Molecular Complexity: A Simplified Formula Adapted to Individual Atoms

JAMES B. HENDRICKSON,* PING HUANG, and A. GLENN TOCZKO

Edison Chemical Laboratories, Brandeis University, Waltham, Massachusetts 02254

Received March 21, 1985

The Bertz formula for calculating molecular complexity is a sum of bond connectivities. This is converted to a simpler form based on the number of hydrogens attached to each atom. Also, the calculation of symmetry terms is derived from simple equations applied to atoms of the same equivalence class. We outline here a simple program (CPXCAL) in FORTRAN for microcomputer which yields the same values as Bertz's formulas. It is applied to a number of examples as well as to all four-, five-, and six-carbon skeletons to evaluate its validity. The formulas are well adapted for use in locating synthesis pathways with least molecular complexity in our synthesis program.

The concept of molecular complexity has been advanced by Bertz[1] and shown to be relevant in the design of syntheses through minimizing the sum of molecular complexities of the synthetic intermediates. His formula is derived from information theory and is based on the sum of bond connectivities on non-hydrogen atoms as well as on the variety of kinds of these atoms. The concept represents a useful test of the relative simplicity of different synthetic pathways to a target molecule and, as such, interests us for use in ranking the different syntheses produced on our computer program (SYNGEN)[2] for generating syntheses.

Bertz's measure of molecular complexity (C) is a sum of two parts (eq 1): the first and major term, C_η, measures skeletal complexity as a function of bond connectivities (η); the second term, C_E, is a function of the diversity of elements, or kinds of atoms, present. Each of these terms also is composed of two parts: first, an overall complexity term; and second, a symmetry term subtracted from it so as to reduce the complexity to the extent that symmetry is present. The formulas are shown as eq 1–3 (lg is used for $\log_2$).

$$C = C_\eta + C_E \tag{1}$$

$$C_\eta = 2\eta \lg \eta - \sum_i \eta_i \lg \eta_i \tag{2}$$

$$C_E = E \lg E - \sum_j E_j \lg E_j \tag{3}$$

In the elements term (eq 3), E is the total number of non-hydrogen atoms and E_j is the number of type j. Thus, the second term represents a concept of the "symmetry" of like atoms: if all atoms are the same kind, $C_E = 0$; if there are atoms of many kinds, the first term for C_E will be much larger than the second and the overall complexity will incorporate this measure of atomic diversity.

The central feature of the complexity calculation, in eq 2, is the measure η, which represents the sum of all *bond connectivities*, i.e., the number of pairs of bonds connected to each other. In chemical terms, it is the number of ways a linear three-atom (e.g., propane) skeleton[3] can be extracted from the whole molecular skeleton. In the second, or symmetry, term η_i represents the number of symmetrically identical bond pairs of type i.

It is both visually simpler and more adaptable to our system[2] of molecular description to compute the values of η from the characteristics of individual atoms instead of connected bond pairs. Any pair of connected bonds intersects at a particular atom. In Figure 1 the bond connectivities (η) around both saturated and unsaturated atoms are illustrated; the only bonds counted are those to non-hydrogen atoms. For the saturated atoms, it will be observed that η at that atom is simply a function of the number of hydrogens (h) on that atom,[4] as in eq 4.

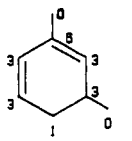
$$\eta = \tfrac{1}{2}(4 - h)(3 - h) \tag{4}$$

For unsaturated atoms the formula is the same, i.e., the sum of η so calculated at each atom, except that the π-bond is counted twice and so must be separately discounted. Hence, the formula for η for a whole molecule is given in eq 5, in which

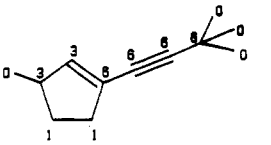
$$\eta = \tfrac{1}{2}\sum_i (4 - h_i)(3 - h_i) - D - 3T \tag{5}$$

D is the number of double bonds, T is the number of triple bonds, and i refers to the ith atom. This equation represents a much simpler way of calculating η for a molecule by just counting the hydrogens (and unshared electron pairs[4]) on each atom and the number of double and triple bonds in the molecule.

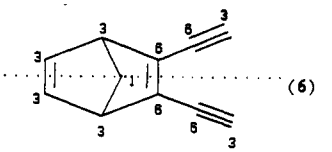
These values for η are easily summed by hand from eq 5 as illustrated in the three examples of eq 6.



$\eta = 19 - 2 = 17$



$\eta = 32 - 1 - 3(1) = 28$



(6)

$\eta = 43 - 2 - 3(2) = 35$

In a similar fashion, the second, or symmetry, term of eq 2 may be calculated from the symmetrically equivalent atoms, i.e., those atoms which are automorphic or of the same equivalence class.[5] In the third example of eq 6, there is a symmetry element (dotted line) through the molecule, creating five pairs of equivalent atoms with η_i of 3, 3, 6, 6, and 3 each, respectively.

We can operate the second terms for eq 2 in a general way by collecting (as in Figure 1) all possible symmetry types and deriving the appropriate symmetry equation for each type, so that symmetrical atoms can then be recognized by type and their symmetry terms summed. Thus, any bond-pair con-

0095-2338/87/1627-0063$01.50/0

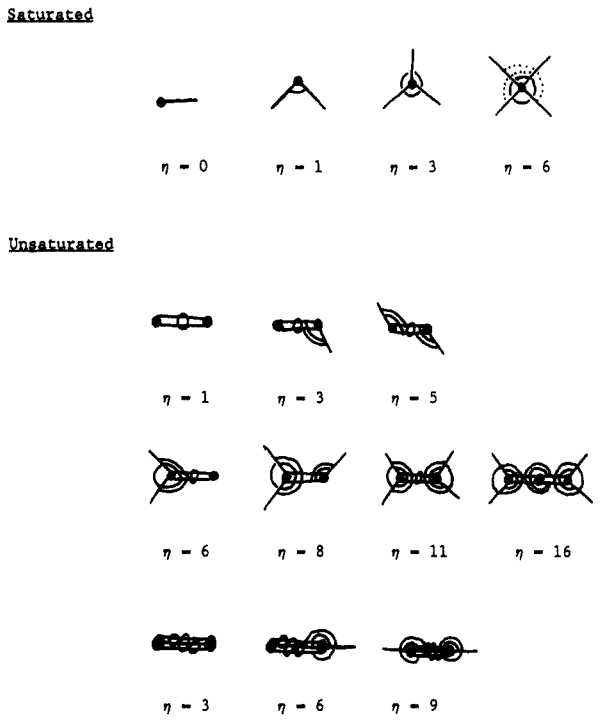


Figure 1. Bond connectivities at saturated and unsaturated atoms.

Table I. Numerical Values[a]

N	N lg N	N	N lg N	N	N lg N
1	0	18	75.1	35	179.5
2	2	19	80.7	36	186.1
3	4.8	20	86.4	37	192.7
4	8	21	92.2	38	199.4
5	11.6	22	98.1	39	206.1
6	15.5	23	104.0	40	212.9
7	19.7	24	110.0	41	219.7
8	24	25	116.1	42	226.5
9	28.5	26	122.2	43	233.3
10	33.2	27	128.4	44	240.2
11	38.1	28	134.6	45	247.1
12	43.0	29	140.9	46	254.1
13	48.1	30	147.2	47	261.1
14	53.3	31	153.6	48	268.1
15	58.6	32	160	49	275.1
16	64	33	166.5	50	282.2
17	69.5	34	173.0		

[a] lg = $\log_2$.

nectivity centers on an atom A, which in turn is connected to atoms B, C, D, etc. If there are N such symmetrically equivalent atoms A, the attached atoms must also be the same (B, C, D,...) on each other equivalent atom A. The possible combinations of such equivalent atom sets are shown in Figure 2, and the symmetry term (S_k) is shown for each, a function of the number (N) of such equivalent sets in the molecule.

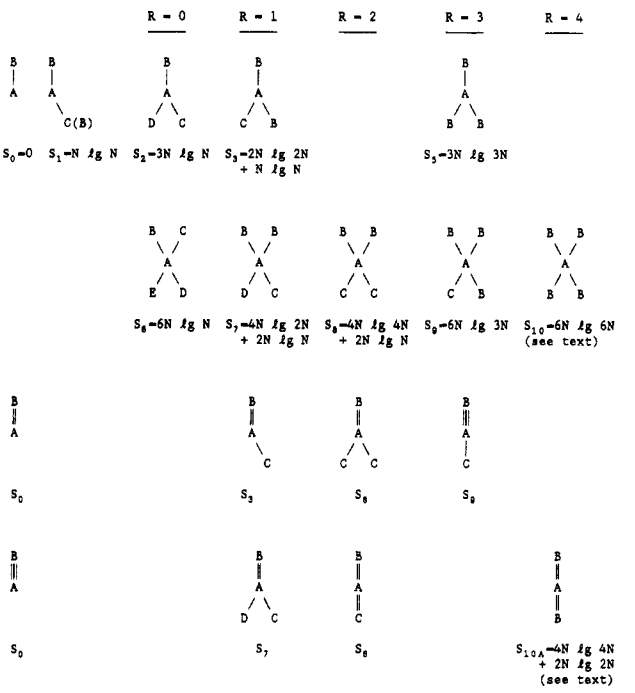


Figure 2. Symmetry classes at central atom A. (Molecule contains N such symmetry equivalent units; lg = $\log_2$.)

These symmetry terms are derived from the duplication of bond connectivities in the atom assemblies shown, understood as themselves duplicated N times in the molecule. These symmetry terms (S_k) are then added for all such equivalent sets and subtracted, as the second term in eq 2, from the overall complexity (first term in eq 2); in other words, eq 2 becomes eq 7.

$$C_\eta = 2\eta \lg \eta - \sum_i (S_k)_i \tag{7}$$

The value of the symmetry type, k, for a given set of N equivalent atoms A can then also be derived from a simple formula (eq 8) based on the characteristics of atom A. Such

$$k = (3 - h)(2 - h) + \mathrm{R} \tag{8}$$

an equation is needed for the computer to identify the symmetry terms. Here the term R designates the number of repetitions of bonds to identical attached atoms on atom A, either σ- or π-bonds. Thus, $R = 0$ if all bonds from A are to different atoms, $R = 1$ if two bonds go to the same or to two identical attached atoms (B=ACD or AB_2CD), $R = 2$ if there are two repetitions (AB_2C_2 or B=AC_2), etc. Values of R are shown for each symmetry type in Figure 2. The formula (eq 8) is simply an empirical derivation designed to yield an increasing series for increasing symmetry. It does not serve for atoms A with only one attached other atom, for which there is no symmetry term: $k = 0$ and $S_0 = 0$. For atoms A with

Table II. Numerical Values of Symmetry Terms (S_k)[a]

	N							
	1	2	3	4	5	6	7	8
$S_1 = N \lg N$	0	2	4.8	8	11.6	15.5	19.7	24
$S_2 = 3N \lg N$	0	6	14.4	24	34.8	46.5	59.1	72
$S_3 = 2N \lg 2N + N \lg N$	2	10	20.3	32	44.8	58.5	73	88
$S_5 = 3N \lg 3N$	4.8	15.5	28.5	43	58.6	75.1	92.2	110
$S_6 = 6N \lg N$	0	12	28.8	48	69.6	93	118.2	144
$S_7 = 4N \lg 2N + 2N \lg N$	4	20	40.6	64	89.6	117	146	176
$S_8 = 4N \lg 4N + 2N \lg N$	8	28	52.6	80	109.6	141	174	208
$S_9 = 6N \lg 3N$	9.6	31	57	86	117.2	150.2	184.4	220
$S_{10} = 6N \lg 6N$	15.5	43	75.1	110	147.2	186.1	226.5	268.1
$S_{10A} = 4N \lg 4N + 2N \lg 2N$	10	32	58.5	88	119.6	153	187.9	224
$S_{10B} = 6N \lg 2N$	6	24	46.5	72	99.6	129	159.9	192

[a] lg ≡ $\log_2$.

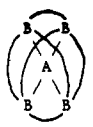
only two attached atoms and no π-bonds (ABC or AB_2; alternatively, $h = 2$), we designate $k = 1$ as shown for Figure 2 (the repetition in AB_2 is ignored).

For convenience in simple hand calculations, values of N lg N are listed in Table I and values for the symmetry terms, S_k, are listed in Table II. Although the symmetry values are described for N equivalent atoms A, in many cases there are equivalent bond connectivities even for $N = 1$. This can be seen, for example, in S_3 (for which there are two equivalent B–A–C connectivities: a symmetry term of 2 lg 2), which is satisfied by the formula for $S_3 = 2N$ lg $2N + N$ lg N when $N = 1$. In fact, only S_1, S_2, and S_6 give $S_k = 0$ for $N = 1$. Since even single atoms ($N = 1$) can thus exhibit duplicated bond connectivities for the symmetry term, the computer simply calculates the appropriate S_k for every equivalence class of atom in the molecule, i.e., for all atoms. Those with no symmetry (S_1, S_2, and S_6) then contribute nothing to the subtracted symmetry term of eq 7.

Although they are very rare in common molecules, there are a few bond-symmetry situations that further expand on these terms when symmetrical bridges link the four attached atoms on tetravalent atoms A. The program must go further then to identify ring symmetry in these cases. Briefly, symmetrical bridges between B and C in S_8 (9) become S_7, although B–B and/or C–C bridges only remain S_8. In the case of S_{10}, the formula is correct for no bridges or for six equivalent bridges connecting atoms B all ways (10). Lesser symmetry characterizes other bridging. Of the six possible B–B links in S_{10}, if two are of one kind and four another, the formula (S_{10A}) is less (11), and if three kinds of pairs exist, the formula is S_{10B} (12). Allenes of the same type (B=A=B) also require

(9)

S_7

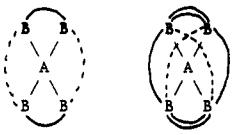


(10)

$S_{10} = 6N$ lg $6N$

(11)

$S_{10A} = 4N$ lg $4N + 2N$ lg $2N$



(12)

$S_{10B} = 6N$ lg $2N$

term S_{10A}. In virtually all real cases, these do not appear, but in cases identified as S_{10}, the computer must examine the equivalency of rings about atom A as well.

In the process of converting from a basis of bond connectivities of one of atom attributes, we find that π-bonds are counted twice, once for each connected atom. This is easily corrected with subtractive terms D and T in eq 5 for η itself, but the same duplication must be corrected in the symmetry terms which involve multiple bonds. These terms for double bonds are all shown in Figure 2; from all these symmetry terms may be separated an N lg N term that contains the duplication.

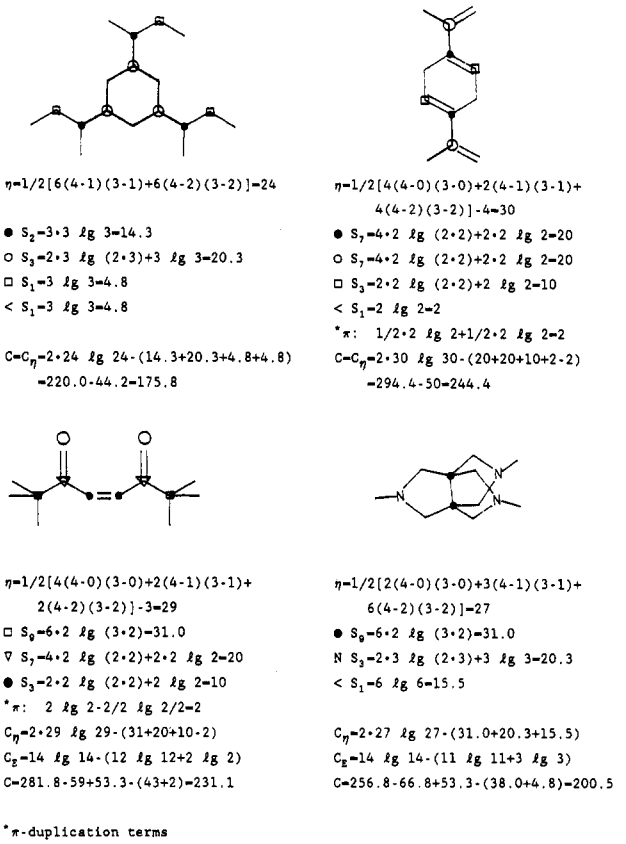
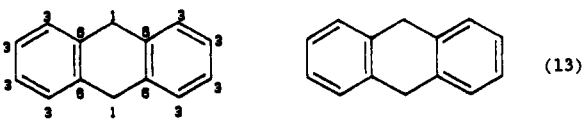


Figure 3. Examples of molecular complexity calculation.

Hence, for cases of double bonds A=B, subtraction of $^1/_2N$ lg N for each atom, A and B, serves to remove the duplication. This is only so if both atoms have $S_k \neq 0$. Thus in cases of $>A{=}BH_2$, since the B atom contributes no symmetry term ($S_k = 0$), no subtraction of $^1/_2N$ lg N is necessary. For triple bonds $-A{\equiv}B-$, the "triplication" term subtracted is $^3/_2N$ lg $3N$ (unless again one atom is $\equiv$BH). In cases of A=A, double bonds with both atoms equivalent, the subtracted term will be N lg $N - N/2$ lg $N/2$) and for triple bonds, A$\equiv$A, it is ($3N$ lg $3N - {}^3/_2N$ lg $^3/_2N$). For unsymmetrical allenes, C=A=B, one subtracts two $^1/_2N$ lg N terms for A and one each for B and C, but in symmetrical ones, B=A=B, one subtracts for atom A half the second term in S_{10A} (eq 11), i.e., $-^1/_2(2N$ lg $2N)$.

Aromatic molecules are, however, much more simply handled by atom attributes rather than bond connectivities since the program treats each π-bonded atom the same way without distinguishing its partner. In this way, all Kekulé forms give the same result. This may be illustrated with the two forms of dihydroanthracene (13), which would give different sym-

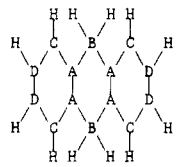


(13)

$\eta = 50 - 6 = 44$

metry terms on strict bond connectivity treatment but which give the same result in the present atom treatment since it registers the presence of a π-bond on each *atom* without noting the atom to which it is bonded.

The calculation of dihydroanthracene is delineated in (14) which shows four equivalence classes of atoms (A–D), of which all but B bear double bonds. The four S_k equations and their double-bond corrections for eq 7 are all shown, making it clear that the particular Kekulé forms are not perceived. Other



(14)

		π-duplication
A:	$S_7 = 4{\cdot}4\ \ell g\ 2{\cdot}4 + 2{\cdot}4\ \ell g\ 4$	$-(4\ \ell g\ 4 - 2\ \ell g\ 2)$
B:	$S_1 = 2\ \ell g\ 2$	-0
C:	$S_3 = 2{\cdot}4\ \ell g\ 2{\cdot}4 + 4\ \ell g\ 4$	$-2\ \ell g\ 4$
D:	$S_3 = 2{\cdot}4\ \ell g\ 2{\cdot}4 + 4\ \ell g\ 4$	$-2\ \ell g\ 4$

$$C_\eta = 2{\cdot}44\ \ell g\ 44 - [32\ \ell g\ 8 + 8\ \ell g\ 4 + 4\ \ell g\ 2] = 480 - 116 = 364$$

examples are treated in Figure 3; the symmetry terms for each atom (or equivalence class of N atoms) are calculated from the S_k equations in Figure 2.

There is one other subtle (and very rare) instance in which the treatment by atoms creates an error. These are cases in which all atoms in a set are equivalent but their connecting bonds are not. This is exemplified by prismane. Here, a

prismane

treatment by atoms shows all atoms to be identical, but the bond connectivity approach would recognize the distinction between the six 3-ring bonds and the three 4-ring bonds. The maximization program[5] that we use to define equivalence classes of atoms sees all six atoms as equivalent, but if one is fixed, the others are no longer equivalent. This secondary equivalence class now allows the differences to be observed in the S_k terms for symmetry and thus affords the same overall value of complexity as that derived by the Bertz method.

The concept of molecular complexity has heretofore been an intangible one, seldom addressed in any quantitative terms by chemists. A molecule can be treated as a graph expressing atom connectivity, but even in the quantitative treatments of graph theory, there has not been a satisfactory index of complexity. Bertz makes a strong case[1] for the validity of his choice of mathematical expression in that (a) it is parallel to the similar analysis of information theory, (b) it affords different values for different molecules, and (c) it correlates well with various kinds of "simplicity" as expressed by overall yields in synthesis.[1d]

We have written a simple FORTRAN program, named CPXCAL, to calculate the complexity of any molecule using eq 1, 3, 5, and 7 as well as the numerical values from Tables I and II. This allows a facile examination of many molecules and so a broad, systematic examination of complexities in a variety of molecular families, or indeed, of graphs generally. Thus, the family of all possible saturated molecules of 4–6 carbons consists of 78 graphs on 6 points (atoms) with maximum degree (valence) of four, 21 graphs from 5 points, and 6 graphs from 4 points. Examination of this family of 105 molecules, or graphs, shows a rather small incidence of duplication. There are only four pairs of exact duplicates in the 105 cases, shown in Figure 4. There are several other pairings with fortuitously identical values of C which, however, differ in number of atoms or rings or in the value of η or C_η. These are also appended in Figure 4. A summary of the ranges of complexities for the 105 graphs is appended in Table III.

It can also be seen that the use of η alone is a relatively indiscriminate measure of molecular complexity. In each of

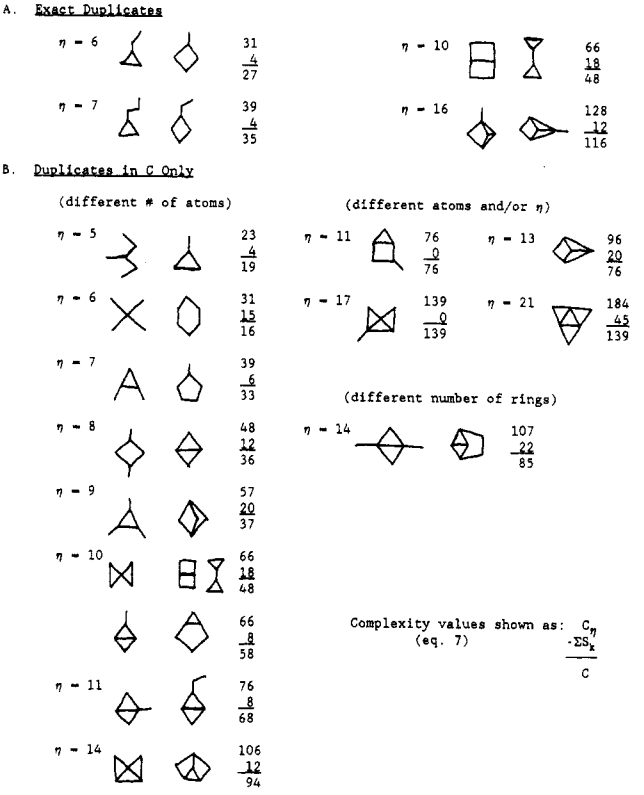


Figure 4. Duplicate complexities.

Table III. Summary of Molecular Complexities for n = 4–6 Atoms[a]

b	r	no. of cases	η	$\sum S_k$	C_η
		n = 4 atoms			
6	3	1	12	43	43
5	2	1	8	12	36
4	1	2	4–5	4–8	8–19
3	0	2	2–3	2–5	2–5
		tot = 6			
		n = 5 Atoms			
10	6	1	30	147	147
9	5	1	24	68	152
8	4	2	18–19	30–42	108–131
7	3	4	13–15	12–36	76–95
6	2	5	9–11	8–20	37–68
5	1	5	5–8	4–12	12–38
4	0	3	3–6	2–16	8–16
		tot = 21			
		n = 6 Atoms			
12	7	1	36	186	186
11	6	3	30–31	50–84	220–244
10	5	8	24–27	20–66	156–212
9	4	14	18–22	0–75	75–178
8	3	18	14–18	0–51	73–139
7	2	17	10–14	0–22	48–96
6	1	12	6–10	0–20	16–62
5	0	5	4–7	2–10	12–30
		tot = 78			

[a] b = number of bonds; r = number of rings; $b = n + r - 1$.

the three groups the maximal, fully connected case has the full symmetry term equal to η lg η, as do the simple symmetrical cycloalkanes in each group, which accordingly also have the least C_η for any of their isomeric monocycles. As Bertz points out, this fact leads to the factor of 2 used in the total complexity term, 2η lg η, in order that C_η shall not be zero. As a result, the total complexity term in general considerably outweighs the subtraction symmetry terms ($\sum S_k$). If it is deemed more appropriate to allow the symmetry terms more relative weight, the factor in the total complexity could be

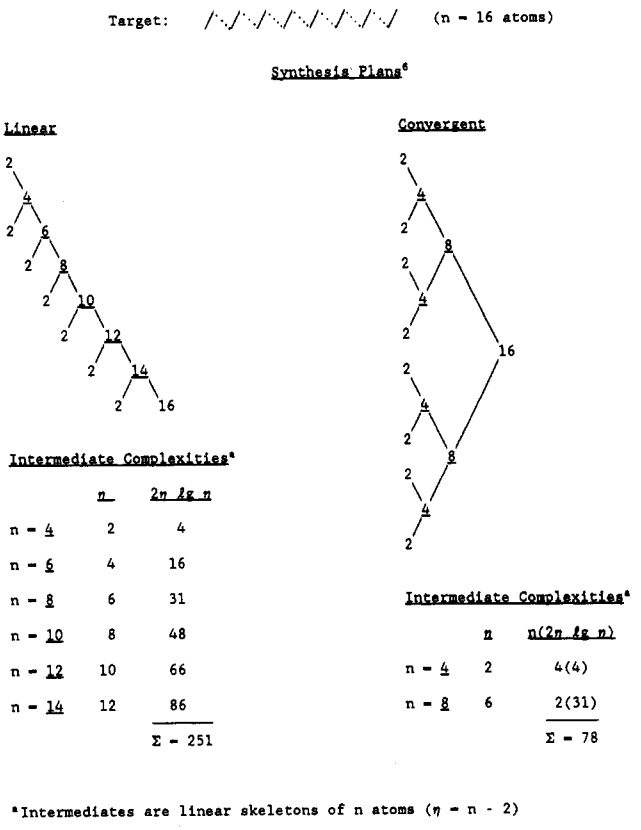


Figure 5. Complexity comparison of linear and convergent syntheses.

lessened, i.e., $Q\eta \lg \eta$, with $2 > Q > 1$.

The application of complexity calculations to assess synthesis efficiency has been illustrated by Bertz, who has used the sum of complexities of intermediates as a comparative measure for the efficiency of synthesis.[1] In comparisons of syntheses for the same target, however, this will generally simply prefer the synthesis with the fewest steps, hence fewest intermediates to sum. However, in a comparison of two syntheses with the same starting fragments and the same number of steps, this measure should show a preference for one. This amounts to two different *orders* of assembly of the same units and is characteristic of the difference between a convergent and a linear synthesis.[6] If we apply this method to the assembly of a linear target skeleton of 16 atoms from 8 2-carbon starting units, the results for the two orders are shown in Figure 5 (functional groups and the minor symmetry terms are not included). The results show a clear preference for the convergent order, as expected from other considerations.[6] This procedure points up another interesting conclusion about synthesis. In a real convergent synthesis with added refunctionalization reactions, the total complexity of intermediates will be much less if these added reactions precede the final coupling of the two skeletal halves. The conclusion for synthesis planning is clear: the final coupling of intermediates should come near the end rather than the beginning of a convergent synthetic route. In more general terms, any step that exhibits a large increase in complexity is more efficiently positioned near the end rather than the beginning of a synthetic sequence.

The calculation of molecular complexity is rendered much easier to carry out by hand when the variations in eq 1, 3, 5, and 7 are used and the values from Tables I and II are applied. The method presented here, used by hand or computer, yields the same complexity values as the Bertz method. Also, it is less liable to error than identifying and counting bond connectivities, when used by hand, and much more amenable to computerization in this way as well. The program CPXCAL[7] is available to anyone with an interest in comparing molecular complexities in any molecular families of interest.

ACKNOWLEDGMENT

We are grateful to the National Science Foundation for a grant (CHE-9102972) that has generously supported this work.

REFERENCES AND NOTES

(1) (a) Bertz, S. H. *Chem. Commun.* **1981**, 818. (b) Bertz, S. H. *J. Am. Chem. Soc.* **1981**, *103*, 3599; **1982**, *104*, 5801. (c) Bertz, S. H. *Bull. Math. Biol.* **1983**, *45*, 849. (d) Bertz, S. H. In *Chemical Applications of Topology and Graph Theory*; King, R. B., Ed.; Elsevier: New York, 1983; p 206.
(2) Hendrickson, J. B.; Grier, D. L.; Toczko, A. G. *J. Am. Chem. Soc.* **1985**, *107*, 5228.
(3) The three-atom skeleton extracted may be of any three linked, nonhydrogen atoms, i.e., C–C–C, C–O–C, C–N–S, etc.
(4) In eq 4, *h* is the number of attached hydrogens plus the number of unshared electron pairs, regarded as the conjugate bases of potentially attached hydrogens.
(5) Hendrickson, J. B.; Toczko, A. G. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 171.
(6) Hendrickson, J. B. *J. Am. Chem. Soc.* **1977**, *99*, 5439.
(7) CPXCAL is written for a DEC VAX computer with an input module for direct graphic input from a Tektronix graphics terminal.

An Algorithm To Identify and Count Coplanar Isomeric Molecules Formed by the Linear Fusion of Cyclopentane Modules

SEYMOUR B. ELK
Elk Technical Associates, New Milford, New Jersey 07646

Received September 18, 1985

Because each of the various possible isomers formed by the coplanar linear fusion of cyclopentane modules may be represented by a binary sequence, the reverse technique of examining each binary sequence of a specified length underlies the formation of an algoithm to identify and count such isomeric molecules. The algorithm involves the specification of a set of three binary operations—which correspond to allowable physical transformations. This may be expressed in the form of a formal algebraic table of operations. Application of this algorithm, with Patterson's drawing convention, produces a canonical representation for each such isomer.

Despite the presence of a certain amount of noncoplanarity in the cyclopentane molecule (which is caused by the relieving of the strain that would result if the hydrogen atoms were allowed to remain eclipsed),[1] the simplified geometrical skeleton model formed by the successive "straight line"[2] concatenation of regular pentagonal modules gives a fairly good

0095-2338/87/1627-0067$01.50/0