On the Structure of Total Synthesis of Complex Natural Products

H. W. Whitlock

Department of Chemistry, University of Wisconsin, Madison, Wisconsin 53706

Received July 23, 1998

A chemically intuitive measure (metric) for molecular complexity is described. It is then applied to several syntheses of complex natural products. We conclude that the metric corresponds well with chemical ideas of complexity, and that different syntheses do in fact have appreciably different logical structures as measured by this. Possible implications of this interesting observation are discussed.

Introduction

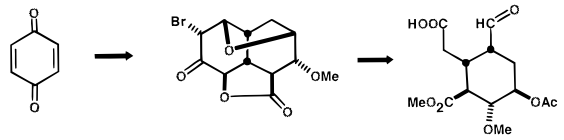
For some time I have been struck by the differences in the logical structure of published total syntheses of natural products. The original impetus for this thought was Woodward's reserpine synthesis,[1] wherein one observes the initial construction of a complex multibridged intermediate followed by its unraveling to the CDE ring system of the alkaloid (Scheme 1).

Subsequently we have seen the tremendous impact of Corey's synthon concept[2] on organic synthesis and the flowering of synthetic methods, culminating perhaps in Kishi's recent halichondrin synthesis.[3]

We divide this paper into three sections, each with a specific subject: (1) We can reduce the chemical idea of structural complexity to a number. This is defined, illustrated, and compared with the Bertz[4] topological definition of molecular complexity. (2) Application of the metric to Woodward's reserpine synthesis shows this synthesis to go through a numerical complexity maximum as in Scheme 1. Application to other total syntheses shows them to have considerable variation from one to another. This seems to be a reflection of their underlying structure. (3) We finally offer several conclusions and conjectures based on the set of total syntheses examined.

The crypto-mathematical nature of this paper involves several symbols. They will be defined below but are briefly stated here: H is a measure of structure size. It is the number of bonds. S is a measure of structural complexity. It is the count of various structural features. ΔS is the change in complexity of a synthetic step. It may be positive or negative. ΔS_m measures the complexity of an entire synthesis or synthetic sequence.

Scheme 1



Part 1. The Structural Complexity Metric *S*, and the Size Metric *H*.

The purpose of this work is to quantitate our chemist's intuition of molecular complexity as encountered in synthesis. We define two metrics, S for *molecular complexity*, and H for *molecular size*. These symbols were chosen to reflect their similarity to traditional thermodynamic quantities. Other mathematically based metrics are available but were not used.[4a−d,20]

1a. The Size Metric. We measure molecular size as simply the number of bonds in the molecule, one for single, two for double, and three for triple. This is designated as H, but is discussed only briefly.

1b. The Complexity Metric. We define molecular complexity (designated by the symbol S) to be the sum of the following terms: (1) A constant RINGVAL times the number of rings. (2)A constant UNSATVAL times the number of unsaturations. (3) A constant HETEROVAL times the number of heteroatoms. (4) A constant CHIRALVAL times the number of chiral centers.

In the present parametrization, RINGVAL is 4, UNSATVAL is 2, HETEROVAL is 1, and CHIRALVAL is 2. While certainly ad hoc, these parameters seem to be chemically reasonable and satisfactorally characterize the reserpine synthesis (see below).

The complexity metric S is easily calculated, either manually or by computer. We use a program fm.exe.[5] This simple program relieves us of the necessity of counting hydrogens, and permits numerous notational contractions (e.g. substitution of Bz for the C_7H_7 structural fragment). Being of no theoretical interest, the program is of crucial importance in rendering the counting metric manageable.

The program fm.exe takes as its input line the following arguments: (1) An optional leading comment. (2) The

(1) Woodward, R. B.; Bader, F. E.; Bickel, H.; Frey, A. J.; Kierstead, R. *Tetrahedron* **1958**, *2*, 1−57.

(2) Corey, E. J.; Long, a. K.; Rubenstein, D. *Science* **1985**, *228*, 408−418.

(3) Acher, T. D.; Buszek, K. R.; Fang, F. G.; Forsyth, C. J.; Jung, S. H.; Kishi, Y.; Matelich, M. J.; Scola, P. M.; Spero, D. M.; Yoon, S. K. *J. Am. Chem. Soc.* **1992**, *114*, 3162.

(4) (a) Bertz, S. H. *J. Am. Chem. Soc.* **1981**, *103*, 3599−3601. Bertz, S. H. *J. Am. Chem. Soc.* **1982**, *104*, 5801−5803. Bertz, S. H. *J. Chem. Soc., Chem. Commun.* **1986**, 1627−1628 Bertz, S. H. *Abstracts of Papers*, National Meeting of the American Chemical Society; American Chemical Society: Washington, DC, 1996, ORGN, 282.

(5) Available on request as source and executable from the author.

10.1021/jo9814546 CCC: $15.00
Published on Web 10/06/1998

Scheme 2

```
Input line:

Taxol(4.1)      C23O10Ac2Bz2NPh   r=4   u=3  c=11


Output:

Value:  C47H51O14N, 4 rings, 7 unsat, 11 chirals, 3

AromRing, 9 AromUnsat

  H-nbond=  135

  S-nring=  16

  S-numsat= 14

  S_hetero= 15

  S-chiral= 22

H=135, S=67, (G=202)  S:H =0.50
```

molecular formula without hydrogen. (3) The following in any order: (4) The number of alicyclic rings (r = ...). (5) The number of aromatic rings (a = ...). (6) The number of aromatic unsaturations (z = ...). (7) The number of unsaturations (u = ...). (8) The number of chiral centers (c = ...).

As output, the program produces a brief summary and the *H* (size) and *S* (complexity) metrics. The program accepts a file name as its argument. This file may have as many lines of arguments as desired, so a sequence of synthetic conversions is easily processed.

A sample of the program's input and output for the molecule taxol is shown in Scheme 2. In this case the complexity metric *S* is calculated to have the value 67 (no units).

1c. Discussion. Any attempt to quantitate intuitive knowledge[6] must state clearly the assumptions and techniques used. Our procedure involves three steps.

First we identify the protecting groups present on the structure and manually replace them with shorthand notation. For example, we reduce all benzyl type protecting groups (substituted benzyl, trityl, methoxytrityl, etc.) to a single generic benzyl. All silyl protecting groups are replaced by a single generic TMS group. This step is not required by our treatment of the complexity metric, but simplifies the interpretation of the *S* value. Not being required, this abstraction should be applied consistently if at all.

Next, a skeletal formula without hydrogens, together with the above structural features, is generated manually. This step can be easily automated. The structural features (number of rings etc.) are appended to the formula line as shown in the top line of Scheme 2.

The skeletal formula is submitted to the program fm.exe. Output of the program is the complete molecular formula, the bond count (*H*) and the complexity metric *S* (Scheme 2). Computation of a molecular formula is extremely tedious if done manually.

The program fm.exe first expands pseudo-atom macros[7] (e.g. "Et", "Bn", "E", etc.) to chemical symbols. It then computes[8] the molecular formula, and hence the bond count. The complexity metric is computed from the molecular formula and the above structural features that were manually counted.

Table 1. Comparison of *H*, *S*, and η

structure	Bertz[4] η	*H*	*S*
n-pentane	3	16	0
2-methylbutane	4	16	0
neopentane	6	16	0
cyclohexane	6	18	4
bicyclobutane	8	11	8
cyclobutene	7	11	6
tetrahedrane	12	10	12
modhephene	34	42	20

Since our focus in devising this metric was simplicity and intuition, it is important to apply this simplification in a consistent manner. The complexity values used for the various structural features (e.g. $\Delta S = 4$ for a ring) were chosen simply because they gave a reasonable picture. We are reasonably satisfied with the values used and do not anticipate any major changes, but we note their empirical nature.

Aromatic rings are not counted toward complexity. Parameters for aromatic rings and unsaturations are used only in calculating molecular formula. The number of unsaturations is the number of nonaromatic unsaturations. All aromatic rings may be treated equally. The pseudo-atom "Ph" is accepted as a generic aromatic ring and is converted to a six-carbon fragment.

All input and output values are included in the Supporting Information, and on the author's web page at http://organic.chem.wisc.edu/supplement.

With some notable exceptions, *S* tracks Bertz η value fairly closely (Table 1), but not surprisingly there are differences. *S* is zero for any acyclic alkane, while η varies as to the branching.

1d. Comments. Our metric has several characteristics that follow from its derivation. First it is *linear* in structure and not, in any reasonable sense, recursive or context sensitive. This means that it simply counts locally defined structural features such as number of heteroatoms, chiral centers, and rings. An alternative would be to define complexity as a measure of interacting functionalities. This could lead to a metric spanning the entire molecule. While the idea that complexity in some sense should encompass global molecular features has some appeal we wanted the *simplest* metric that would capture our ideas.

Second, *S* is a function of structure and *not* of the synthetic reactions available. Certainly the synthetic complexity of a molecule would be very different if the set of synthetic transformations was the classic Grignard cycle (interconversions of alcohols, alkyl halides, aldehydes/ketones and alkenes, and the Grignard coupling) instead of those presently available. We will touch on this issue below.

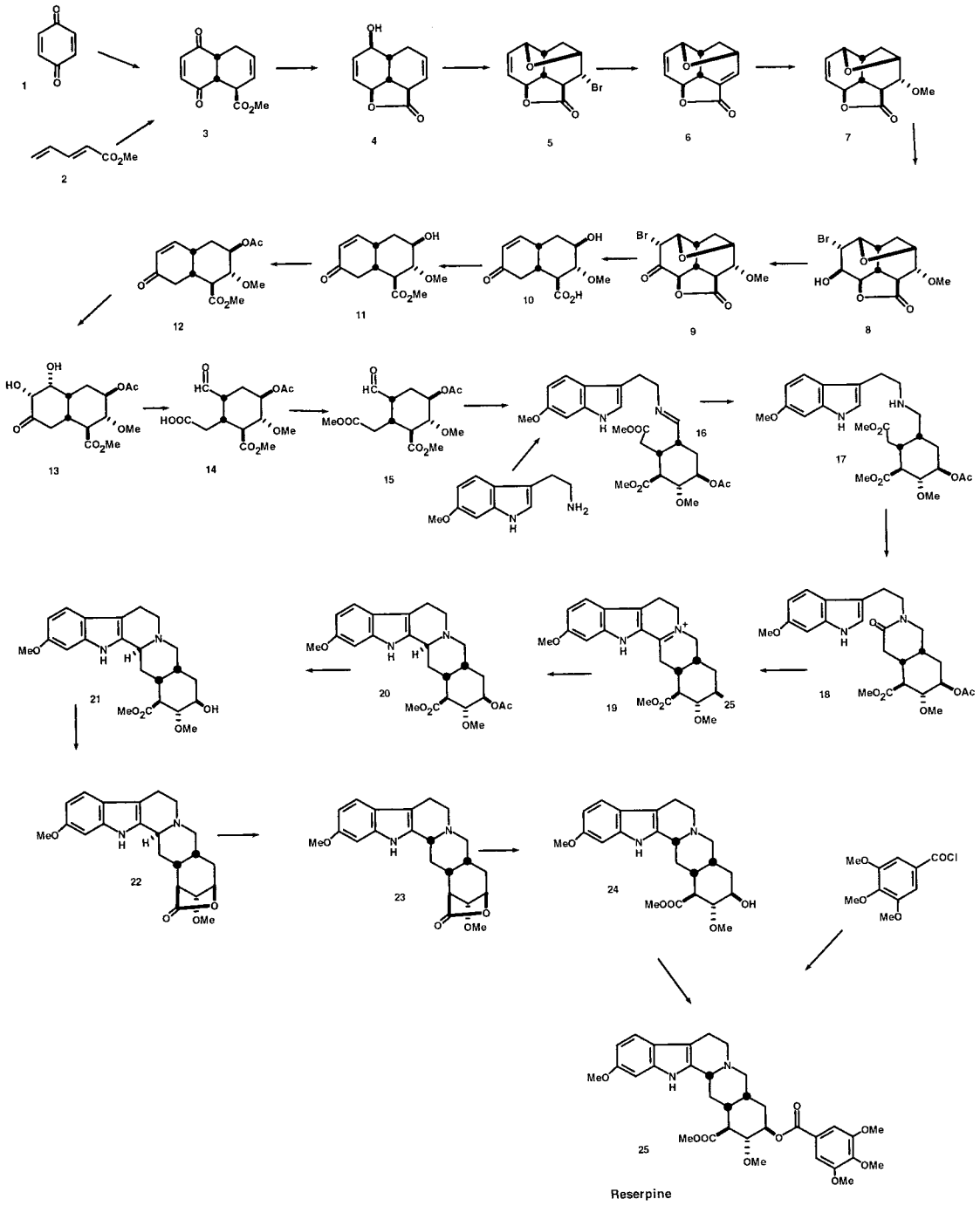
Third, *S* is a simple counting metric, and is *not* a mathematically correct measure of structural complex-

(6) A similar procedure has been used by Corey in evaluating synthetic transforms. Corey, E. J., Long,. A. K.; Lotto, I.; Rubenstein, S. D. *Recl. Trav. Chim. Pay.* **1992**, *111*, 304.

(7) Pseudo-atoms currently incorporated into fm.cpp are "Et" for ethyl group, "Pr" for propyl, "R" for generic alkyl, "Cx" for carboxyl, "E" for methyl or other simple ester, "Bn" for benzyl, "Ph" for phenyl or simply substituted phenyl, "Ac" for acetyl, "Q" for generic trisubstituted silyl protecting group, and "Bz" for benzoyl or simply substituted benzoyl group.

(8) The molecular formula calculation uses characteristic valency information. Fm.exe "understands the following elements: "X" (monovalent), "I", "Li", "Se", "Bz", "R", "O", "Cl", "Br", "N", "P", "S", "S(O2)", "N(+)" (quadravalent ammonium ion), "Si", and "Sn". These are present as a data array and are readily modified and added to.

Scheme 3. Structure Numbering for Reserpine Synthesis



ity.[4] This is by design; the situation in total syntheses of the type considered here is *so* complex that we felt it imperative to keep the metric as simple and chemically intuitive as possible. Simple counting measures are considered to be best for comparisons of complicated systems.[9] Metric *S* resembles the Cyclomatic[10] measure of computer program complexity. The Cyclomatic metric simply counts loops and branches in programs, but is considered to be an outstanding indicator of program complexity.[10] The fact that we are addressing the behavior of synthetic chemists, and not the abstract nature of molecular structure was also an important consideration in our choice of measures.

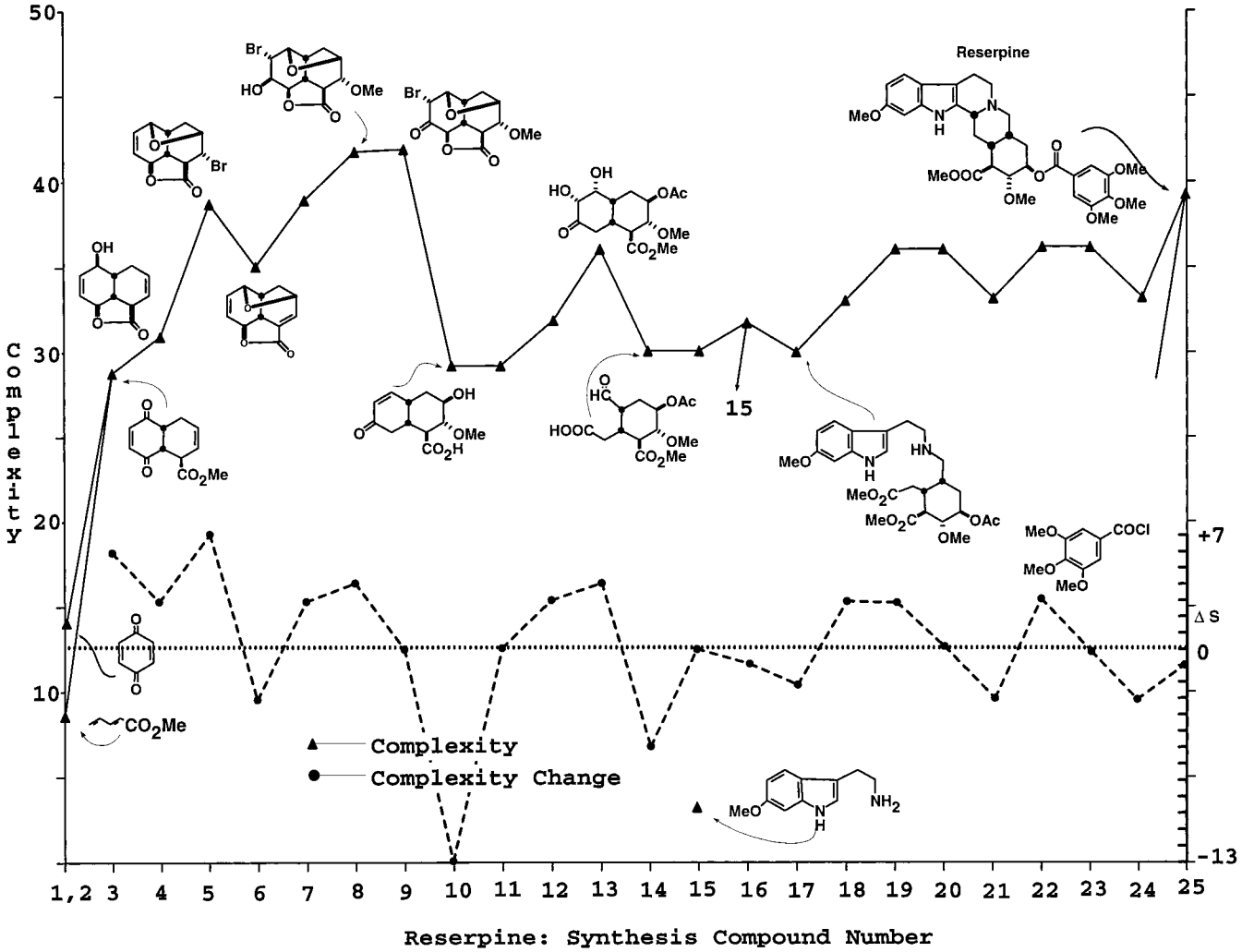
Part 2. Application to Several Syntheses. We now apply *S* to several syntheses. We conclude that *S* both reflects our biases concerning synthetic complexity and reveals distinct differences between syntheses of even the same target molecule.

2a. Woodward's Reserpine Synthesis. In Scheme 4 we present a plot of the metric *S* vs synthesis steps for Woodward's reserpine synthesis.[1] Structure numbering is shown in Scheme 3. Two things are plotted: the solid line represents the course of complexity *S* during the course of the synthesis, and the dotted line labeled

(9) Landes, D. S. *The Wealth and Poverty of Nations*; W. W. Norton, **1998**; p 416.

(10) (a) McCabe, T. *IEEE Trans. Software Eng.*, SE-2 **1976**, *4*, 308. (b) McCabe, T. *IEEE Software* **1996**, May. (c) Feghali, I.; Watson, A. *Communications of the ACM*, April **1994**. (d) McConnell, S. *Code Complete*; Microsoft Press: Redmond, CA, 1993.

Scheme 4. Complexity Plot for Reserpine Synthesis



"complexity change" (ΔS) represents the change in complexity per reaction. The term ΔS is defined as the difference in S between a structure and the sum of S values for its precursor(s). S is always positive, while ΔS may be positive or negative.

The following characteristics of this synthesis are revealed by this presentation.

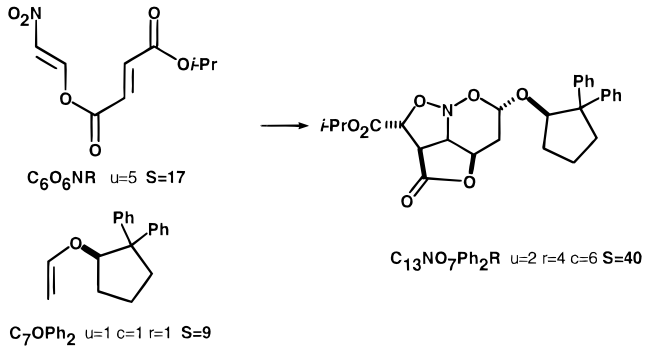
(1) The impetus for this work, the idea that this synthesis goes through an initial burst of complexity[11] associated with the Diels–Alder halo-etherification sequence is supported. Rapid construction of the tetracyclic structure **9** is followed by its opening and cleavage leading to a sharp drop in S (to **10**) and the monocyclic aldehyde **14**.

(2) On a per-reaction basis, ΔS bobbles around zero. When summed over the synthesis, it of course gives the difference in S between reserpine and the "starting" materials quinone and methyl pentadienoate.

(3) As expected, cleavage reactions (e.g. **9** → **10**) have a negative ΔS, while ring closures such as **4** → **5** have positive. This is to be expected.

(4) One observes that the Diels–Alder reaction has a substantial positive ΔS value (structure 18). However this is *not* a feature of condensation reactions in general. It arises from the combination of ring and chiral center formation. Other electrocyclic reactions of course share this property[12] (Scheme 5, $\Delta S = 14$). The final step in the reserpine synthesis, condensation of **24** with 3,4,5-trimethoxybenzoyl chloride to produce the reserpine target has a slight negative ΔS value, even though the main route complexity jumps substantially.

Scheme 5. Another Complexity-Increasing Cycloaddition

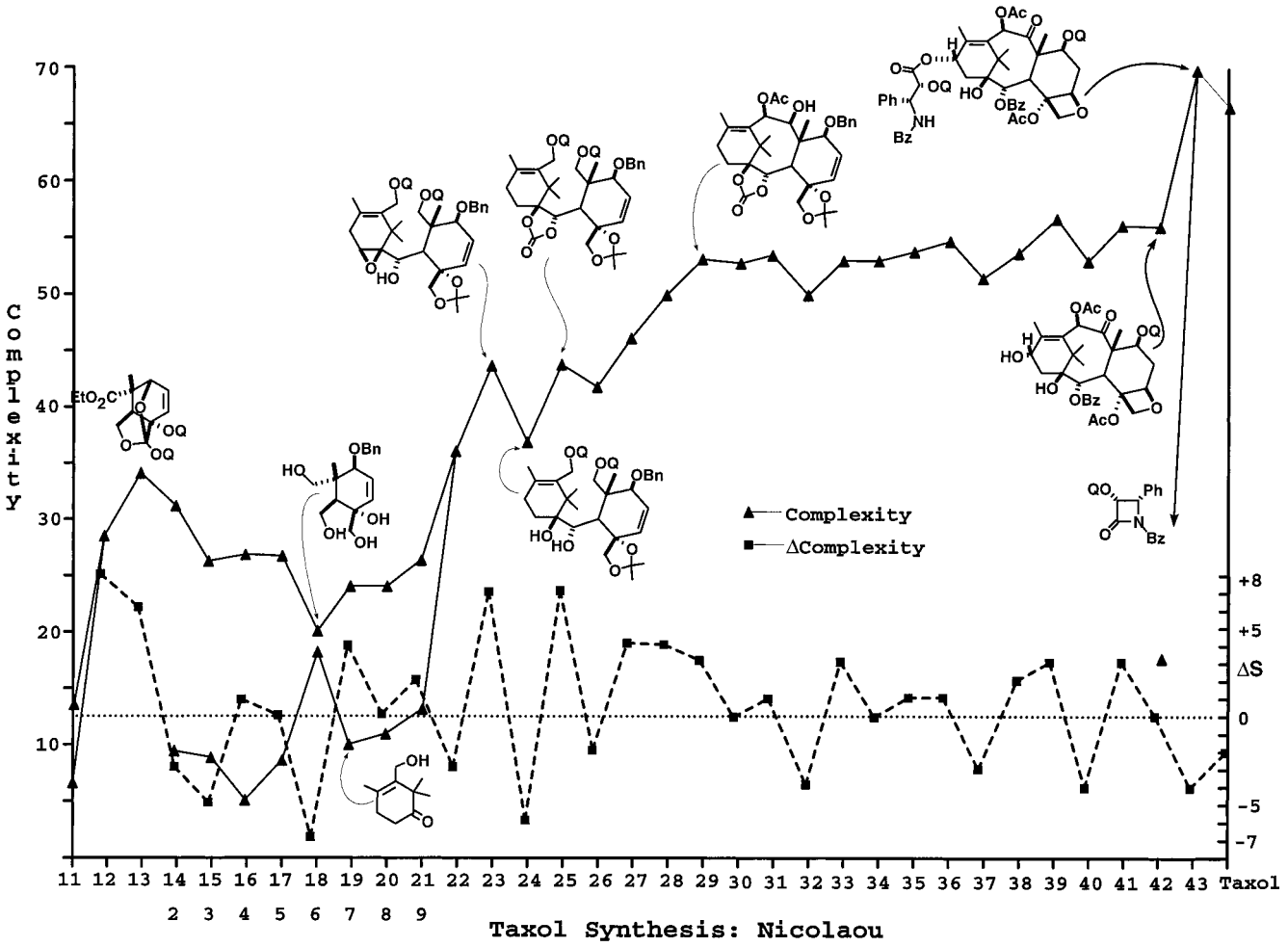


2b. The Complexity Plot Depends on the Synthesis and Not on the Target. In Scheme 6 we show the corresponding complexity plot for Nicolaou's synthesis of taxol. In Scheme 7 we illustrate the plot for Holton's

(11) Bertz (see ref 3) has observed a similar phenomenon using a graph-theoretic measure of complexity.

(12) Denmark, S. E.; Thorarensen, A.; Middleton, D. S. *J. Am. Chem. Soc.* **1996**, *118*, 8266.

Scheme 6. Complexity Plot for Taxol Synthesis



taxol synthesis. These two examples were chosen to compare two total syntheses of the same target molecule. We see that they are appreciably different. The Nicolaou synthesis shows the gyrations of *S* that were characteristic of the reserpine synthesis. Again, the Diels–Alder reaction followed by fragmentative simplification is the cause. The result of this, structure 18 (Scheme 6), lies at a complexity minimum and is followed by a continuous climb in *S* and a long sequence of structural modifications involving small ΔS. The penultimate acylation of **42** with the β-lactam jumps *S* substantially, but this condensation has a ΔS of zero.

In contrast the Holton synthesis shows an almost straight line (see below)! While jumps in complexity appear, for example cyclic carbonate formation **21** → **22**, and **15** → **16**, the synthesis as a whole is a smooth linear complexity increase, with no Diels–Alder lumps. The largest (negative) ΔS is found in the fragmentation of tetracyclic **8**.

2c. Halichondrin B. We now turn briefly to Kishi's herculean synthesis of halichondrin B.[3,13] Rather than present a complexity plot of this highly convergent[13] approximately 133-step synthesis we produce a three-dimensional representation[14] showing the linear synthetic runs and condensations (see Scheme 8).

Scheme 8 is a perspective drawing where the vertical lines are proportional to the *S* of the attached intermediate. Synthetic runs (linear sequences of transformations) go from left to right and are represented by gray ribbons. Condensations are represented by simple black lines connecting the two reactants (left) with the condensate (right). Scheme 9 illustrates this: Intermediate **87** (see ref 3 and the Supporting Information[14]) has $S = 92$, intermediate **83** has $S = 36$, and intermediate **86** has $S = 56$. Condensation of **83** and **86** to form **87** has a ΔS of zero, even though **87** is more complex than either of its precursors.

The remarkable feature of this total synthesis is the extremely linear nature of the fourteen precondensation synthetic runs employed. Some runs (e.g. **16** → **27**) even have negative ΔS's!

2d. Complexity of a Total Synthesis. We define the quantity ΔS_m as the mean[15] of the absolute values of ΔS over the entire synthesis (or a linear run) to get a quantitative picture of this linearity issue. ΔS may be negative or positive, so we take its absolute value. Table 2 summarizes the mean ΔS_m of the four syntheses examined above, together with Corey's recent synthesis of the potent anti-HIV terpene neotripterifordin.[16,17]

(13) Velluz, L.; Valls, J.; Mathieu, J. *Angew Chem., Int. Ed. Engl.* **1967**, *6*, 778.

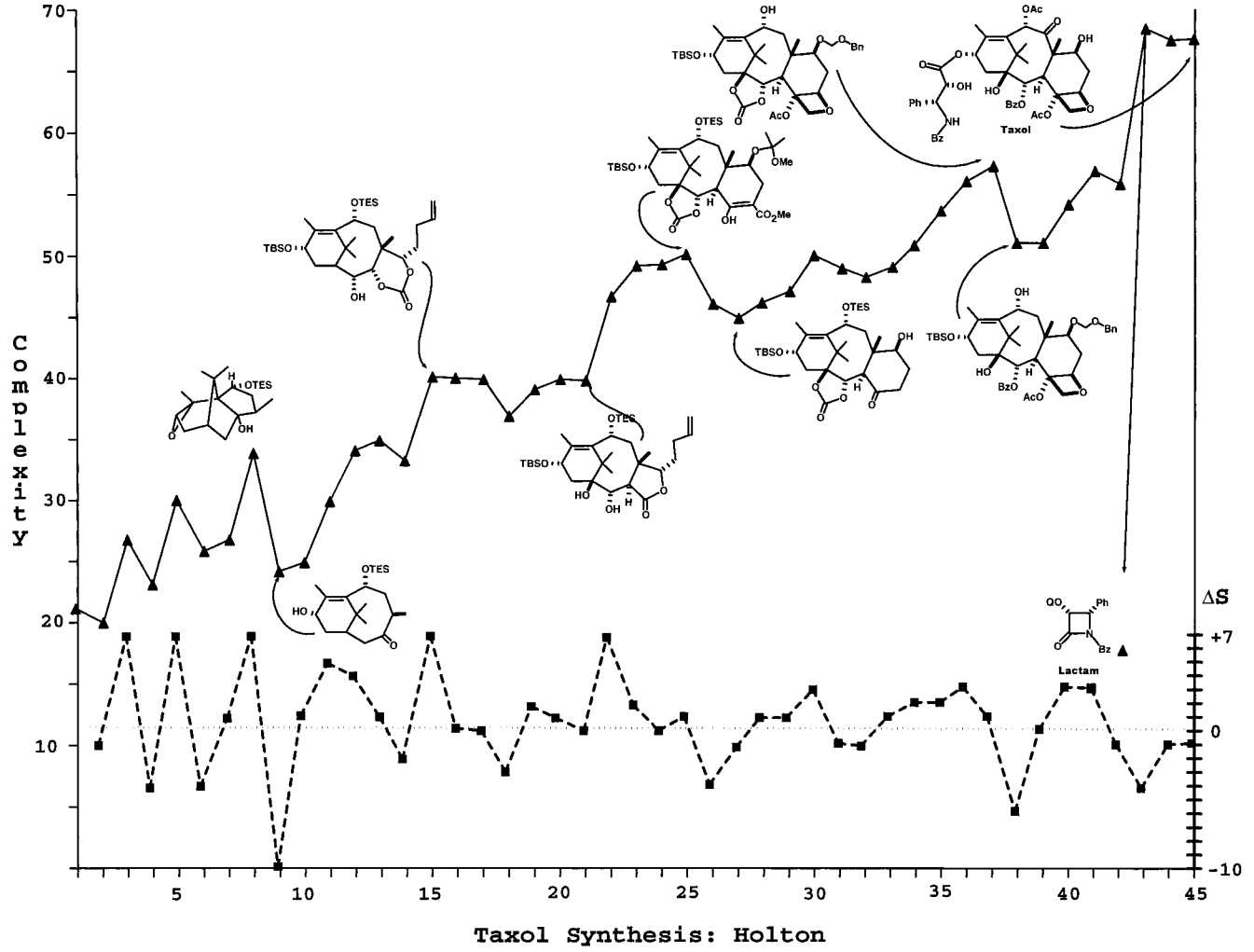
(14) The structure numbering used in this Scheme is that contained in the Supporting Information. This is also on the author's web site http://organic.chem.wisc.edu.

(15) $|\Delta S_m|$ is the sum of the absolute values of the individual steps divided by the number of steps.

(16) We thank Ms. Karin Maxson for this test case.

(17) Corey, E. J.; Liu, K. *J. Am. Chem. Soc.* **1997**, *119*, 9929.

Scheme 7. Complexity Plot for Taxol Synthesis (Holton)



Looking at Table 2, we first see that the above qualitative arguments as to relative complexity changes derived from examination of the complexity plots are fully supported by the ΔS_m values. The reserpine synthesis has the largest ΔS_m followed closely by Nicolaou's taxol synthesis. Then comes neotripterifordin followed by Holton's taxol synthesis; Kishi's halichondrin B has the smallest value of ΔS_m over the synthesis. The *most* complex synthesis (reserpine) is greater than 40% more complex than the *least complex* synthesis (halichondrin). That this is an integral feature of the synthesis is supported by the standard deviations, which fall in the same order. The larger ΔS_m, the larger the standard deviation. A large ΔS_m arises from the large positive and negative ΔS reactions as discussed above.

The above set of data (Table 2) may be fit to the equation

$$\Delta S_m = 3.2415 - 8.0649 \times 10^{-3} \times (\text{number of steps}), \quad r = -0.96$$

As the length of the synthesis increases, the mean ΔS per step (ΔS_m) decreases. A glaring exception to this linear dependence is Corey's recent synthesis of the sesterperpene scalarenedial,[18,16] an exceptionally short synthesis involving a biomimetic polyene cyclization. Its ΔS_m is 4.0 for 11 steps. While it conforms qualitatively to the idea that short syntheses have large ΔS_m, it does not fall on the regression line (predicted $\Delta S_m = 3.153$). Another example would be the cyclization of squalene oxide to protolanosterol. This has a ΔS_m of 21 for one step (calculated 3.23).

Conclusions

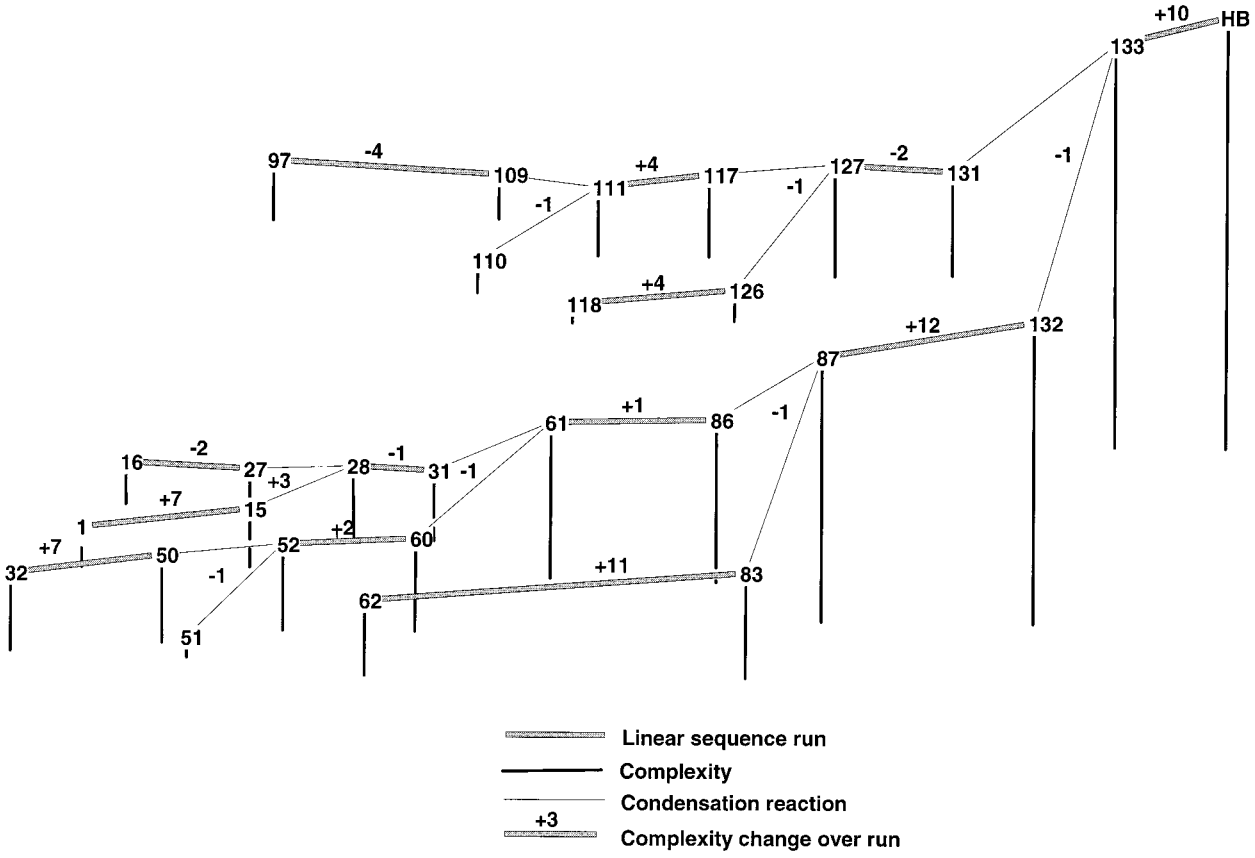
This work has some major assumptions built into it. The metric *S* was designed not from a theoretical but from a chemical basis. The linearity metric ΔS_m was not derived from a model or theory[4b,19] but from observation of published work. We define neither "total synthesis" nor "natural product", assuming these to be contextually clear. We assume that syntheses differing in some 40 years in time are comparable in form, even though the technology is not.

(1) The original impetus of this work was to see if the Woodward reserpine synthesis could be reduced to a plot that reflected our chemical biases of complexity, and this has been done. The metric *S*, while quite simple, tracks molecular complexity changes in total syntheses in an extremely gratifying manner. There is no reason to

(18) Corey, E. J.; Luo, G.; Lin, L. S. *J. Am. Chem. Soc.* **1997**, *119*, 9927.

(19) A particularly thorough consideration of the formal properties of synthesis trees may be found in Hendrickson, J. *Topics Current Chem.* **1976**, *62*, 49−172.

Scheme 8. Alternative Complexity Plot for Halichondrin B Synthesis



Scheme 9. Clarification of Scheme 8

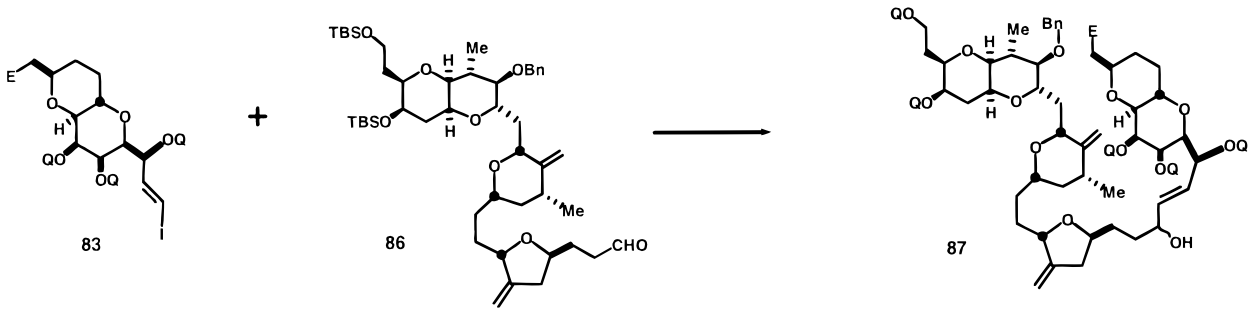


Table 2. Complexity of Several Syntheses

synthesis	no. of steps	ΔS_m	std deviation
reserpine	23	3.0869	2.9374
neotripterifordin	24	2.9583	2.2662
taxol (Nicolaou)	40	3.075	2.2689
taxol (Holton)	44	2.7954	2.5112
halichondrin B	128	2.203	1.9737

expect that S is unique in this respect.[20] Our suspicion is that pretty much any estimate of structural complexity would do the job, but the conclusion that chemically intuitive complexity metrics may be applied to total synthesis seems valid. That one can produce a "reasonable" numerical plot of something as complex as a total synthesis strikes the author as a peculiar but potentially useful conclusion.

(2) We have applied S to several total syntheses as discussed above and find that again it tracks complexity changes in a visual and chemically satisfying manner. But now we find that there are striking visual differences between various syntheses of the same target. We draw the following conclusion. *Total syntheses of complex natural products have a substantial underlying structure or form.*[21] This form is revealed by use of the metric S.

(3) The syntheses examined support the following conjecture. With respect to the metric ΔS_m, *A shorter synthesis will have a larger* ΔS_m *than a longer. It will be "more complex".* Since we have not defined "total synthesis", or "complex natural product", this rule easily falls to *reductio ad absurdum*.[22] Exceptions should not be hard to find, but why do all of the target molecules examined conform to this conjecture? The *obvious* answer is that short syntheses must assemble complexity quickly, while extremely long syntheses will fall victim to what is essentially the phenomenon of regression to the mean.

(20) Luzanov, A. V.; Babich, E. N. *J. Mol. Struct. (THEOCHEM)* **1995**, *333*, 279.

(21) We rely heavily on the fact that two syntheses of the *same* target differ.

(22) For example the two-step synthesis of glycol diacetate, $|\Delta S_m|$ is 1.

(4) Where does the increase in *S* (and *H*) during a synthesis come from? Our model is extremely simple; all reactions are either operators or condensations. Since most condensations have a small or slightly negative ΔS, we suggest that *the purpose of reagents is to transfer mass and complexity*. It is for this reason that we use the symbols *S* and *H*. There is clearly conservation of complexity at work.

Supporting Information Available: Complete structure numbering schemes, complexity data, and the source code (C++) of the program FM.EXE (48 pages). This material is contained in libraries on microfiche, immediately follows this article in the microfilm version of the journal, and can be ordered from the ACS; see any current masthead page for ordering information.

JO9814546