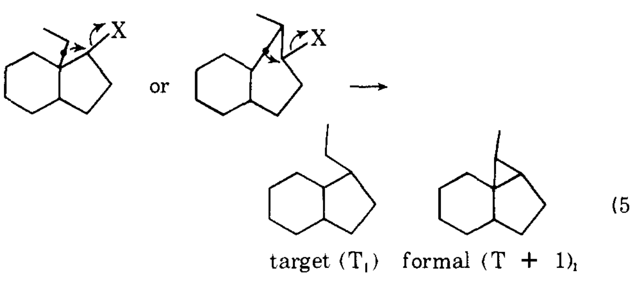
$$\begin{aligned} e_1' &= \beta_{01} + \beta_{02} &= 9 \\ e_2' &= \beta_{03} + \beta_{04} &= 8 \\ e_3' &= \beta_{23} &= 1 \end{aligned} \qquad A_{2,k} = \binom{9}{2} + \binom{8}{2} = 64$$

Monocycles		Bicycles				Tricycles	
[1]	15	[12]	20	[1 (23)]	46	[123]	26
[2]	15	[23]	20	[(23) 4]	42	[124]	30
[3]	15	[34]	16	[1 (34)]	42	[134]	51
[4]	10	[13]	30	[2 (34)]	38	[234]	22
[(12)]	45	[24]	25	[(123) 4]	84	[(12) 34]	48
[(23)]	45	[14]	25	[1 (234)]	76	[1 (23) 4]	44
[(34)]	36	[(12) 3]	46	[(12) (34)]	64	[12 (34)]	44
[(123)]	91	[(12) 4]	65				
[(234)]	78						
[(1234)]	136						

Figure 13. Numbers of annelations to steroid r-cycles.

synthetic sequence, of course, so that (T + 1) structures are considered for enumeration but the (T + 1) stage need never be a point in a given sequence.

Enumeration of rearrangement possibilities is by contrast quite simple since the allowed formal structures in $(T + 1)_I$ must be only those in which the extra bond to be cleaved is one which creates a three-membered ring on the product structure. Thus the problem is the graphical one of determining how many triangles can be created by adding lines (bonds) to the skeleton graph. Considering that such a bond bridges an atom by bonding two atoms adjacent to it, the possible number must be the sum of binary combinations $\binom{\sigma_i}{2}$ of the σ values for each skeletal atom, i.e., zero for a primary site ($\sigma = 1$), 1 for secondary ($\sigma = 2$), 3 for tertiary ($\sigma = 3$), 6 for quaternary ($\sigma = 4$).[1a] The extra bond which rearranges can, however, rearrange so as to create either of the two "real" bonds of the skeleton which constitute its triangle in $(T + 1)_I$ as illustrated in transformation 5.



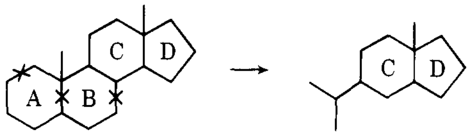
(5)

Hence the number of possible rearrangements which can be used (at any step in the sequence) will be:

$$\text{rearr} = 2 \sum_{i=1}^{n_Q} \binom{\sigma_i}{2}$$

for 21-carbon steroids (Figure 2), this is $2[3\binom{1}{2} + 11\binom{2}{2} + 5\binom{3}{2} + 2\binom{4}{2}] = 76$ rearrangements. These may be broken down as applied to ring expansions or contractions, etc., and some will be structurally impossible, invoking pentacovalent ($\sigma = 5$) carbons in the precursors.

It may be noted that the grid as defined takes no account of possible syntheses involving synthons en route which have more skeletal atoms than the target.[19,24] The grid concept allows in principle more rings than there are in the product so that prestruct sets (T + 1, T + 2, etc.) are possible (with $r > r_0$), and cleavable to T, and the product (T_1) is no longer the extreme lower right-hand point on the grid. If extra, cleavable skeletal atoms were accepted, then the product could be T_n, not T_1; i.e., there would be more than one component at the end of the synthesis, the other components being the extra atoms removed, and the product would appear in the midst of the grid rather than at the lower right corner. This makes a more general but more complicated formulation which does not appear important for the few occasions in which the sacrifice of skeletal atoms seems a viable synthetic approach.[26] However, there are such synthetic possibilities which cannot be ignored. Stork's synthesis of cantharidin (ref 4, p 59) involves discarding two of the four carbons of the butadiene used in a Diels–Alder annelation, and it could be argued that a rational synthesis of the C_{15} picrotoxin skeleton could be a carving down of a suitable, available steroid of 20 or more carbons since the absolute stereochemistry is the same (transformation 6).[27] Indeed the biosynthesis of cholesterol (n_0 = 27) or estrone (n_0 = 18) proceeds via squalene and lanosterol (n_0 = 30). The biosynthesis of lanosterol itself involves $3C_2 \rightarrow C_6 \rightarrow C_5$ via a one-carbon loss from mevalonic acid. In terms of the present treatment, lanosterol biosynthesis is $(C_2 + C_2 + C_1 \rightarrow C_5) \times 6$, or $O_{18} \rightarrow \ldots\ldots \rightarrow O_1$ (squalene) $\rightarrow D_1$ (compare Figure 6).



(6)

In summary, while the synthesis tree is enormous, it is possible through definition of its construction subtree as the construction grid to examine its size and so to survey the scope of the problem of systematic synthesis design, at least in terms of construction. As a reflection of the great proliferation in combinatorials, these grids are still impressive in size even for small molecules. They can be calculated, however, and such calculations can offer new insights into modes of synthesis for particular target molecules, particularly with respect to defining combinations of rings in polycyclic targets.

Appendix. General Enumeration Formulas for Intermediates and Routes

Given the definitions and relating equations in Figure 5, let x = ring cuts (into β_{ij} bonds) so that $(\lambda - x)$ = cuts in $\alpha(\beta_{ii})$, or acyclic bonds. From the face graph (F) and ring matrix (R), several lists are first constructed in which the β_{ij} values of R are redefined as e_i, t_{ii}, etc., according to their participation in particular subgraphs (potential cutset subgraphs) of F, as follows:

(1) e_i = order list of nonzero β_{ij} values from R taken in sequence from the right half of the matrix above the diagonal. The list will contain e β values. If every face is adjacent to the outside face (all $\beta_{oi} > 0$), $e_1 = \beta_{01}$, $e_2 = \beta_{02}$, $e_3 = \beta_{03}, \ldots . e_{r_0} = \beta_{or_0}$, etc.

(2) t_{ci} = matrix ($t \times e$) of triangles in F; there are t triangles, each expressed in one row of the matrix as a list of the three e_i values composing the triangle followed by all other $(e - 3)e_i$ values in F, as values $t_{c4} \rightarrow t_{ce}$. The triangles are arbitrarily numbered with the row number (cycle number), c ($1 \leq c \leq t$). Each triangle list includes all other e_i values so that the matrix may be used to enumerate triangles for $\Delta r = 2$, $e' = 3$ as well as $\Delta r = 3$, $e' = 4$ (▷— in Table III).

(3) s_{ci} = matrix ($s \times 5$) of open squares (□) and simple crossed squares (⧄) in F; there are s squares whether crossed once or not, yielding s corresponding matrix rows of five e_i values each, labeled as above with the matrix row number (cycle number), c ($1 \leq c \leq s$). The four e_i values for the open square are listed first and the fifth item (s_{c5}) will be the e_i corresponding to a simple cross in ⧄, if present, or to zero if not. Note, however, that, in a collection of simple crossed squares, the number of open squares contained is the same except when the K_4 subgraph (double-crossed square (⊠) or ▲) is present: this has three open (□) but six simple crossed squares (⧄). Combining these in one matrix allows it to be used for evaluation of modes of cutting to $\Delta r = 3$ for both $e' = 4$ and 5. With any K_4 subgraph, however, the three extra simple crossed squares must be added to the list.

(4) s'_{ci} = similar matrix ($s' \times 6$) of double-crossed squares (K_4 subgraphs) in F, a set of s' lists of double-crossed squares with six e_i values each. Lists are again numbered c ($1 < c < s'$). These lists, for cutting to $\Delta r = 3$ for $e' = 6$, will be uncommon, occurring with three skeletal faces all mutually fused, as in tricyclenes (cf. Figure 1) or tetrahedrane, or with structures like the second in Figure 3 which includes a K_4 subgraph in the incomplete face graph, $\bar{F}$.

Define an operator

$$Q_x(a, b, c, \ldots) = \sum_{u,v,w\ldots}^{x} \binom{a}{u}\binom{b}{v}\binom{c}{w}\ldots;$$

$$x = u + v + w + \ldots; u \leq a; v \leq b; w \leq c; \text{ etc.}$$

This will be the sum of the products of combinations of a, b, c ... taken u, v, w ... at a time, respectively, wherein u, v, w ... represent all integral partitions of x. Description and use of the Q_x operator is further amplified below.

The enumeration formulas for the construction grid are then the following, expressed as functions of given values for r, k.

$$\Delta r = 0\text{: } N_{r_0,k} = |T_k| = \binom{\alpha}{k-1}$$

$$\Delta r = 1\text{: } N_{r_0-1,k} = |(T-1)_k| = \sum_{x=1}^{k} \binom{\alpha}{k-x} \sum_{i=1}^{e} \binom{e_i}{x}$$

$$\Delta r = 2\text{: } N_{r_0-2,k} = |(T-2)_k| = \sum_{x=2}^{k+1} \binom{\alpha}{k-x+1} [(N'_{20})_x + (N'_{21})_x]$$

$$\Delta r = 3\text{: } N_{r_0-3,k} = |(T-3)_k| = \sum_{x=3}^{k+2} \binom{\alpha}{k-x+2} \times [(N'_{30})_x - (N'_{21})_x + (N'_{31})_x + (N'_{32})_x + (N'_{33})_x]$$

where

$$(N'_{20})_x = \sum_{i=1}^{e-1} \sum_{j=i+1}^{e} Q_x(e_i,e_j)$$

$$(N'_{21})_x = \sum_{c=1}^{t} Q_x(t_{c1},t_{c2},t_{c3})$$

$$(N'_{30})_x = \sum_{i=1}^{e-2} \sum_{j=i+1}^{e-1} \sum_{k=j+1}^{e} Q_x(e_i,e_j,e_k)$$

$$(N'_{31})_x = \sum_{c=1}^{t} \sum_{i=4}^{e} Q_x(t_{c1},t_{c2},t_{c3},t_{ci}) + \sum_{c=1}^{s} Q_x(s_{c1},s_{c2},s_{c3},s_{c4})$$

$$(N'_{32})_x = \sum_{c=1}^{s} Q_x(s_{c1},s_{c2},s_{c3},s_{c4},s_{c5})$$

$$(N'_{33})_x = \sum_{c=1}^{s'} Q_x(s'_{c1},s'_{c2},s'_{c3},s'_{c4},s'_{c5},s'_{c6})$$

The two digits in the N' subscript are respectively Δr and c' ($= \Delta k_{\min}$), and the sum of those digits is the *minimum* value for x in the Q_x combination. The N' terms represent special enumerations required which correspond to the presence in F of the subgraphs in Table III, N'_{20} and N'_{21} for $\Delta r = 2$ and N'_{30}, N'_{31}, N'_{32}, and N'_{33} for $\Delta r = 3$. If these subgraphs are absent in the face graph, the corresponding N' term is zero.

The following identities are understood for the combination terms of n things taken p at a time:

$$\binom{n}{p} = 0 \text{ for } p > n; p < 0; n = 0 \text{ (except } n = p = 0)$$

$$\binom{n}{p} = 1 \text{ for } p = n; p = 0; n = p = 0$$

$$\binom{n}{p} = n \text{ for } p = 1; p = n - 1$$

The enumerations may be extended to $r_0 = 4$ via the total number of prestructs in the family of sets λ; thus, in a tetracyclic skeleton, $|A_k|$, $|B_k|$, $|C_k|$, and $|D_k|$ are obtained, and $|O_k|$ arises by substraction from $|S_\lambda|$.

$$|S_\lambda| = \binom{b_0}{\lambda}; |S_0| = |T_1| = 1; |S_1| = b_0; |S_{b_0}| = |O_{n_0}| = 1$$

$$|S_\lambda| = |O_{\lambda-r_0+1}| + |A_{\lambda-r_0+2}| + |B_{\lambda-r_0+3}| + |C_{\lambda-r_0+4}| + \cdots$$

and

$$|O_k| = \binom{b_0}{n_0-k} \text{ for } k \geq M_1 \; (M_1 = k_{\max} \text{ for monocycles})$$

The same general formulas for enumeration may also be used for defined r-cycle subsets of R_k in which the e_i values (β_{ij}) which are used in calculation are the complement of those characteristic of the particular r-cycle, along with any lists of triangles (t_{ci}) or squares (s_{ci}) also contained in the complement of the r-cycle.

Horizontal (cyclization) ways out ($\bar{H}_{r,k}$) from the prestructs in R_k may be directly obtained by enumeration.

$$\bar{H}_{r_0-1,k} = \binom{\alpha}{k-1}(b_0 - \alpha)$$

$$\bar{H}_{r_0-2,k} = \sum_{x=2}^{k+1} \binom{\alpha}{k-x+1} \sum_{i=1}^{e} e_i \sum_{j\neq i}^{e} \binom{e_j}{x-1}$$

$$\bar{H}_{r_0-3,k} = \sum_{x=3}^{k+2} \binom{\alpha}{k-x+2} [(H'_{30})_x - (H'_{21})_x + (H'_{31})_x]$$

where

$$(H'_{30})_x = \sum_{i=1}^{e} e_i \left[\sum_{j=1}^{e-1} \sum_{k=j+1}^{e} Q_{x-1}(e_j,e_k)\right]; (j \neq i \neq k)$$

$$(H'_{21})_x = \sum_{c=1}^{t} \Big[t_{c1}Q_{x-1}(t_{c2},t_{c3}) + t_{c2}Q_{x-1}(t_{c1},t_{c3}) + t_{c3}Q_{x-1}(t_{c1},t_{c2}) \Big]$$

$$(H'_{31})_x = \sum_{c=1}^{t} Q_{x-1}(t_{c1},t_{c2},t_{c3}) \left[\sum_{i=4}^{e} t_{ci}\right]$$

As above, these enumerations may also be used for defined r-cycle subsets of R_k by disallowing in the calculation those e_i values (β_{ij}) which are characteristic of the particular r-cycle.

The α term is always $\binom{\alpha}{k+\Delta r-1-x}$ and represents the number of possible nonring cut combinations. The enumeration values for $|R_k|$ are not a function of the location of acyclic bond appendages to the ring system but only of their number. For an unsubstituted ring skeleton, $\alpha = 0$, and the α term only exists (and equals one) when $k = x + 1 - \Delta r$.

Definition of the operator Q_x for partition combination may be further amplified as follows (note that x must be at least as large as the number of parentheses):

$$Q_x(a) = \binom{a}{x}; Q_x(a,b) = \sum_{u=1}^{x-1} \binom{a}{u}\binom{b}{x-u}$$

$$Q_x(a,b,c) = \sum_{u=1}^{x-2}\sum_{v=1}^{x-2} \binom{a}{u}\binom{b}{v}\binom{c}{x-(u+v)}$$

in general:

$$Q_x(a_i)_{i=1}^{z+1} = Q_x(a_1,a_2,a_3, \ldots a_z,a_{z+1})$$

$$= \sum_{u_1=1}^{y_1} \sum_{u_2=1}^{y_2} \sum_{u_3=1}^{y_3} \cdots \sum_{u_z=1}^{y_z} \prod_{i=1}^{z} \binom{a_i}{u_i} \binom{a_{z+1}}{x - \Sigma u_i}$$

where

$$a_i \geq y_i \leq (x - z + 1) \text{ and } \left(\sum_{i=1}^{z} u_i + 1\right) \leq x \geq (z + 1)$$

Thus

$$Q_x(5, 4, 3) \text{ for } x = 3 \text{ is: } \binom{5}{1}\binom{4}{1}\binom{3}{1} = 60$$

$$\text{for } x = 4 \text{ is: } \binom{5}{1}\binom{4}{1}\binom{3}{2} + \binom{5}{1}\binom{4}{2}\binom{3}{1} + \binom{5}{2}\binom{4}{1}\binom{3}{1} = 270$$

$$\text{for } x = 5 \text{ is: } \binom{5}{1}\binom{4}{1}\binom{3}{3} + \binom{5}{1}\binom{4}{2}\binom{3}{2} + \binom{5}{1}\binom{4}{3}\binom{3}{1} + \binom{5}{2}\binom{4}{1}\binom{3}{2} + \binom{5}{2}\binom{4}{2}\binom{3}{1} + \binom{5}{3}\binom{4}{1}\binom{3}{1} = 590$$

It may be noted that Q_x is a commutative operator, i.e., $Q_x(a, b, c) = Q_x(a, c, b) = Q_x(b, a, c)$, etc., and that unit terms may be omitted, i.e., $Q_x(a, b, 1) = Q_{x-1}(a, b)$

$$Q_x(a, b, 1, 1) = Q_{x-2}(a, b)$$

$$Q_x(a, b, c, 1, 1, 1, 1) = Q_{x-4}(a, b, c)$$

Since many $e_i = 1$ for common skeletons and since other e_i (β_{ij}) values may be duplicated, some ordering of the combinations data vastly simplifies these enumerations. Let d_i represent the *unique* e_i values in the e list, in decreasing order, and n_i the corresponding numbers of times each d_i appears in the e list. Hence, if d = number of d_i values, i.e., unique e_i values, then

$$\sum_{i=1}^{d} n_i d_i = \sum_{i=1}^{e} e_i = (b_0 - \alpha) \text{ and } \sum_{i=1}^{d} n_i = e \text{ where } (e \geq d)$$

Combinations of two d_i values may also be listed as d_i' and n_i', referring to the numbers (n_i') of possible binary combinations of d_id_i or d_id_j (the pairs are listed as d_i'). Similarly, combinations of three d_i values (same or different) may be listed as d_i'' and n_i''. Owing to the frequency of unit terms and duplication of e values, it is rarely necessary to have lists of larger combinations. The number of n_i/d_i terms is of course d, while the number of binary combinations, n_i'/d_i', is $\binom{d+1}{2}$, and that of ternary n_i'/d_i' terms is $\binom{d+2}{3}$. The sum of the numbers of combinations n_i' or n_i'' is given by these expressions, each term in which is one n_i' or n_i'' value in the list.

$$\Sigma n_i' = \sum_i n_i \left(\frac{n_i - 1}{2} + \sum_{j=i+1} n_j\right)$$

$$\Sigma n_i'' = \sum_i \left[\binom{n_i}{3} + \sum_{j\neq i} n_i \binom{n_j}{2} + \sum_{j=i+1}\sum_{k=j+1} n_in_jn_k\right]$$

Table V. Steroid Skeleton Enumeration

e list	n_i	d_i	n_i'	d_i'	n_i''	d_i''		Q_x (5, 4)	Q_x (5, 4, 4)
$e_1 = 5 = \beta_{01}$	1	5	0	5, 5	0	5, 5, 5	$x = 2$	20	
$e_2 = 4 = \beta_{02}$	3	4	3	5, 4	0	5, 5, 4	3	70	80
$e_{3'} = 4 = \beta_{03}$	3	1	3	5, 1	0	5, 5, 1	4	120	400
$e_4 = 4 = \beta_{04}$			3	4, 4	3	5, 4, 4	5	125	980
$e_5 = 1 = \beta_{12}$			9	4, 1	9	5, 4, 1	6	84	1520
$e_6 = 1 = \beta_{23}$			3	1, 1	3	5, 1, 1	7	36	1636
$e_7 = 1 = \beta_{34}$	$\Sigma n_i' = 21$				1	4, 4, 4	8	9	1268
					9	4, 4, 1	9	1	713
					9	4, 1, 1	10		286
					1	1, 1, 1	11		78
			$\Sigma n_i'' = 35$				12		13
							13		1

In this way, combinations of Q_x terms (especially in the general cutting enumerations N'_{20}, N'_{30}, N'_{31}) may be grouped as $\Sigma_{ij}Q_x(e_i,e_j) = \Sigma_i n_i'Q_x(d_i')$ and $\Sigma_{ijk}Q_x(e_i,e_j,e_k) = \Sigma_i n_i''Q_x(d_i'')$ and the possible binary and ternary sums for the unique combinations of e values assembled in advance, i.e., $Q_x(d_i,d_j)$ and $Q_x(d_i,d_j,d_k)$ as a f(x). These lists are illustrated for the steroid skeleton (face graph, Figure 2) in Table V. Hence

$$(N'_{20})_3 = \Sigma_{ij}Q_3(e_i,e_j) = \Sigma_k n_i'Q_3(d_i') = 0 \times Q_3(5, 5) + 3Q_3(5, 4) + 3Q_3(5, 1) + 3Q_3(4, 4) + 9Q_3(4, 1) + 3Q_3(1, 1) = (0 \times 100) + (3 \times 70) + (3 \times 10) + (3 \times 48) + (9 \times 6) + (3 \times 0) = 438$$

The subgraph matrixes required for other N' terms for the steroid are shown in Chart I.

Chart I

Triangles (t_{ci}): (N'_{21} and N'_{31})

$$\begin{array}{r|ccc:cccc} & \multicolumn{3}{c}{\Delta} & \multicolumn{4}{c}{\text{others}} \\ c = 1 & e_1 & e_2 & e_5 & e_3 & e_4 & e_6 & e_7 \\ 2 & e_2 & e_3 & e_6 & e_1 & e_4 & e_5 & e_7 \\ 3 & e_3 & e_4 & e_7 & e_1 & e_2 & e_5 & e_6 \end{array}$$

$(t = 3)$

$$\text{or}\quad \begin{array}{r|ccc:cccc} c = 1 & 5 & 4 & 1 & 4 & 4 & 1 & 1 \\ 2 & 4 & 4 & 1 & 5 & 4 & 1 & 1 \\ 3 & 4 & 4 & 1 & 5 & 4 & 1 & 1 \end{array}$$

Squares (s_{ci}): (N'_{31} and N'_{32})

$$\begin{array}{r|ccccc} c = 1 & e_1 & e_3 & e_6 & e_5 & e_2 \\ 2 & e_2 & e_4 & e_7 & e_6 & e_3 \end{array}$$

$(s = 2)$

$$\text{or}\quad \begin{array}{r|ccccc} c = 1 & 5 & 4 & 1 & 1 & 4 \\ 2 & 4 & 4 & 1 & 1 & 4 \end{array}$$

Double-crossed squares (s'_{ci}): none; hence $(N'_{33})_x = 0$

$$(N'_{21})_5 = \sum_{c=1}^{2} Q_5(t_{c1}, t_{c2}, t_{c3}) = Q_5(5, 4, 1) + Q_5(4, 4, 1) + Q_5(4, 4, 1) = Q_4(5, 4) + 2Q_4(4, 4) = 120 + 2(68) = 256$$

References and Notes

(1) Previous papers in the series: (a) J. B. Hendrickson, *J. Am. Chem. Soc.*, **93**, 6647 (1971); (b) *ibid.*, **93**, 6654 (1971).

(2) The first approaches to abstract discussion of synthesis were made by E. J. Corey, *Pure Appl. Chem.*, **14**, 19 (1967); *Q. Rev., Chem. Soc.*, **25**, 455 (1971); E. J. Corey and W. T. Wipke, *Science*, **166**, 178 (1969).

(3) Indirect routes include one or more skeletal cleavages and include skeletal rearrangements as formal construction + cleavage. Such routes are important to many actual successful syntheses, about 30% of the syntheses in a recent survey[4] containing C–C cleavages, although rarely more than one per synthesis.

(4) Excellent summaries of syntheses may be found in N. Ayand, J. S. Bindra, and S. Ranganathan, "Art in Organic Synthesis", Holden-Day, San Francisco, Calif., 1970.

(5) To divide an acyclic molecule (of n carbons) in two requires only one cut at any of (n − 1) bonds, but a monocycle requires two cuts. A bicyclic molecule requires three cuts if the two resultant pieces are to be acyclic, and not any three will serve.

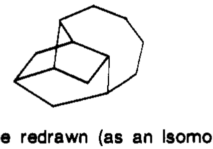
(6) The skeleton may include other atoms if desired, particularly nitrogen in rings, and the graphical analysis which follows is the same.

(7) A good introduction is found in F. Harary, "Graph Theory", Addison-Wesley, Boston, Mass., 1969.

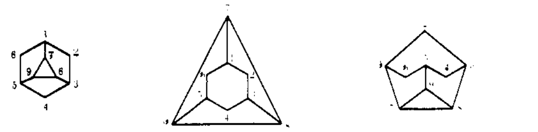
(8) The adjacency matrix of a chemical structure was first described by L. Spialter, *J. Am. Chem. Soc.*, **85**, 2012 (1963).

(9) (a) J. T. Welch, Jr., *J. Assoc. Comput. Mach.*, **13**, 205 (1966); N. E. Gibbs, *ibid.*, **16**, 564 (1969); C. C. Gotlieb and D. S. Cornell, *Commun. ACM*, **10**, 760 (1967); K. Paton, *ibid.*, **12**, 514 (1969); (b) R. Fugmann, U. Dölling, and H. Nickelsen, *Angew. Chem., Int. Ed. Engl.*, **6**, 723 (1967); (c) M. Plotkin, *J. Chem. Doc.*, **11**, 60 (1971); (d) E. J. Corey and G. A. Petersson, *J. Am, Chem. Soc.*, **94**, 460 (1972); (e) acyclic structures enumerated by J. Lederbeng, G. L. Sutherland, B. G. Buchanan, E. A. Feigenbaum, A, V. Robertson, A. M. Duffield, and C. Djerassi, *ibid.*, **91**, 2973, 2977 (1969).

(10) Highly articulated polycyclic molecules may serve to illustrate the usual possibility of plane graph representation: copaene, twistane, longifolene (ref 4, pp 117, 347, 233, respectively), as well as the tricyclane in Figure 1 or congressane (Figure 2). The structure below cannot be represented by a plane graph:



(11) Any plane graph can be redrawn (as an isomorphic graph) with any specified monocycle or face becoming the exterior face. The formal procedure is to place the graph on a sphere with the specified face over the north pole, to stretch and spread the graph south over the sphere without allowing any line to cross the north pole, and then to project the graph onto a plane tangent to the south pole. Three isomorphs of one structure are illustrated below; they are all equivalent representations of the same molecular skeleton. Ten isomorphs of the steroid skeleton are possible since the skeleton has ten rings, any of which can bound the exterior face.



(12) The 18-methyl of the steroids, for example, can be drawn projecting into the exterior face (β_{00} = 1) into the faces of rings 3 or 4 (β_{33} or β_{44} = 1).

(13) C_{r_0} = 1 only if $\bar{\text{F}}$ is fully connected; if it is not connected, then C_{r_0} = 0, and the largest monocycle is defined by the largest connected subgraph of r points in $\bar{\text{F}}$ (C_r = 1). $\bar{\text{F}}$ is not connected if the rings in the skeleton are not fused (i.e., no bonds in common), as in spiro compounds and separated rings like biphenyl. The fourth example in Figure 3 illustrates this. The complete face graph, F, however, is always connected.

(14) The six four-point connected graphs are (ref 7, p 215): ⊔; ∠; ⧅; □; ⧄; ⊠ ≡ ▲ (latter two are designated as *crossed squares*).

(15) The line graph L represents a reduction of $\bar{\text{F}}$ such that one point (ij) in L corresponds to two adjacent points in $\bar{\text{F}}$ and a line (ijk) in L is thus three points in $\bar{\text{F}}$. Therefore, counting the lines in L is equivalent to counting pairs of adjacent lines in $\bar{\text{F}}$ as in enumerating C_3. The number of lines in L is $l = \Sigma\binom{d_i}{2}$ and then $C_3 = l - 2\Delta$. This reduction procedure is analogous to that used by D. Cartwright and T. C. Gleason, *Psychometrika*, **31**, 179 (1966).

(16) There is one exception to the enumeration formulas of Tables I and II. If a skeleton contains a completely surrounded ring (β_{oi} = 0), the monocycle described by fusing all the rings which surround the central ring is an annulus, not a ring, as may be seen in the first structure of ref 10, in which disconnecting bonds 17, 36, and 59 leave such an improper ring. Hence, in such cases, one must be subtracted from C_r if r rings surround the central one. Hence for the tetracycle in ref 10, C_3 = 3 rather than 4, and for cubane, C_4 = 4 not 5. The second example of Figure 3 also illustrates this reduction of C_3.

(17) It would also be desirable to sort out synthons by size. This requires enumerating all the connected n-point subgraphs[7,14] in the skeleton, and this in turn is a very complex procedure owing to the number of forms such subgraphs can take, i.e., the number of possible structural isomers of C_n synthon skeletons. Thus there is a problem of all partitions of n_0 atoms into k parts just to define the possible categories to enumerate. The problem is parallel to that described for partitioning the face graph into categories in Table II but much more complex for the larger n_0-point graph of the full skeleton.

(18) Such a prestruct is a "tree" in graph theory, i.e., a graph without cycles. A graph with cycles contains a particular number of spanning trees linking all its points; this set of spanning trees for the product skeleton would be O_1 on the grid in Figure 5.

(19) The grid considers only those skeletal atoms in a prestruct which ultimately are incorporated into the final target so that the final target is necessarily T_1, with only one component. Examination of syntheses in ref 3 shows that, in most instances, synthon skeletal atoms cleaved *and lost* are only C_1 units usually lost as CO_2 or an equivalent. They are ignored in this treatment.

(20) F. Harary,[7] private communication.

(21) The procedure is of course amenable to simple Fortran programming, but this has not been done here.

(22) Such determinants, cofactors of M, are all equal to each other.

(23) As an extension, removal of any n linked atoms (rows and columns)

gives all the spanning trees incorporating those atoms with their original links intact. If two linked atoms are removed from M, one enumerates all acyclic precursors containing the bond which links them. The normal procedure for evaluating all $|O_1|$ removes any one atom, which is tantamount to counting all acyclic precursors containing that one atom, i.e., all acyclic precursors. If all atoms are removed from matrix M, the evaluation is taken as 1.

(24) As noted in ref 19, cleavages which remove carbons not ultimately incorporated in the skeleton (usually as CO_2) are not included in the grid; they may be regarded as functionalizing (or defunctionalizing) reactions. Thus the acceptable bonds indicated for cleavages in this discussion constitute a new ring, and their cleavage is a horizontal ring-opening line on the grid ($\Delta r = -1$; $\Delta k = 0$).

(25) Nearly 30 of the 100 syntheses in ref 4 exhibit such indirect routes in which a skeletal bond, not in the product but useful at an early stage of the sequence, is ultimately cleaved. As examples, in Corey's caryophyllene (p 70), a large ring is formed by cleaving a more accessible bicycle; in the Syntex cecropia horomone synthesis (p 79), two cleavages of a bicycle to an acyclic skeleton are used to create stereochemical control; in Johnson's progesterone (p 266), two ring sizes are changed at the same time by cleavage and recyclization (cf. Figure 6).

(26) It would be misleading, for example, to consider the Barbier–Wieland degradation as an affixation of two six-carbon skeletal synthons followed by cleavage of a 13-carbon unit. The present conception sees it as merely a functionalization of R–CH_2COOR′ → R–COOH in which the only skeletal carbons are R–C, i.e., those appearing in the final product.

(27) The idea was proposed for picrotoxin *bio*synthesis years ago by H. Conroy.

Systematic Synthesis Design. IV. Numerical Codification of Construction Reactions

James B. Hendrickson

Contribution from the Edison Chemistry Laboratories, Brandeis University, Waltham, Massachusetts 02154. *Received January 22, 1974*

Abstract: A simple but rigorous system of codification for construction reactions is developed from structural fundamentals, free of mechanistic preconception. The system allows all constructions to be represented with a numerical representation of the involved functionality and skeletal requirements of substrate and product and their interrelation. The scheme is valuable in systematic searching for synthetic routes as well as in cataloging construction reactions and developing new ones.

An essential requirement for the development of systematic synthesis design must be a simple but rigorous numerical codification of the reactions used. Such a system must be free from prejudice about present capabilities or reaction yields. This paper develops such a system for construction reactions from the numerical characterization of structure previously presented.[1]

That constructions are the central reactions of synthesis may be seen from consideration of the ideal synthesis. The ideal synthesis creates a complex skeleton from simpler starting materials[2] and so must link several such synthon molecules via construction reactions. Ideally, the synthesis would start from available small molecules so functionalized as to allow constructions linking them together directly, in a sequence only of successive construction reactions involving no intermediary refunctionalizations, and leading directly to the structure of the target, not only its skeleton but also its correctly placed functionality. If available, such a synthesis would be the most economical, and it would contain only construction reactions. The previous paper in this issue[3] develops mathematically the enumeration of the possible modes of construction of target skeletons. Here the actual chemistry which can be used to effect these constructions will be codified to define all possibilities in terms of their related substrate and product functionalities. Restrictive preconceptions about reaction mechanism are avoided in this development in favor of the more neutral and general conception of the *net structural change* occurring in any reaction.

The net structural change at any single carbon site was previously characterized[1] in terms of four kinds of attachment to that carbon: H for hydrogen, R for σ bond to carbon, Π for π bond to carbon, Z for any bond to heteroatom. In any reaction, the change from one attachment to another was characterized by two letters, the first showing the bond made, the second showing that broken. Thus, of the 16 possible reactions so characterized, the construction reactions are RH, RZ, and RΠ,[4] with respect to either one of the two carbons forming the carbon–carbon σ bond.

A construction requires two partners, the linking carbon of each being characterized by RH, RZ, or RΠ, and these show oxidation state changes of $\Delta x = +1, -1$, and 0, respectively.[1] The RΠ construction necessarily changes the character of a least one other carbon as well, the other carbon of the Π bond undergoing addition, and the oxidation state changes of all must be added to find the net change (Δx) for RΠ constructions. Thus the net change in RΠ constructions is always $\Delta x = \pm 1$. (For C═C → R—C—C—Z, RΠ·ZΠ, $\Delta x = +1$ but, for C═C—C—Z → R—C—C═C, RΠ·$\Pi\Pi$·ΠZ, $\Delta x = -1$). The overall oxidation state change (the sum of both involved components) can be either oxidative or reductive, or isohypsic,[1] with $\Sigma\Delta x = +2, -2$, or 0, respectively. Oxidative and reductive couplings, however, are rarely useful in synthesis since they are only effective for creating symmetrical dimers in intermolecular reactions (although they can unite dissimilar functionalities in cyclizations). The present treatment largely focuses on isohypsic constructions of one oxidative and one reductive partner. Each partner in a construction will be categorized by reaction type as RH, RZ, or RΠ, depending on the change at the carbon forming the construction link.

The numerical characterization[1] concerns the numbers of each kind of attachment to a single carbon, as summarized in Figure 1. The skeletal value (σ) shows the number of σ bonds to other carbons, i.e., σ = 0–4, and the functional value (f) shows the functionality level at that carbon site, f = 0–4. Since $f = \Pi + Z$, the sum of functional π bonds to adjacent carbon and the number (Z) of heteroatom bonds, a distinction is made by placing one or two bars over an f value to denote the number (Π) of π bonds to adjacent carbon. Thus an enol ether carbon is $f = \bar{2}$, the same functional level as the parent ketone (f = 2), and a chloroacetylene carbon is $f = \bar{\bar{3}}$, while a dichlorovinyl carbon is $f = \bar{3}$, both at the functional level of carboxyl, f = 3.